SHRINKAGE ESTIMATION OF NONLINEAR MODELS AND COVARIANCE MATRIX

JIANG QIAN

NATIONAL UNIVERSITY OF SINGAPORE

2012

SHRINKAGE ESTIMATION OF NONLINEAR MODELS AND COVARIANCE MATRIX

JIANG QIAN

(B.Sc. and M.Sc. Nanjing University)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2012

ACKNOWLEDGEMENTS

I would like to give my sincere thanks to my supervisor, Professor Xia Yingcun, who accepted me as his student at the beginning of my PhD study at NUS. Thereafter, he offered me so much advice and brilliant ideas, patiently supervising me and professionally guiding me in the right direction. This thesis would not have been possible without his active support and valuable comments. I truly appreciate all the time and effort he has spent on me.

I also want to thank other faculty members and support staffs of the Department of Statistics and Applied Probability for teaching me and helping me in various ways. Special thanks to my friends Ms. Lin Nan, Mr. Tran Minh Ngoc,

Mr. Jiang Binyan, Ms. Li Hua, Ms. Luo Shan, for accompanying me on my PhD journey.

Last but not least, I would like to take this opportunity to say thank you to my family. My dear parents, who encouraged me to pursue a PhD abroad. My devoted husband, Jin Chenyuan, who gives me endless love and understanding.

CONTENTS

SUMMARY

Recent developments in shrinkage estimation are remarkable. Being capable of shrinking some coefficients to exactly 0, the L_1 penalized approach combines continuous shrinkage with automatic variable selection. Its application to the estimation of sparse covariance matrix also gains a lot of interest. The thesis makes some contributions to this area by proposing to use the L_1 penalized approach for the selection of threshold variable in a Smooth Threshold Autoregressive (STAR) model, applying the L_1 penalized approach to a proposed varying coefficient model and extending a clustered Lasso (cLasso) method as a new way of covariance matrix estimation in high dimensional case.

After providing a brief literature review and the objectives for the thesis, we will study the threshold variable selection problem of the STAR model in Chapter

2. We apply the adaptive Lasso approach to this nonlinear model. Moreover, by penalizing the direction of the coefficient vector instead of the coefficients themselves, the threshold variable is more accurately selected. Oracle properties of the estimator are obtained. Its advantage is shown with both numerical and real data analysis.

A novel varying coefficient model, called the Principal Varying Coefficient Model (PVCM), will be proposed and studied in Chapter 3. Compared with the conventional varying coefficient model, PVCM reduces the actual number of nonparametric functions thus having better estimation efficiency and becoming more informative. Compared with the Semi-Varying Coefficient Model (SVCM), PVCM is more flexible but with the same estimation efficiency as SVCM when they have same number of varying coefficients. Moreover, we apply the L_1 penalty approach to identify the intrinsic structure of the model and improve the estimation efficiency as a result.

Covariance matrix estimation is important in multivariate analysis with a wide area of applications. For high dimensional covariance matrix estimation, assumptions are usually imposed such that the estimation can be done in one way or another, of which the sparsity is the most popular one. Motivated by the theories in epidemiology and finance, in Chapter 4, we will consider a new way of covariance matrix estimation through variate clustering.

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Background of the Thesis

1.1.1 Penalized Approaches

Modeling the relationship between a dependent variable and its associated independent variables is a very common problem in statistics. Moreover, many covariates which are initially available for inclusion may not be significant and should be excluded from the model. Given a sample of size n, variable selection can help improve the prediction performance of the fitted model by removing the redundant

independent variables. In recent years, an enormous amount of research has been done on algorithms and theory for variable selection.

Classical variable selection procedures include best subset selection and greedy subset selection. Exhaustive subset selection needs to evaluate all subsets of covariates, which is quite computationally expensive when there are a large number of predictors. For the three popularly used greedy subset selection methods: forward selection, backward elimination and stepwise selection, selecting or deleting one independent variable through some criteria is needed. However, it has been recognized that small changes in data would result in widely discrepant models from these methods. Moreover, Breiman (1996) showed that the subset selection procedures are unstable which costs large predictive loss.

Local curvature can be captured as more variables are chosen but the coefficient estimates suffer from high variance simultaneously. By observing that the unconstrained coefficients can explode, various penalized approaches have been proposed in the past few decades to regularize the coefficients thus controlling the variance.

Consider the linear regression model $y = X\beta + \varepsilon$, where y is an $n \times 1$ vector of responses, X is an $(n \times d)$-design matrix, β is a d-vector of parameters and ε is an $n \times 1$ vector of IID random errors. The penalized least squares estimates are

obtained by minimizing the residual squared error plus a penalty function, i.e.,

$$\hat{\beta}_{\text{penalized}} = \arg\min_{\beta} ||y - X\beta||^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

where $p_\lambda(\cdot)$ is a penalty function and the non-negative λ is a tuning parameter.

The ridge penalty function, introduced by Hoerl and Kennard (1970), is

$$p_\lambda(|\beta_j|) = \lambda|\beta_j|^2$$

and the bridge penalty function, introduced by Frank and Friedman (1993), is

$$p_\lambda(|\beta_j|) = \lambda|\beta_j|^q \, , q > 0.$$

The ridge regression utilizes the L_2-penalty and has good performance in the presence of collinearity. However, it shrinks the OLS estimates proportionally thus using all the predictors. The bridge regression shrinks smaller regression parameters to zero thus producing sparse models when $0 < q \leq 1$. However, if $0 < q < 1$, the penalty function is not convex, which will make the minimization problems hard to deal with.

Recent developments of penalized methods are noteworthy. Least absolute shrinkage and selection operator (Lasso), proposed by Tibshirani (1996), utilizes

$p_\lambda(|\beta_j|) = \lambda|\beta_j|$, i.e., it imposes an L_1-penalty on the regression coefficients. Because of the nature of the L_1-penalty, the Lasso does both continuous shrinkage and automatic variable selection at the same time. This approach is particularly promising not only because the resulting model is interpretable but also because it achieves the sparseness goal of variable selection. Fan and Li (2001) proposed the Smoothly clipped absolute deviation (Scad) penalty where

$$p'_\lambda(|\beta_j|) = \lambda\{I(|\beta_j| \leq \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} I(|\beta_j| > \lambda)\}$$

for some $a > 2$, where $I(A) = 1$ if the condition A is satisfied and $I(A) = 0$ otherwise. They further advocated using penalty functions which can result in an estimator with properties of sparsity, continuity and unbiasedness. As discussed in Fan and Li (2001), penalized methods should ideally satisfy the "oracle properties": that is, asymptotically

- zero coefficients and only zero coefficients are estimated as exactly 0, that is, the right subset model is identified;
- the non-zero coefficients are estimated as well as the correct subset model is known and the optimal estimation rate $1/\sqrt{n}$ is obtained.

The Scad penalty function can result in sparse, continuous and unbiased solutions, and the oracle estimator. However, it is limited to the non-convex penalty

function which increases the difficulty of finding a global solution to the optimization problem. Zou and Hastie (2005) proposed the elastic net estimator which is defined as

$$\hat{\beta}_{\text{enet}} = \arg\min_{\beta} ||y - X\beta||^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2.$$

The L_1 part of the penalty generates a sparse model while the L_2 part of the penalty can handle the highly correlated predictors thus overcoming the drawback of the Lasso.

On the other hand, the Lasso method is shown to be inconsistent in variable selection thus lacking the oracle property; see for example, Zhao and Yu (2006). To overcome this drawback, Zou (2006) proposed the adaptive Lasso estimator

$$\hat{\beta}_{\text{AdaLasso}} = \arg\min_{\beta} ||y - X\beta||^2 + \lambda \sum_{j=1}^{d} \hat{\omega}_j |\beta_j|,$$

where $\hat{\omega}_j = |\hat{\beta}_j^{\text{ini}}|^{-\gamma}$, $j = 1, \ldots, d$ with $\hat{\beta}_j^{\text{ini}}$ being an initial root-n consistent estimate of β_j. It allows an adaptive amount of shrinkage for each regression coefficient which can result in an estimator with oracle properties.

1.1.2 Threshold Variable Selection

Tong's threshold autoregressive (TAR) model (see, e.g., Tong and Lim (1980)) is one of the most popular models in the analysis of time series in biology, finance, economy and many other areas. It assumes different AR model in different regions of the state space divided according to some threshold variable y_{t-d}, $d \geq 1$. A typical two-regime threshold autoregressive (TAR) model is

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + (b_0 + \sum_{j=1}^{p} b_j y_{t-j}) I_r(y_{t-d}) + \varepsilon_t,$$

where I_r is an indicator function such that

$$I_r(x) = \begin{cases} 1 & \text{if } x > r \\ 0 & \text{if } x \leq r. \end{cases}$$

In order to estimate the model, it is necessary to specify the threshold variable. Tong and Lim (1980) used AIC (Akaike (1974)) to select d. Tsay (1989) proposed to use the F-statistic in the nonlinearity test $\hat{F}(p, d)$ to find the estimate of d such that $\hat{d} = \arg \max_{v \in \{1,\ldots,p\}} \{\hat{F}(p, v)\}$. This direct approach is not applicable when considering linear combination of several variables as the threshold variable.

Chen (1995) proposed two classification algorithms: discarding algorithm and

Bayesian algorithm to search for the most suitable threshold variable in the general situation. In the discarding algorithm, finding good initial parameter values is the first and important step where the data range of p-dimensional explanatory space is partitioned into k^p blocks with range of each explanatory variable partitioned into k equal intervals. Therefore, large sample is needed to provide reasonable initial values. The proposed Bayesian algorithm is automatic but relies on the information of prior distribution and Gibbs sampling method. From the review of van Dijk, Teräsvirta and Franses (2002), most existing studies focus on either model specification or parameter estimation with the delay parameter d chosen by hypothesis testing.

Wu and Chen (2007) proposed a k-state threshold variable driven switching AR (TD-SAR) model as follows

$$y_t = \mathrm{y_{t-1}}\phi^{(\mathrm{J_t})} + \varepsilon_{\mathrm{t}}^{(\mathrm{J_t})},$$

where $\mathrm{y}_{t-1} = (1, y_{t-1}, \ldots, y_{t-p})^\top$ and the switching mechanism is determined by the hidden state variable J_t with $p_{jt} = P(J_t = j) = g_j(Z_t), j = 1, \ldots, k$. The threshold variable $Z_t = \beta_0 + \beta_1 X_{1t} + \ldots + \beta_m X_{mt}$ where $X_{it}, i = 1, \ldots, m$ may be lag variables, observable exogenous variables or their transformations.

A three-stage algorithm is proposed to build the TD-SAR model in their paper.

First, the probabilities of the states J_t are estimated through a classification algorithm based on Bayesian approach. Second, the threshold variables are searched or constructed to provide the best fit of $\hat{p}_{jt}$. Three methods: CUSUM, SVM and SVM-CUSUM are provided in this step to select the candidates of threshold variables. The cumulative sum (CUSUM) method originated from statistical quality control is used to measure the agreement between the preliminary classification $\hat{p}_{jt}$ and a threshold variable candidate. The support vector machine (SVM) as a powerful tool for classification is applied to find the optimal linear combination $\beta = (\beta_0, \beta_1, \ldots, \beta_m)^\top$ for the threshold variable Z_t. The SVM-CUSUM is a combined method of CUSUM and SVM to find the potential candidates of threshold variables. Last, using Bayesian approach, the full model is fitted to the selected small number of threshold variable candidates based on some posterior BIC (PBIC) which is defined as the average BIC value given the posterior parameter distribution.

The link function $g_j(\cdot)$ in Wu and Chen (2007) is chosen to be the logistic function

$$P(J_t = j) = \frac{e^{Z_{jt}}}{1 + e^{Z_{jt}}}.$$

Actually, this idea of using a smooth link function to replace the step function $I(\cdot)$ originates from Chan and Tong (1986, esp., P187). They proposed to use this soft thresholding and introduced a more data driven model, smooth threshold

autoregressive (STAR) model of the form

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + (b_0 + \sum_{j=1}^{p} b_j y_{t-j}) F\Big(\frac{y_{t-d} - r}{c}\Big) + \varepsilon_t.$$

Here, $F(\cdot)$ is any sufficiently smooth function with a rapidly decaying tail. For example, $F(\cdot)$ can be chosen to be logistic distribution function or cumulative normal distribution function. This model includes the TAR model as a limiting case when $c \to 0$ and attracts lots of applications in econometrics, finance and biology. See, e.g., Chapter 3 of Franses and van Dijk (2000).

1.1.3 Varying Coefficient Model

As a hybrid of parametric and nonparametric model, semi-parametric model has recently gained much attention in econometrics and statistics. It retains the advantages of both parametric and nonparametric model and improves the estimation performance in high dimensional data analysis. Parametric model often imposes some assumptions in the form of the functional such as linear or polynomial, which are not always realistic in applications. Nonparametric model relaxes the assumptions on model specification and is more adequate in exploring the hidden relationship between response variable and covariates. However, the local

smoothing method used by nonparametric modeling has the problem of increasing variance for increasing dimensionality. This is often referred to as the "curse of dimensionality". Therefore, the application of the nonparametric model is not highly successful. Great effort has been made to reduce the complexity of high dimensional problems. Partly parametric modeling is allowed and the resulting models belong to semi-parametric models.

Semi-parametric models can reduce the dimension of the estimation by examining a lower dimension structure although different semi-parametric models explore the prior information from different angles. Varying Coefficient Model (VCM), introduced by Cleveland, Grosse and Shyu (1991), assumes that

$$Y = X^T \beta(U) + \varepsilon = \sum_{i=1}^{p} X_i \beta_i(U) + \varepsilon,$$

where $Y \in \mathbb{R}^1$ is the response of interest, $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ is the associated p-dimensional predictor, $U \in \mathbb{R}^1$ is the so-called univariate *index* variable, ε is the random noise and $\beta(U) \in \mathbb{R}^p$ is a vector of unknown smooth functions in $u \in \mathbb{R}^1$, called the varying coefficients. From its mathematical expression, we can see that the VCM only relies on the index variable and allows the coefficients to be fully nonparametric. It thus provides a powerful tool for the study of dimension reduction because the model is easy to interpret and free of the "curse of dimensionality" of nonparametric modeling.

As for the estimation of the VCM model, Hastie and Tibshirani (1993) proposed a one-step estimate for $\beta_i(U)$ based on a penalized least squares criterion. This algorithm can estimate the models flexibly. However, it is limited to the assumption that all the coefficient functions have the same degree of smoothness which is quite strong. Without this assumption, Fan and Zhang (1999) showed that the one-step method is not optimal. They also proposed a two-step method to repair this drawback. However, the two-step estimation is numerically unstable. This is because the two-step estimation adopts the kernel smoothing approach to estimate the functional coefficients and the kernel approach needs dramatically increasing sample size to improve the numerical stability when the predictor's dimension is large; see, Silverman (1986).

1.2 Research Objectives and Organization of the Thesis

As can be seen from the above review, the following research gaps still exist:

- Selection of the threshold variable is essential in building a Smooth Threshold Autoregressive (STAR) model. However, determining an appropriate threshold variable is not easy in practice. Current approaches either focus

on hypothesis testing methods or some classification algorithms. The hypothesis testing methods are feasible for univariate threshold variable but tedious for the linear combination of variables. The classification algorithms either require a good initial fit or rely on some Bayesian algorithm which may be computationally expensive.

- Varying coefficient models can be used to model multivariate nonlinear structure flexibly and partly solve the "curse of dimensionality" issue. However, the numerical stability of the estimation methods has yet to be improved. Small error in the initial condition will result in large discrepancy in the prediction results due to the numerical instability of the method.
- Currently, studies of high dimensional covariance matrix estimation mainly focus on the sparse assumption where the shrinkage approaches are applied to shrink the off-diagonal elements of covariance matrix to exactly 0. However, it is well known that in many biological and financial cases, the sparsity assumption amongst all the coefficients is inappropriate. Grouping the variables if their coefficients are the same is a natural way of solving this issue as well as achieving the goal of dimension reduction.

In the following Chapter 2 to Chapter 4, we aim to make some contributions to the above-mentioned three gaps.

In Chapter 2, we will study the threshold variable selection problem of the STAR model. We will propose to select the threshold variable by the recently developed L_1 penalizing approach. Meanwhile, noticing that the norm of the coefficient vector implies the threshold shape, which should not be penalized, this thesis will propose a direction adaptive Lasso method by penalizing the direction of the coefficient vector instead of the coefficients themselves. This study would provide insights into the threshold variable selection problem and should offer a better understanding on the application of the penalizing approaches to nonlinear models.

In Chapter 3, we will propose a novel varying coefficient model, called Principal Varying Coefficient Model (PVCM). By characterizing the varying coefficients through linear combinations of a few principal functions, the PVCM reduces the actual number of nonparametric functions, which may contribute to the improvement of the numerical stability, estimation efficiency and practical interpretability of the traditional varying coefficient model. Moreover, incorporating the nonparametric smoothing with the L_1 penalty, the intrinsic structure can be identified automatically and hence the estimation efficiency can be further improved.

In Chapter 4, we will consider a way of simplifying a model through variate clustering. Extension of the approach to the estimation of covariance matrix will

also be studied. Numerical studies will be performed, suggesting that the clustering idea has better prediction performance than the sparsity assumption in some situations.

We will conclude the thesis in Chapter 5 with the summarization and discussion on future research.

CHAPTER 2

Threshold Variable Selection via a L_1 Penalty

2.1 Introduction

In this chapter, we study the following STAR(p, q) model

$$y_t = (a_0 + \sum_{j=1}^{p} a_j y_{t-j}) + (b_0 + \sum_{j=1}^{p} b_j y_{t-j})\Phi(\theta_0 + \sum_{j=1}^{q} \theta_j y_{t-j}) + \varepsilon_t, \quad (2.1)$$

where we set the smooth link function $F(\cdot)$ in Chan and Tong's STAR model to be the standard Gaussian distribution for simplicity of discussion although this is not essential. $\{\varepsilon_t\}$ is assumed to be a white noise with finite variance σ^2, and be independent of the past observations $\{y_s, s < t\}$.

We also choose the threshold variable $z_t = \theta_0 + \sum_{j=1}^{q} \theta_j y_{t-j}$ which is a linear function of lagged endogenous variables. One advantage of the proposed model is in the selection of threshold variable. For example, if θ_k are all zeros except for $k = j$, then the selected threshold variable is y_{t-j}. We have the following result about the stationarity of the model.

Lemma 2.1. *If*

$$\sup_{0 \le u \le 1} \sum_{j=1}^{p} |a_j + b_j u| < 1, \tag{2.2}$$

there exists a strictly stationary solution $\{y_t\}$ *from the model (2.1).*

We propose to use the recently developed L_1 regularization approaches which tend to produce a parsimonious number of nonzero coefficients for z_t, thus leading to a simple way of selecting the significant/threshold variables without testing the $2^q - 1$ subsets of $\{y_{t-1}, y_{t-2}, \ldots, y_{t-q}\}$. The lasso penalty can perform model selection as well as estimation. However, its variable selection may be inconsistent; see, e.g., Zou (2006). In this Chapter, we adopt the adaptive lasso penalty proposed

in Zou (2006), which is convex and leads to a variable selection estimator with the oracle properties. Moreover, we propose a direction adaptive lasso method. By penalizing the direction of the coefficient vector instead of the coefficients themselves, the threshold variable is more accurately selected, especially when the sample size is not large enough. Note that the norm of the coefficient vector implies the threshold shape, which should not be penalized. Our penalization on the direction can achieve this goal while the direct penalization on the coefficient cannot. Both numerical and real data analysis are provided to illustrate its advantage. The oracle properties of the resulting estimators are also obtained.

2.2 Estimation

2.2.1 The Conditional Least Squares Estimator

Let $a = (a_0, a_1, \ldots, a_p)^\top$, $b = (b_0, b_1, \ldots, b_p)^\top$, $\theta = (\theta_0, \theta_1, \ldots, \theta_q)^\top$, we rewrite model (2.1) as

$$y_t = x_t^\top a + (x_t^\top b)\Phi(s_t^\top \theta) + \varepsilon_t, \tag{2.3}$$

where

$$x_t^\top = (1, y_{t-1}, \ldots, y_{t-p}),\ s_t^\top = (1, y_{t-1}, \ldots, y_{t-q}),$$

for $t = m+1, \ldots, T$ and $m = \max(p, q)$.

The unknown parameter vector $\eta = (a^\top, b^\top, \theta^\top)^\top = (\eta_1, \ldots, \eta_L)^\top$ $(L = 2p + q + 3)$ is assumed to be in an open set Θ of $\mathbb{R}^{\otimes(2p+q+3)}$. Denote $\theta = (\theta_0, \vartheta^\top)^\top = (\theta_0, \theta_1, \ldots, \theta_q)^\top$ with $\vartheta = (\theta_1, \ldots, \theta_q)^\top \in \mathbb{R}^q$ and the true value $\vartheta_0 = (\theta_{10}, \ldots, \theta_{q0})^\top$. Denote the true value of η by $\eta_0 = (\boldsymbol{a}_0^\top, \boldsymbol{b}_0^\top, \boldsymbol{\theta}_0^\top)^\top$. For ease of exposition, we use the boldfaced letter to denote a vector if there exists the same notation for a scalar. For example, $\boldsymbol{a}_0$ denotes the true value of the vector $a = (a_0, a_1, \ldots, a_p)^\top$ and $\boldsymbol{\theta}_0$ denotes the true value of vector $\theta = (\theta_0, \theta_1, \ldots, \theta_q)^\top$. Let K be the index set of those $j \in I \equiv \{1, \ldots, q\}$ with $\theta_{j0} \neq 0$ and κ be the number of components of K and denote $\bar{K} = I \backslash K$.

For each t, we refer to the lagged variables of y_t in the set $\{y_{t-j}, j \in K\}$ as the significant threshold variables and define the transition variable z_t as

$$z_t = s_t^\top \theta = \theta_0 + \theta_1 y_{t-1} + \ldots + \theta_q y_{t-q}. \tag{2.4}$$

Denote by $\mathscr{F}_t = \sigma(y_1, \ldots, y_t)$ $(t \geq 1)$ the $\sigma-$field generated by $y_s, 1 \leq s \leq t$ and

by $\mathscr{F}_0$ the trivial $\sigma-$field. Define

$$l_t = (1, \tilde{l}_t^\top)^\top, \;\; \tilde{l}_t = (y_{t-1}, \ldots, y_{t-m})^\top \tag{2.5}$$

and

$$g(\eta, \tilde{l}_t) \;\; = \;\; g(\eta, \mathscr{F}_{t-1}) \equiv E_\eta(y_t|\mathscr{F}_{t-1}) = x_t^\top a + (x_t^\top b)\Phi(s_t^\top \theta), \;\; t \geq 1.$$

Given a set of observations $\{y_1, \ldots, y_T\}$, the conditional least squares (LS) estimator minimizes the objective function

$$\begin{aligned} Q_T(\eta) &= \sum_{t=m+1}^{T} (y_t - E_\eta(y_t|\mathscr{F}_{t-1}))^2 \\ &= \sum_{t=m+1}^{T} \left\{ y_t - x_t^\top a - (x_t^\top b)\Phi(s_t^\top \theta) \right\}^2, \end{aligned} \tag{2.6}$$

with respect to η. Let η_T^{LS} denote the least squares estimator.

Theorem 2.1. *If $\{y_t\}$ is a stationary ergodic sequence of integrable variables and $\tilde{l}_0$ has a positive density function almost everywhere, then as $T \to \infty$,*

$$\eta_T^{LS} \to \eta_0, \;\; a.s. \tag{2.7}$$

and

$$T^{1/2}(\eta_T^{LS} - \eta_0) \Rightarrow N(0, \sigma^2 U^{-1}), \tag{2.8}$$

where

$$\begin{aligned} U &\equiv E_{\eta_0}\left(\frac{\partial g(\tilde{l}_t, \eta_0)}{\partial \eta} \cdot \frac{\partial g(\tilde{l}_t, \eta_0)}{\partial \eta^\top}\right) \\ &= E_{\eta_0}\left(\frac{\partial g(\tilde{l}_0, \eta_0)}{\partial \eta} \cdot \frac{\partial g(\tilde{l}_0, \eta_0)}{\partial \eta^\top}\right) \end{aligned} \tag{2.9}$$

is positive definite.

Remark 2.1. Using the Fisher information matrix $I(\eta)$,

$$\begin{aligned} I(\eta) &= E_\eta\left\{\frac{\partial \log f}{\partial \eta} \cdot \frac{\partial \log f}{\partial \eta^\top}\right\} \\ &= \frac{1}{\sigma^2} E_\eta\left\{\frac{\partial g(\tilde{l}_t, \eta)}{\partial \eta} \cdot \frac{\partial g(\tilde{l}_t, \eta)}{\partial \eta^\top}\right\}, \end{aligned} \tag{2.10}$$

where $f = (\sqrt{2\pi}\sigma)^{-1} \exp\{-\frac{\varepsilon_t^2}{2\sigma^2}\}$, the result of the Theorem 2.1 can be written as

$$T^{1/2}(\eta_T^{LS} - \eta_0) \Rightarrow N(0, I^{-1}(\eta_0)). \tag{2.11}$$

2.2.2 The Adaptive Lasso Estimator

In this section, we shrink the unnecessary coefficients of the transition variable z_t to 0 and select the true threshold variables by the adaptive lasso approach proposed by Zou (2006). We use η_T^{ADL} to denote the adaptive lasso estimator of η which minimizes

$$Q_T^{ADL}(\eta) \;=\; Q_T(\eta) + \lambda_T \sum_{j=1}^{q} \hat{w}_j |\theta_j|, \tag{2.12}$$

where the weight $\hat{w}_j$ is the reciprocal of an increasing function of the corresponding LS estimate of θ_j, i.e., $\hat{w}_j = 1/|\theta_j^{LS}|^\gamma$, $\lambda_T > 0$, $\gamma > 0$ are two nonnegative tuning parameters.

Let $K_T^{ADL} = \{j : \theta_j^{ADL} \neq 0, 1 \leq j \leq q\}$, where θ_j^{ADL} is the adaptive lasso estimate of θ_j. Recall that $K = \{1 \leq j \leq q : \theta_{j0} \neq 0\}$ and $\kappa = |K|$. That is, the correct model has κ significant threshold variables. For any vector/matrix A, denote by $A_{(K)}$ a sub-vector/sub-matrix of A formed by the elements at K'th rows (and K'th columns) of A. For example, if $A = (a_{ij})_{1\leq i,j\leq 5}$ and $K = \{1, 3\}$, then $A_{(K)} = (a_{ij})_{i,j=1,3}$.

Theorem 2.2. *Suppose that* $\frac{\lambda_T}{\sqrt{T}} \to 0$, *and* $\lambda_T T^{\frac{\gamma-1}{2}} \to \infty$. *Then the adaptive lasso estimates* η_T^{ADL} *satisfy the following oracle properties:*

1. *Consistency in variable selection:*

$$\lim_{T\to\infty} P(K_T^{ADL} = K) = 1.$$

2. *Asymptotic normality:*

$$\sqrt{T}(\eta_{T,(K)}^{ADL} - \eta_{0,(K)}\Big) \Rightarrow N_{2p+\kappa+3}\big(\mathbf{0}, I^{-1}(\eta_{0,(K)})\big).$$

The second part of Theorem 2.2 implies that the final estimator can achieve the efficiency of the estimator when the true threshold variables are known and estimated with irrelevant variables being removed. Thus, as in the literature estimator η_T^{ADL} has the so-called oracle property.

2.2.3 The Direction Adaptive Lasso Estimator

As $c \to +\infty$, the function $\Phi(c(x-r))$ approaches to the indicator function

$$I_r(x) = \begin{cases} 1 & \text{if } x > r, \\ 0 & \text{if } x \leq r, \end{cases}$$

which is the threshold principle of the classical two-regime TAR model. However, in the STAR(p,q) model (2.1), when the length of the vector $\vartheta = (\theta_1, \ldots, \theta_q)^\top$ is

large, penalizing $\tilde{\theta}_j \equiv \theta_j/\|\vartheta\|$ instead of θ_j seems more desirable ($j = 1, 2, \ldots, q$) than penalizing the coefficient vector since the latter also penalizes the length of the coefficients, which plays the role of c.

We call the estimator by adaptively penalizing the direction of coefficient vector the direction adaptive lasso estimator and denote it as η_T^{DAL}, which minimizes

$$Q_T(\eta) + \lambda_T \sum_{j=1}^{q} \tilde{w}_j |\tilde{\theta}_j| = Q_T(\eta) + \frac{\lambda_T}{l(\vartheta)} \sum_{j=1}^{q} \tilde{w}_j |\theta_j|, \tag{2.13}$$

where $l(\vartheta) = \sqrt{\theta_1^2 + \ldots + \theta_q^2}$, the new weight $\tilde{w}_j$ is the reciprocal of an increasing function of the corresponding LS estimate of $\tilde{\theta}_j$, i.e.,

$$\tilde{w}_j = 1/|\tilde{\theta}_j^{LS}|^\gamma = \frac{l^\gamma(\theta_T^{LS})}{|\theta_j^{LS}|^\gamma},$$

and $\lambda_T > 0$, $\gamma > 0$ are two nonnegative tuning parameters.

The oracle properties of η_T^{ADL} are provided by the following theorem.

Lemma 2.2. *As $T \to \infty$, $\tilde{\vartheta}_T^{LS}$, the LS estimator of $\tilde{\vartheta}$ satisfies*

$$\tilde{\vartheta}_T^{LS} \to \tilde{\vartheta}_0, \ \ a.s.$$

and

$$T^{1/2}(\tilde{\vartheta}_T^{LS} - \tilde{\vartheta}_0) \Rightarrow N(0, \tilde{\Sigma}),$$

where $\tilde{\vartheta}_0 = \vartheta_0/l(\vartheta_0)$ *and*

$$\tilde{\Sigma} = (\vartheta_0^\top \vartheta_0)^{-1}(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)$$

is a non-negative definite matrix with rank $q-1$. *Here,* I_q *is the* $q \times q$ *identity matrix,* $I^{-1}(\vartheta_0)$ *is submatrix composed of the last* q *rows and the last* q *columns of the inverse matrix of* $I(\eta_0)$ *defined in (2.10).*

Denote $K_T^{DAL} = \{j : \tilde{\theta}_j^{DAL} \neq 0, 1 \leq j \leq q\}$, where $\tilde{\theta}_j^{DAL}$ is the adaptive lasso estimate of $\tilde{\theta}_j$.

Theorem 2.3. *Suppose that* $\frac{\lambda_T}{\sqrt{T}} \to 0$, *and* $\lambda_T T^{\frac{\gamma-1}{2}} \to \infty$. *Then the direction adaptive lasso estimates* η_T^{DAL} *satisfy the following oracle properties:*

1. *Consistency in variable selection:*

$$\lim_{T\to\infty} P(K_T^{DAL} = K) = 1.$$

2. *Asymptotic normality:*

$$\sqrt{T}(\eta^{DAL}_{T,(K)} - \eta_{0,(K)}\Big) \Rightarrow N_{2p+\kappa+3}\big(\mathbf{0}, I^{-1}(\eta_{0,(K)})\big).$$

Under the same condition as the adaptive lasso method, Theorem 2.3 indicates that the proposed direction adaptive lasso also selects the correct subset of threshold variables consistently. From the asymptotic normality, the method can estimate the non-zero parameters efficiently as if we knew in advance which variables were uninformative and were removed.

2.3 Numerical Experiments

2.3.1 Computational Issues

For the adaptive lasso and direction adaptive lasso estimator, we apply the Local Quadratic Approximation (LQA) proposed in Fan and Li (2001) to our implementation. Suppose we have an initial value $\boldsymbol{\theta}_0 = (\theta_{00}, \theta_{01}, \ldots, \theta_{0q})^\top$ that is close to the optimization solution, except for a constant, we can equivalently get

the adaptive lasso estimator through minimizing

$$Q_T(\eta) + \frac{\lambda_T}{2}\theta^\top \Sigma \theta,$$

and get the direction adaptive lasso estimator through minimizing

$$Q_T(\eta) + \frac{\lambda_T}{2l(\theta)}\theta^\top \Sigma \theta,$$

where $\Sigma \equiv \Sigma(\boldsymbol{\theta}_0) = \text{diag}(v)$ with $\boldsymbol{\theta}_0$ being the value of the last step, and for the adaptive lasso,

$$v = (0, w_1/|\theta_{01}|, \ldots, w_q/|\theta_{0q}|)^\top, \;\; w_i = 1/|\theta_i^{LS}|^\gamma,$$

for the direction adaptive lasso,

$$v = (0, \tilde{w}_1/|\theta_{01}|, \ldots, \tilde{w}_q/|\theta_{0q}|)^\top, \;\; \tilde{w}_i = 1/|\tilde{\theta}_i^{LS}|^\gamma.$$

Remark 2.2. Under the assumption that $\theta_0 \neq 0$, the transition variable

$$z_t = \theta_0 + \theta_1 y_{t-1} + \ldots + \theta_q y_{t-q} \tag{2.14}$$

can also be equivalently written as

$$z_t = \frac{1 + \tau_1 y_{t-1} + \ldots + \tau_q y_{t-q}}{c} \tag{2.15}$$

with

$$c = 1/\theta_0, \;\; \tau_j = \theta_j/\theta_0, \;\; j = 1, \ldots, q.$$

In the numerical experiments, we use this form to evaluate the estimation accuracy.

Specifically, when we evaluate the MSE of the estimate of $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_q)^\top$, we use $(\hat{\tau}, \hat{c}) = (\hat{\tau}_1, \ldots, \hat{\tau}_q, \hat{c})$ instead. That is, we evaluate the deviation of $(\hat{\tau}, \hat{c})$ from the true value (τ_0, c_0) with $\tau_0 = (\tau_{10}, \ldots, \tau_{q0}) = (\theta_{10}/\theta_{00}, \ldots, \theta_{q0}/\theta_{00})$ and $c_0 = 1/\theta_{00}$.

M-folder Cross Validation (CV) and Bayesian Information Criterion (BIC) are used to select the tuning parameter $\rho = (\lambda, \gamma)$ and $\gamma \in \{0.5, 1, 2\}$ which is consistent with the choice of γ in Zou (2006). For the BIC, the criterion is

$$\mathrm{BIC}_\rho = \log(\mathrm{RSS}_\rho) + \mathrm{df}(\rho) \times \frac{\log(T-m)}{T},$$

where

$$\mathrm{RSS}_\rho = T^{-1} \sum_{t=m+1}^{T} \left\{ y_t - x_t^\top a - (x_t^\top b)\Phi(s_t^\top \theta) \right\}^2$$

and $\mathrm{df}(\tau) = 2p+3+\hat{q}$ with $\hat{q}$ being the number of nonzero coefficients identified by the estimate. For the M-folder CV, denote the full data set by T, and denote the cross-validation training and test set by $T - T^{\nu}$ and $T^{\nu}, \nu = 1, \ldots, M$, respectively. For each ρ and ν, we find the estimate using the training set and find a ρ that minimizes

$$CV(\rho) = \sum_{\nu=1}^{M} \sum_{y_k \in T^{\nu}} (y_k - \hat{y}_k)^2,$$

where $\hat{y}$ is the corresponding fitted value.

2.3.2 Numerical Results

Our aim of numerical experiments is to show the performance of using the L_1-penalization to select the threshold variables. Moreover, the finite sample performance of the LS estimator, adaptive lasso estimator and the proposed direction adaptive lasso estimator are also compared. We summarize the results in the following aspects. (1) Estimation accuracy. Mean Squared Error (MSE) is examined. For $r = 1, \ldots, R$, let

$$\begin{aligned} \mathrm{MSE}_r &= \sum_{i=0}^{p} (\hat{a}_i^r - a_{i0})^2 + \sum_{i=0}^{p} (\hat{b}_i^r - b_{i0})^2 \\ &\quad + \sum_{i=1}^{q} (\hat{\tau}_i^r - \tau_{i0})^2 + (\hat{c}^r - c_0)^2. \end{aligned}$$

and MSE= $\sum_{r=1}^{R} \text{MSE}_r/R$. The standard deviation MSE_r over the R simulation replications is also measured. (2) The average number of correctly selected 0 coefficients of the threshold variable.

For the tuning parameter selection, we use one of the following three setups for tuning parameter selection.

Setup 1 Two folder CV.

λ is taken to be a set of values with exponentially increasing gaps, say, $\lambda = n^{db}, db = lb + (N-1)d$, with $lb > 0, d = \frac{ub-lb}{N-1}, ub < 0.5$, where the integer N is the number of choices of λ, and lb and ub are chosen such that (λ, γ) satisfies

$$\frac{\lambda}{\sqrt{n}} \to 0, \ \frac{\lambda}{\sqrt{n}} \cdot n^{\gamma/2} \to \infty.$$

as $n \to \infty$.

Setup 2 Five folder CV and $\lambda = 0.5i, i = 1, 2, \ldots, 20$.

Setup 3 BIC and $\lambda = 0.5i, i = 1, 2, \ldots, 20$.

In the Example 2.1 and 2.2, a total of 50 simulation replications are conducted for each model. For every simulated data, we find the least squares (LS), adaptive lasso (AL) and the direction adaptive lasso (DAL) estimates. The calculation results are summarized in below tables. We can see from Tables 2.1, 2.2, 2.3 and

2.4 that the DAL method can indeed improve the estimation efficiency and is more powerful in eliminating the unimportant variables.

Example 2.1. *In the simulation, we consider the following STAR model.*

Model 1 *:* $p = 2, q = 2$, *the true threshold variable set is* $\{y_{t-2}\} \subseteq \{y_{t-1}, y_{t-2}\}$.

The model is

$$y_t = (8 - 0.4y_{t-1} + 0.5y_{t-2}) + (-10 + 0.3y_{t-1} - 0.4y_{t-2})\Phi(-5 + 6y_{t-2}) + \varepsilon_t,$$

where ε_t *is simulated from* $N(0, 1)$.

Table 2.1 Estimation results for Example 2.1 under Setup 1

n	Method	MSE	S.d.	Avg. no. of 0 coeff.
50	LS	5.0885	230.7513	0
	AL	1.3200	46.6351	0.56
	DAL	0.6407	27.2812	0.76
100	LS	1.1944	58.5926	0
	AL	0.1322	1.3404	0.58
	DAL	0.2261	5.9606	0.66
200	LS	0.0446	0.4138	0
	AL	0.0353	0.2849	0.74
	DAL	0.0401	0.3846	0.84
500	LS	0.0113	0.0962	0
	AL	0.0108	0.0946	0.76
	DAL	0.0111	0.0964	0.78

Example 2.2. *In this example, we let the order* $q = 4$ *which is bigger than the largest lag of the true threshold variables.*

Model 2: $p = 2, q = 4$, *true threshold variable set is* $\{y_{t-1}, y_{t-3}\} \subseteq \{y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}\}$.

Table 2.2 Estimation results for Example 2.1 under Setup 2

n	Method	MSE	S.d.	Avg. no. of 0 coeff.
50	LS	4.2117	224.9379	0
	AL	1.1895	35.3070	0.58
	DAL	0.4380	8.1961	0.72
100	LS	45.8695	2908.6	0
	AL	0.2163	6.7845	0.62
	DAL	0.1372	2.5903	0.70
200	LS	0.0499	0.8037	0
	AL	0.0398	0.3985	0.60
	DAL	0.0427	0.5044	0.64

The model is

$$y_t = (2 + 0.5y_{t-1} - 0.4y_{t-2})$$

$$+(-1.5 - 0.4y_{t-1} + 0.2y_{t-2})\Phi(-10 + 5y_{t-1} + 3y_{t-3}) + \varepsilon_t,$$

Table 2.3 Estimation results for Example 2.2 under Setup 1

n	Method	Estimation accuracy MSE	S.d.	Model complexity Avg. 0 no.	$\hat{\theta}_2 = 0$ and $\hat{\theta}_4 \neq 0$	$\hat{\theta}_4 = 0$ and $\hat{\theta}_2 \neq 0$	$\hat{\theta}_2 = 0$ and $\hat{\theta}_4 = 0$
50	LS	0.1136	1.0532	0	-	-	-
	AL	0.5348	14.2598	0.88	0.30	0.14	0.22
	DAL	0.1828	4.9394	1.44	0.14	0.14	0.58
100	LS	0.0677	0.7656	0.02	0.02	0	0
	AL	0.2207	5.3545	0.92	0.28	0.16	0.24
	DAL	0.0710	0.9065	1.3	0.08	0.10	0.56
200	LS	0.0274	0.2856	0	-	-	-
	AL	0.0882	1.6219	1.32	0.26	0.10	0.48
	DAL	0.0302	0.3619	1.68	0.10	0.06	0.76
500	LS	0.0098	0.0795	0.02	0	0.02	0
	AL	0.0124	0.1393	1.50	0.14	0.08	0.64
	DAL	0.0103	0.1007	1.82	0.04	0.10	0.84

Example 2.3 (The Canadian Lynx Data)**.** To further illustrate the performance

Table 2.4 Estimation results for Example 2.2 under Setup 3

n	Method	Estimation accuracy MSE	S.d.	Model complexity Avg. 0 no.	$\hat{\theta}_2 = 0$ and $\hat{\theta}_4 \neq 0$	$\hat{\theta}_4 = 0$ and $\hat{\theta}_2 \neq 0$	$\hat{\theta}_2 = 0$ and $\hat{\theta}_4 = 0$
50	LS	0.1531	2.6703	0.02	0	0.02	0
	AL	9.2426	596.30	1.28	0.22	0.18	0.44
	DAL	0.1932	3.2533	1.62	0.16	0.06	0.70
100	LS	0.0678	0.7654	0	-	-	-
	AL	0.0801	1.0342	1.28	0.12	0.24	0.46
	DAL	0.0683	0.8363	1.72	0.06	0.10	0.78
200	LS	0.0293	0.3022	0	-	-	-
	AL	0.0302	0.3418	1.52	0.16	0.12	0.62
	DAL	0.0299	0.3301	1.82	0.12	0.02	0.84

of the proposed method in selecting the threshold variable set, we examine one popular studied real data set. Following Tong (1990), we transform the data by taking base-10 logarithm to the original data, and denoted the transformed time series by y_t. Now assume that the time series follows the STAR(p,q) model. Applying different estimation methods to the data, we have the results listed in Table 2.5.

Both biological facts and previous statistical data analysis suggest that the significant threshold variable can be y_{t-2} or y_{t-3} or both. See, e.g., Tong (1990) section 7.2, and Fan and Yao (2003). Both the adaptive Lasso and the direction adaptive Lasso tend to lend support to the above suggestion.

Table 2.5 Estimation results for Example 2.3 under Setup 1

p	q	Method	threshold variable(s)	p	q	Method	threshold variable(s)
2	2	AL	y_{t-2}	3	2	AL	y_{t-2}
		DAL	y_{t-2}			DAL	y_{t-2}
	3	AL	$y_{t-1}, y_{t-2}, y_{t-3}$		3	AL	y_{t-2}
		DAL	y_{t-1}, y_{t-3}			DAL	y_{t-2}
	4	AL	y_{t-2}, y_{t-4}		4	AL	$y_{t-1}, y_{t-2}, y_{t-3}$
		DAL	y_{t-2}			DAL	y_{t-3}
	5	AL	y_{t-2}, y_{t-4}		5	AL	$y_{t-2}, y_{t-3}, y_{t-4}$
		DAL	y_{t-2}			DAL	y_{t-2}
4	2	AL	y_{t-2}	5	2	AL	y_{t-2}
		DAL	y_{t-2}			DAL	y_{t-2}
	3	AL	y_{t-3}		3	AL	y_{t-2}
		DAL	y_{t-3}			DAL	y_{t-2}
	4	AL	y_{t-3}		4	AL	y_{t-3}
		DAL	y_{t-3}			DAL	y_{t-3}
	5	AL	y_{t-3}		5	AL	y_{t-3}
		DAL	y_{t-3}			DAL	y_{t-3}

2.4 Proofs

Proof of Lemma 2.1: For $x = (x_1, \ldots, x_m)^\top$, $m = \max(p, q)$, denote $\Phi(x) = \Phi(\theta_0 + \sum_{j=1}^{q} \theta_j x_j)$ thus $0 \leq \Phi(x) \leq 1$ and we have

$$\begin{aligned}
|g(\eta, x)| &= |(a_0 + \sum_{j=1}^{p} a_j x_j) + (b_0 \Phi(x) + \sum_{j=1}^{p} b_j \Phi(x) x_j)| \\
&= |(a_0 + b_0 \Phi(x)) + \sum_{j=1}^{p} (a_j + b_j \Phi(x)) x_j| \\
&\leq |a_0 + b_0 \Phi(x)| + \Big| \sum_{j=1}^{p} (a_j + b_j \Phi(x)) x_j \Big| \\
&\leq \sum_{j=1}^{p} |a_j + b_j \Phi(x)||x_j| + C
\end{aligned}$$

$$\leq \sum_{j=1}^{p} \Big| a_j + b_j \Phi(x) \Big| \max\{|x_1|, \ldots, |x_p|\} + C$$

When

$$\sup_{0 \leq u \leq 1} \sum_{j=1}^{p} |a_j + b_j u| < 1,$$

the model is geometrically ergodic by the Theorem 3.2 of An and Huang (1996). Hence, there exists a stationary distribution F such that the time series y_t given by (2.1) and initiated at $\tilde{l}_0 = (y_{-1}, \ldots, y_{-m+1})^\top \sim F$ is strictly stationary. □

Proof of Theorem 2.1:

The proof that U is positive definite is the same as the proof given by Chan and Tong (1986) in its Appendix II, we thus omit it here.

To show the consistency and asymptotic normality, we follow from the standard method proposed in Klimko and Nelson (1978).

First, note that η_T^{LS} is actually obtained by solving the equations

$$\frac{\partial Q_T(\eta)}{\partial \eta_j} = 0, \;\; j = 1, 2, \ldots, L, \tag{2.16}$$

and if we denote the difference $u_t(\eta)$ by

$$u_t(\eta) = y_t - g(\eta, \mathscr{F}_{t-1}),$$

then $\{u_t(\eta_0)\}$ is a sequence of martingale differences.

Now, we expand $T^{-1/2}\partial Q_T(\eta)/\partial\eta$ in a Taylor series at η_0 and suppose that η_T^{LS} satisfies (2.16), we have

$$\begin{aligned} 0 &= T^{-\frac{1}{2}}\frac{\partial Q_T(\eta_T^{LS})}{\partial\eta} \\ &= T^{-\frac{1}{2}}\frac{\partial Q_T(\eta_0)}{\partial\eta} + T^{-1}(U_T + D_T(\eta^*))\cdot T^{\frac{1}{2}}(\eta_T^{LS} - \eta_0), \end{aligned} \tag{2.17}$$

where

$$U_T \equiv \frac{\partial^2 Q_T(\eta_0)}{\partial\eta\partial\eta^\top},$$

$$\begin{aligned} D_T(\eta^*) &\equiv \frac{\partial^2 Q_T(\eta^*)}{\partial\eta\partial\eta^\top} - U_T \\ &= \frac{\partial^2 Q_T(\eta^*)}{\partial\eta\partial\eta^\top} - \frac{\partial^2 Q_T(\eta_0)}{\partial\eta\partial\eta^\top}, \end{aligned} \tag{2.18}$$

and η^* being an appropriate intermediate point between η_0 and η_T^{LS}.

We claim that

$$(2T)^{-1}U_T \to U, \text{ a.s. .} \tag{2.19}$$

In fact, denote $(U_T)_{ij}$ as the (i,j)-th element of the matrix U_T, we have

$$\begin{aligned}\frac{1}{2}(U_T)_{ij} &= \left(\sum_{t=m+1}^{T}\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_i}\cdot\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_j}\right)\\ &\quad-\left(\sum_{t=m+1}^{T}\frac{\partial^2 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j}u_t(\eta_0)\right).\end{aligned}$$

By the strong law of large numbers for martingales, we get

$$\frac{1}{T}\sum_{t=m+1}^{T}\frac{\partial^2 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j}u_t(\eta_0)\to 0,\ \ a.s., \tag{2.20}$$

and by the ergodic theorem we have

$$\frac{1}{T}\sum_{t=m+1}^{T}\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_i}\cdot\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_j}\to U_{ij}\ \ a.s.,$$

thus

$$\frac{1}{2T}(U_T)_{ij}\to U_{ij},\ \ \text{a.s.}$$

Similar to (2.20), we have

$$\frac{1}{T}\frac{\partial Q_T(\eta_0)}{\partial\eta}=-\frac{2}{T}\sum_{t=m+1}^{T}\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta}u_t(\eta_0)\to 0,\ \ \text{a.s.}.$$

Next, we show that for any $\delta > 0$ such that $\|\eta^* - \eta_0\| \leq \delta$,

$$\lim_{T\to\infty} \sup_{\delta\to 0} \frac{|D_T(\eta^*)_{ij}|}{T\delta} < \infty,\ 1 \leq i, j \leq L,\ \text{a.s. .} \tag{2.21}$$

In fact,

$$\begin{aligned} |D_T(\eta^*)_{ij}| &= \Big|\frac{\partial^2 Q_T(\eta^*)}{\partial\eta_i\partial\eta_j} - \frac{\partial^2 Q_T(\eta_0)}{\partial\eta_i\partial\eta_j}\Big| \\ \leq\ & |\sum_{t=m+1}^{T} \{\frac{\partial g(\tilde{l}_t,\eta^*)}{\partial\eta_i}\cdot\frac{\partial g(\tilde{l}_t,\eta^*)}{\partial\eta_j} - \frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_i}\cdot\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_j}\}| \\ & + \Big|\sum_{t=m+1}^{T} \Big\{\frac{\partial^2 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j}u_t(\eta_0) - \frac{\partial^2 g(\tilde{l}_t,\eta^*)}{\partial\eta_i\partial\eta_j}u_t(\eta^*)\Big\}\Big|. \end{aligned}$$

And from the Taylor expansion,

$$u_t(\eta^*) = u_t(\eta_0) + \frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta^\top}(\eta_0 - \eta^*)(1 + o_p(1)),$$

$$\frac{\partial g(\tilde{l}_t,\eta^*)}{\partial\eta_i} = \frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_i} + \frac{\partial g^2(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta^\top}(\eta^* - \eta_0)(1 + o_p(1)),$$

and

$$\frac{\partial^2 g(\tilde{l}_t,\eta^*)}{\partial\eta_i\partial\eta_j} = \frac{\partial^2 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j} + \frac{\partial^3 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j\partial\eta^\top}(\eta^* - \eta_0)(1 + o_p(1)).$$

Note that

$$\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta} = \big(x_t, x_t\Phi_t, (x_t^\top b)\varphi_t s_t\big)^\top,$$

where $\Phi_t \equiv \Phi(s_t^\top\theta), \varphi_t \equiv \varphi(s_t^\top\theta)$ with $\varphi(\cdot)$ being the standard normal pdf are

both continuous for all $\eta \in \Theta$. Since $\{y_t\}$ is a stationary ergodic sequence of integrable variables, $u_t(\eta_0)$ is a sequence of martingale differences, by the martingale convergence theorem, it is easy to see that (2.21) is satisfied.

The conditions of the Theorem 2.1 of Klimko and Nelson (1978) are satisfied. We thus get the strong consistency (2.7) from (2.19), (2.20) and (2.21).

Next, we prove the asymptotic normality: $T^{1/2}(\eta_T^{LS} - \eta_0) \Rightarrow N(0, \sigma^2 U^{-1})$.

In view of (2.17), (2.19) and the proved consistency result, we only need to show that

$$\frac{1}{2} T^{-\frac{1}{2}} \frac{\partial Q_T(\eta_0)}{\partial \eta} \Rightarrow N(0, \sigma^2 U). \tag{2.22}$$

In fact, using the Cramer-Wold method, to show (2.22), it suffices to prove that $\forall\, h = (h_1, \ldots, h_L)^\top \in \mathbb{R}^L$,

$$\frac{1}{2} T^{-\frac{1}{2}} h^\top \frac{\partial Q_T(\eta_0)}{\partial \eta} \Rightarrow N(0, v), \tag{2.23}$$

where

$$v = \sigma^2 E_{\eta_0} \left(\sum_{k=1}^{L} h_k \frac{\partial g(\tilde{l}_t, \eta_0)}{\partial \eta_k} \right)^2 .$$

Note that $\partial Q_T(\eta_0)/\partial\eta = -2\sum_{t=m+1}^{T} u_t(\eta_0)\partial g(\tilde{l}_t,\eta_0)/\partial\eta$, let

$$f_1(\tilde{l}_t, h, \eta) \equiv -\sum_{k=1}^{L} h_k \frac{\partial g(\tilde{l}_t,\eta)}{\partial \eta_k},$$

it follows that

$$\frac{1}{2}T^{-\frac{1}{2}}h^\top \frac{\partial Q_T(\eta_0)}{\partial\eta} = T^{-\frac{1}{2}}\sum_{t=m+1}^{T} f_1(\tilde{l}_t, h, \eta_0)u_t(\eta_0). \tag{2.24}$$

Define

$$Y_t = \frac{f_1(\tilde{l}_t, h, \eta_0)u_t(\eta_0)}{\sigma\sqrt{E_{\eta_0}(f_1^2(\tilde{l}_t, h, \eta_0))}} = \frac{f_1(\tilde{l}_t, h, \eta_0)u_t(\eta_0)}{\sqrt{v}}$$

$$V_T^2 = \sum_{t=m+1}^{T} E(Y_t^2|\mathscr{F}_{t-1}) \ , \sigma_T^2 = EV_T^2,$$

we claim that

(1) $V_T^2/\sigma_T^2 \to 1$ in probability. This is shown by

$$V_T^2 = \sum_{t=m+1}^{T} E(Y_t^2|\mathscr{F}_{t-1}) = \big(\sum_{t=m+1}^{T} f_1^2\big)/Ef_1^2, \ \ \sigma_T^2 = T-m$$

and the ergodic theorem.

(2) Lindeberg condition: for any $\epsilon > 0$,

$$\frac{1}{\sigma_T^2} \sum_{t=m+1}^{T} E(Y_t^2 I(|Y_t| \geq \epsilon \sigma_T)) \to 0.$$

is satisfied. This is shown by noting that

$$\begin{aligned} Y_{T,t} &\equiv \frac{Y_t}{\sigma_T} = \frac{Y_t}{\sqrt{T-m}} \\ &= \frac{f_1(\tilde{l}_t, h, \eta_0) u_t(\eta_0)}{\sqrt{T-m}\sigma \sqrt{E(f_1^2(\tilde{l}_t, h, \eta_0))}} \leq \frac{C}{\sqrt{T-m}} \to 0 \end{aligned}$$

as $T \to \infty$ where $C > 0$ is some finite constant.

By the martingale CLT, we have

$$\sum_{t=m+1}^{T} Y_t / \sqrt{T} \Rightarrow N(0,1) \tag{2.25}$$

and (2.22) is proved.

We therefore complete the proof of consistency and asymptotic normality of η_T^{LS}. □

Remark 2.3. The result (2.19) can be written as

$$(2T)^{-1} \left(\frac{\partial^2 Q_T(\eta_0)}{\partial \eta \partial \eta^\top} \right) \to \sigma^2 I(\eta_0), \text{ a.s.} \tag{2.26}$$

and the result (2.22) can be written as

$$\frac{1}{2}T^{-\frac{1}{2}}\frac{\partial Q_T(\eta_0)}{\partial \eta} \Rightarrow N(0, \sigma^4 I(\eta_0)). \tag{2.27}$$

Proof of Theorem 2.2:

The proof is an application of the same method used to show the oracle properties of the adaptive lasso estimator in Zou (2006) to our case.

Step 1. We first show the asymptotic normality.

Let $\eta = \eta_0 + u/\sqrt{T}$, $u = (u_1, \ldots, u_L)^\top$, $L = 2p + 3 + q$, and

$$\Psi_T(u) = Q_T(\eta_0 + u/\sqrt{T}) + \lambda_T \sum_{j=1}^{q} \hat{w}_j \Big| \theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}} \Big|.$$

Suppose $\hat{u}_T = \arg\min_u \Psi_T(u)$, then

$$\eta_T^{ADL} = \eta_0 + \hat{u}_T/\sqrt{T} \text{ or } \hat{u}_T = \sqrt{T}(\eta_T^{ADL} - \eta_0)$$

since

$$\eta_T^{ADL} = \arg\min Q_T(\eta) + \lambda_T \sum_{j=1}^{q} \hat{w}_j |\theta_j|.$$

Denote $V_T(u) \equiv \Psi_T(u) - \Psi_T(\mathbf{0})$, we have

$$\begin{aligned} V_T(u) &= \{Q_T(\eta_0 + u/\sqrt{T}) - Q_T(\eta_0)\} \\ &\quad + \Big\{\lambda_T \sum_{j=1}^{q} \hat{w}_j \big(|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}| - |\theta_{j0}|\big)\Big\} \\ &\equiv H_T(u) + P_T(u), \end{aligned} \tag{2.28}$$

where the loss function term

$$H_T(u) = Q_T(\eta_0 + u/\sqrt{T}) - Q_T(\eta_0)$$

and the penalty term

$$P_T(u) = \lambda_T \sum_{j=1}^{q} \hat{w}_j \big(|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}| - |\theta_{j0}|\big).$$

Note that

$$\begin{aligned} & Q_T(\eta_0 + u/\sqrt{T}) - Q_T(\eta_0) \\ = \; & \frac{1}{\sqrt{T}} u^\top \frac{\partial Q_T(\eta_0)}{\partial \eta} + \frac{1}{2T} u^\top \frac{\partial^2 Q_T(\eta_0)}{\partial \eta \partial \eta^\top} u(1 + o_p(1)). \end{aligned}$$

From the results (2.26) and (2.27), we know that as $T \to \infty$,

$$\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_0)}{\partial \eta} \Rightarrow W \sim N(\mathbf{0}, 4\sigma^4 I(\eta_0))$$

and

$$\frac{1}{2T}\frac{\partial^2 Q_T(\eta_0)}{\partial \eta \partial \eta^\top} \to \sigma^2 I(\eta_0) \text{ a.s..}$$

Thus the loss function term

$$H_T(u) \Rightarrow u^\top W + \sigma^2 u^\top I(\eta_0) u.$$

Now we consider the limiting behavior of the penalty term.

If $j \in K$, i.e., $\theta_{j0} \neq 0$, from the result of the Theorem 2.1,

$$\hat{w}_j = 1/|\theta_j^{LS}|^\gamma \to |\theta_{j0}|^{-\gamma}, \text{ a.s.}$$

and

$$\sqrt{T}\big(|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}| - |\theta_{j0}|\big) \to u_{2p+3+j}\text{sgn}(\theta_{j0}).$$

Since $\lambda_T/\sqrt{T} \to 0$, we have

$$\frac{\lambda_T}{\sqrt{T}}\hat{w}_j\sqrt{T}\big(|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}| - |\theta_{j0}|\big) \to 0.$$

If $j \in \bar{K}$, i.e., $\theta_{j0} = 0$, then $\sqrt{T}\big(|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}| - |\theta_{j0}|\big) = |u_{2p+3+j}|$. Since $\sqrt{T}\theta_j^{LS} = O_p(1)$ and $\lambda_T T^{(\gamma-1)/2} \to \infty$, we have

$$\frac{\lambda_T}{\sqrt{T}}\hat{w}_j = \lambda_T T^{\frac{\gamma-1}{2}}|\sqrt{T}\theta_j^{LS}|^{-\gamma} \to \infty.$$

Therefore, by Slutsky's theorem, we have $V_T(u) \Rightarrow V(u)$ for every u, where

$$V(u) = \begin{cases} (u_{(K)})^\top W_{(K)} + \sigma^2 (u_{(K)})^\top I(\eta_{0,(K)}) u_{(K)}, \\ \qquad\qquad \text{if } u_{2p+3+j} = 0, \forall j \in \bar{K} \\ \infty, \qquad \text{otherwise}, \end{cases}$$

where $u_{(K)}$ and $W_{(K)}$ are the j-th ($j \in \{2p+3+k : k \in \bar{K}\}$) elements deleted from u and W respectively. Note that the unique minimum of $V(u)$ is

$$u_{min} = \begin{pmatrix} -\frac{1}{2\sigma^2} I^{-1}(\eta_{0,(K)}) W_{0,(K)} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ denotes that the other corresponding components $u_{2p+3+j}, j \in \bar{K}$ are all 0

in the vector u. Following the epi-convergence property of Knight (1999), we have

$$\hat{u}_{T,(K)} \Rightarrow -\frac{1}{2\sigma^2} I^{-1}(\eta_{0,(K)}) W_{(K)}$$

and the other components $\to \mathbf{0}$, i.p..

Finally, recall that $W_{(K)} \sim N(\mathbf{0}, 4\sigma^4 I(\eta_{0,(K)}))$, we get

$$\sqrt{T}(\eta^{ADL}_{T,(K)} - \eta_{0,(K)}) \Rightarrow N\big(\mathbf{0}, I^{-1}(\eta_{0,(K)})\big). \tag{2.29}$$

Step 2. Now we prove the consistency.

If $j \in K$, then $\theta^{ADL}_j \to \theta_{j0}$ i.p., thus $P(j \in K^{ADL}_T) \to 1$. Thus we only need to show that $\forall j \in \bar{K},\ P(j \in K^{ADL}_T) \to 0$.

By the KKT optimality conditions,

$$\frac{1}{\sqrt{T}} \frac{\partial Q_T(\eta^{ADL}_T)}{\partial \theta_j} + \frac{\lambda_T}{\sqrt{T}} \hat{w}_j \mathrm{sgn}(\theta^{ADL}_j) = 0.$$

Note that

$$\big|\frac{\lambda_T}{\sqrt{T}} \hat{w}_j \mathrm{sgn}(\theta^{ADL}_j)\big| = \frac{\lambda_T}{\sqrt{T}} T^{\gamma/2} |\sqrt{T}\theta^{LS}_j|^{-\gamma} \to \infty, \text{ i.p.},$$

whereas

$$\begin{aligned} & \frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{ADL})}{\partial \theta_j} \\ = \quad & \frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_0)}{\partial \theta_j} + \frac{1}{T}\frac{\partial^2 Q_T(\eta_0)}{\partial \theta_j^2}\sqrt{T}(\theta_j^{ADL} - \theta_{j0})(1 + o_p(1)) \\ \Rightarrow \quad & \text{some normal distribution} \end{aligned}$$

by (2.29) and Slutsky's theorem. Thus, for $j \in \bar{K}$,

$$P(j \in K_T^{ADL}) \leq P\Big(|\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{ADL})}{\partial \theta_j}| = \frac{\lambda_T}{\sqrt{T}}\hat{w}_j\Big) \to 0.$$

This completes the proof. □

Proof of Lemma 2.2:

Recall that $\vartheta = (\theta_1, \ldots, \theta_q)^\top$, denote

$$g(\vartheta) = (\vartheta^\top \vartheta)^{-1/2} = \frac{1}{\sqrt{\theta_1^2 + \ldots + \theta_q^2}}, \tag{2.30}$$

then

$$\tilde{\vartheta} = \frac{\vartheta}{l(\theta)} = \frac{\vartheta}{(\vartheta^\top \vartheta)^{1/2}} \equiv \vartheta g(\vartheta).$$

From the asymptotic result of ϑ_T^{LS}, we have

$$g(\vartheta_T^{LS}) \to g(\vartheta_0).$$

Thus,

$$\tilde{\vartheta}_T^{LS} = \vartheta_T^{LS} g(\vartheta_T^{LS}) \to \tilde{\vartheta}_0 = \vartheta_0 g(\vartheta_0) \text{ a.s..}$$

Next we will show the asymptotic normality. From (2.11), we know that

$$\sqrt{T}(\theta_T^{LS} - \theta_0) \Rightarrow N(0, I^{-1}(\theta_0)),$$

where $I^{-1}(\vartheta_0)$ is submatrix composed of the last q rows and the last q columns of the inverse matrix of $I(\eta_0)$ defined in (2.10). Thus,

$$\begin{aligned}
& \sqrt{T}(\tilde{\vartheta}_T^{LS} - \tilde{\vartheta}_0) = \sqrt{T}(\vartheta_T^{LS} g(\vartheta_T^{LS}) - \vartheta_0 g(\vartheta_0)) \\
= \;& \sqrt{T}\big(\vartheta_T^{LS} g(\vartheta_T^{LS}) - \vartheta_0 g(\vartheta_T^{LS}) + \vartheta_0 g(\vartheta_T^{LS}) - \vartheta_0 g(\vartheta_0)\big) \\
= \;& \sqrt{T}(\vartheta_T^{LS} - \vartheta_0) g(\vartheta_T^{LS}) + \vartheta_0 \sqrt{T}(g(\vartheta_T^{LS}) - g(\vartheta_0)) \\
\Rightarrow \;& \text{some normal distribution}
\end{aligned}$$

by the Slutsky theorem and the continuous mapping theorem.

It is easy to see that the mean of the asymptotic normal distribution is $\mathbf{0}$. We

now provide the asymptotic covariance matrix $\tilde{\Sigma}$ and show that its rank is $q-1$.

Note that $\partial g(\vartheta)/\partial\vartheta = -(\vartheta^\top\vartheta)^{-3/2}\vartheta$ and $\vartheta_T^{LS} - \vartheta_0 = O_p(T^{-1/2})$, we have

$$\begin{aligned}
& \vartheta_0\sqrt{T}(g(\vartheta_T^{LS}) - g(\vartheta_0)) \\
= \;& \vartheta_0\sqrt{T}\frac{\partial g(\vartheta_0)}{\partial\vartheta^\top}(\vartheta_T^{LS} - \vartheta_0) + O_p(T^{-1/2}) \\
= \;& -\vartheta_0\vartheta_0^\top\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)(\vartheta_0^\top\vartheta_0)^{-3/2} + O_p(T^{-1/2}).
\end{aligned}$$

Denote $Z_{T,1} = \sqrt{T}(\vartheta_T^{LS} - \vartheta_0)g(\vartheta_T^{LS})$ and $Z_{T,2} = -\vartheta_0\vartheta_0^\top\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)(\vartheta_0^\top\vartheta_0)^{-3/2}$, we next calculate the covariance matrix of $Z_{T,1} + Z_{T,2}$.

$$\begin{aligned}
& \mathrm{Var}(Z_{T,1} + Z_{T,2}) = \mathrm{E}(Z_{T,1} + Z_{T,2})(Z_{T,1} + Z_{T,2})^\top \\
= \;& \mathrm{E}\Big(\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)^\top g^2(\vartheta_T^{LS})\Big) \\
& -\mathrm{E}\Big(\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)^\top\vartheta_0\vartheta_0^\top(\vartheta_0^\top\vartheta_0)^{-3/2}g(\vartheta_T^{LS})\Big) \\
& -\mathrm{E}\Big(\vartheta_0\vartheta_0^\top\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)^\top(\vartheta_0^\top\vartheta_0)^{-3/2}g(\vartheta_T^{LS})\Big) \\
& +\mathrm{E}\Big(\vartheta_0\vartheta_0^\top\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)^\top\vartheta_0\vartheta_0^\top(\vartheta_0^\top\vartheta_0)^{-3}\Big).
\end{aligned}$$

Since as $T\to\infty$, $\sqrt{T}(\vartheta_T^{LS} - \vartheta_0) \Rightarrow N(0, I^{-1}(\vartheta_0))$ and $g(\vartheta_T^{LS}) \to g(\vartheta_0)$, a.s., we thus get the limiting covariance matrix

$$\tilde{\Sigma} \;=\; I^{-1}(\vartheta_0)(\vartheta_0^\top\vartheta_0)^{-1} - I^{-1}(\vartheta_0)\vartheta_0\vartheta_0^\top(\vartheta_0^\top\vartheta_0)^{-2}$$

$$-\vartheta_0\vartheta_0^\top I^{-1}(\vartheta_0)(\vartheta_0^\top\vartheta_0)^{-2} + \vartheta_0\vartheta_0^\top I^{-1}(\vartheta_0)\vartheta_0\vartheta_0^\top(\vartheta_0^\top\vartheta_0)^{-3}.$$

Recall that $\tilde{\vartheta}_0 = \vartheta_0(\vartheta_0^\top\vartheta_0)^{-1/2}$, we have

$$\begin{aligned}
\tilde{\Sigma} &= \{I^{-1}(\vartheta_0)(\vartheta_0^\top\vartheta_0)^{-1} - I^{-1}(\vartheta_0)\tilde{\vartheta}_0\tilde{\vartheta}_0^\top(\vartheta_0^\top\vartheta_0)^{-1}\} \\
&\quad -\{\tilde{\vartheta}_0\tilde{\vartheta}_0^\top I^{-1}(\vartheta_0)(\vartheta_0^\top\vartheta_0)^{-1} - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top I^{-1}(\vartheta_0)\tilde{\vartheta}_0\tilde{\vartheta}_0^\top(\vartheta_0^\top\vartheta_0)^{-1}\} \\
&= (\vartheta_0^\top\vartheta_0)^{-1}I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top) \\
&\quad -(\vartheta_0^\top\vartheta_0)^{-1}\tilde{\vartheta}_0\tilde{\vartheta}_0^\top I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top) \\
&= (\vartheta_0^\top\vartheta_0)^{-1}(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top).
\end{aligned}$$

Notice that the $q \times q$ matrix $I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top$ is an idempotent matrix due to the relationship $\tilde{\vartheta}_0^\top\tilde{\vartheta}_0 = 1$. That is, $(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)^2 = I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top$ given $\tilde{\vartheta}_0^\top\tilde{\vartheta}_0 = 1$. We thus have

$$\text{rank}(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top) = q - 1.$$

Denote $A = I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top = A^\top, B = I^{-\frac{1}{2}}(\vartheta_0)$ and $C = AB$ then

$$\tilde{\Sigma} = (I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top) = CC^\top.$$

From the Sylvester's inequality, we get

$$\begin{aligned}
\text{rank}(\tilde{\Sigma}) &= \text{rank}(CC^\top) = \text{rank}(C) \\
&= \text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\} = q - 1 \\
\text{rank}(\tilde{\Sigma}) &= \text{rank}(AB) \geq \text{rank}(A) + \text{rank}(B) - q = q - 1.
\end{aligned}$$

Therefore, we show that the rank of the matrix $\tilde{\Sigma}$ is $q - 1$. □

Proof of Theorem 2.3:

The proof is very similar to that of the Theorem 2.2 and the only difference concerns the treatment of the penalty term.

Let $\eta = \eta_0 + u/\sqrt{T}$, $u = (u_1, \ldots, u_L)^\top$, $L = 2p + 3 + q$, and

$$\begin{aligned}
&\Psi_T(u) = Q_T(\eta_0 + u/\sqrt{T}) \\
&+\lambda_T \sum_{j=1}^{q} \tilde{w}_j \Big| (\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}) g(\vartheta_0 + \frac{u_{2p+4:2p+3+q}}{\sqrt{T}}) \Big|,
\end{aligned}$$

where $g(\vartheta)$ is defined in (2.30) and the q-dimensional sub-vector $u_{2p+4:2p+3+q}$ is composed of the components $u_{2p+4}, u_{2p+5}, \ldots, u_{2p+3+q}$ of the vector u. We denote $u_{2p+4:2p+3+q}$ as $\tilde{u}$.

It follows that the penalty term

$$P_T(u) = \lambda_T \sum_{j=1}^{q} \tilde{w}_j \Big(\Big| \theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}} \Big| g\Big(\vartheta_0 + \frac{\tilde{u}}{\sqrt{T}}\Big) - |\theta_{j0}| g(\vartheta_0) \Big).$$

Since $g'(\vartheta) = -(\vartheta^\top \vartheta)^{-3/2} = -(g(\vartheta))^3$, from the Taylor expansion of g, we have

$$g\Big(\vartheta_0 + \frac{\tilde{u}}{\sqrt{T}}\Big) = g(\vartheta_0) - (g(\vartheta_0))^3 \frac{\tilde{u}^\top \vartheta_0}{\sqrt{T}} (1 + o_p(1)).$$

If $j \in K$, i.e., $\tilde{\theta}_{j0} \neq 0$, from the result of the Lemma 2.2,

$$\tilde{w}_j = 1/|\tilde{\theta}_j^{LS}|^\gamma \to |\tilde{\theta}_{j0}|^{-\gamma}, \quad \text{a.s.}$$

and

$$\begin{aligned}
& \sqrt{T} \Big(\Big| \theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}} \Big| g\Big(\vartheta_0 + \frac{\tilde{u}}{\sqrt{T}}\Big) - |\theta_{j0}| g(\vartheta_0) \Big) \\
= \; & \sqrt{T} \Big(\Big| \theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}} \Big| - |\theta_{j0}| \Big) g(\vartheta_0) \\
& - \Big| \theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}} \Big| (g(\vartheta_0))^3 \tilde{u}^\top \vartheta_0 (1 + o_p(1)) \\
\to \; & u_{2p+3+j} \mathrm{sgn}(\theta_{j0}) g(\vartheta_0) - |\theta_{j0}| (g(\vartheta_0))^3 \tilde{u}^\top \vartheta_0 \\
= \; & g(\vartheta_0) \Big(u_{2p+3+j} \mathrm{sgn}(\tilde{\theta}_{j0}) - |\tilde{\theta}_{j0}| \tilde{u}^\top \tilde{\vartheta}_0 \Big).
\end{aligned}$$

Since $\lambda_T/\sqrt{T} \to 0$, we have

$$\frac{\lambda_T}{\sqrt{T}}\tilde{w}_j\sqrt{T}\Big(\Big|\theta_{j0}+\frac{u_{2p+3+j}}{\sqrt{T}}\Big|g\Big(\vartheta_0+\frac{\tilde{u}}{\sqrt{T}}\Big)-|\theta_{j0}|g(\vartheta_0)\Big)\to 0.$$

If $j \in \bar{K}$, i.e., $\tilde{\theta}_{j0}=0$, then

$$\begin{aligned}&\sqrt{T}\Big(\Big|\theta_{j0}+\frac{u_{2p+3+j}}{\sqrt{T}}\Big|g\Big(\vartheta_0+\frac{\tilde{u}}{\sqrt{T}}\Big)-|\theta_{j0}|g(\vartheta_0)\Big)\\ =\;& |u_{2p+3+j}|g\Big(\vartheta_0+\frac{\tilde{u}}{\sqrt{T}}\Big)\to |u_{2p+3+j}|g(\vartheta_0).\end{aligned}$$

When $\tilde{\theta}_{j0}=0$, we have $\sqrt{T}\tilde{\theta}_j^{LS}=\sqrt{T}(\tilde{\theta}_j^{LS}-\tilde{\theta}_{j0})=O_p(1)$ from the asymptotical normality result of Lemma 2.2. It follows that

$$\frac{\lambda_T}{\sqrt{T}}\tilde{w}_j=\lambda_T T^{\frac{\gamma-1}{2}}|\sqrt{T}\tilde{\theta}_j^{LS}|^{-\gamma}\to\infty$$

since $\lambda_T T^{(\gamma-1)/2}\to\infty$.

Therefore, using the same notations as in the proof of Theorem 2.2 and by Slutsky's theorem, we have $V_T(u)\Rightarrow V(u)$ for every u, where

$$V(u)=\begin{cases}(u_{(K)})^\top W_{(K)}+\sigma^2(u_{(K)})^\top I(\eta_{0,(K)})u_{(K)},\\ \qquad\qquad\qquad \text{if } u_{2p+3+j}=0,\forall j\in\bar{K}\\ \infty, \qquad \text{otherwise},\end{cases}$$

and get the same asymptotic normality result.

As for the variable selection consistency, we only need to show that

$$\forall j \in \bar{K},\ P(j \in K_T^{DAL}) \to 0.$$

Recall that the objective function of the direction adaptive lasso estimator is

$$Q_T(\eta) + \lambda_T \sum_{i=1}^{q} \tilde{w}_i |\tilde{\theta}_i| = Q_T(\eta) + \lambda_T g(\vartheta) \sum_{i=1}^{q} \tilde{w}_i |\theta_i|.$$

For $j \in \bar{K}$, consider the event $j \in K_T^{DAL}$. By the KKT optimality conditions, we have

$$\begin{aligned}
0 &= \frac{1}{\sqrt{T}} \frac{\partial Q_T(\eta_T^{DAL})}{\partial \theta_j} - \frac{\lambda_T}{\sqrt{T}} (g(\vartheta_T^{DAL}))^3 \theta_j^{DAL} \sum_{i=1}^{q} \tilde{w}_i |\theta_i^{DAL}| \\
&\quad + \frac{\lambda_T}{\sqrt{T}} \tilde{w}_j g(\vartheta_T^{DAL}) \mathrm{sgn}(\tilde{\theta}_j^{DAL}) \\
&= \frac{1}{\sqrt{T}} \frac{\partial Q_T(\eta_T^{DAL})}{\partial \theta_j} - \frac{\lambda_T}{\sqrt{T}} g(\vartheta_T^{DAL}) \tilde{\theta}_j^{DAL} \sum_{i=1}^{q} \tilde{w}_i |\tilde{\theta}_i^{DAL}| \\
&\qquad + \frac{\lambda_T}{\sqrt{T}} \tilde{w}_j g(\vartheta_T^{DAL}) \mathrm{sgn}(\tilde{\theta}_j^{DAL}) \\
&= \Big\{ \frac{1}{\sqrt{T}} \frac{\partial Q_T(\eta_T^{DAL})}{\partial \theta_j} \\
&\quad - \frac{\lambda_T}{\sqrt{T}} g(\vartheta_T^{DAL}) \tilde{\theta}_j^{DAL} \sum_{i \in K} \tilde{w}_i |\tilde{\theta}_i^{DAL}| \Big\} \\
&\quad + \Big\{ \frac{\lambda_T}{\sqrt{T}} \tilde{w}_j g(\vartheta_T^{DAL}) \mathrm{sgn}(\tilde{\theta}_j^{DAL})
\end{aligned}$$

$$-\frac{\lambda_T}{\sqrt{T}} g(\vartheta_T^{DAL}) \tilde{\theta}_j^{DAL} \sum_{i \in \bar{K}} \tilde{w}_i |\tilde{\theta}_i^{DAL}| \Big\} \tag{2.31}$$

$$\equiv \quad S_{T1} + S_{T2}$$

We first claim that the term

$$\begin{aligned} S_{T1} \quad &= \quad \frac{1}{\sqrt{T}} \frac{\partial Q_T(\eta_T^{DAL})}{\partial \theta_j} \\ &\qquad -\frac{\lambda_T}{\sqrt{T}} g(\vartheta_T^{DAL}) \tilde{\theta}_j^{DAL} \sum_{i \in K} \tilde{w}_i |\tilde{\theta}_i^{DAL}| \\ &\Rightarrow \quad \text{some normal distribution} \end{aligned} \tag{2.32}$$

In fact,

$$\frac{1}{\sqrt{T}} \frac{\partial Q_T(\eta_T^{DAL})}{\partial \theta_j} \Rightarrow \quad \text{some normal distribution}$$

and

$$\frac{\lambda_T}{\sqrt{T}} g(\vartheta_T^{DAL}) \tilde{\theta}_j^{DAL} \sum_{i \in K} \tilde{w}_i |\tilde{\theta}_i^{DAL}| \to 0$$

as for $i \in K$, $\tilde{w}_i \to |\theta_{i0}|^{-\gamma}, \tilde{\theta}_j^{DAL} \to_p 0, \tilde{\theta}_i^{DAL} \to_p \tilde{\theta}_{i0}$ and $\lambda_T/\sqrt{T} \to 0$. By Slutsky's theorem, we get (2.32).

We next show that $S_{T2} \to_p \infty$. Note that

$$\begin{aligned} S_{T2} &= \frac{\lambda_T}{\sqrt{T}}\tilde{w}_j g(\vartheta_T^{DAL})\mathrm{sgn}(\tilde{\theta}_j^{DAL}) \\ &\quad -\frac{\lambda_T}{\sqrt{T}} g(\vartheta_T^{DAL})\tilde{\theta}_j^{DAL} \sum_{i\in\bar{K}} \tilde{w}_i |\tilde{\theta}_i^{DAL}| \\ &= \lambda_T T^{\frac{\gamma-1}{2}} g(\vartheta_T^{DAL})\Big\{ \frac{1}{|\sqrt{T}\tilde{\theta}_j^{LS}|^\gamma}\mathrm{sgn}(\tilde{\theta}_j^{DAL}) \\ &\quad -\tilde{\theta}_j^{DAL} \sum_{i\in\bar{K}} \frac{1}{|\sqrt{T}\tilde{\theta}_i^{LS}|^\gamma}|\tilde{\theta}_i^{DAL}|\Big\} \\ &\to_p \infty \end{aligned}$$

since $\lambda_T T^{\frac{\gamma-1}{2}} \to \infty$ and $\forall j \in \bar{K}, \sqrt{T}\tilde{\theta}_j^{LS} = O_p(1)$.

Therefore, for $j \in \bar{K}$,

$$P(j \in K_T^{DAL}) \leq P(|S_{T1}| = |S_{T2}|) \to 0.$$

This completes the proof. □

CHAPTER 3

On a Principal Varying Coefficient Model (PVCM)

3.1 Introduction of PVCM

Let (Y, X, U) be a random triplet, where $Y \in \mathbb{R}^1$ is the response of interest, $X = (X_1, \cdots, X_p)^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor, and $U \in \mathbb{R}^1$ is the so-called *index* variable. The conventional varying coefficient model (Hastie and Tibshirani (1993)) assumes that $Y = X^\top \boldsymbol{\beta}(U) + \varepsilon$, where ε is the random noise and $\boldsymbol{\beta}(u) = (\beta_1(u), \ldots, \beta_p(u))^\top \in \mathbb{R}^p$ is a vector of unknown smooth functions in

u, called the varying coefficients. Ever since Hastie and Tibshirani (1993), VCM has gained a lot of popularity in the literature attributing to the following three facts. Firstly, VCM is easy to interpret. This is because, conditioned on the index variable $U = u$, VCM reduces to a standard linear regression model which has been well understood in practice. Secondly, VCM allows the varying coefficient $\boldsymbol{\beta}(u)$ to be fully nonparametric. Thus, it has much stronger modeling capability than a standard linear regression model. Lastly, because the index variable $U \in \mathbb{R}^1$ is typically a univariate variable, VCM is free of the *curse of dimensionality*. VCM and its variants have been extensively studied in the literature during the past two decades. See, for example, Fan and Zhang W. (1999), Cai *et al* (2000), Fan and Zhang W. (2000), Fan and Zhang J. (2000), Huang *et al* (2002), Zhang *et al* (2002), Fan and Huang (2005), and Fan and Zhang W. (2008).

It is remarkable that, although the estimation of VCM requires only one dimensional kernel smoothing, it is still very unstable. The model cannot be estimated well when the predictor's dimension p is large even moderately. There are two approaches to improve the estimation efficiency. The first approach is to employ a more efficient estimation method. It is generally believed that the polynomial splines especially the penalized polynomial splines are more efficient than the kernel smoothing approach. See Wood (2006) for a comprehensive review. Another way to improve the efficiency is through further model specification without losing

much information. The semi-varying coefficient model (SVCM) proposed by Zhang W. *et al* (2002) and Fan and Huang (2005) is a good example for this purpose. SVCM confines some coefficients to be constant but allows the others to vary with the index variable U.

In this Chapter, we consider an extension of the SVCM by allowing different varying coefficients to be linearly dependent. To further illustrate the idea, let us revisit the Boston house price data. The response of interest is the median value of owner-occupied homes (MEDV, in \$1000) with 13 covariates, denoted by $X_1, \ldots, X_{13}$ respectively. As noticed by Fan and Huang (2005), the following varying coefficient model with the lower status of the population ($U = LSTAT$) being the index variable is appropriate for the data,

$$MEDV = \beta_1(U)X_1 + \ldots + \beta_{13}(U)X_{13} + \varepsilon. \tag{3.1}$$

In the below figure 3.1, the first panel shows all the coefficients, of which with big variation are selected and labeled. The selected coefficients are redrawn in the second panel. After liner transformation and standardization, those selected coefficients are shown in panel 3.

In (3.1), the varying coefficients can be estimated by the method based on the local linear smoothing; see for example Wu and Liang (2004). The estimated

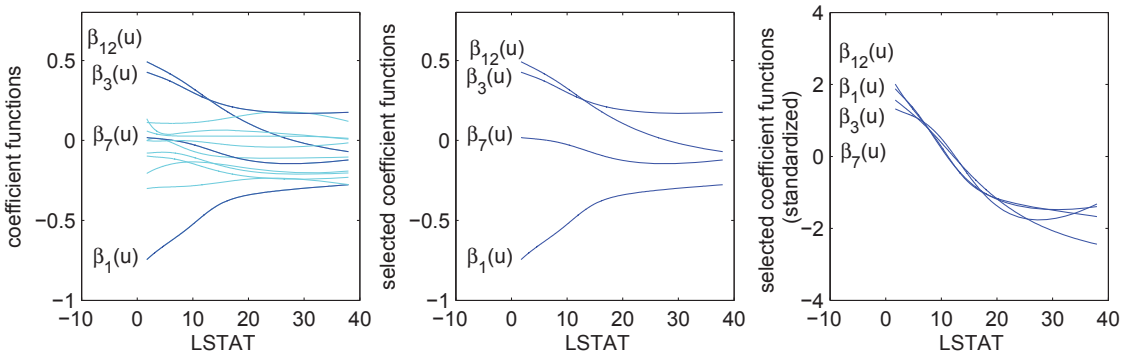

Figure 3.1 The estimated varying coefficients for the Boston House Price Data.

coefficients are shown in the first panel of Figure 3.1. For those coefficients with big variation as labeled and redrawn in the second panel, remarkably similar shapes are shared after some linear transformations as shown in the third panel, which implies that different varying coefficients are likely to be linearly dependent and that the index affects those coefficients in a similar way. To quantify such a linear dependency, we estimate $\Sigma_\beta = \text{cov}\{(\beta_1(U), \beta_2(U), \ldots, \beta_{13}(U))^\top\}$ and find that its eigenvalues are 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0004, 0.0006, 0.0126, 0.0370, 0.1445, 0.5668, 25.8584 respectively. The largest eigenvalue 25.8584 by itself can explain 97% of Σ_β's total variation, suggesting that the coefficients $\beta_\ell(u), \ell = 1, 2, \ldots, 13$ have a common principal function $\gamma(u)$, i.e.

$$\beta_k(U) = \theta_k + \phi_k \gamma(U) + \text{other terms with less contribution},$$

where both θ_k and ϕ_k $(1 \leq k \leq 13)$ are constant parameters. As a consequence,

model (3.1) can be further simplified as

$$\text{MEDV} = \Big(\theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_{13} X_{13}\Big) + \gamma(U)\Big(\phi_1 X_1 + \phi_2 X_2 + \ldots + \phi_{13} X_{13}\Big) + \varepsilon. \tag{3.2}$$

Theoretically, the estimators produced by (3.2) are more efficient than those by (3.1), because only one nonparametric function $\gamma(\cdot)$ needs to be estimated in (3.2), but a total of $p = 13$ functions need to be done in (3.1). Furthermore, model (3.2) identifies two important factors given by $\theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_{13} X_{13}$ and $\phi_1 X_1 + \phi_2 X_2 + \ldots + \phi_{13} X_{13}$ respectively. The first factor is linearly related to the response, and the second one nonlinearly in the sense that it has a nontrivial interaction with the index variable U. Thus, model (3.2) is also more informative as compared with model (3.1).

In this Chapter, we shall discuss a more general model of (3.2). For convenience, we refer to the new model as the Principal Varying Coefficient Model (PVCM). Compared with the conventional varying coefficient model, PVCM discovers the possible linear dependence structure amongst the varying coefficients. As one can see from (3.2), such a linear dependence structure can reduce the actual number of nonparametric functions, and thus further improves estimation efficiency. On the other hand, separating the coefficients into liner and nonlinear parts is more informative in data analysis. Moreover, PVCM is more flexible and allows a predictor to appear in both linear and nonlinear parts simultaneously.

3.2 Model Representation and Identification

Let (Y_i, X_i, U_i) be the observation collected from the ith subject, $i = 1, 2, \ldots, n$, where $Y_i \in \mathbb{R}^1$ is the response, $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ is the p-dimensional predictor, and $U_i \in \mathbb{R}^1$ is the univariate index variable. The conventional VCM model assumes

$$Y_i = \beta_1(U_i)X_{i1} + \beta_2(U_i)X_{i2} + \ldots + \beta_p(U_i)X_{ip} + \varepsilon_i,$$

where $\beta_k(\cdot), k = 1, \ldots, p$, are unknown coefficient functions and $E(\varepsilon_i|X_i, U_i) = 0$ almost surely. Let $\boldsymbol{\beta}_0(u) = (\beta_1(u), \beta_2(u), \ldots, \beta_p(u))^\top$. Motivated by the example above, we further assume the following principal component structure for the coefficient functions

$$\boldsymbol{\beta}_0(u) = \theta_0 + B_0\boldsymbol{\gamma}_0(u),$$

where $\theta_0 \in \mathbb{R}^p$ and $B_0 = (b_1, \ldots, b_{d_0}) \in \mathbb{R}^{p\times d_0}$, with $\text{rank}(B_0) = d_0 < p$, are parameters and $\boldsymbol{\gamma}_0(u) = (g_1(u), \ldots, g_{d_0}(u))^\top$ are unknown principal functions. As a consequence, we come up with the following Principal Varying Coefficient Model (PVCM)

$$Y_i = \theta_0^\top X_i + g_1(U_i)b_1^\top X_i + \ldots + g_{d_0}(U_i)b_{d_0}^\top X_i + \varepsilon_i. \tag{3.3}$$

For convenience, we refer to d_0 as the number of principal functions, $\theta_0^\top X_i$ the linear part, and $X_i^\top B_0\boldsymbol{\gamma}_0(U_i)$ the nonlinear part. We further assume that the principal functions $\boldsymbol{\gamma}_0(u) \in \mathbb{R}^{d_0}$ satisfy $\text{rank}\{\text{cov}(\boldsymbol{\gamma}_0(U_i))\} = d_0$. Otherwise, functional elements in $\boldsymbol{\gamma}_0(u)$ are linearly dependent, and the rank of B_0 can be

further reduced. Obviously, model (3.3) becomes a standard linear regression model if $d_0 = 0$ and a full VCM if $d_0 = p$. PVCM also includes the semi-varying coefficient model (SVCM) of Zhang W. *et al* (2002) as a special case if the last $p-q$ elements in θ_0 are zeros and the first q elements in all $b_k, k = 1, \ldots, d_0$, are zeros.

Model (3.3) is not uniquely identifiable. For example, let C be an arbitrary $d_0 \times d_0$ orthonormal matrix. Then, we can re-define $B_0 := B_0 C$ and $\boldsymbol{\gamma}_0(u) := C^\top \boldsymbol{\gamma}_0(u)$. Model (3.3) still holds with these newly defined B_0 and $\boldsymbol{\gamma}_0(u)$. Parameter vector θ_0 is also not unique even if B_0 is fixed. For example, let $c \in \mathbb{R}^{d_0}$ be an arbitrary constant vector. We can re-define $\theta_0 := \theta_0 - B_0 c$ and $\boldsymbol{\gamma}_0(\cdot) := \boldsymbol{\gamma}_0(\cdot) + c$, then model (3.3) is still correct. To fix the identification problem, we can always appropriately select the constant c such that $E\boldsymbol{\gamma}_0(U) = 0$. Next proposition can be easily proved by noting $\boldsymbol{\beta}_0(u) = \theta_0 + B_0 \boldsymbol{\gamma}_0(u)$ and the Sylvester's rank inequality.

Proposition 3.1. *With cov$\{\boldsymbol{\gamma}_0(U_i)\}$ being of full rank, the linear subspaces spanned by B_0 and cov$\{\boldsymbol{\beta}_0(U_i)\}$ are the same, i.e. $\mathscr{S}(\text{cov}\{\boldsymbol{\beta}_0(U_i)\}) = \mathscr{S}(B_0)$. If we further rewrite the model such that $E\{\boldsymbol{\gamma}_0(U_i)\} = 0$, then $E\{\boldsymbol{\beta}_0(U_i)\} = \theta_0$.*

Because $\mathscr{S}(B_0) = \mathscr{S}(\Sigma_\beta)$ with $\Sigma_\beta = \text{cov}\{\boldsymbol{\beta}_0(U)\}$, we can define $B_0 = (b_1, \cdots, b_{d_0}) \in \mathbb{R}^{p \times d_0}$, where b_j $(1 \le j \le d_0)$ are the eigenvectors associated with Σ_β's d_0 largest eigenvalues. As long as Σ_β's first d_0 eigenvalues are mutually different, B_0 is uniquely identifiable. For convenience, we assume throughout the rest of this Chapter that the non-zero eigenvalues of Σ_β are all different from one another.

As an alternative of model identification, we can also rewrite the model in such

a way that

$$B_0^\top B_0 = I_{d_0}, \quad \theta_0 \perp B_0 \text{ and } \operatorname{var}(g_1(U)) \geq \ldots \geq \operatorname{var}(g_{d_0}(U)) > 0. \tag{3.4}$$

By Proposition 3.1, it is easy to see that any PVCM satisfying (3.4) is identifiable. This way of identifying the model is more preferable because it has less parameters when $E\{\boldsymbol{\beta}_0(U)\} \in \mathscr{S}(\operatorname{cov}\{\boldsymbol{\beta}_0(U_i)\})$, in which case $\theta_0 = 0$. This fact will be used in our test for whether there exists a linear combination of X whose coefficient does not change with U. This fact can also be used to test whether there are constant coefficients in SVCM.

We end this section by mentioning relevant ideas of principal functions. Factor models or principal component analysis that extracts the main informative variables from a large number of variables are powerful approaches towards multivariate analysis. However, most of the models are under linear settings or under nonlinear framework; see for example Stock and Watson (2002) and Hastie and Stuetzle (1989). Our approach is under a functional framework.

3.3 Model Estimation

3.3.1 Profile Least-square Estimation of PVCM

PVCM is a semiparametric model, thus the popular nonparametric smoothing methods such as kernel smoothing and splines can be used for its estimation. In

this section, we will investigate the model estimation using the kernel smoothing approach. Estimation based on splines can be investigated similarly.

We firstly consider the estimation of θ_0 and B_0 under the assumption d_0 is known in advance. The estimation of d_0 will be addressed later. Proposition 3.1 motivates a very convenient way to estimate B_0 and θ_0. Specifically, by the local linear estimation (see, e.g., Fan and Gijbels (1996)) we can estimate $\boldsymbol{\beta}_0(u)$ by $\hat{\boldsymbol{\beta}}(u)$, where $\hat{\boldsymbol{\beta}}(u)$ is the minimizer of a in

$$\min_{a\in\mathbb{R}^p, b\in\mathbb{R}^p} n^{-1}\sum_{i=1}^{n}\Big\{Y_i - a^\top X_i - b^\top X_i(U_i-u)\Big\}^2 K_h(U_i-u), \tag{3.5}$$

where $K_h(u) = K(u/h)/h$ and $K(\cdot)$ is a kernel function. Consequently, we estimate Σ_β by $\hat{\Sigma}_\beta = n^{-1}\sum\{\hat{\boldsymbol{\beta}}(U_i) - \bar{\beta}\}\{\hat{\boldsymbol{\beta}}(U_i) - \bar{\beta}\}^\top$, where $\bar{\beta} = n^{-1}\sum\hat{\boldsymbol{\beta}}(U_i)$. We then estimate θ_0 by $\theta^{(0)} \stackrel{def}{=} \bar{\beta}$ and B_0 by $B^{(0)} \stackrel{def}{=} (\hat{b}_1^{(0)}, \cdots, \hat{b}_{d_0}^{(0)})$, where $\hat{b}_j^{(0)}$ is the eigenvector associated with the jth largest eigenvalue of $\hat{\Sigma}_\beta$ for $1 \le j \le d_0$. Let A be an arbitrary matrix and $\vec{A}$ stand for a vector constructed by stacking A's columns. Denote by $\|A\|$ the operation norm, i.e., the maximal absolute singular value of A. The estimation error for $B^{(0)}$ can be then defined as $\|\hat{B}^{(0)}(\hat{B}^{(0)})^\top - B_0B_0^\top\|$. We have the following consistency for the estimators.

Theorem 3.1. *Under the conditions (C.1)–(C.4) in the section 3.6, we have* $\|\theta^{(0)} - \theta_0\| = O_p\{h^2 + (nh/\log(n))^{-1/2}\}$ *and* $\|\hat{B}^{(0)}(\hat{B}^{(0)})^\top - B_0B_0^\top\| = O_p\{h^2 + (nh/\log(n))^{-1/2}\}$.

If parameters B_0 and θ_0 are temporarily known, we can then estimate the nonparametric functions in $\boldsymbol{\gamma}_0(u)$ easily by the standard estimation methods for

varying coefficient models; see for example Fan and Zhang W. (1999). The resulted estimators will be functions of the parameters θ_0 and B_0. Substituting the estimators into the model, we get a parametric model nominally, in which the parameters θ_0 and B_0 can then be estimated using the standard nonlinear parametric regression methods. Specifically, consider the local linear smoother of model (3.3)

$$\min_{a(u)\in\mathbb{R}^p, b(u)\in\mathbb{R}^p} \sum_{i=1}^{n} \{Y_i - X_i^\top\theta - a(u)^\top B^\top X_i - b(u)^\top B^\top X_i(U_i-u)/h\}^2 K_h(U_i-u).$$

If B and θ are close to the true values, then the minimizer of $a(u)$ is a local linear estimator of the coefficient functions $\boldsymbol{\gamma}_0(u)$, denoted by

$$\begin{aligned}\hat{\boldsymbol{\gamma}}(u|B,\theta) &= \{S_n(u,B)\}^{-1}B^\top[L_{n,0}(u) - S_{n,0}(u)\theta \\ &\quad -S_{n,1}(u)B(B^\top S_{n,2}(u)B)^{-1}B^\top(L_{n,1}(u) - S_{n,1}(u)\theta)],\end{aligned} \tag{3.6}$$

where

$$S_{n,k}(u) = \sum_{i=1}^{n} K_h(U_i-u)\{(U_i-u)/h\}^k X_i X_i^\top,$$

$$L_{n,k}(u) = \sum_{i=1}^{n} K_h(U_i-u)\{(U_i-u)/h\}^k X_i y_i,$$

for $k = 0, 1, 2$, and

$$S_n(u,B) = B^\top\{S_{n,0}(u) - S_{n,1}(u)B(B^\top S_{n,2}(u)B)^{-1}B^\top S_{n,1}(u)\}B.$$

Let $\bar{\boldsymbol{\gamma}}(B,\theta) = n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{\gamma}}(U_i|B,\theta)$ and $\tilde{\boldsymbol{\gamma}}(u|B,\theta) = \hat{\boldsymbol{\gamma}}(u|B,\theta) - \bar{\boldsymbol{\gamma}}(B,\theta)$.

Substituting $\tilde{\boldsymbol{\gamma}}(U_i|\theta,B)$ into the model, we have $Y_i \approx X_i^\top\theta + X_i^\top B\tilde{\boldsymbol{\gamma}}(U_i|\theta,B) + \varepsilon_i$,

$i = 1, 2, \ldots, n$. Thus, we consider

$$Q(\theta, B) = n^{-1} \sum_{i=1}^{n} \Big\{ Y_i - X_i^\top \theta - X_i^\top B \tilde{\boldsymbol{\gamma}}(U_i|\theta, B) \Big\}^2,$$

and estimate θ_0 and B_0 by

$$(\hat{\theta}, \hat{B}) = \arg\min_{\theta, B} Q(\theta, B). \tag{3.7}$$

Although the minimization is searched over the whole space, as in many model estimations an initial estimator is sometimes essential. The initial estimator $\theta^{(0)}$ and $B^{(0)}$ can be used for this purpose. Other robust estimation method such as the back-fitting method of Wu and Liang (2004) is also helpful. To facilitate the theoretical investigation, Theorem 3.1 allows us to restrict the parameter space in a small range of the true parameters $\Theta_n = \{(\theta, B) : ||\theta - \theta_0|| + ||B - B_0|| \leq M(h^2 + \delta_n)\}$ for some constant $M > 0$.

Theorem 3.2. *Suppose the conditions (C.1)–(C.4) in the section 3.6 hold. Let* $(\hat{\theta}, \hat{B}) = arg\min_{(\theta,B)\in\Theta_n} Q_n(\theta, B)$. *Then*

$$\sqrt{n} \begin{pmatrix} \hat{\theta} - \theta_0 \\ vec(\hat{B} - B_0) \end{pmatrix} \xrightarrow{D} N\{0, \Sigma_0^{-1}(\Sigma_1 + \Sigma_2)\Sigma_0^{-1}\}$$

in distribution, where

$$\Sigma_0 = E\Big\{ \begin{pmatrix} X \\ \boldsymbol{\gamma}_0(U) \otimes X \end{pmatrix} \begin{pmatrix} X \\ \boldsymbol{\gamma}_0(U) \otimes X \end{pmatrix}^\top \Big\},$$

$$\Sigma_1 = E\left\{\Big[\Big\{\begin{pmatrix} I_p \\ \boldsymbol{\gamma}_0(U)\otimes I\end{pmatrix} + \Big(\begin{pmatrix} W(U) \\ \boldsymbol{\gamma}_0(U)\otimes W(U)\end{pmatrix} - E\begin{pmatrix} W(U) \\ \boldsymbol{\gamma}_0(U)\otimes W(U)\end{pmatrix}\Big)V(U)\Big\}X\varepsilon\Big]^{\otimes 2}\right\},$$

$$\Sigma_2 = E\begin{pmatrix} W(U) \\ \boldsymbol{\gamma}_0(U)\otimes W(U)\end{pmatrix} B_0 E\{\boldsymbol{\gamma}_0(U)\boldsymbol{\gamma}_0^\top(U)\}B_0^\top E\begin{pmatrix} W(U) \\ \boldsymbol{\gamma}_0(U)\otimes W(U)\end{pmatrix}^\top,$$

with $W(U) = E(XX^\top|U)$ *and* $V(U) = B_0(B_0^\top W(U)B_0)^{-1}B_0^\top$.

After θ_0 and B_0 are estimated, we can estimate $\boldsymbol{\gamma}_0(u)$ immediately by $\hat{\boldsymbol{\gamma}}(u|\hat{\theta},\hat{B})$ and have the following limiting distribution.

Theorem 3.3. *Under the regularity condition (C.1)-(C.4) in the section 3.6, we have in distribution*

$$\sqrt{nh\hat{f}(u)}\{\hat{\boldsymbol{\gamma}}(u|\hat{B},\hat{\theta}) - \boldsymbol{\gamma}_0(u) - \frac{1}{2}\mu_2\boldsymbol{\gamma}_0''(u)h^2\} \xrightarrow{D} N\Big(0, \{B_0^\top W(u)B_0\}^{-1}B_0^\top W_2(u)B_0\{B_0^\top W(u)B_0\}^{-1}\Big),$$

where $W_2(u) = \int K^2(v)dvE\{XX^\top\varepsilon^2|U=u\}$, $\mu_2 = \int v^2K(v)dv$ *and* $\hat{f}(u) = n^{-1}\sum_{i=1}^n K_h(U_i-u)$.

Writing the model as a VCM, the estimated coefficient functions are $\hat{\beta}_{PVCM}(u) = \hat{\theta} + \hat{B}\hat{\boldsymbol{\gamma}}(u|\hat{B},\hat{\theta})$. It follows from Theorems 3.2 and 3.3 that

$$\sqrt{nh\hat{f}(u)}\{\hat{\beta}_{PVCM}(u) - \boldsymbol{\beta}_0(u) - \frac{1}{2}\mu_2\boldsymbol{\beta}_0''(u)h^2\} \xrightarrow{D} N\{0, \Sigma_{PVCM}(u)\},$$

where $\Sigma_{PVCM}(u) = B_0\{B_0^\top W(u)B_0\}^{-1}B_0^\top W_2(u)B_0\{B_0^\top W(u)B_0\}^{-1}B_0$. However, if we treat the model as a VCM and estimate it by the method in Fan and Zhang

W. (1999), then the estimator $\hat{\beta}_{VCM}(u)$ has

$$\sqrt{nh\hat{f}(u)}\{\hat{\beta}_{VCM}(u) - \boldsymbol{\beta}_0(u) - \frac{1}{2}\mu_2\boldsymbol{\beta}_0''(u)h^2\} \xrightarrow{D} N\{0, \Sigma_{VCM}(u)\},$$

where $\Sigma_{VCM}(u) = \{W(u)\}^{-1}W_2(u)\{W(u)\}^{-1}$; see Fan and Zhang W. (1999). If $d_0 < p$, it is easy to see that

$$\Sigma_{PVCM}(u) < \Sigma_{VCM}(u),$$

indicating that the estimator based on a PVCM is indeed more efficient than that based on a VCM. The smaller d_0 is, the more efficient is PVCM compared with VCM.

To make statistical inference, we also need to estimate the variance-covariance matrices in the limiting distributions. These matrices can be estimated simply by their sample versions with the unknown functions and parameters being replaced by their estimators respectively. By the local linear kernel smoothing, $W(u)$ can be estimated consistently by

$$\hat{W}(u) = \sum_{i=1}^{n} w_{n,h}(U_i - u)X_iX_i^\top / \sum_{i=1}^{n} w_{n,h}(U_i - u),$$

where $w_{n,h}(U_i-u) = K_h(U_i-u)\sum_{i=1}^{n} K_h(U_i-u)\{(U_i-u)/h\}^2 - K_h(U_i-u)\{(U_i-u)/h\}\sum_{i=1}^{n} K_h(U_i-u)\{(U_i-u)/h\}$, and $E\{XX^\top\varepsilon^2|U=u\}$ by

$$\sum_{i=1}^{n} w_{n,h}(U_i - u)X_iX_i^\top\{Y_i - X_i^\top\hat{\theta} - \hat{\boldsymbol{\gamma}}(U_i)\hat{B}^\top X_i\}^2 / \sum_{i=1}^{n} w_{n,h}(U_i - u).$$

As an example of hypothesis testing, we consider whether there is a separate linear

part in the model under identification (3.4), i.e. whether there exists a linear combination $\theta_0^\top X$ such that $\theta_0^\top B_0 = 0$ and $\theta_0 \neq 0$. The corresponding hypothesis is

$$H_0 : (I - B_0 B_0^\top)\theta_0 \neq 0.$$

With the identification of (3.4), we can construct a test statistic

$$ST = n(\hat{\theta} - \theta_0)^\top \hat{P}(\hat{P} S_{00} \hat{P})^+ \hat{P}(\hat{\theta} - \theta_0),$$

where $\hat{P} = (I - \hat{B}\hat{B}^\top)$ and S_{00} is the submatrix of estimated $\Sigma_0^{-1}(\Sigma_1 + \Sigma_2)\Sigma_0^{-1}$ in its first p rows and first p columns, and A^+ denotes the Moore-Penrose inverse of matrix A. We get the following corollary from $\text{rank}(I_p - B_0 B_0^\top) = p - d_0$, the identification (3.4) and Theorem 3.2.

Corollary 1. *Under the model assumption (C.1) and (C.4) and H_0, with identification (3.4) we have $ST \xrightarrow{D} \chi^2(p - d_0)$ as $n \to \infty$.*

By Corollary 1, we reject H_0 if $ST > \chi^2_{1-\alpha}(p - d_0)$ with significance level α.

Next, we consider the estimation of d_0. To this end, we propose here the following BIC-type criterion:

$$\text{BIC}(d) = \log \hat{\sigma}_d^2 + d \times \frac{\log(nh)}{nh}, \tag{3.8}$$

where d is the working number of principal functions, nh is the effective sample

size in nonparametric regression, and $\hat{\sigma}_d^2$ is given by

$$\hat{\sigma}_d^2 = n^{-1} \sum_{k=1}^{n} \left\{ Y_k - X_k^\top \hat{\theta} - \hat{\boldsymbol{\gamma}}^\top(U_k) \hat{B}^\top X_k \right\}^2,$$

where estimators $\hat{\theta}$, $\hat{B}$ and $\hat{\boldsymbol{\gamma}}(U_i)$ are all obtained under the working number, d, of principal functions. For the purpose of completeness, define $\mathrm{BIC}(0) = n^{-1} \sum (Y_i - \bar{Y})^2$ with $\bar{Y} = n^{-1} \sum Y_i$. Then d_0 is estimated by $\hat{d} = \mathrm{argmin}_{0 \le d \le p} \mathrm{BIC}(d)$.

Theorem 3.4. *Assuming the technical conditions (C.1)–(C.4) as given in the section 3.6, we have* $P(\hat{d} = d_0) \to 1$.

By Theorem 3.4, it is also easy to see that Theorems 3.1 - 3.3 still hold if we replace d_0 by $\hat{d}$.

3.3.2 Refinement of Estimation Based on the Adaptive Lasso Penalty

In this section, we estimate the model by incorporating the kernel smoothing with the L_1 penalty. As well demonstrated in the literature, the L_1 penalty approach has several advantages. Specifically for PVCM, the L_1 penalty can achieve the following goals simultaneously. (1) To identify variables that have cross effect with the index variable on the response, and those that only have simple linear effect. (2) To identify abundant variables and automatically remove them from the model. (3) To improve the estimation efficiency. Moreover, the L_1 penalty approach can estimate the model well even when the number of covariates is large.

Let $\alpha = (\alpha_1, \ldots, \alpha_{p(d_0+1)})^\top = (\theta^\top, vec(B)^\top)^\top$. Let $S = \{1, 2, \ldots, p(d_0+1)\}$ and $\mathscr{A} = \{s \in S : \alpha_s \neq 0\}$. Then $\mathscr{A}$ is the index set that contains only non-zeros elements in α. Following Zou (2006), consider the following adaptive Lasso estimation,

$$\begin{aligned}\tilde{\alpha}^{(n)} = (\tilde{\theta}_n^\top, vec(\tilde{B}_n)^\top)^\top &= \arg\min_{(\theta,B)} \Big\{ Q(\theta, B) + \lambda_n \sum_{i=1}^{p} (\hat{w}_k |\theta_k| + \sum_{j=1}^{d_0} \hat{w}_{ij} |B_{ij}|) \Big\} \\ &= \arg\min_{\alpha \in \mathscr{R}^S} \Big\{ Q(\theta, B) + \lambda_n \sum_{s=1}^{p(d_0+1)} \hat{w}_s |\alpha_s| \Big\}, \qquad (3.9)\end{aligned}$$

where $\hat{w}_s = 1/|\hat{\alpha}_s|^\tau$ with $\tau > 0$ and $\hat{\alpha}_s$ is the estimator of α defined in (3.7). Let $\mathscr{A}_n = \{s \in S : \tilde{\alpha}_s^{(n)} \neq 0\}$. Then $\mathscr{A}_n$ is the variables that are selected in either the linear part or nonlinear part of PVCM or both. If a variable is not selected either in the linear or the nonlinear part, the variable is abundant and will be removed automatically from the model.

Theorem 3.5. *Under the conditions of Theorem 3.2 and* $\lambda_n/\sqrt{n} \to 0$, $\lambda_n n^{\frac{\tau-1}{2}} \to \infty$, *we have the following asymptotic properties for the estimators* $\tilde{\theta}_n$ *and* $\tilde{B}_n$.

(1) The coefficients with nonzero values in both θ_0 *and* B_0 *can be consistently identified, i.e.*

$$\lim_{n\to\infty} P(\mathscr{A}_n = \mathscr{A}) = 1.$$

(2) The estimated parameters achieve the oracle efficiency where the zero coefficients are known and removed in advance, i.e.

$$\sqrt{n} \begin{pmatrix} \tilde{\theta} - \theta_0 \\ vec(\tilde{B} - B_0) \end{pmatrix}_{\mathscr{A}} \xrightarrow{D} N\Big\{0, \Big((\Sigma_0)_{\mathscr{A}}\Big)^{-1} \Big(\Sigma_1 + \Sigma_2\Big)_{\mathscr{A}} \Big((\Sigma_0)_{\mathscr{A}}\Big)^{-1}\Big\}, \qquad (3.10)$$

where notation $M_{\mathscr{A}}$ denotes the submatrix of M with jth row (and jth column if M is a matrix) being removed from matrix M, where $j \in \mathscr{A}$.

In practice, the selection of the tuning parameter λ_n is essential in the estimation. We found the commonly used BIC criterion works well, which is stated below. To indicate the dependence of the estimators on the tuning parameter λ, write the estimators of (3.9) as $\tilde{\theta}_\lambda$ and $\tilde{B}_\lambda$ respectively. Define

$$\mathrm{BIC}(\lambda) = \log\{Q(\tilde{\theta}_\lambda, \tilde{B}_\lambda)\} + \log(n)\frac{p_n}{n},$$

where p_n is the number of nonzero values in $\tilde{\theta}_\lambda$ and $\tilde{B}_\lambda$. The asymptotic performance of the BIC in selecting λ can be similarly discussed as in Wang and Xia (2009). The details are omitted here.

3.4 Simulation Studies

Consider two varying coefficient models where the covariates $X_{i1} \equiv 1$, and X_{ij}s $(1 < j \leq p)$ are simulated from a multivariate normal distribution with $\mathrm{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$ for any $j_1, j_2 \geq 2$, and U_i is simulated from $U[0,1]$, and ε from $N(0,1)$. The coefficients and principal functions are respectively as follows.

Model 1. $\theta_0 = b_0, \quad B_0 = b_1, \quad \boldsymbol{\gamma}_0(u) = 10u(1-u) - 5/3,$

Model 2. $\theta_0 = b_0, \quad B_0 = (b_2, b_3), \quad \boldsymbol{\gamma}_0(u) = \{\cos(2\pi u), \sin(2\pi u)\}^\top,$

where

$$b_0 = (\underbrace{1, 1, \ldots, 1}_{7}, 0, \ldots, 0)^\top, \qquad b_1 = (\underbrace{1, -1, \ldots, 1, -1}_{[(p-1)/3]}, 0, \ldots, 0)^\top,$$

$b_2 = (\underbrace{1, \ldots, 1}_{[(p-1)/3]}, 0, \ldots, 0)^\top$ and $b_3 = (\underbrace{0, \ldots, 0}_{[(p-1)/3]}, \underbrace{1, \ldots, 1}_{[(p-1)/3]}, 0 \ldots, 0)^\top$. As one can see, Model 1 has 1 principal function ($d_0 = 1$) and Model 2 has 2 ($d_0 = 2$).

In the following calculation, we use the Newton-Rahpson algorithm to solve the minimization problem in (3.7). For the minimization in (3.9), we use the quadratic norm to approximate the L_1 norm and then the Newton-Rahpson algorithm to solve the minimization numerically.

For each model setting, a total of 500 simulation replications are conducted. For each simulation replication, we first compute the conventional varying coefficient estimator $\hat{\boldsymbol{\beta}}(u)$ according to (3.5). See, e.g., Fan and Zhang (1999) for more details. The bandwidth h is selected by leave-one-out cross-validation. The same bandwidth is then used throughout the rest of the entire computational process, except for the estimation of B_0 and θ_0 where the bandwidth is multiplied by $n^{-0.1}$ for the purpose of undersmoothing; see Carroll *et al* (1997). We apply the proposed BIC criterion (3.8) to estimate the number of principal functions, $\hat{d}$. The percentage of replications in which the number of principal functions is correctly estimated is summarized in the third column of Table 1. In the column, as sample size increases the percentage of replications with $\hat{d} = d_0$ converges 100% quickly, confirming that $\hat{d}$ is indeed a consistent estimator for d_0.

As shown in Theorem 3.5, the proposed estimation with L_1 penalty can do the

variable selection. To check the variable selection in the linear part and nonlinear part, we count in each estimation the number of zero rows (i.e. the rows in which all elements are zeros) in the estimated θ and B respectively. Note that if a row of estimated θ is zero, it means the corresponding variable is not selected in the linear part. Similarly, if all the elements in a row of B are zero, it means the corresponding variable is removed from the nonlinear part. From the fourth and fifth columns of Table 1, by comparing the numbers with those in the square brackets that correspond to true number of zeros, we see that as sample size increases, the adaptive L_1 penalty is consistent in selecting the variables in the linear part or nonlinear part.

We evaluate the overall model estimation performance by checking the estimation error of the coefficients. With estimated d_0, we then compute $\hat{\theta}$ and $\hat{B}$ and thus $\hat{\boldsymbol{\beta}}(u_i) = \hat{\theta} + \hat{B}\hat{\boldsymbol{\gamma}}(u_i)$. The estimation error is evaluated by

$$n^{-1} \sum_{i=1}^{n} |\hat{\boldsymbol{\beta}}(u_i) - \theta_0 - B_0 \boldsymbol{\gamma}_0(u_i)|,$$

where $|\ell| = (|\ell_1| + |\ell_2| + \ldots + |\ell_p|)/p$ for any vector $\ell = (\ell_1, \ldots, \ell_p)^\top$. The average error of estimators across the 500 simulation replications are summarized in columns 6, 7 and 8 of Table 3.1. In these columns, as the sample size increases, the error steadily shrinks towards 0. This trend confirms that all the estimator are consistent. However, treating a PVCM as a VCM, the estimation efficiency will be very much adversely affected by comparing column 7 with column 6. By comparing the eighth column with the seventh column, we can see that imposing the adaptive L_1 penalty, the estimation efficiency can be substantially improved, especially when the number of covariates is large.

Below figure 3.2 shows the simulation results of the testing hypothesis H_0 with significance level 0.05. In each panel, the black, blue, green and red lines correspond to sample sizes 100, 200, 500 and 1000 respectively. The horizontal dash line marks the significance level 0.05.

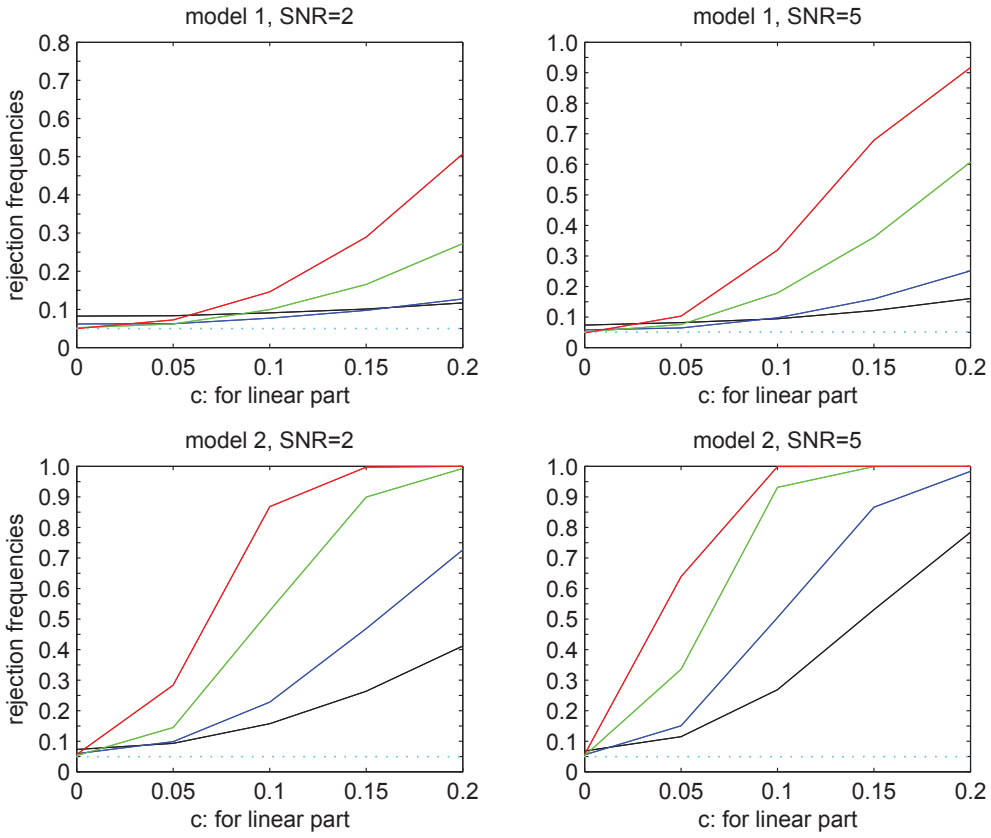

Figure 3.2 The simulation results based on 5000 replications under each model setting.

Next, we check the performance of the proposed statistical testing for the hypothesis on the linear part. However, the following simulations are done with small p, since the testing method may not work well when p is large as being understood in the literature. We allow the linear part θ_0 to change with c, i.e. $\theta = c \times b_0$. The

bigger c is, the more influential the linear part is. If $c = 0$, the models have no linear part. We also change the signal-to-noise ratios (SNR) by changing the variance of ε. With significance level $\alpha = 0.05$, we calculate the rejection frequencies for $H_0 : |\theta_0| = 0$ under model specification (3.4). In both models, when $c = 0$ there is not linear part, and thus the rejection frequency should be around 0.05. As c increases, the influence of the linear parts increases. As a consequence, the rejection frequencies should also increase. Our simulation results for $c = 0, 0.05, 0.1, 0.15$ and 0.2 reported in Figure 3.2 support our theory quite well, indicating that the hypothesis testing statistic has reasonable power with roughly correct significant level. It is also reasonable to see that as the number of principal components increases, the power of hypothesis testing decreases because the freedom of the linear part is reduced.

3.5 A Real Example

The Boston House Price Data of Harrison and Rubinfeld (1978) has attracted lots of attention of statisticians. Various models have been applied to it, such as the linear regression model (Belsley *et al* (1980)), the additive model (Fan and Jiang (2005)) and the varying coefficient model (Fan and Huang (2006)). The response of interest is the median value of owner-occupied homes (MEDV, in $1000) with 13 covariates: lower status of the population (LSTAT), per capita crime rate (CRIM) by town, average number of rooms per dwelling (RM), full-value property-tax rate per $10,000 (TAX), nitrogen dioxides concentration (NOX, parts per 10 million), pupil-teacher ratio by town (PTRATIO), proportion of owner-occupied units built

prior to 1940 (AGE), proportion of residential land zoned for lots over 25,000 square feet (ZN), proportion of non-retail business acres per town (INDUS), Charles River dummy variable (1 if tract bounds river; 0 otherwise; CHAS), weighted distances to five Boston employment centres (DIS), index of accessibility to radial highways (RAD), $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town (B).

Fan and Huang (2005) proposed to fit the data with a semi-varying coefficient model with $U = \sqrt{LSTAT}$ as the index variable. However, as the number of covariates $p = 13$ is too big for a varying coefficient model to be estimated well, Fan and Huang (2005) only used 6 variables in their model. With the superior estimation efficiency of PVCM over CVM, we can include all the variables into the PVCM. Next, we fit the PVCM to the data with all the variables. We standardize all the variables before fitting the model.

As we mentioned in the first section, remarkably similar shapes are shared among different estimated varying coefficients. The eigenvalues of the estimated Σ_β suggest that the number of principal functions is $d_0 = 1$. Such a conclusion is more formally confirmed by our BIC criterion (3.8). The BIC values for $d_0 = 0$(linear model), $d_0 = 1, \ldots,$ and $d_0 = 10$ are respectively -1.1593, -1.7199, -1.6950, -1.5482, -1.4933, -1.2018, -0.8020, -0.5044, -0.2011 and -0.1034. Therefore, the number of principal functions is selected as 1. The corresponding parameters in the model are estimated and listed in Table 2. It is interesting to see that some of the covariates are eliminated from the model such as AGE, INDUS and CHAS because they do not appear in either the linear part or the nonlinear part. In a different model that only includs the variables in the top panel of Table 2, AGE was also removed by Fan and Huang (2006) based on a statistical testing approach. Some other

covariates have no cross effect with LSTAT on the response, such as TAX, NOX, ZN and DIS.

The principal function $\hat{\gamma}(u)$ is estimated and shown in Figure 3.3 together with its centralized pointwise 95% confidence band based on Theorem 3.3. Its 95%

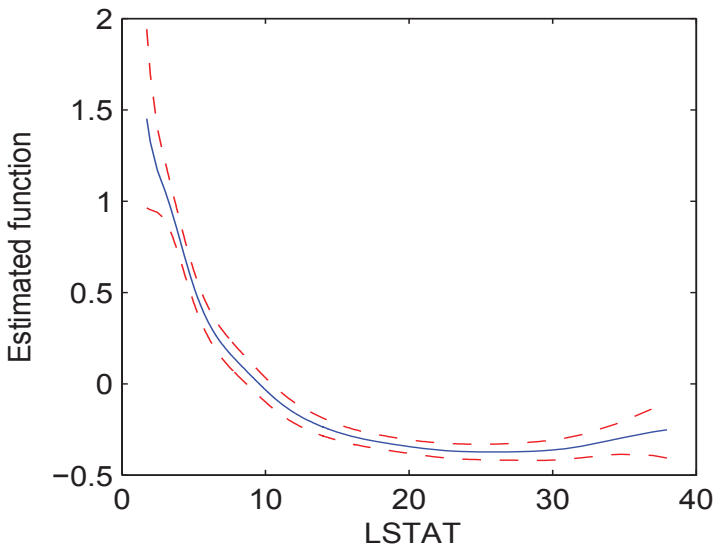

Figure 3.3 The estimated principal function (in the middle) for the real dataset.

centralized point wise confidence band is denoted by the dash lines.

To further verify the model appropriateness to the data, we consider the prediction of the PVCM and compare it with linear regression model and the conventional varying coefficient model (VCM). We randomly partition all the 506 observations into a training set and a prediction set. We estimate the PVCM based on the training set, and use the estimated model to make prediction for the prediction set. With different size of training set and prediction set, the average prediction errors based on 1000 random partitions are listed on Table 3.3. It is easy to see from the table that the conventional VCM has very poor prediction capability, and

is much worse than the simple linear regression model. However, PVCM with one principal function as identified by the proposed method has much better prediction ability than VCM and even substantially better than the linear regression model. The prediction ability can be further improved when the L_1 penalty is imposed in the estimation, though the primary purpose of imposing the L_1 penalty is for variable selection.

3.6 Proofs

To establish the asymptotic theory for the proposed estimation methods, we need the following technical assumptions.

(C.1) (*The Index Variable*). The *index* variable U has a bounded compact support $\mathscr{D}$ and a probability density function $f(u)$, which is Lipschitz continuous and bounded away from 0 on $\mathscr{D}$.

(C.2) (*Smoothness Assumptions*). Every component of $W(u) = E(XX^\top|U = u)$ and $L(u) = E(XY^\top|U = u)$ is Lipschitz continuous. In addition to that, we assume $\beta_0(u)$ has continuous second order derivatives in $u \in \mathscr{D}$. The matrix $W(u)$ is positive definite for all $u \in \mathscr{D}$.

(C.3) (*Moment Conditions*). There exist $s > 2$ and $\delta < 2 - s^{-1}$, such that $E\|X\|^s < \infty$ with $n^{2\delta-1}h \to \infty$, where $\|\cdot\|$ stands for a typical L_2 norm.

(C.4) (*The Kernel and Bandwidth*). We assume that the kernel function $K(\cdot)$ is a symmetric density function with a compact support. Moreover, we assume

$h \propto n^{-c}$ with $c > 0$ such that $\sqrt{n}h^2 \to 0$ and $nh/\log n \to \infty$.

We remark that the above regularity conditions are rather standard. Similar assumptions have been used in, for example, Zhang W. Y. *et al* (2002) and Fan and Huang (2005). Let $\mu_k = \int t^k K(t)$. Then by (C.4) we have $\mu_0 = 1$ and $\mu_1 = 0$. For ease of exposition, we further standardize $K(\cdot)$ such that $\mu_2 = 1$ in the following proofs.

Lemma 3.1. *Under the regularity conditions (C.1)-(C.4), for the estimator defined in (3.6) we have the following expansion*

$$\begin{aligned}\hat{\boldsymbol{\gamma}}(u|B,\theta) &= \boldsymbol{\gamma}_0(u) + \frac{1}{2}\mu_2\boldsymbol{\gamma}_0''(u)h^2 + \{B^\top W(u)B\}^{-1}\{nf(u)\}^{-1}B^\top \sum_{i=1}^n K_h(U_{iu})X_i\varepsilon_i \\ &\quad + \{B^\top W(u)B\}^{-1}B^\top W(u)(B_0 - B)\boldsymbol{\gamma}_0(u) + \{B^\top W(u)B\}^{-1}B^\top W(u)(\theta_0 - \theta) \\ &\quad + O_p(h^3 + h\delta_n + \delta_n^2)\end{aligned}$$

uniformly for any $u \in \mathscr{U}$ and $(\theta, B) \in \Theta_n$.

Proof. Write $Y_i - X_i^\top\theta = \varepsilon_i + X_i^\top B\boldsymbol{\gamma}_0(U_i) + X_i^\top(B_0 - B)\boldsymbol{\gamma}_0(U_i) + X_i^\top(\theta_0 - \theta)$. Thus

$$\begin{aligned}\sum_{i=1}^n K_h(U_{iu})X_i\{Y_i - X_i^\top\theta\} &= \sum_{i=1}^n K_h(U_{iu})X_i\varepsilon_i + \sum_{i=1}^n K_h(U_{iu})X_iX_i^\top B\boldsymbol{\gamma}_0(U_i) \quad (3.11) \\ &\quad + \sum_{i=1}^n K_h(U_{iu})X_iX_i^\top(B_0 - B)\boldsymbol{\gamma}_0(U_i) + S_n(u)(\theta_0 - \theta).\end{aligned}$$

Let $s_n(u) = \sum_{i=1}^n K_h(U_{iu})$. By Mack and Silverman (1982), we have uniformly for $u \in \mathscr{U}$, $s_n^{-1}(u) = (nf(u))^{-1}(1 + O_p(h^2 + \delta_n))$, and

$$\frac{1}{n}\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top = f(u)W(u)(1 + O_p(h^2 + \delta_n)), \quad \frac{1}{n}\sum_{i=1}^n K_h(U_{iu})X_i\varepsilon_i = O_p(\delta_n).$$

Thus,

$$s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top = W(u) + O_p(h^2 + \delta_n),$$

$$s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top\boldsymbol{\gamma}_0(U_i) = W(u)\boldsymbol{\gamma}_0(u) + O_p(h^2 + \delta_n),$$

$$s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_i\varepsilon_i = \{nf(u)\}^{-1}\sum_{i=1}^n K_h(U_{iu})X_i\varepsilon_i + O_p(h^2\delta_n + \delta_n^2),$$

and

$$s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top(B_0-B)\boldsymbol{\gamma}_0(U_i) = W(u)(B_0-B)\boldsymbol{\gamma}_0(u)+\|B_0-B\|O_p(h^2+\delta_n)$$

uniformly for $u \in \mathscr{U}$. Combining the above results yields that uniformly in $u \in \mathscr{U}$,

$$\begin{aligned}
& s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top B\boldsymbol{\gamma}_0(U_i) \\
= \; & s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top B\boldsymbol{\gamma}_0(u) + s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top B\{\boldsymbol{\gamma}_0(U_i) - \boldsymbol{\gamma}_0(u)\} \\
= \; & s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top B\boldsymbol{\gamma}_0(u) \\
& + s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top B\{\boldsymbol{\gamma}_0'(u)(U_{iu}) + \frac{1}{2}\mu_2\boldsymbol{\gamma}_0''(u)(U_{iu})^2 + O_p(U_{iu}^3)\} \\
= \; & s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_iX_i^\top B\boldsymbol{\gamma}_0(u) + \{f^{-1}(u)f'(u)W'(u)B\boldsymbol{\gamma}_0'(u) + \frac{1}{2}\mu_2 W(u)B\boldsymbol{\gamma}_0''(u)\}h^2 \\
& + O_p(h^3).
\end{aligned}$$

For $(\theta, B) \in \Theta_n$, we have

$$\begin{aligned}
\hat{\boldsymbol{\gamma}}(u|B,\theta) &= (B^\top S_n(u)B)^{-1}B^\top \sum_{i=1}^n K_h(U_{iu})X_i\{Y_i - X_i^\top\theta\} \\
&= (B^\top s_n^{-1}(u)S_n(u)B)^{-1}B^\top\Big(s_n^{-1}(u)\sum_{i=1}^n K_h(U_{iu})X_i\{Y_i - X_i^\top\theta\}\Big) \\
&= \boldsymbol{\gamma}_0(u) + \frac{1}{2}\mu_2\boldsymbol{\gamma}_0''(u)h^2 + \{B^\top W(u)B\}^{-1}\{nf(u)\}^{-1}B^\top\sum_{i=1}^n K_h(U_{iu})X_i\varepsilon_i \\
&\quad + \{B^\top W(u)B\}^{-1}B^\top W(u)(B_0 - B)\boldsymbol{\gamma}_0(u) + \{B^\top W(u)B\}^{-1}B^\top W(u)(\theta_0 - \theta) \\
&\quad + O_p(h^3 + h\delta_n + \delta_n^2).
\end{aligned}$$

As a special case,

$$\begin{aligned}
\hat{\boldsymbol{\gamma}}(u|B_0,\theta_0) &= \boldsymbol{\gamma}_0(u) + \frac{1}{2}\mu_2\boldsymbol{\gamma}_0''(u)h^2 + \{B_0^\top W(u)B_0\}^{-1}\{nf(u)\}^{-1}B_0^\top\sum_{i=1}^n K_h(U_{iu})X_i\varepsilon_i \\
&\quad + O_p(h^3 + h\delta_n + \delta_n^2). \qquad (3.12)
\end{aligned}$$

We have completed the proof. □

Proof of Theorems 3.1.

By Theorem 1 of Fan and Zhang (2000) or Lemma 3.1, we have

$$\sup_{u\in\mathscr{D}}|\hat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}_0(u)| = O_p(h^2 + \delta_n), \qquad (3.13)$$

where $\delta_n = \{nh/\log(n)\}^{-1/2}$. Theorems 3.1 follows immediately from (3.13). □

Proof of Theorem 3.2.

Let $\alpha = (\theta^\top, vec(B)^\top)^\top$, $\alpha_0 = (\alpha_{0,1}, ..., \alpha_{0,p(d_0+1)})^\top = (\theta_0^\top, vec(B_0)^\top)^\top$, $\hat{\alpha} = (\hat{\theta}^\top, vec(\hat{B})^\top)^\top$ and $Q(\alpha) = Q(\theta, B)$. By Taylor expansion about α_0, we have

$$0 = \frac{\partial Q(\hat{\alpha})}{\partial \alpha} = \frac{\partial Q(\alpha_0)}{\partial \alpha} + \frac{\partial^2 Q(\alpha^*)}{\partial \alpha \partial \alpha^\top}(\hat{\alpha} - \alpha_0),$$

where α^* lies on the line segment between α_0 and $\hat{\alpha}$. Let $\Delta_i(\alpha) = Y_i - X_i^\top \theta - X_i^\top B\tilde{\boldsymbol{\gamma}}(U_i)$, $\eta_i(\alpha) = Y_i - X_i^\top \theta - X_i^\top B\boldsymbol{\gamma}_0(U_i)$, then $\Delta_i(\alpha) = \eta_i(\alpha) - X_i^\top B(\tilde{\boldsymbol{\gamma}}(U_i) - \boldsymbol{\gamma}_0(U_i))$, $\eta_i(\alpha_0) = \varepsilon_i$, and

$$Q(\alpha) = \sum_{i=1}^n \Delta_i^2(\alpha).$$

Let $Q_0(\alpha) = \sum_{i=1}^n \eta_i^2(\alpha)$. From Lemma 3.1, when $||\alpha - \alpha_0|| = O_p(h^2 + \delta_n)$ we have

$$\sup_{u \in \mathscr{U}} \|\tilde{\boldsymbol{\gamma}}(u) - \boldsymbol{\gamma}_0(u)\| = O_p(h^2 + \delta_n) = o_p(1).$$

Thus $\Delta_i(\alpha) = \eta_i(\alpha) - X_i^\top B(\tilde{\boldsymbol{\gamma}}(U_i) - \boldsymbol{\gamma}_0(U_i)) = \eta_i(\alpha) + o_p(1)$, $\partial\Delta_i(\alpha)/\partial\alpha = \partial\eta_i(\alpha)/\partial\alpha + o_p(1)$. It follows that

$$\begin{aligned}
\frac{1}{2n}\frac{\partial^2 Q(\alpha)}{\partial\alpha\partial\alpha^\top} &= \frac{1}{n}\sum_{i=1}^n \frac{\partial\Delta_i(\alpha)}{\partial\alpha}\frac{\partial\Delta_i(\alpha)}{\partial\alpha^\top} + \frac{1}{n}\sum_{i=1}^n \Delta_i(\alpha)\frac{\partial^2\Delta_i(\alpha)}{\partial\alpha\partial\alpha^\top} \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\partial\eta_i(\alpha)}{\partial\alpha}\frac{\partial\eta_i(\alpha)}{\partial\alpha^\top} + \frac{1}{n}\sum_{i=1}^n \eta_i(\alpha)\frac{\partial^2\eta_i(\alpha)}{\partial\alpha\partial\alpha^\top} + o_p(1) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\frac{\partial\eta_i(\alpha_0)}{\partial\alpha^\top} + \frac{1}{n}\sum_{i=1}^n \eta_i(\alpha_0)\frac{\partial^2\eta_i(\alpha_0)}{\partial\alpha\partial\alpha^\top} + o_p(1) \\
&\to E\Big\{\frac{\partial\eta_1(\alpha_0)}{\partial\alpha}\frac{\partial\eta_1(\alpha_0)}{\partial\alpha^\top}\Big\} = W, \quad \text{in probability.}
\end{aligned}$$

In the last step, $\partial^2\eta_i(\alpha_0)/(\partial\alpha\partial\alpha^\top)=0$ is used. Write

$$\frac{1}{2\sqrt{n}}\frac{\partial Q(\alpha_0)}{\partial\alpha}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\eta_i(\alpha_0)-X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i)-\boldsymbol{\gamma}_0(U_i))\}\{\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}+\frac{\partial\Delta_i(\alpha_0)}{\partial\alpha}-\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\}. \tag{3.14}$$

Let $Z_{n0}=Z_{n1}+Z_{n2}$ with $Z_{n1}=n^{-1/2}\sum_{i=1}^n\eta_i(\alpha_0)\partial\eta_i(\alpha_0)/\partial\alpha$ and

$$Z_{n2}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i(\alpha_0)\Big(\frac{\partial\Delta_i(\alpha_0)}{\partial\alpha}-\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\Big)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i)-\boldsymbol{\gamma}_0(U_i))\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}.$$

By Lemma 3.1, we have

$$\begin{aligned}\Big|\frac{1}{2\sqrt{n}}\frac{\partial Q(\alpha_0)}{\partial\alpha}-Z_{n0}\Big| &= \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i)-\boldsymbol{\gamma}_0(U_i))\Big(\frac{\partial\Delta_i(\alpha_0)}{\partial\alpha}-\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\Big)\Big|\\ &\le \sqrt{n}\max_{1\le i\le n}|X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i)-\boldsymbol{\gamma}_0(U_i))|\max_{1\le i\le n}\Big\|\frac{\partial\Delta_i(\alpha_0)}{\partial\alpha}-\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\Big\|\\ &= \sqrt{n}O_p(h^2+\delta_n)O_p(h^2+\delta_n)=o_p(1).\end{aligned}$$

It is easy to check that

$$Z_{n1} = -n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix}X_i\\ \boldsymbol{\gamma}_0(U_i)\otimes X_i\end{pmatrix}\varepsilon_i.$$

Let $\ell(U)=(1,\boldsymbol{\gamma}_0(U)^\top)^\top\otimes W(U)$ and $\bar{\ell}=E\ell(U)$. Write $Z_{n2}=n^{1/2}(E_{n1}-E_{n2})$, where

$$E_{n1}=n^{-1/2}\sum_{i=1}^{n}\eta_i(\alpha_0)\Big(\frac{\partial\Delta_i(\alpha_0)}{\partial\alpha}-\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}\Big),$$

$$E_{n2}=n^{-1/2}\sum_{i=1}^{n}X_i^\top B_0(\tilde{\boldsymbol{\gamma}}(U_i)-\boldsymbol{\gamma}_0(U_i))\frac{\partial\eta_i(\alpha_0)}{\partial\alpha}.$$

Under assumptions (C.1)-(C.4), we can show that

$$E_{n1} = o_p(1) \tag{3.15}$$

and

$$\begin{aligned} E_{n2} &= \frac{1}{2}E\{(\ell(U)-\bar{\ell})B_0\boldsymbol{\gamma}_0''(U)\}n^{1/2}h^2 + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\ell(U_j)-\bar{\ell})V(U_j)X_j\varepsilon_j \\ &\quad +\bar{\ell}B_0\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\gamma}_0(U_i) + o_p(1). \end{aligned} \tag{3.16}$$

Thus, we have

$$\begin{aligned} Z_{n2} &= \frac{1}{2}E\{(\ell(U)-\bar{\ell})B_0\boldsymbol{\gamma}_0''(U)\}n^{1/2}h^2 + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\ell(U_j)-\bar{\ell})V(U_j)X_j\varepsilon_j \\ &\quad + \begin{pmatrix} EW(U) \\ E\{\boldsymbol{\gamma}_0(U)\otimes W(U)\} \end{pmatrix} B_0\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\gamma}_0(U_i) + o_p(1), \end{aligned}$$

where $W(u)$ and $V(u)$ are defined in Theorem 3.2. By the Central Limit Theorem (CLT), we have

$$Z_{n1} + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\ell(U_j)-\bar{\ell})V(U_j)X_j\varepsilon_j \to N(0,\Sigma_1),$$

where Σ_1 is given in Theorem 3.2. On the other hand, since $E\boldsymbol{\gamma}_0(U)=0$, we have

$$n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\gamma}_0(U_i) \to N\Big(0, E\{\boldsymbol{\gamma}_0(U)\boldsymbol{\gamma}_0^{\top}(U)\}\Big).$$

Theorem 3.2 follows from last three equations and (3.14).

Now, we turn to prove (3.15) and (3.16). We only give the details for the latter. Decompose E_{n2} into two terms.

$$E_{n2} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i^\top B_0(\hat{\boldsymbol{\gamma}}(U_i) - \boldsymbol{\gamma}_0(U_i)) \frac{\partial \eta_i(\alpha_0)}{\partial \alpha} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i^\top B_0 \bar{\boldsymbol{\gamma}} \frac{\partial \eta_i(\alpha_0)}{\partial \alpha} \triangleq E_{n2}^1 - E_{n2}^2, \tag{3.17}$$

where $\hat{\boldsymbol{\gamma}}(U_i) = \hat{\boldsymbol{\gamma}}(U_i|\theta_0, B_0)$ and $\bar{\boldsymbol{\gamma}} = n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\gamma}}(U_i)$. From Lemma 3.1, we have

$$E_{n2}^1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ \boldsymbol{\gamma}_0(U_i) \otimes X_i \end{pmatrix} X_i^\top B_0 \{\frac{1}{2}\boldsymbol{\gamma}_0''(U_i)h^2 + R_n(U_i) + O_p(h^3 + h\delta_n + \delta_n^2)\},$$

where $R_n(U_i) = \{nf(U_i)B_0^\top W(U_i)B_0\}^{-1} B_0^\top \sum_{j=1}^n K_h(U_{ij})X_j\varepsilon_j$. It follows from the laws of large numbers

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ \boldsymbol{\gamma}_0(U_i) \otimes X_i \end{pmatrix} X_i^\top B_0 \boldsymbol{\gamma}_0''(U_i)h^2 = E\{\ell(U)B_0\boldsymbol{\gamma}_0''(U)\}n^{1/2}h^2 + o_p(1). \tag{3.18}$$

As $f(u)$ is bounded away from 0, we then have

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ \boldsymbol{\gamma}_0(U_i) \otimes X_i \end{pmatrix} X_i^\top B_0 R_n(U_i) &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \Big\{ \sum_{i=1}^{n} \begin{pmatrix} X_i \\ \boldsymbol{\gamma}_0(U_i) \otimes X_i \end{pmatrix} X_i^\top V(U_i) \\ &\qquad \times \frac{1}{nf(U_i)} K_h(U_{ij}) \Big\} X_j\varepsilon_j \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \ell(U_j)V(U_j)X_j\varepsilon_j + \Delta_n, \end{aligned} \tag{3.19}$$

where

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \Big\{ \sum_{i=1}^{n} \begin{pmatrix} X_i \\ \boldsymbol{\gamma}_0(U_i) \otimes X_i \end{pmatrix} X_i^\top V(U_i) \frac{1}{nf(U_i)} K_h(U_{ij}) - \ell(U_j) V(U_j) \Big\} X_j \varepsilon_j.$$

By simple calculation, we have $Var(\Delta_n) = O\{(h^2 + \delta_n)^2\}$ and thus

$$\Delta_n \;=\; O_p(h^2 + \delta_n). \tag{3.20}$$

For E_{n2}^2, by Lemma 3.1 we have $\bar{\boldsymbol{\gamma}} = O_p(h^2 + \delta_n)$,

$$\bar{\boldsymbol{\gamma}} \;=\; \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\gamma}_0(U_i) + \frac{1}{2} E\boldsymbol{\gamma}_0''(U) h^2 + \frac{1}{n} \sum_{i=1}^{n} (B_0^\top W(U_i) B_0)^{-1} B_0^\top X_i \varepsilon_i + o_p(n^{-1/2})$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \eta_i(\alpha_0)}{\partial \alpha} X_i^\top = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} X_i X_i^\top \\ \boldsymbol{\gamma}_0(U_i) \otimes X_i X_i^\top \end{pmatrix} = \bar{\ell} + O_p(n^{-1/2}).$$

It follows from Lemma 3.1 that

$$E_{n2}^2 = \bar{\ell} \Big\{ B_0 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\gamma}_0(U_i) + \frac{1}{2} B_0 E\boldsymbol{\gamma}_0''(U) \sqrt{n} h^2 + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V(U_i) X_i \varepsilon_i \Big\} + o_p(1). \tag{3.21}$$

Equation (3.16) follows from (3.17)-(3.21) and the following fact

$$\bar{\ell} B_0 \frac{1}{2} E\boldsymbol{\gamma}_0''(U) h^2 - E\{\ell(U) B_0 \frac{1}{2} \boldsymbol{\gamma}_0''(U)\} n^{-1/2} h^2 = -\frac{1}{2} E\{(\ell(U) - \bar{\ell}) B_0 \boldsymbol{\gamma}_0''(U)\} h^2.$$

This completes the proof. □

Proof of Theorem 3.3. The result of theorem 3.3 can be easily seen from (3.12).

Proof of Theorem 3.4.

For ease of exposition, denote $U_i - u$ by U_{iu} and $U_i - U_j$ by U_{ij}. For any fixed d, denote the estimators of θ_0, B_0 and $\boldsymbol{\gamma}_0(u)$ by $\hat{\theta}_d$, $\hat{B}_d$ and $\hat{\boldsymbol{\gamma}}_d(u)$ respectively. By the proof of Theorem 1, $\hat{\theta}_d - \theta_0 = O_p(h^2 + \delta_n)$, and that there exist nonrandom matrix B_d and function $\boldsymbol{\gamma}_d(u)$ such that

$$\hat{B}_d - B_d = O_p(h^2 + \delta_n), \quad \hat{\boldsymbol{\gamma}}_d(u) - \boldsymbol{\gamma}_d(u) = O_p(h^2 + \delta_n)$$

uniformly for $u \in \mathscr{D}$. By the definition of d_0, if $d \geq d_0$ then $B_d\boldsymbol{\gamma}_d(u) = B_0\boldsymbol{\gamma}_0(u)$, and if $d < d_0$ then $E||B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U)|| > 0$. It is easy to see by the above facts and the CLT that

$$\begin{aligned}
\hat{\sigma}_d^2 &= n^{-1}\sum_{i=1}^{n}\{Y_i - (\theta_0 + B_d\boldsymbol{\gamma}_d(U_i))^\top X_i\}^2 + O_p(h^2 + \delta_n) \\
&= n^{-1}\sum_{i=1}^{n}\{\varepsilon_i - (B_0\boldsymbol{\gamma}_0(U_i) - B_d\boldsymbol{\gamma}_d(U_i))^\top X_i\}^2 + O_p(h^2 + \delta_n) \\
&= n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 - 2n^{-1}\sum_{i=1}^{n}\varepsilon_i(B_0\boldsymbol{\gamma}_0(U_i) - B_d\boldsymbol{\gamma}_d(U_i))^\top X_i \\
&\quad + n^{-1}\sum_{i=1}^{n}\{(B_0\boldsymbol{\gamma}_0(U_i) - B_d\boldsymbol{\gamma}_d(U_i))^\top X_i\}^2 + O_p(h^2 + \delta_n) \\
&= \sigma^2 + E\{(B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U))^\top X\}^2 + O_p(h^2 + \delta_n + n^{-1/2}). \quad (3.22)
\end{aligned}$$

Therefore, as a special case we have $\hat{\sigma}_{d_0} = \sigma^2 + O_p(h^2 + \delta_n + n^{-1/2})$. Note that

$$\begin{aligned} E\{(B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U))^\top X\}^2 &= E\{(B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U))^\top W(U)(B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U))\} \\ &\geq \lambda_1(W(u))E||B_0\boldsymbol{\gamma}_0(U) - B_d\boldsymbol{\gamma}_d(U)|| \stackrel{def}{=} c_0 > 0. \end{aligned}$$

Therefore, for $d < d_0$ we have $\hat{\sigma}_d^2 \geq \sigma_{d_0}^2 + c_0 + O_p(h^2 + \delta_n + n^{-1/2})$. Therefore

$$P\Big\{\text{BIC}(d) > \text{BIC}(d_0)\Big\} \to 1 \text{ for any } d < d_0. \tag{3.23}$$

Case 2. ($d \geq d_0$, overfitted model) For ease of exposition, we only consider the case that ε_i is independent of (X_i, U_i). If $d > d_0$, following the same argument of Theorem 3.2 and Lemma 3.1 we have $\hat{\theta}_d - \theta_0 = O_p(n^{-1/2})$ and

$$\begin{aligned} B_d\boldsymbol{\gamma}_d(u) - B_0\boldsymbol{\gamma}_0(u) &= \frac{1}{2}\mu_2 B_d\boldsymbol{\gamma}_d''(u)h^2 + B_d\{nf(u)B_d^\top W(u)B_d\}^{-1}B_d^\top \sum_{i=1}^n K_h(U_{iu})X_i\varepsilon_i \\ &\quad + O_p(n^{-1/2} + h^3 + h\delta_n + \delta_n^2). \end{aligned}$$

where $O_p(n^{-1/2} + h^3 + h\delta_n + \delta_n^2)$ are independent of ε_i. Thus, by CLT we have

$$\begin{aligned} \hat{\sigma}_d^2 &= n^{-1}\sum_{j=1}^n \Big(\varepsilon_j - \frac{1}{2}\mu_2 B_d\boldsymbol{\gamma}_d''(U_j)h^2 - B_d\{nf(U_j)B_d^\top W(U_j)B_d\}^{-1}B_d^\top \sum_{i=1}^n K_h(U_{ij})X_i\varepsilon_i\Big)^2 \\ &\quad + O_p\{n^{-1/2}(n^{-1/2} + h^3 + h\delta_n + \delta_n^2)\} \\ &= n^{-1}\sum_{i=1}^n \varepsilon_i^2 - 2n^{-1}\sum_{j=1}^n \varepsilon_i B_d\{nf(U_j)B_d^\top W(U_j)B_d\}^{-1}B_d^\top \sum_{i=1}^n K_h(U_{ij})X_i\varepsilon_i \\ &\quad + \frac{1}{4}\mu_2^2 E\{B_d\boldsymbol{\gamma}_d''(U)\}^2 h^4 + O_p((nh)^{-1} + n^{-1/2}h^2 + n^{-1}). \end{aligned}$$

It is easy to see that

$$Var(n^{-1}\sum_{j=1}^{n}\varepsilon_i B_d\{nf(U_j)B_d^\top W(U_j)B_d\}^{-1}B_d^\top\sum_{i=1}^{n}K_h(U_{ij})X_i\varepsilon_i) = O(\frac{1}{n^2h}).$$

Note that $B_d\boldsymbol{\gamma}_d''(U)$ are the same for different $d \geq d_0$. Thus, we have

$$\hat{\sigma}_d^2 = \hat{\sigma}_{d_0}^2 + O_p\{(nh)^{-1} + n^{-1/2}h^2\}.$$

It follows that $\log\hat{\sigma}_d^2 - \log\hat{\sigma}_{d_0}^2 = O_p\{(nh)^{-1} + n^{-1/2}h^2\}$. As a consequence, we have

$$\mathrm{BIC}(d) - \mathrm{BIC}(d_0) = (d-d_0)\frac{\log(nh)}{nh} + O_p\{(nh)^{-1} + n^{-1/2}h^2\},$$

where the first term on the right hand side dominates under the condition (C.4). Hence,

$$P\Big\{\mathrm{BIC}(d) > \mathrm{BIC}(d_0)\Big\} \to 1 \text{ for any } d > d_0. \tag{3.24}$$

Equations (3.23) and (3.24) together imply that $P\{\mathrm{BIC}(d) > \mathrm{BIC}(d_0)\} \to 1$. This further implies that $P(\hat{d} = d_0) = 1$. □

Proof of Theorem 3.5 .

The proof is an adaption to our case of Zou (2006). We first show (3.10).

Let $\tilde{\alpha}^{(n)} = \alpha_0 + u/\sqrt{n}$ where $u = (u_1, \ldots, u_S)^\top \in \mathscr{R}^S$, the objective function (3.9) can be written as a function of u as

$$\tilde{Q}_n(u) = Q_n(\alpha_0 + \frac{u}{\sqrt{n}}) + \lambda_n\sum_{s=1}^{S}\hat{w}_s|\alpha_{0,s} + \frac{u}{\sqrt{n}}|.$$

Let $\tilde{u} = \arg\min_{u \in \mathscr{R}^S} \tilde{Q}_n(u)$ and obviously $\tilde{Q}_n(u)$ is minimized at $\tilde{u}_n = \sqrt{n}(\tilde{\alpha}^{(n)} - \alpha_0)$. Next, write

$$\begin{aligned} D_n(u) &= \tilde{Q}_n(u) - \tilde{Q}_n(0) \\ &= \big(Q_n(\alpha_0 + \frac{u}{\sqrt{n}}) - Q_n(\alpha_0)\big) + \lambda_n \sum_{s=1}^{S} \hat{w}_s\big(|\alpha_{0,s} + \frac{u_s}{\sqrt{n}}| - |\alpha_{0,s}|\big) \\ &\equiv I_{1,n}(u) + I_{2,n}(u), \end{aligned}$$

where $I_{1,n}(u) = Q_n(\alpha_0 + \frac{u}{\sqrt{n}}) - Q_n(\alpha_0)$ is due to the loss function and $I_{2,n}(u)$ is due to the penalty term. From the proof of theorem 2, we know that

$$\begin{aligned} \frac{1}{2n}\frac{\partial^2 Q(\alpha_0)}{\partial\alpha\partial\alpha^\top} &\to \Sigma_0 \text{ in probability,} \\ \frac{1}{2}n^{-\frac{1}{2}}\frac{\partial Q(\alpha_0)}{\partial\alpha} &\xrightarrow{D} Z = N(0, \Sigma_1 + \Sigma_2). \end{aligned}$$

Thus the loss function term

$$I_{1,n}(u) = \frac{1}{\sqrt{n}}u^\top\frac{\partial Q(\alpha_0)}{\partial\alpha} + \frac{1}{2n}u^\top\frac{\partial^2 Q(\alpha_0)}{\partial\alpha\partial\alpha^\top}u(1 + o_p(1)) \xrightarrow{D} 2u^\top Z + u^\top\Sigma_0 u.$$

Now, we consider the limiting behavior of the penalty term $I_{2,n}(u)$. If $s \in \mathscr{A}$, that is $\alpha_{0,s} \neq 0$, then $\hat{w}_s \to |\alpha_{0,s}|^{-\tau}$ in probability and $\sqrt{n}(|\alpha_{0,s} + u_s/\sqrt{n}| - |\alpha_{0,s}|) \to u_s\text{sgn}(\alpha_{0,s})$. Since $\lambda_n/\sqrt{n} \to 0$, we have

$$\frac{\lambda_n}{\sqrt{n}}\hat{w}_s\sqrt{n}(|\alpha_{0,s} + u_s/\sqrt{n}| - |\alpha_{0,s}|) \to 0.$$

If $s \notin \mathscr{A}$ then $\sqrt{n}(|\alpha_{0,s} + u_s/\sqrt{n}| - |\alpha_{0,s}|) = |u_s|$. Since $\sqrt{n}\hat{\alpha}_n = O_p(1)$ and $\lambda_n n^{\frac{\tau-1}{2}} \to \infty$, we have $\frac{\lambda_n}{\sqrt{n}}\hat{w}_s = \lambda_n n^{\frac{\tau-1}{2}}|\sqrt{n}\hat{\alpha}_s^{(n)}|^{-\tau} \to \infty$ in probability. It follows

that

$$D_n(u) \Rightarrow D(u) = \begin{cases} 2(u_{\mathscr{A}})^\top Z_{\mathscr{A}} + (u_{\mathscr{A}})^\top (\Sigma_0)_{\mathscr{A}}(u_{\mathscr{A}}), & \text{if } u_s = 0, \forall s \notin \mathscr{A} \\ \infty, & \text{otherwise} , \end{cases}$$

where $u_{\mathscr{A}}$ and $Z_{\mathscr{A}}$ are the j-th ($j \in \mathscr{A}^{\bar{}}$) elements deleted from u and Z respectively. Note that $D_n(u)$ is convex, and the unique minimum of $D(u)$ is

$$u_{min} = \begin{pmatrix} -\Big((\Sigma_0)_{\mathscr{A}}\Big)^{-1} Z_{\mathscr{A}} \\ 0 \end{pmatrix},$$

where 0 denotes a vector of zeros. Following the result of epi-convergence, we have

$$\tilde{\alpha}^{(n)}_{\mathscr{A}} \xrightarrow{D} \Big((\Sigma_0)_{\mathscr{A}}\Big)^{-1} Z_{\mathscr{A}} = N\left(0, \Big((\Sigma_0)_{\mathscr{A}}\Big)^{-1} (\Sigma_1 + \Sigma_2)_{\mathscr{A}} \Big((\Sigma_0)_{\mathscr{A}}\Big)^{-1}\right) \tag{3.25}$$

and $\tilde{\alpha}^{(n)}_{\mathscr{A}^{\bar{}}} \to 0$. Now we prove the consistency part. It suffices to show that $\forall s \in \mathscr{A}^{\bar{}}$, $P(s \in \mathscr{A}_n) \to 0$. By the KKT optimality conditions,

$$\frac{1}{\sqrt{n}} \frac{\partial Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s} + \frac{\lambda_n}{\sqrt{n}} \hat{w}_s \text{sgn}(\tilde{\alpha}^{(n)}_s) = 0.$$

If $s \in \mathscr{A}^{\bar{}}$, then

$$\frac{\lambda_n}{\sqrt{n}} \hat{w}_s = \lambda_n n^{\frac{\tau-1}{2}} |\sqrt{n}\hat{\alpha}^{(n)}_s|^{-\tau} \to \infty$$

in probability, whereas

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s} &= \frac{1}{\sqrt{n}} \frac{\partial Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s} + \frac{1}{n} \frac{\partial^2 Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s^2} \sqrt{n}(\tilde{\alpha}^{(n)}_s - \alpha_{0,s})(1 + o_p(1)) \\ &\xrightarrow{D} \text{ some normal distribution} \end{aligned}$$

by (3.25) and Slutsky's theorem. Thus, for $s \in \bar{\mathscr{A}}$,

$$P(s \in \mathscr{A}^{(n)}) \leq P\Big(|\frac{1}{\sqrt{n}}\frac{\partial Q_n(\tilde{\alpha}^{(n)})}{\partial \alpha_s}| = \frac{\lambda_n}{\sqrt{n}}\hat{w}_s\Big) \to 0.$$

We have completed the proof. □

Table 3.1 Estimation results based on 500 replications

Model and (p)	sample Size	correct d_0	corr. (incorr.) zeros in the rows of θ	corr. (incorr.) zeros in the rows of B	estimation errors (and their standard error) VCM	PVCM	PVCM+L_1
	100	98%	0.0(0.0)	4.5(0.0)	0.2287 (0.0411)	0.1741 (0.0457)	0.1409 (0.0340)
I($p = 7$)	200	100%	0.0(0.0)	4.9(0)	0.1578 (0.0270)	0.1121 (0.0249)	0.0910 (0.0243)
	500	100%	0.0(0.0)	5.0(0.0)	0.0972 (0.0149)	0.0742 (0.0140)	0.0576 (0.0139)
			[0(0)]	[5(0)]			
	100	90%	0(0)	2.9(0.1)	0.2584 (0.0494)	0.2129 (0.0399)	0.1887 (0.0363)
II($p = 7$)	200	100%	0.0(0.0)	3.0(0.0)	0.1721 (0.0275)	0.1407 (0.0271)	0.1243 (0.0273)
	500	100%	0(0)	3.0(0.0)	0.1117 (0.0137)	0.0873 (0.0132)	0.0861 (0.0135)
			[0(0)]	[3(0)]			
	100	93%	5.9(0.0)	8.8(0.1)	0.2796 (0.0530)	0.2114 (0.0449)	0.0998 (0.0378)
I($p = 13$)	200	100%	6.0(0.0)	9.0(0.0)	0.1749 (0.0216)	0.1327 (0.0226)	0.0617 (0.0151)
	500	100%	6.0(0.0)	9.0(0.0)	0.1030 (0.0130)	0.0694 (0.0110)	0.0365 (0.0081)
			[6(0)]	[9(0)]			
	100	86%	5.3(0.4)	4.9(1.8)	0.3701 (0.0782)	0.2651 (0.0476)	0.2273 (0.0393)
II($p = 13$)	200	97%	5.6(0.0)	5(0.3)	0.2094 (0.0298)	0.1478 (0.0201)	0.1161 (0.0211)
	500	100%	6.0(0.0)	5.0(0.0)	0.1241 (0.0122)	0.0884 (0.0101)	0.0759 (0.0105)
			[6(0)]	[5(0)]			
	100	72%	13.8(0.1)	14.4(2)	0.3409 (0.0867)	0.2919 (0.0931)	0.1180 (0.0551)
I($p = 21$)	200	%	14.0(0.0)	15.0(0.0)	0.1878 (0.0231)	0.1400 (0.0217)	0.0485 (0.0191)
	500		14.0(0.0)	15.0(0.0)	0.1124 (0.0099)	0.0719 (0.0102)	0.0298 (0.0064)
			[14(0)]	[15(0)]			
	100	84%	12.0(0.4)	9.0(5.7)	0.5395 (0.1045)	0.4305 (0.1025)	0.3197 (0.0510)
II($p = 21$)	200	92%	13.5(0.0)	9.0(1.0)	0.2559 (0.0300)	0.1704 (0.0214)	0.1242 (0.0173)
	500	100%	13.9(0.1)	9.0(0.2)	0.1334 (0.0104)	0.0876 (0.0082)	0.0584 (0.0082)
			[14(0)]	[9(0)]			

Table 3.2 Estimation results for the Boston House Price Data

coefficient	LSTAT	CRIM	RM	TAX	NOX	PTRATIO	AGE
θ_0	-0.5478 (0.0586)	0 (—)	0.1968 (0.0701)	-0.2335 (0.0482)	-0.1325 (0.0526)	-0.1756 (0.0235)	0 (—)
B_0	-0.2683 (0.1924)	0.3026 (0.1999)	0.6068 (0.1762)	0 (—)	0.2743 (0.1292)	0 (—)	0 (—)
	ZN	INDUS	CHAS	DIS	RAD	B	
θ_0	0.1003 (0.0390)	0 (—)	0 (—)	-0.2373 (0.0413)	0.3749 (0.0602)	0.1381 (0.0526)	
B_0	0 (—)	0 (—)	0 (—)	0 (—)	0.6245 (0.2303)	0 (—)	

Table 3.3 Average prediction errors of 1000 partitions

Size of training set	Linear model	VCM	PVCM	PVCM + penality
200	0.3028	0.9312	0.2514	0.2434
300	0.2918	0.8210	0.2349	0.2262
400	0.2866	0.8661	0.2274	0.2215

CHAPTER 4

Shrinkage Estimation on Covariance Matrix

4.1 Introduction

Recent development in applying L_1 penalty for the estimation and variable selection of parametric models is an effective way towards the challenging "small-n-large-P" problems that are encountered often in data analysis, especially in genetic analysis, finance study and other disciplines; see, for example, Fan and Li (2001), Wang *et al.* (2007) and Zou (2006). Compared with traditional estimation methods, its major advantage is the simultaneous execution of both parameter estimation and variable selection. In particular, allowing an adaptive amount of shrinkage for

each regression coefficient results in an estimator as efficient as oracle. Furthermore, the L_1 penalty approach has very good computational properties; see, for example, Osborne *et al* (2000) and Efron *et al* (2004) for more details.

However, current penalty approaches mainly aim to shrink some coefficients to exactly zero thus obtaining the sparsity in the unknown parameter and reducing the dimension. She (2010) proposed a clustered Lasso which incorporates the L_1-penalty for clustering into the objective function, where the motivation is to group relevant variables based on their effects in the microarray study. In this Chapter, we will study the rationality of clustering effect through several examples, and extend the clustering idea to the covariance matrix estimation after giving a clear framework of clustered Lasso regression. Simulation studies and real data analysis will then be provided, suggesting that the clustering idea applied to both covariance matrix estimation and regression can have better prediction performance than the sparsity assumption in some situations.

Consider model

$$Y_i = \beta_0^\top X_i + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{4.1}$$

where $X_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{ip})^\top$ and $\beta_0 = (\beta_{10}, ..., \beta_{p0})^\top$. When p is large, the estimation of model (4.1) is very unstable or even impossible for example when $n < p$. Therefore some constraints must be imposed such that the model can be simplified and estimated.

A fundamental assumption for the estimation to have better efficiency is the sparsity assuming that many of the coefficients $\beta_{10}, ..., \beta_{p0}$ are exactly 0. Statistically, sparsity is equivalent to variable selection, which is one of the reasons for

sparsity to get the most attention of statisticians, though its appropriateness is debatable as in a survey study recently by Siegfried (2010). Alternative ways of constraining the model complexity have been proposed. For example, the fused Lasso of Tibishrani *et al* (2005) assumes that the coefficients changes according to a specified index. The group Lasso of Yuan and Lin (2006) assumes that the coefficients in a group are either all 0 or all nonzero.

One of the motivations for the research in this Chapter is the well-known fact in epidemiology and other disciplines that there are often cumulative effect which is "individually minor, but collectively significant" (http://ceq.hss.doe.gov/). Simply speaking, it is the summation of a set of variables that generates effect. As an example, air pollutants are cumulated in human body and thus cause health problems. More precisely, let $NO_{2,t}$ be the average concentration of NO_2 on day t. To investigate its effect on public health y_t, say the number of hospital admission on day t, one can use a linear regression model

$$y_t = a_0 + a_1 NO_{2,t-1} + ... + a_p NO_{2,t-p} + \varepsilon_t,$$

where p is very large due to the fact that pollutants cannot be cleaned easily from human body especially from the lungs. In terms of statistical modelling, direct estimation of the model will result in very unstable estimate that has negative coefficients which is hard to be interpreted, and has very bad prediction ability. See the calculation in Figure 4.3 with the penalty parameter $\lambda = 0$. Actually, it is known in epidemiology that the effect of air pollutant is through cumulation of

pollutants in a period of time, i.e.

$$y_t = a_0 + a_1 \sum_{k=1}^{q} NO_{2,t-k} + a_2 \sum_{k=q+1}^{p} NO_{2,t-k} + \varepsilon_t,$$

where $a_1 > 0$ but $a_2 = 0$. See Xia and Tong (2006) for more details. In this model, some of coefficients are the same and are clustered, which is what we are going to investigate in this Chapter.

Another motivation for this research is the well-known factor models in finance study, such as the Capital Asset Pricing Model (CAPM) and the three-factor model of Famma and French (1993). CAPM states that the returns of any individual security or a portfolio, denoted by $\mathbf{x}_1, ..., \mathbf{x}_p$ and their values at time t by $\mathbf{x}_{t,1}, ..., \mathbf{x}_{t,p}$, have a common factor namely the market return, denoted by R_m or $R_{t,m}$ at time t, which is roughly the overall performance of all the individual securities, i.e. $R_{t,m} = \sum_{i=1}^{p} \mathbf{x}_{t,i}$. Therefore, when we try to make prediction of one portfolio based on all the other portfolios, the theory in finance suggests the following model

$$\mathbf{x}_{t,k} = \alpha_k + b_k R_{t-1,m} + \varepsilon_{t,k} = \alpha_k + b_k \sum_{i=1}^{p} \mathbf{x}_{t-1,i} + \varepsilon_{t,k},$$

where $\varepsilon_{t,k}$ are random noise. In other words, all the coefficients in the above model are clustered to 1 group with the same values. In other cases, Famma and French (1993) suggested that there are 3 clusters of coefficients corresponding to the 3 factors in finance.

Figure 4.1 shows the calculations based on the data provided by Professor Kenneth French for the monthly returns of 100 portfolios in the past years. See,

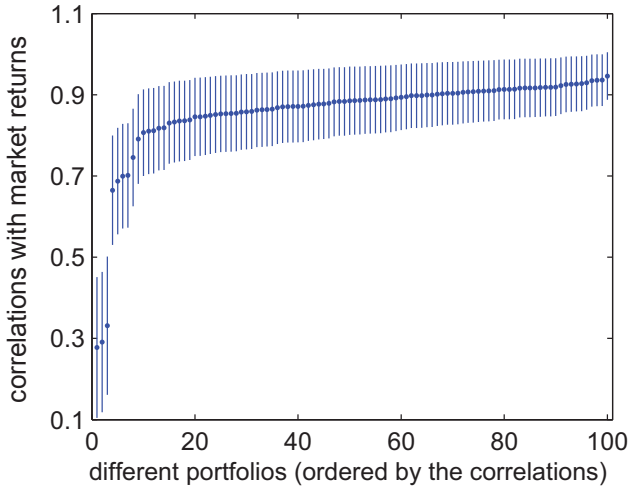

Figure 4.1 The correlation coefficients between each individual of 100 portfolios and the market performance.

e.g., http://mba.tuck.dartmouth.edu /pages/faculty/ken.french/data_library.html. The 95% confidence interval for each coefficient is also plotted by the vertical bar, showing that none of the coefficient is close to 0. It suggests that there are clusters of portfolios, for example $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, that have the same loading in the common factor. As a consequence, many elements, for example $\sigma_{k1} = cov(\mathbf{x}_k, \mathbf{x}_1), \sigma_{k2} = cov(\mathbf{x}_k, \mathbf{x}_2)$ and $\sigma_{k3} = cov(\mathbf{x}_k, \mathbf{x}_3)$ for any $k > 3$, in the covariance matrix $\Sigma_0 = (\sigma_{ij})_{1 \leq i,j \leq p}$ should have same value.

This kind of patterns in the covariance matrix were frequently observed in finance analysis. Tsay (2011) calculated the correlation coefficients of the daily returns of 9 companies, with stock symbols AIC, BA, BAC, GS, INTC, JPM, MS, PG and WFC respectively in New York stock exchange market, from 2000 to 2009 for 2515 observations. Based on statistical hypothesis testing, he clustered the correlation coefficient matrix in blocks as shown in Table 1. It is interesting to see that many of the correlation coefficients have the same values, and there is no zero

values in the matrix.

Table 4.1 Correlation coefficient matrix for the daily returns of 9 stocks

	AIG	GS	MS	BAC	JPM	WFC	BA	INTC	PG
AIG	1.00	0.58	0.58	0.56	0.56	0.56	0.35	0.35	0.35
GS	0.58	1.00	0.58	0.56	0.56	0.56	0.35	0.35	0.35
MS	0.58	0.58	1.00	0.56	0.56	0.56	0.35	0.35	0.35
BAC	0.56	0.56	0.56	1.00	0.69	0.69	0.35	0.35	0.35
JPM	0.56	0.56	0.56	0.69	1.00	0.69	0.35	0.35	0.35
WFC	0.56	0.56	0.56	0.69	0.69	1.00	0.35	0.35	0.35
BA	0.35	0.35	0.35	0.35	0.35	0.35	1.00	0.27	0.27
INTC	0.35	0.35	0.35	0.35	0.35	0.35	0.27	1.00	0.27
PG	0.35	0.35	0.35	0.35	0.35	0.35	0.27	0.27	1.00

The above examples suggest that the parameters in many statistical estimations have clusters in each of which the parameters take the same values. In this Chapter, we are going to incorporate the clusters into the estimation method and improve the estimation efficiency.

4.2 Coefficients Clustering of Regression

Motivated by the examples and discussions above, we consider a model with several clusters of coefficients, in each cluster the coefficients are the same, i.e.

$$Y = \phi_0 + \phi_1 \sum_{\ell \in A_1} \mathbf{x}_\ell + ... + \phi_L \sum_{\ell \in A_L} \mathbf{x}_\ell + \varepsilon. \tag{4.2}$$

We call model (4.2) the clustered model. It is easy to see that the sparsity assumption is a special case of the clustered model. Another motivation for this model is based on the approximation of a general model by a reduced model with a smaller number of coefficient values. It is known that with sample size n, we can only

estimate at most $p = n - 1$ parameters. When there are more than n parameters in the model, a straightforward approach is to find a model with $L(< n)$ parameters values that can best approximate the original model. The clustered model tries to group the parameters into $L < \min\{p, n\}$ clusters, and approximate the parameters in each cluster by one single value, in order to minimize the difference of the approximation, i.e.

$$\min_{\phi_0,\phi_1,...,\phi_L} \Big\{ \min_{\substack{\beta_k \in \{\phi_1,...,\phi_L\} \\ k=1,2,...,p}} E\{Y - (\phi_0 + \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p)\}^2 \Big\}.$$

For ease of exposition, after rearranging the order of $\mathbf{x}_i, i = 1, ..., p$, we can assume that

$$A_1 = \{1, ..., p_1\}, A_2 = \{p_1 + 1, ..., p_1 + p_2\}, ..., A_L = \{\sum_{i=1}^{L-1} p_i + 1, ..., \sum_{i=1}^{L-1} p_i + p_L\} \quad (4.3)$$

where $\sum_{i=1}^{L} p_i = p$.

The model (4.2) has strong link with the fused Lasso of Tibshirani *et al* (2005), where the clusters are usually based on the orders of a specified index. In this setting, the clustering is allowed to be more flexible and the index may not exist. This is more realistic because in many cases variates are arranged randomly. The difference between this model and the grouped Lasso by Yuan and Lin (2006) is obvious. The grouped Lasso cares about whether a group of variates can be removed from the model as a whole, while model (4.2) cares about whether those variates share the same coefficients.

Since clusters $A_1, ..., A_L$ in model (4.2) are unknown, we can only start with

the general model (4.1) and identify the clusters later. Suppose that $\hat{\beta}$ is a root-n consistent estimator of β_0, which will be taken as the initial value. More details about the initial value will be discussed later. Define weights $\hat{w}_{jk} = 1/|\hat{\beta}_j - \hat{\beta}_k|^\alpha$ for some $\alpha \geq 0$, and weight matrix $W = (\hat{w}_{jk})$. Consider the following objective function

$$Q(\beta) = \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 + \lambda_n \sum_{k=1}^{p}\sum_{j=k+1}^{p} \hat{w}_{jk}|\beta_j - \beta_k|, \tag{4.4}$$

and let

$$\hat{\beta}^{*(n)} = \arg\min_{\beta} Q(\beta). \tag{4.5}$$

We call $\hat{\beta}^{*(n)}$ the clustered Lasso estimator (cLasso).

To make the idea clear and state the theory easily, we introduce the following variable transformation. For (4.3), without loss of generality we assume that the variables are rearranged such that values in β_0 are in a descending order, i.e.

$$\beta_0 = (\beta_{1,0}, \ldots, \beta_{p_1,0}, \beta_{p_1+1,0}, \ldots, \beta_{p_1+p_2,0}, \ldots, \beta_{p,0})^\top,$$

where $\beta_{1,0} \geq \beta_{2,0} \geq ... \geq \beta_{p,0}$; otherwise we can apply a permutation matrix to β_0. The design matrix X and the weight matrix W are also arranged accordingly. Let

$\Gamma = \text{diag}(\Gamma_1,\ \Gamma_2,\ ...,\ \Gamma_L)$ with

$$\Gamma_j = \begin{pmatrix} 1/p_j & 1/p_j & 1/p_j & ... & 1/p_j \\ 1 & -1 & 0 & ... & 0 \\ 1 & 0 & -1 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & ... & -1 \end{pmatrix}_{p_j \times p_j}, \text{ for } j = 1, \ldots, L.$$

Define $M_l = p_0 + p_1 + \ldots + p_{l-1}$ where $p_0 = 0, l = 1, \ldots, L, L+1$. We have

$$\gamma_0 := \Gamma\beta_0 := (\gamma_1^0, \gamma_{12}^0, \ldots, \gamma_{1p_1}^0, \gamma_2^0, \gamma_{22}^0, \ldots, \gamma_{2p_2}^0, \ldots, \gamma_L^0, \gamma_{L2}^0, \ldots, \gamma_{Lp_L}^0)^\top, \quad (4.6)$$

where for $l = 1, \ldots, L$,

$$\gamma_l^0 = \frac{1}{p_l}\sum_{i=1}^{p_l} \beta_{M_l+i,0},\ \ \gamma_{lk}^0 = \beta_{M_l+1,0} - \beta_{M_l+k,0} = 0,\ \ k = 2, \ldots, p_l. \quad (4.7)$$

Finally, using a permutation matrix P, we get

$$\begin{aligned} P\Gamma\beta_0 &= (\gamma_1^0, \ldots, \gamma_L^0, \gamma_{12}^0, \ldots, \gamma_{1p_1}^0, \ldots, \gamma_{L2}^0, \ldots, \gamma_{Lp_L}^0)^\top \quad (4.8) \\ &= (\gamma_1^0, \gamma_2^0, \ldots, \gamma_L^0, \underbrace{0, \ldots, 0}_{p-L})^\top. \end{aligned}$$

We use the operator $\mathscr{G}(\cdot)$ to denote these transformations, that is, $\mathscr{G}(\beta) = P\Gamma\beta$ for any vector $\beta \in \mathbb{R}^p$. Accordingly, for the clustered Lasso estimator $\hat{\beta}^{*(n)}$ defined in (4.5),

$$\mathscr{G}(\hat{\beta}^{*(n)}) = P\Gamma\hat{\beta}^{*(n)} := (\hat{\phi}_1^{*(n)}, \hat{\phi}_2^{*(n)}, \ldots, \hat{\phi}_p^{*(n)})^\top. \quad (4.9)$$

Let

$$\tilde{X}_i = P(\Gamma^{-1})^\top X_i, \quad i = 1, \ldots, n,$$

then model $Y_i = X_i^\top \beta + \varepsilon_i$ is transformed to

$$Y_i = \tilde{X}_i^\top \mathscr{G}(\beta) + \varepsilon_i. \tag{4.10}$$

Let $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n)^\top$, then $\tilde{X} = X\Gamma^{-1}P^\top$, where the design matrix $X = (X_1, \ldots, X_n)^\top$. Define the index sets $\mathscr{A}^+ = \{1, 2, \ldots, L\}$, $\mathscr{A}^- = \{L+1, \ldots, p\}$ and

$$\mathscr{A}_n^+ = \{j \in \{1, \ldots, p\}, \hat{\phi}_j^{*(n)} \neq 0\}, \ \mathscr{A}_n^- = \{1, \ldots, p\} - \mathscr{A}_n^+.$$

The corresponding design matrix $\tilde{X}$ can also be written as

$$\tilde{X} := (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_p) = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_L, \tilde{\mathbf{x}}_{L+1}, \ldots, \tilde{\mathbf{x}}_p) := (\tilde{X}_{\mathscr{A}^+}, \ \tilde{X}_{\mathscr{A}^-}),$$

where $\tilde{X}_{\mathscr{A}^+}$ denotes an $n \times L$ matrix composed of the elements of $\tilde{X}$ with the column index belonging to $\mathscr{A}^+$, and $\tilde{X}_{\mathscr{A}^-}$ denotes an $n \times (p-L)$ matrix composed of the elements of $\tilde{X}$ with the column index belonging to $\mathscr{A}^-$.

Assume that the design matrix X satisfies

$$C = \lim_{n \to \infty} n^{-1} X^\top X \tag{4.11}$$

where C is a positive definite matrix. Then

$$\tilde{C} := \lim_{n\to\infty} n^{-1}\tilde{X}^\top \tilde{X} = P(\Gamma^{-1})^\top C \Gamma^{-1} P^\top \tag{4.12}$$

is also positive definite.

Theorem 4.2.1. Suppose $\varepsilon_1, ..., \varepsilon_n$ are IID with mean 0 and variance σ^2, and that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\alpha-1)/2} \to \infty$. Then the cLasso estimator $\hat{\beta}^{*(n)}$ must satisfy

(1) Consistency in variable clustering: $\lim_{n\to\infty} P(\mathscr{A}_n^+ = \mathscr{A}^+) = 1$.

(2) Asymptotic normality: $\sqrt{n}\left(\mathscr{G}(\hat{\beta}^{*(n)})_{\mathscr{A}^+} - \mathscr{G}(\beta_0)_{\mathscr{A}^+}\right) \xrightarrow{D} N(0, \sigma^2 \tilde{C}^{-1}_{\mathscr{A}^+})$, where $\tilde{C}_{\mathscr{A}^+}$ denotes the sub-matrix of $\tilde{C}$ formed by the elements at $\mathscr{A}^+$'s row and $\mathscr{A}^+$'s column of $\tilde{C}$.

To investigate the estimation efficiency, we consider a special case where $\beta_0 = \mathbf{1}_p\beta_{10}$. The simple LSE, denoted by $\hat{\beta}^{(n)}$, satisfies $\sqrt{n}(\hat{\beta}^{(n)} - \mathbf{1}_p\beta_{10}) \to N(0, C^{-1}\sigma^2)$. By the definition of Γ, the estimated β_{10} by cLasso is actually $\hat{\beta}_1^{*(n)} = p^{-1}\mathbf{1}_p^\top \hat{\beta}^{(n)}$. The most efficient and unbiased estimator of β_{10} among the linear combinations of $\hat{\beta}^{(n)}$ is $\ell^\top \hat{\beta}^{(n)}$ where $\ell = (\ell_1, ..., \beta_p)^\top$ with $\ell_1 + ... + \ell_p = 1$ and $\ell^\top C^{-1}\ell = \min$! It is known from the theory of quadratic programming with linear equality constraints that ℓ satisfies

$$\begin{pmatrix} C^{-1} & \mathbf{1}_p \\ \mathbf{1}_p^\top & 0 \end{pmatrix} \begin{pmatrix} \ell \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

where λ is a Lagrange multiplier. The solution is $\ell = \{\mathbf{1}_p^\top C \mathbf{1}_p\}^{-1} C \mathbf{1}_p$. It is easy to see that only when the sums of every row of C are the same, the cLasso estimator achieves the optimal efficiency. Noting that the fused Lasso also has similar problem in its estimation efficiency. The estimator $\ell^\top \hat{\beta}^{(n)}$ is actually the LSE of

model $y = \beta_{10} \sum_{k=1}^{p} \mathbf{x}_k + \varepsilon$. After the blocks are identified by cLasso as shown in Theorem 4.2.1(1), we can achieve optimal estimation efficiency by estimating the reduced model.

When the true parameter β_0 is sparse, i.e., some of the elements are exactly 0, the penalty function (4.4) cannot penalize them to exactly 0. In the above notation, this sparsity means $\gamma_l^0 = 0$ for some $l \in \{1, \ldots, L\}$. Without loss of generality, we assume $\gamma_L^0 = 0$. Otherwise we can use a permutation matrix to rearrange it. To achieve sparsity, we need to impose another penalty. Denote the weight vector by $\hat{\omega}$ with $\hat{\omega}_j = 1/|\hat{\beta}_j|^\alpha, j = 1, \ldots, p$. We add a penalty for sparsity to the penalty function in (4.4) and obtain the following estimator

$$\tilde{\beta}^{*(n)} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 + \lambda_n \Big\{ \sum_{k=1}^{p} \sum_{j=k+1}^{p} \hat{w}_{jk} |\beta_j - \beta_k| + \sum_{j=1}^{p} \hat{\omega}_j |\beta_j| \Big\}. \quad (4.13)$$

Let $\tilde{\mathscr{A}}^+ = \{1, 2, \ldots, L-1\}$, $\tilde{\mathscr{A}}^- = \{L, L+1, \ldots, p\}$ and

$$\tilde{\mathscr{A}}_n^+ = \{j \in \{1, \ldots, p\}, \tilde{\phi}_j^{*(n)} \neq 0\}, \ \tilde{\mathscr{A}}_n^- = \{1, \ldots, p\} - \tilde{\mathscr{A}}_n^+.$$

We have the following theorem.

Theorem 4.2.2. Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\alpha-1)/2} \to \infty$. Then $\tilde{\beta}^{*(n)}$ must satisfy

(1) Consistency in variable clustering: $\lim_{n\to\infty} P(\tilde{\mathscr{A}}_n^+ = \tilde{\mathscr{A}}^+) = 1$.

(2) Asymptotic normality: $\sqrt{n}\left(\mathscr{G}(\tilde{\beta}^{*(n)})_{\tilde{\mathscr{A}}^+} - \mathscr{G}(\beta_0)_{\tilde{\mathscr{A}}^+}\right) \xrightarrow{D} N(0, \sigma^2 \tilde{C}^{-1}_{\tilde{\mathscr{A}}^+})$, where $\tilde{C}_{\tilde{\mathscr{A}}^+}$ denotes the sub-matrix of $\tilde{C}$ formed by the elements at $\tilde{\mathscr{A}}^+$'s row and $\tilde{\mathscr{A}}^+$'s column of $\tilde{C}$.

4.3 Extension to the Estimation of Covariance Matrix

As mentioned in the introduction, the covariance matrix sometimes has clusters of common elements as well. In this section, we extend cLasso to the estimation of covariance matrices. Under sparsity assumption, many methods have been proposed for the estimation of covariance matrix. See, for example, Wu and Pourahmadi (2003), Huang *et al* (2006), Bickel and Levina (2008), Levina *et al* (2008) and Lam and Fan (2009). Without loss of generality, we assume that

$$E(X) = 0,\ cov(X) = \Sigma_0 = (\sigma_{ij})_{i,j=1,\ldots,p},\ \text{and } X \text{ has finite fourth moments.}$$

Given random samples $X_1, \ldots, X_n$ from X, the sample covariance matrix

$$S = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$$

is an unbiased estimator of Σ_0. Let $Z_i = X_i X_i^\top$. Consider the half-vectorization

$$Vech(\Sigma_0) = (\sigma_{11}, \sigma_{12}..., \sigma_{1p}, \sigma_{22}, \sigma_{23}, ..., \sigma_{2p}, ..., \sigma_{pp})^\top$$

and accordingly

$$\nu_0 = Vech(\Sigma_0) \in \mathscr{R}^m, \qquad \hat{\boldsymbol{\nu}}_n = Vech(S), \tag{4.14}$$

where $m = p(p+1)/2$. It follows from the central limit theorem,

$$\sqrt{n}(\hat{\boldsymbol{\nu}}_n - \nu_0) \xrightarrow{D} N(0, W), \tag{4.15}$$

where $W = cov(Vech(Z_i))$. After the vectorization of the covariance matrix, we can view the covariance matrix estimation problem in the following way. Let

$$Y = (\mathbf{x}_1^2, \mathbf{x}_1\mathbf{x}_2, \ldots, \mathbf{x}_1\mathbf{x}_p, \mathbf{x}_2^2, \mathbf{x}_2\mathbf{x}_3, \ldots, \mathbf{x}_2\mathbf{x}_p, \ldots, \mathbf{x}_p^2)^\top \in \mathscr{R}^m \tag{4.16}$$

and $Y_i = (\mathbf{x}_{i1}^2, \mathbf{x}_{i1}\mathbf{x}_{i2}, \ldots, \mathbf{x}_{i1}\mathbf{x}_{ip}, \mathbf{x}_{i2}^2, \mathbf{x}_{i2}\mathbf{x}_{i3}, \ldots, \mathbf{x}_{i2}\mathbf{x}_{ip}, \ldots, \mathbf{x}_{ip}^2)^\top \in \mathscr{R}^m$, where $X = (\mathbf{x}_1, ..., \mathbf{x}_p)^\top$ and $X_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{ip})^\top$. Then estimator $\hat{\boldsymbol{\nu}}_n$ is the same as the LSE of the following regression model

$$Y = \nu_0 + \varepsilon, \tag{4.17}$$

where ε is an m-dimensional random error with zero mean and positive definite covariance matrix. In other words,

$$\hat{\boldsymbol{\nu}}_n = \arg\min_{\nu \in \mathscr{R}^m} \sum_{i=1}^{n} \|Y_i - \nu\|^2 = \arg\min_{\nu \in \mathscr{R}^m} \|\mathscr{Y} - (1_{\mathrm{n}} \otimes \mathrm{I_m})\nu\|^2, \tag{4.18}$$

where $\mathscr{Y} = (Y_1^\top, \ldots, Y_n^\top)^\top \in \mathscr{R}^{mn}$, $1_{\mathrm{n}} = (1, \ldots, 1)^\top \in \mathscr{R}^{\mathrm{n}}$. We propose to use the following cLasso estimation to estimate ν by

$$\hat{\nu}^{*(n)} = \arg\min_{\nu \in \mathscr{R}^m} \left\{ \|\mathscr{Y} - (1_{\mathrm{n}} \otimes \mathrm{I_m})\nu\|^2 + \lambda_{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{m}} \sum_{\mathrm{k}=\mathrm{j}+1}^{\mathrm{m}} \hat{\mathrm{w}}_{\mathrm{jk}} |\nu_{\mathrm{j}} - \nu_{\mathrm{k}}| \right\}, \tag{4.19}$$

where $\hat{w}_{jk} = |\hat{\nu}_j - \hat{\nu}_k|^{-\alpha}$ with $\hat{\boldsymbol{\nu}}_n = (\hat{\nu}_1, \ldots, \hat{\nu}_m)^\top$ given in (4.15), and $\alpha \geq 0$.

To present our theoretical result, we introduce the following variable transformations. Suppose that $\nu_0 = Vech(\Sigma_0)$ has L groups of distinct values and $\tilde{\nu}_0$ is the rearrangement of ν_0 such that the distinct values are arranged in a descending order. That is, let Q be the $m \times m$ permutation matrix such that

$$\tilde{\nu}_0 \quad = \quad Q\nu_0 := (\underbrace{\varphi_1, ..., \varphi_1}_{m_1}, \underbrace{\varphi_2, ..., \varphi_2}_{m_2}, ..., \underbrace{\varphi_L, ..., \varphi_L}_{m_L})^\top, \tag{4.20}$$

where $\varphi_1 > \varphi_2 > \ldots > \varphi_L$. Correspondingly, let $\mathscr{M} = \{1, \ldots, m\}$ and

$$\mathscr{M}_l = \{\sum_{i=0}^{l-1} m_i + 1, \sum_{i=0}^{l-1} m_i + 2, ..., \sum_{i=0}^{l-1} m_i + m_l\}$$

such that $\cup_{l=1}^{L} \mathscr{M}_l = \mathscr{M}$, $\mathscr{M}_l \cap \mathscr{M}_s = \emptyset$, for $l \neq s$, where $l, s = 1, 2, ..., L$ and $m_0 = 0$. Moreover, denote $M_j = m_0 + m_1 + \ldots + m_{j-1}$, $j = 1, \ldots, L, L+1$, and

$$\mathscr{M}^+ = \{M_1 + 1, M_2 + 1, \ldots, M_L + 1\}, \ \mathscr{M}^- = \mathscr{M} - \mathscr{M}^+. \tag{4.21}$$

Let $\tilde{Y} = QY$, $\tilde{\varepsilon} = Q\varepsilon$. It follows that the linear regression model in (4.17) can be written as

$$\tilde{Y} = \tilde{\nu}_0 + \tilde{\varepsilon}. \tag{4.22}$$

Let $\Gamma = \text{diag}(\Gamma_1,\ \Gamma_2,\ ...,\ \Gamma_L)$, where

$$\Gamma_l = \begin{pmatrix} 1/m_l & 1/m_l & 1/m_l & ... & 1/m_l \\ 1 & -1 & 0 & ... & 0 \\ 1 & 0 & -1 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & ... & -1 \end{pmatrix}_{m_l \times m_l}, \text{ for } l = 1, \ldots, L \tag{4.23}$$

then

$$\Gamma\tilde{\nu}_0 \ = \ (\varphi_1, \underbrace{0, \ldots, 0}_{m_1-1}, \varphi_2, \underbrace{0, \ldots, 0}_{m_2-1}, \ldots, \varphi_L, \underbrace{0, \ldots, 0}_{m_L-1})^\top.$$

There is an $m \times m$ permutation matrix P such that

$$P\Gamma\tilde{\nu}_0 = (\phi_1, \ldots, \phi_L, \underbrace{0, \ldots, 0}_{m-L})^\top. \tag{4.24}$$

Finally, we use $\mathscr{T}(\nu_0)$ to denote the combination of the above transformations, that is,

$$\mathscr{T}(\nu_0) = P\Gamma Q\nu_0 = (\phi_1, \ldots, \phi_L, \underbrace{0, \ldots, 0}_{m-L})^\top. \tag{4.25}$$

The linear regression model in (4.17) thus becomes

$$\mathscr{T}(Y) = \mathscr{T}(\nu_0) + \mathscr{T}(\varepsilon). \tag{4.26}$$

For the cLasso estimator $\hat{\nu}^{*(n)}$ defined in (4.19), denote

$$\mathscr{T}(\hat{\nu}^{*(n)}) = P\Gamma Q\hat{\nu}^{*(n)} := (\hat{\tau}_1^{*(n)}, \hat{\tau}_2^{*(n)}, \ldots, \hat{\tau}_m^{*(n)})^\top. \tag{4.27}$$

Let $\mathscr{A}^+ = \{1, 2, \ldots, L\}$, $\mathscr{A}^- = \{L+1, \ldots, m\}$ and

$$\mathscr{A}_n^+ = \{j \in \{1, \ldots, m\}, \hat{\tau}_j^{*(n)} \neq 0\}, \ \mathscr{A}_n^- = \{1, \ldots, m\} - \mathscr{A}_n^+.$$

We have the following theorem.

Theorem 4.3.1. Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\alpha-1)/2} \to \infty$. Then the adaptive Lasso estimator $\hat{\phi}^{*(n)} = \mathscr{T}(\hat{\nu}^{*(n)})$ must satisfy

(1) Consistency in variable clustering: $\lim_{n\to\infty} P(\mathscr{A}_n^+ = \mathscr{A}^+) = 1$.

(2) Asymptotic normality:

$$\sqrt{n}(\mathscr{T}(\hat{\nu}^{*(n)})_{\mathscr{A}^+} - \mathscr{T}(\nu_0)_{\mathscr{A}^+}) \xrightarrow{D} N(0, G_{\mathscr{A}^+}),$$

where the asymptotic covariance matrix $G = P\Gamma QWQ^\top\Gamma^\top P^\top$ with W given in (4.15).

Again as in fused Lasso and the discussion after Theorem 4.2.1, the cLasso estimator may not always be the most efficient. If $\sigma_{i_1j_1}, ..., \sigma_{i_kj_k}$ are identified to be in one cluster with the same value $\tilde{\sigma}$. Let $U = (\mathbf{x}_{i_1}\mathbf{x}_{j_1}, ..., \mathbf{x}_{i_k}\mathbf{x}_{j_k})^\top$, then $EU = \mathbf{1}_k\tilde{\sigma}$. Let $V = cov(U)$. Following the same argument after Theorem 4.2.1, a more efficient estimator of $\tilde{\sigma}$ is $\ell^\top U$, where $\ell = (\mathbf{1}_k^\top V\mathbf{1}_k)^{-1}V\mathbf{1}_k$. Since V is unknown, the optimal efficient estimator cannot be obtained easily. Consider a

simple factor model as an example. Suppose $\mathbf{x}_k = aF_1 + \epsilon_k, k = 1, 2$ and F_1 is independent of ϵ_k. If ϵ_1 and ϵ_2 are IID with variance σ_e^2, and are independent of $\mathbf{x}_3$, then $U = (\mathbf{x}_1\mathbf{x}_3, \mathbf{x}_2\mathbf{x}_3)^\top, cov(\mathbf{x}_1, \mathbf{x}_3) = cov(\mathbf{x}_2, \mathbf{x}_3)$ and

$$V = cov(U) = \begin{pmatrix} a^2E(F_1^2\mathbf{x}_3^2) + \sigma_e^2 var(\mathbf{x}_3^2) & a^2E(F_1^2\mathbf{x}_3^2) \\ a^2E(F_1^2\mathbf{x}_3^2) & a^2E(F_1^2\mathbf{x}_3^2) + \sigma_e^2 var(\mathbf{x}_3^2) \end{pmatrix}.$$

Note that the sum of each rows of V are the same. Thus the cLasso achieve the optimal estimation efficiency. In other words, for factor models, cLasso estimators can achieve the optimal efficiency under some weak conditions.

4.4 Simulations

To implement the estimation, we propose the following quadratic approximation algorithm; see Fan and Li (2001) for more details. For ease of exposition, we only give the details for cLasso with $\alpha = 0$, i.e.

$$Q(\beta, \lambda_n) = \sum_{i=1}^{n}(Y_i - X_i^\top\beta)^2 + \lambda_n \sum_{k=1}^{p}\sum_{j=k+1}^{p}|\beta_j - \beta_k|.$$

With an initial estimator of β, denoted by $\tilde{\beta}$, we approximate $Q(\beta, \lambda_n)$ by

$$\tilde{Q}(\beta, \lambda_n) = \sum_{i=1}^{n}(Y_i - X_i^\top\beta)^2 + \lambda_n \sum_{k=1}^{p}\sum_{j=k+1}^{p}\frac{(\beta_j - \beta_k)^2}{|\tilde{\beta}_j - \tilde{\beta}_k|}.$$

We have

$$\frac{\partial \tilde{Q}}{\partial \beta} = 2X^\top X\beta - 2X^\top Y + 2\lambda_n(W_0(\tilde{\beta}) - W_1(\tilde{\beta}))\beta,$$

where

$$W_0(\tilde{\beta}) = diag(\sum_{i\neq 1}|\tilde{\beta}_1 - \tilde{\beta}_i|^{-1},\ \sum_{i\neq 2}|\tilde{\beta}_2 - \tilde{\beta}_i|^{-1},\ ...,\ \sum_{i\neq p}|\tilde{\beta}_p - \tilde{\beta}_i|^{-1})$$

and

$$W_1(\tilde{\beta}) = \begin{pmatrix} 0 & |\tilde{\beta}_1 - \tilde{\beta}_2|^{-1} & ... & |\tilde{\beta}_1 - \tilde{\beta}_p|^{-1} \\ |\tilde{\beta}_2 - \tilde{\beta}_1|^{-1} & 0 & ... & |\tilde{\beta}_2 - \tilde{\beta}_p|^{-1} \\ \vdots & & & \\ |\tilde{\beta}_p - \tilde{\beta}_1|^{-1} & |\tilde{\beta}_p - \tilde{\beta}_{p-1}|^{-1} & ... & 0 \end{pmatrix}.$$

By solving $\partial\tilde{Q}/\partial\beta = 0$, we update the estimator by

$$\tilde{\beta} := \{X^\top X + \lambda_n(W_0(\tilde{\beta}) - W_1(\tilde{\beta}))\}^{-1}X^\top Y. \tag{4.28}$$

Repeat (4.28) until convergence. The final value is our cLasso estimator.

In the following calculations, we use the estimator from minimizing

$$Q(\beta, \lambda_n) = \sum_{i=1}^{n}(Y_i - X_i^\top\beta)^2 + \lambda_n\sum_{k=1}^{p}\sum_{j=k+1}^{p}(\beta_j - \beta_k)^2$$

as the initial values for which we has a closed form. We use randomized 3-fold cross-validation method to select the penalty parameter λ_n based on 100 times of random splitting. However, all the selection methods for the penalty parameter λ_n are applicable here. See for example Wang *et al* (2007).

Since any estimator $\hat{\beta}$ is obtained for further prediction, for a new subject X the prediction error is $X^\top\hat{\beta} - X^\top\beta$. We thus measure of estimation error by

$$ERR(\hat{\beta}) = (\hat{\beta} - \beta)^\top Var(X)(\hat{\beta} - \beta)/\beta^\top Var(X)\beta.$$

We also evaluate the estimation methods by checking their number of clusters, i.e. the number of different values in the estimated β.

Example 4.4.1. In the following examples, covariates $X = (\mathbf{x}_1, ..., \mathbf{x}_p)^\top$ are normally distributed with mean 0 and variance covariance matrix $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$ and $\sigma_{ij} = 0.5^{|i-j|}$. In model $Y = \beta^\top X + \varepsilon$, we consider two settings of coefficients,

$$\begin{aligned} (I) \qquad & \beta = (\underbrace{2, -1}, \underbrace{2, -1}, ...)^\top, \\ (II) \qquad & \beta = -p/4 + (0, 0.5, 1, 1.5, ...)^\top. \end{aligned}$$

For the first setting (I), the segment underbraced is repeated in β. In the second setting, the parameter increases by 0.5 each element. For each setting, different signal-noise ratios (SNR) are considered. The first setting is a desired model for cLasso. We use this example to check the theory of the estimation method such as the number of clusters in the estimated coefficients and the estimation efficiency as well. Our simulation results listed in Table 4.2 suggest that as $p < n$ and SNR is big or n/p is big, cLasso can indeed cluster the coefficients appropriately; see the last column of Table 4.2. The estimation efficiency is much better than Lasso as shown in columns 3-5. The efficiency is also satisfactory even when p is larger than n as compared with ridge regression and Lasso regression. This superior efficiency over ridge regression or Lasso regression is not surprising because the model is in

Table 4.2 Simulation results for setting (I) based on sample size $n = 40$ and 100 replications

		ERR			no. of clusters	
p	SNR	ridge	Lasso	cLasso	Lasso	cLasso
10	2	0.0919	0.2601	0.0714	7.52	4.99
10	4	0.0248	0.0656	0.0177	9.92	3.88
10	8	0.0045	0.0161	0.0025	10.00	2.71
20	2	0.2131	0.5522	0.2222	10.35	6.89
20	4	0.0626	0.1339	0.0264	18.54	5.69
20	8	0.0145	0.0297	0.0039	20.00	3.54
40	2	0.4484	1.3728	0.5559	9.73	5.37
40	4	0.3218	0.7471	0.3465	12.51	6.97
40	8	0.5497	0.8080	0.1927	12.90	6.74
80	2	0.7029	0.9704	0.6071	5.67	6.22
80	4	0.5919	0.9432	0.5967	4.90	7.17
80	8	0.5607	0.9240	0.5548	5.92	6.72
120	2	0.7818	0.9803	0.6352	3.59	6.77
120	4	0.7000	0.9803	0.5826	4.19	5.67
120	8	0.6806	0.9789	0.5563	4.60	5.20

favor of cLasso. However, calculations for setting (II) below confirms the superior of cLasso in a more general situation.

For setting (II), the coefficients are different from one another and are not in favor of cLasso. However, cLasso still clusters the coefficients as shown in the last column of Table 4.3, and generates much more efficient estimators of the coefficients. It is interesting to see that the clustering is in such a way that the clustered model estimated by cLasso can approximate the true models well as evidenced by the estimation error shown in columns 3-5.

Example 4.4.2. In this example, we consider multivariate random vector $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)^\top$ the following factor model

$$\mathbf{x}_i = 2F_1 + F_3 + \varepsilon_i, i = 1, ..., q$$

Table 4.3 Simulation results for setting (II) based on sample size $n = 40$ and 100 replications

p	SNR	ERR			no. of clusters	
		ridge	Lasso	cLasso	Lasso	cLasso
10	2	0.108	0.1412	0.0998	7.88	6.82
	4	0.0663	0.0838	0.0591	8.38	7.43
	8	0.0419	0.0506	0.0362	8.82	7.87
20	2	0.1654	0.2494	0.1324	13.39	11.76
	4	0.1089	0.1653	0.0859	14.72	12.74
	8	0.0729	0.1111	0.0567	15.92	13.9
40	2	0.2357	0.4536	0.207	27.44	27.36
	4	0.1694	0.2928	0.1301	27.68	26.16
	8	0.1211	0.2071	0.0844	28.94	26.47
80	2	0.4021	0.8595	0.4683	46.82	64.35
	4	0.312	0.5601	0.242	46.08	61.69
	8	0.2435	0.4288	0.1535	46.89	61.31
120	2	0.4947	0.9306	0.4983	57.03	102.05
	4	0.3965	0.6928	0.2858	56.95	97.45
	8	0.3224	0.5727	0.1941	57.87	95.03

and

$$\mathbf{x}_j = F_2 - F_3 + 2\varepsilon_j, i = q+1, ..., p,$$

where F_1, F_2 and F_3 are IID and follow $N(0,1)$ each, $p = 10$ and $q = 5$. We further assume that $\varepsilon_i, i = 1, ..., p$ are IID and independent of F_1, F_2, F_3. Then we have

$$\Sigma_0 = cov(X) = I_{p\times p} + \begin{pmatrix} 5\mathbf{1}_q\mathbf{1}_q^\top & -\mathbf{1}_q\mathbf{1}_{p-q}^\top \\ -\mathbf{1}_{p-q}\mathbf{1}_q^\top & 2\mathbf{1}_{p-q}\mathbf{1}_{p-q}^\top + 3I_{p-q} \end{pmatrix}.$$

Note that the matrix has 3 clusters of values for the non-diagonal elements. To evaluate the estimation efficiency, we define the errors of estimator $\hat{\Sigma}$ by

$$d(\hat{\Sigma}, \Sigma_0) = \mathrm{tr}^{1/2}\{(\hat{\Sigma} - \Sigma_0)^\top(\hat{\Sigma} - \Sigma_0)\}.$$

Our simulation results are listed in Table 4.4, where the number of clusters are counted for elements not in the diagonal. Table 4.4 suggests that cLasso can indeed cluster the elements in the covariance matrix well and can generate more accurate estimators than the MLE method.

Table 4.4 Simulation results for Example 4.4.2

	MLE		cLasso	
Sample size	error	no. of clusters	error	no. of clusters
50	0.1244	435	0.1029	112.11
100	0.0873	435	0.0505	5.76
200	0.0608	435	0.0278	3.43

4.5 Real Data Analysis

Example 4.1. The Leukemia data from high-density affymetrix oligonucleotide arrays have been analyzed in Golub *et al.* (1999) and are available at the following website: http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. There are 7129 genes and 72 samples from two classes: 47 in class ALL (acute lymphocytic leukemia) and 25 in class AML (acute mylogenous leukemia).

Before we apply our method to the data, we employ the screening procedure proposed by Fan and Lv (2008) and select $P = 50, 100, ..., 500$ covariates in a linear regression model. We randomly split the data into training set and testing set containing 48 observations and 24 observations respectively. We use the training set to estimate the linear model and use the estimated model to classify the samples in the testing set. We split the data 100 times and calculate the average relative misclassification error. Table 4.5 lists the misclassification errors with different P

when the penalty parameters are selected by the CV method. Table 4.5 suggests that models estimated by cLasso has much smaller classification error than ridge regression or Lasso regression. On the other hand, cLasso clusters the genes into a few clusters. In genetic analysis, these clusters are called blocks; see for example Zhang, W. H. *et al* (2002).

Table 4.5 Simulation results of the Leukemia Data

	100		200		300		400		500	
Method	error	n.c.	error	n.c.	error	n.c.	error	n.c.	error	n.c.
ridge	4.63	—	4.63	—	4.37	—	4.46	—	4.50	—
Lasso	5.00	38.0	4.75	51.9	4.63	57.2	5.08	59.51	5.13	60.70
cLasso	3.71	14.5	3.54	15.6	3.12	10.5	3.37	17.07	3.72	13.75

To remove the effect of choosing the penalty parameters, we also plotted the classification errors against all values of penalty parameters as shown in Figure 4.2. In terms of the number of clusters, the fitted model by cLasso is the smallest. Compared with the ridge regression and Lasso, cLasso gives much more accurate classification as shown in Figure 4.2 for almost all values of the penalty parameters. It is also interesting to see that Lasso has worse classification than ridge regression, indicating that sparsity is not an appropriate assumption for the data.

In genome analysis, it is well demonstrated that the genes function in blocks called linkage disequilibrium block; see, e.g., Zhang, W. H. *et al* (2002). The cLasso actually is in line with the understanding and thus has better classification than ridge regression and Lasso regression. In each panel, dash line represents the calculations for Lasso, solid line for cLasso, and dash-dot line for ridge regression.

Example 4.2. We revisit the airpollution data in Hong Kong. The daily average concentration of air pollutants such as CO_2, NO_2 and particulate matters were

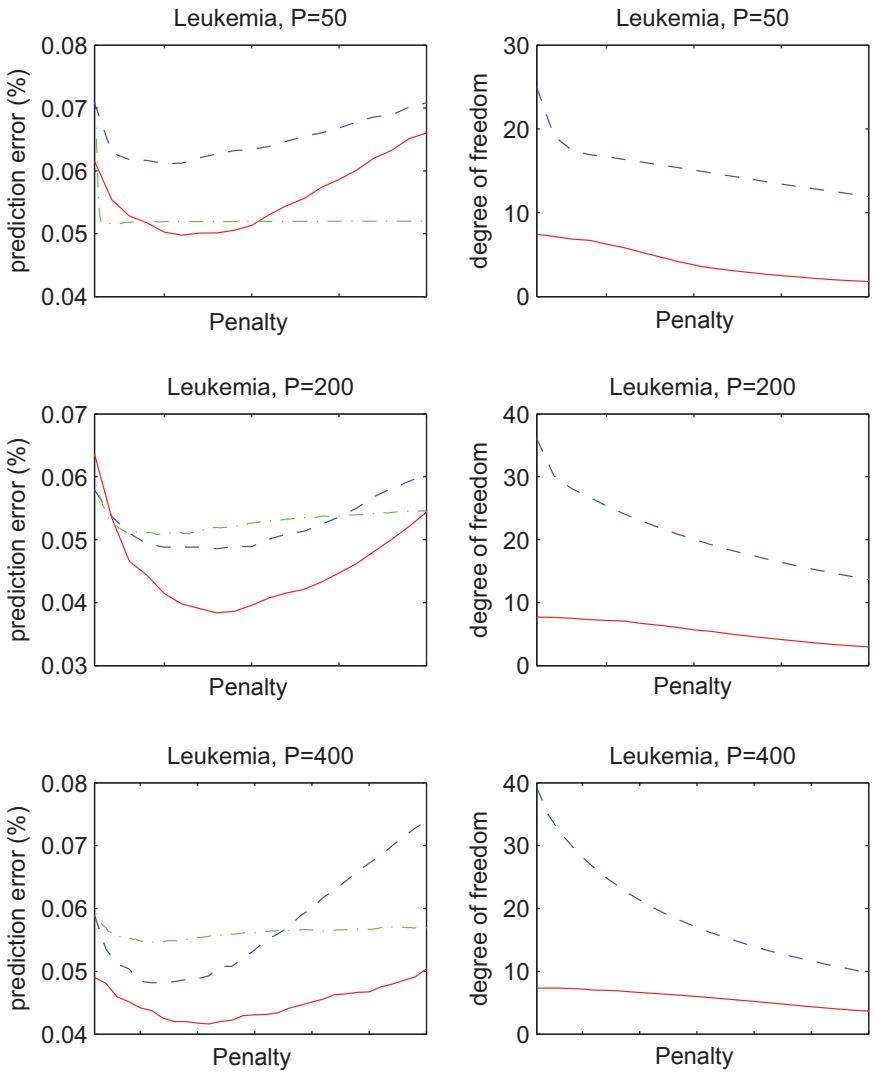

Figure 4.2 Calculation results for the Leukemia Data.

recorded from 1994 to 1997. Denoted by $NO_{2,t}$ the average concentration of NO_2 on day t. To investigate the effect of air-pollution on public health, the daily number of hospital admissions of patients suffering from respiratory diseases in the same period were also recorded, denoted by y_t. The data was investigated in statistical literature for different purpose; see for example Zhang *et al* (2001). Xia and Tong (2004) fitted the data with a model based on the cumulative effect. In

the following, we consider a linear regression model

$$y_t = a_0 + a_1 NO_{2,t-1} + ... + a_1 NO_{2,t-p} + \varepsilon_t,$$

where $p = 365$ to accommodate the pollutants in the past one year. As comparison, we consider four different penalties including ridge regression, Lasso regression, fused Lasso and cLasso. We split the data similarly as in the above example, i.e. 2 thirds of the data are used for the estimation and the other 1/3 for prediction. The prediction errors are shown in Figure 4.3. We can see that the Lasso approach again is not appropriate because its prediction is even worse than the ridge regression. The fused Lasso and cLasso have similar performance especially when their penalty parameters are both large, in which case the two penalties generate the same estimator. However, cLasso is slightly better than fused Lasso when the penalty parameter is properly selected.

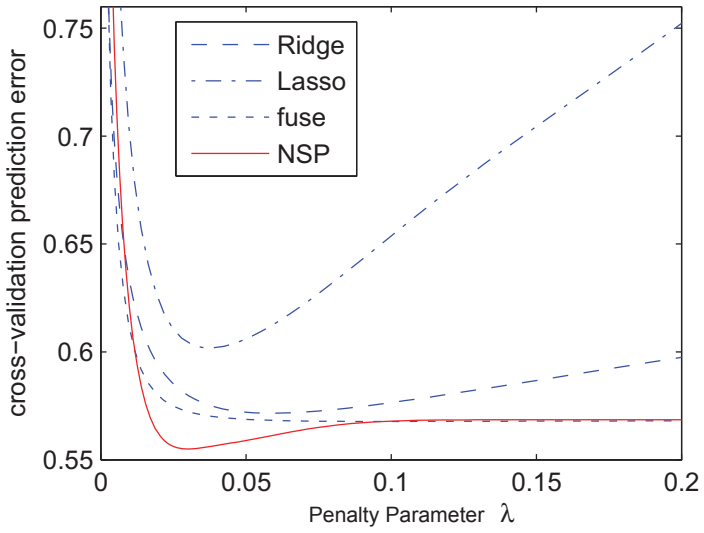

Figure 4.3 The prediction error based on different methods. The penalty parameters for different methods are adjusted for better visualization in the figure.

Example 4.5.1. In the last example, we consider the portfolio data mentioned in the introduction section. We employ cLasso to two problems (1) prediction of returns of any single portfolio and (2) estimation of covariance matrix of all the portfolios. In his website, Professor Kenneth French provides monthly returns of two sets of portfolios. The first set has 25 portfolios and the second 100 portfolios. We shall consider them separately.

Consider prediction of return $\mathbf{x}_{i,t+1}$ based on the past returns of all portfolios $\{\mathbf{x}_{i,s} : s \leq t, \quad i = 1, ..., m\}$. By checking the ACF or PACF of returns of each portfolio, it seems that an autoregressive model with lag 1 is appropriate. Therefore, we consider the simple regression model

$$\mathbf{x}_{i,t+1} = \beta_0 + \beta_1 \mathbf{x}_{1,t} + ... + \beta_m \mathbf{x}_{m,t} + \varepsilon_{i,t}.$$

For every month t, we use the data in its past 10 years to estimate the model and use the estimated model to predict the month's return. Again, we apply ridge regression, Lasso regression and cLasso to estimate the model and check their prediction error by

$$PE = \frac{1}{60} \sum_{t=1}^{5*12} |\mathbf{x}_{i,t} - \hat{\mathbf{x}}_{i,t}|^2.$$

For the three methods, we denote the prediction error respectively by PE(ridge), PE(Lasso) and PE(cLasso). For ease of comparison, we take the prediction error of ridge regression as the benchmark and consider their relative prediction errors, PE(ridge)-PE(Lasso) and PE(ridge)-PE(cLasso). Thus, for example, if PE(ridge)-PE(cLasso)$>$ 0 then the prediction error of cLasso is smaller than the ridge regression, or cLasso has better prediction capability than the ridge regression. The

prediction errors are shown in figure 4.4, suggesting that for most of the portfolios, cLasso is better than the ridge regression as shown in the first panel, but Lasso regression has worse prediction than the ridge regression as shown in the second panel. By employing the paired Z test (David and Gunnunk (1997)), we conclude that the cLasso method gives significantly better prediction than ridge estimation method with Z-value -9.0229 and p-value 9.16×10^{-20}, and that the Lasso method gives significantly worse prediction than ridge estimation method with Z-value 5.16 and p-value 1.23×10^{-7}. In the first panel of figure 4.4 , the ridge regression and

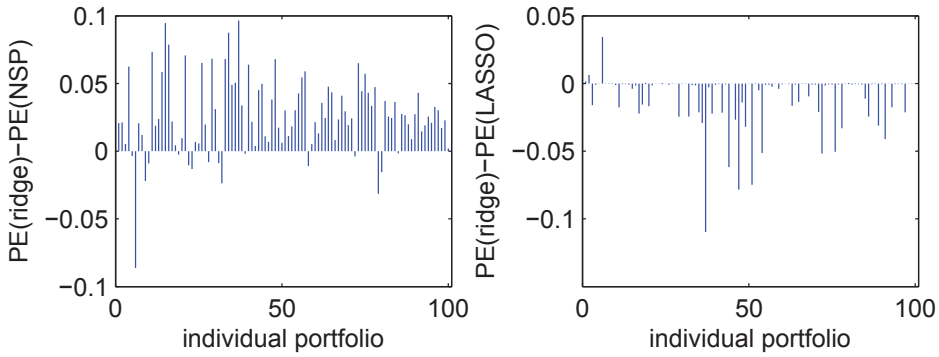

Figure 4.4 Relative prediction errors for the 100 portfolios based on different methods.

the cLasso are compared; in the second panel, the ridge regression and Lasso are compared. In each panel of figure 4.5, the dashed line (top), dash-dot line (middle) and the solid line are respectively the prediction errors of Lasso, ridge regression and cLasso. Next, we turn to investigate the estimation of the covariance matrix. Again, consider the monthly returns of portfolios from 2001 to 2010. To evaluate different estimation methods, we randomly split the data into 2 parts: the training set contains 2/3 of the total observations and the validation set contains 1/3. We use the training sets to estimate the covariance matrix with different penalty

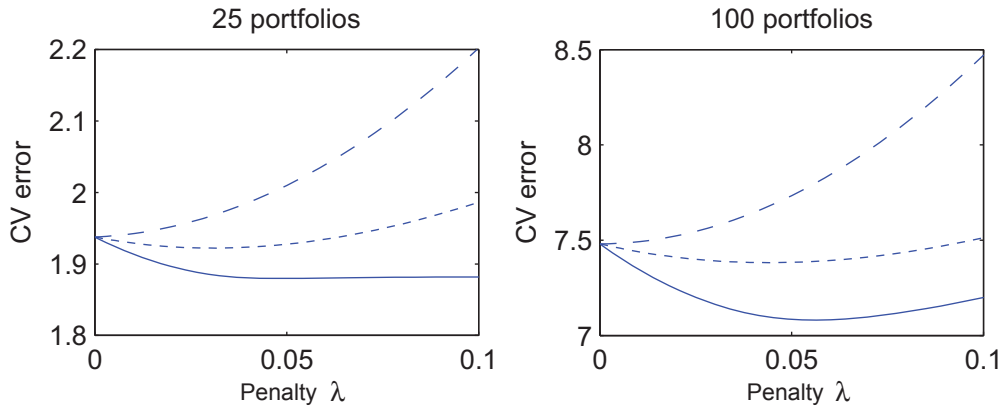

Figure 4.5 The calculation results for the estimation of covariance matrices for two sets of portfolios.

parameter values λ, denote the estimator of the ith splitting by $\hat{\Sigma}_{i,\lambda}$.

We also calculate the sample variance matrix from the testing set, denoted by $\tilde{\Sigma}_i$. We then evaluate the error of the estimator by the Frobenius norm, $d(\hat{\Sigma}_{i,\lambda}, \tilde{\Sigma}_i)$, defined in Example 4.4.2. Based on 100 random splitting, the average of errors, i.e. $\sum_{i=1}^{100} d(\hat{\Sigma}_{i,\lambda}, \tilde{\Sigma}_i)/100$, are shown in the two panels respectively for the two data sets in Figure 4.5. The Lasso method cannot improve the efficiency at all. Moreover, the increasing trend with λ for the prediction error in the estimation of Lasso indicates that the LSE is better than any Lasso estimator. The average error of cLasso is smaller than that of the ridge method, and that of Lasso is bigger than the ridge method, which clearly indicates that cLasso can improve the estimation efficiency over the ridge method and Lasso regression as well.

4.6 Proofs

Proof of Theorem 4.2.1.

Decompose the objective function into two parts: loss function

$$\mathscr{L}(\beta) := \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2 \tag{4.29}$$

and penalty function

$$\mathscr{P}(\beta) := \lambda_n \sum_{k=1}^{p}\sum_{j=k+1}^{p} \hat{w}_{jk}|\beta_j - \beta_k|. \tag{4.30}$$

We first prove the asymptotic normality part.

Let

$$\beta = \beta_0 + u/\sqrt{n} \Longleftrightarrow \mathscr{G}(\beta) = \mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n},$$

where $u = (u_1, \ldots, u_p)^\top \in \mathbb{R}^p$ and denote

$$P^\top \mathscr{G}(u) = (v_1, v_{12}, \ldots, v_{1p_1}, v_2, v_{22}, \ldots, v_{2p_2}, \ldots, v_L, v_{L2}, \ldots, v_{Lp_L})^\top. \tag{4.31}$$

Now let

$$\begin{aligned}\Psi_n(\mathscr{G}(u)) &= \mathscr{L}(\mathscr{G}(\beta)) + \mathscr{P}(\mathscr{G}(\beta)) \\ &= \mathscr{L}\Big(\mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n}\Big) + \mathscr{P}\Big(\mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n}\Big).\end{aligned}$$

Denote $\hat{u}^{(n)} = \arg\min \Psi_{\mathrm{n}}(\mathrm{u})$ then

$$\hat{\phi}^{*(n)} = \mathscr{G}(\hat{\beta}^{*(n)}) = \mathscr{G}(\beta_0) + \mathscr{G}(\hat{u}^{(n)})/\sqrt{n},$$

that is $\sqrt{n}\left(\mathscr{G}(\hat{\beta}^{*(n)}) - \mathscr{G}(\beta_0)\right) = \mathscr{G}(\hat{u}^{(n)})$.

Note that $\Psi_n(\mathscr{G}(u)) - \Psi_n(0) = V_n(\mathscr{G}(u))$, where

$$\begin{aligned} V_n(\mathscr{G}(u)) &= \left\{\mathscr{L}\left(\mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n}\right) - \mathscr{L}\left(\mathscr{G}(\beta_0)\right)\right\} \\ &\quad + \left\{\mathscr{P}\left(\mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n}\right) - \mathscr{P}\left(\mathscr{G}(\beta_0)\right)\right\} \\ &:= \mathscr{L}_n(\mathscr{G}(u)) + \mathscr{P}_n(\mathscr{G}(u)). \end{aligned}$$

For the loss function term

$$\begin{aligned} \mathscr{L}_n(\mathscr{G}(u)) &= \mathscr{L}\left(\mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n}\right) - \mathscr{L}\left(\mathscr{G}(\beta_0)\right) \\ &= \mathscr{G}(u)^\top \frac{\tilde{X}^\top \tilde{X}}{n} \mathscr{G}(u) - 2\mathscr{G}(u)^\top \frac{\tilde{X}^\top \varepsilon}{\sqrt{n}}. \end{aligned} \tag{4.32}$$

For the penalty term $\lambda_n \sum_{k=1}^{p} \sum_{j=k+1}^{p} \hat{w}_{jk} |\beta_j - \beta_k|$, we define

$$S(\beta) = \begin{pmatrix} 0 & \beta_1 - \beta_2 & \beta_1 - \beta_3 & \ldots & \beta_1 - \beta_p \\ \beta_2 - \beta_1 & 0 & \beta_2 - \beta_3 & \ldots & \beta_2 - \beta_p \\ \beta_3 - \beta_1 & \beta_3 - \beta_2 & 0 & \ldots & \beta_3 - \beta_p \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \beta_p - \beta_1 & \beta_p - \beta_2 & \beta_p - \beta_3 & \ldots & 0 \end{pmatrix},$$

and the $p \times p$ matrix

$$W = (w_{ij}) \text{ with } w_{jj} = 0, w_{jk} = \hat{w}_{jk} = (|\hat{\beta}_j - \hat{\beta}_k|)^{-\alpha}, j \neq k, \tag{4.33}$$

then

$$\lambda_n \sum_{k=1}^{p} \sum_{j=k+1}^{p} \hat{w}_{jk} |\beta_j - \beta_k| = \lambda_n \frac{1}{2} \mathbf{1}_p^\top \Big(W \odot |S(\beta)| \Big) \mathbf{1}_p. \tag{4.34}$$

where each element of the matrix $|S(\beta)|$ is the absolute value of the corresponding element of the matrix $S(\beta)$ and $\odot$ denotes the Hadmard product of two matrices.

Denote

$$P^\top \mathscr{G}(\beta) = \Gamma\beta := \gamma := (\gamma_1, \gamma_{12}, \ldots, \gamma_{1p_1}, \ldots, \gamma_L, \gamma_{L2}, \ldots, \gamma_{Lp_L})^\top,$$

where for $l = 1, \ldots, L$,

$$\gamma_l = \frac{1}{p_l} \sum_{i=1}^{p_l} \beta_{M_l+i}, \;\; \gamma_{lk} = \beta_{M_l+1} - \beta_{M_l+k}, \;\; k = 2, \ldots, p_l, \tag{4.35}$$

since $\Gamma = \text{diag}(\Gamma_1,\ \Gamma_2,\ ...,\ \Gamma_L)$ and

$$\Gamma_j = \begin{pmatrix} 1/p_j & 1/p_j & 1/p_j & ... & 1/p_j \\ 1 & -1 & 0 & ... & 0 \\ 1 & 0 & -1 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & ... & -1 \end{pmatrix}_{p_j \times p_j} , j = 1, \ldots, L.$$

On the other hand, we have $\Gamma^{-1} = \text{diag}(\Gamma_1^{-1},\ \Gamma_2^{-1},\ ...,\ \Gamma_L^{-1})$ with

$$\Gamma_j^{-1} = \begin{pmatrix} 1 & 1/p_j & 1/p_j & ... & 1/p_j \\ 1 & -1+1/p_j & 1/p_j & ... & 1/p_j \\ 1 & 1/p_j & -1+1/p_j & ... & 1/p_j \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1/p_j & 1/p_j & ... & -1+1/p_j \end{pmatrix}.$$

It follows that for $l = 1, \ldots, L$,

$$\begin{pmatrix} \beta_{M_l+1} \\ \beta_{M_l+2} \\ \vdots \\ \beta_{M_l+p_l} \end{pmatrix} = \Gamma_l^{-1} \begin{pmatrix} \gamma_l \\ \gamma_{l2} \\ \vdots \\ \gamma_{lp_l} \end{pmatrix} = \gamma_l + \frac{1}{p_l} \sum_{j=2}^{p_l} \gamma_{lj} - \begin{pmatrix} 0 \\ \gamma_{l2} \\ \vdots \\ \gamma_{lp_l} \end{pmatrix}$$

Now, we use the variable γ to equivalently express $S(\beta)$ as

$$\tilde{S}(\gamma) := \begin{pmatrix} \tilde{S}_{1,1}(\gamma) & \tilde{S}_{1,2}(\gamma) & ... & \tilde{S}_{1,L}(\gamma) \\ \tilde{S}_{2,1}(\gamma) & \tilde{S}_{2,2}(\gamma) & ... & \tilde{S}_{2,L}(\gamma) \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{S}_{L,1}(\gamma) & \tilde{S}_{L,2}(\gamma) & ... & \tilde{S}_{L,L}(\gamma) \end{pmatrix} \tag{4.36}$$

where

$$\tilde{S}_{ll}(\gamma) = \begin{pmatrix} 0 & \gamma_{l2} & \gamma_{l3} & \dots & \gamma_{lp_l} \\ -\gamma_{l2} & 0 & \gamma_{l3}-\gamma_{l2} & \dots & \gamma_{lp_l}-\gamma_{l2} \\ -\gamma_{l3} & \gamma_{l2}-\gamma_{l3} & 0 & \dots & \gamma_{lp_l}-\gamma_{l3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\gamma_{lp_l} & \gamma_{l2}-\gamma_{lp_l} & \gamma_{l3}-\gamma_{lp_l} & \dots & 0 \end{pmatrix}$$

and when $s \neq t$,

$$\tilde{S}_{st}(\gamma) = \begin{pmatrix} \Delta_{st} & \Delta_{st}+\gamma_{t2} & \Delta_{st}+\gamma_{t3} & \dots & \Delta_{st}+\gamma_{t,p_t} \\ \Delta_{st}-\gamma_{s2} & \Delta_{st}+\gamma_{t2}-\gamma_{s2} & \Delta_{st}+\gamma_{t3}-\gamma_{s2} & \dots & \Delta_{st}+\gamma_{t,p_t}-\gamma_{s2} \\ \Delta_{st}-\gamma_{s3} & \Delta_{st}+\gamma_{t2}-\gamma_{s3} & \Delta_{st}+\gamma_{t3}-\gamma_{s3} & \dots & \Delta_{st}+\gamma_{t,p_t}-\gamma_{s3} \\ \vdots & \vdots & \vdots & \vdots & \\ \Delta_{st}-\gamma_{s,p_s} & \Delta_{st}+\gamma_{t2}-\gamma_{s,p_s} & \Delta_{st}+\gamma_{t3}-\gamma_{s,p_s} & \dots & \Delta_{st}+\gamma_{t,p_t}-\gamma_{s,p_s} \end{pmatrix}$$

where $\Delta_{st} = \gamma_s + p_s^{-1}\sum_{j=2}^{p_s}\gamma_{s,j} - \gamma_t - p_t^{-1}\sum_{j=2}^{p_t}\gamma_{t,j}$. Correspondingly, the adaptive weights can be written as

$$W = (w_{ij}) = \begin{pmatrix} W_{11} & W_{12} & \dots & W_{1L} \\ W_{21} & W_{22} & \dots & W_{2L} \\ \vdots & \vdots & \vdots & \vdots \\ W_{L1} & W_{L2} & \dots & W_{LL} \end{pmatrix}. \tag{4.37}$$

Define the operators $\langle\cdot\rangle$ and $\ominus$ for any matrix $A = (a_{ij})_{1\le i\le b, 1\le j\le c}$ respectively

as

$$\langle A\rangle = \sum_{i=1}^{b}\sum_{j=1}^{c} a_{ij}, \quad \text{and} \quad A^{\ominus d} = (a_{ij}^{\ominus d})_{1\le i\le b, 1\le j\le c},$$

with $a_{ij}^{\ominus d} = 1/a_{ij}^{d}$ if $a_{ij}\neq 0$; 0 otherwise. The penalty term can be written as

$$\lambda_n \frac{1}{2}\mathbf{1}_p^\top \Big(W \odot |\tilde{S}(\gamma)|\Big)\mathbf{1}_p = \frac{\lambda_n}{2}\left\{\sum_{l=1}^{L}\langle W_{ll}\odot|\tilde{S}_{ll}(\gamma)|\rangle + \sum_{s=1}^{L}\sum_{t=s+1}^{L}\langle W_{st}\odot|\tilde{S}_{st}(\gamma)|\rangle\right\} \quad (4.38)$$

Note that $\tilde{S}_{ll}(P^\top\mathscr{G}(\beta_0)) = \mathbf{0}_{p_l\times p_l}$ and $\tilde{S}_{st}(P^\top\mathscr{G}(\beta_0)) = \Delta_{st}^0 I_{p_s\times p_t}$ for $s\neq t$, where $I_{p_s\times p_t}$ denotes the $p_s\times p_t$ identity matrix and $\Delta_{st}^0 = \gamma_s^0 - \gamma_t^0$. Recall (4.31), we have

$$\begin{aligned}
\mathscr{P}_n(\mathscr{G}(u)) &= \frac{\lambda_n}{2\sqrt{n}}\sum_{l=1}^{L}\langle W_{ll}\odot\sqrt{n}|\tilde{S}_{ll}(P^\top\mathscr{G}(\beta_0)+P^\top\mathscr{G}(u)/\sqrt{n})|\rangle \\
&\quad + \frac{\lambda_n}{2\sqrt{n}}\sum_{s=1}^{L}\sum_{t=s+1}^{L}\langle W_{st}\odot\sqrt{n}\left(|\tilde{S}_{st}(P^\top\mathscr{G}(\beta_0)+P^\top\mathscr{G}(u)/\sqrt{n})| - |\Delta_{st}^0 I_{p_s\times p_t}|\right)\rangle \\
&= \frac{\lambda_n}{2\sqrt{n}}\sum_{l=1}^{L}\langle W_{ll}\odot\sqrt{n}|\tilde{S}_{ll}(P^\top\mathscr{G}(u)/\sqrt{n})|\rangle \quad (4.39)\\
&\quad + \frac{\lambda_n}{2\sqrt{n}}\sum_{s=1}^{L}\sum_{t=s+1}^{L}\langle W_{st}\odot\sqrt{n}\left(|\Delta_{st}^0 I_{p_s\times p_t}+\tilde{S}_{st}(P^\top\mathscr{G}(u)/\sqrt{n})| - |\Delta_{st}^0 I_{p_s\times p_t}|\right)\rangle
\end{aligned}$$

since the matrices $S_{st}(P^\top\mathscr{G}(\beta_0)+P^\top\mathscr{G}(u)/\sqrt{n}) = S_{st}(P^\top\mathscr{G}(\beta_0))+S_{st}(P^\top\mathscr{G}(u)/\sqrt{n})$.

Combined with (4.32), we have

$$\begin{aligned}
V_n(\mathscr{G}(u)) &= \mathscr{G}(u)^\top\frac{\tilde{X}^\top\tilde{X}}{n}\mathscr{G}(u) - 2\mathscr{G}(u)^\top\frac{\tilde{X}^\top\varepsilon}{\sqrt{n}} \quad (4.40)\\
&\quad + \frac{\lambda_n}{2\sqrt{n}}\sum_{l=1}^{L}\langle W_{ll}\odot\sqrt{n}|\tilde{S}_{ll}(P^\top\mathscr{G}(u)/\sqrt{n})|\rangle \\
&\quad + \frac{\lambda_n}{2\sqrt{n}}\sum_{s=1}^{L}\sum_{t=s+1}^{L}\langle W_{st}\odot\sqrt{n}\left(|\Delta_{st}^0 I_{p_s\times p_t}+\tilde{S}_{st}(P^\top\mathscr{G}(u)/\sqrt{n})| - |\Delta_{st}^0 I_{p_s\times p_t}|\right)\rangle.
\end{aligned}$$

It is given in (4.12) that $\tilde{X}^\top \tilde{X}/n \to \tilde{C}$ and $\frac{\tilde{X}^\top \varepsilon}{\sqrt{n}} \xrightarrow{D} \mathscr{N} \sim N(0, \sigma^2 \tilde{C})$. Now we consider the limiting behavior of the penalty term. We consider two cases.

(1) When $\beta_{j0} \neq \beta_{k0}$, $\hat{w}_{jk} \to_p |\beta_{j0} - \beta_{k0}|^{-\alpha}$ and in this case, $\hat{w}_{jk}$ is an element of some W_{st} with $s \neq t$.

That is to say, for the last term of (4.40) , we have $W_{st} \to \{W_{st}^0\}^{\ominus\alpha}$ in probability and

$$\sqrt{n}\left(|\Delta_{st}^0 I_{p_s \times p_t} + \tilde{S}_{st}(P^\top \mathscr{G}(u)/\sqrt{n})| - |\Delta_{st}^0| I_{p_s \times p_t}\right) \to \operatorname{sgn}(\Delta_{st}^0)\tilde{S}_{st}(P^\top \mathscr{G}(u)).$$

Since $\lambda_n/\sqrt{n} \to 0$, we have in probability

$$\frac{\lambda_n}{2\sqrt{n}} \sum_{s=1}^{L} \sum_{t=s+1}^{L} \langle W_{st} \odot \sqrt{n}\left(|\Delta_{st}^0 I_{p_s \times p_t} + \tilde{S}_{st}(P^\top \mathscr{G}(u)/\sqrt{n})| - |\Delta_{st}^0| I_{p_s \times p_t}\right)\rangle \to 0.$$

(2) When $\beta_{j0} = \beta_{k0}$, in this case, $\hat{w}_{jk}$ is an element of some W_{ll}. And $\frac{\lambda_n}{\sqrt{n}}\hat{w}_{jk} = \frac{\lambda_n}{\sqrt{n}} n^{\alpha/2}(\sqrt{n}|\hat{\beta}_j - \hat{\beta}_k|)^{-\alpha} \to \infty$ since $\sqrt{n}\hat{\beta}_n = O_p(1)$.

Thus,

$$\frac{\lambda_n}{2\sqrt{n}} \langle W_{ll} \odot \sqrt{n}|\tilde{S}_{ll}(P^\top \mathscr{G}(u)/\sqrt{n})|\rangle = \frac{1}{2}\langle \lambda_n n^{(\alpha-1)/2}(W_{ll}/\sqrt{n})^{\ominus\alpha} \odot |\tilde{S}_{ll}(P^\top \mathscr{G}(u))|\rangle \to \infty.$$

Therefore, we have $V_n(\mathscr{G}(u)) \to V(\mathscr{G}(u))$ in probability with

$$V(\mathscr{G}(u)) = \begin{cases} \mathscr{G}(u)_{\mathscr{A}^+}^\top \tilde{C}_{\mathscr{A}^+} \mathscr{G}(u)_{\mathscr{A}^+} - 2\mathscr{G}(u)_{\mathscr{A}^+}^\top \mathscr{N}_{\mathscr{A}^+}, & \text{if } \mathscr{G}(u)_l = 0, \text{ for } l \in \mathscr{A}^-, \\ \infty, & \text{otherwise.} \end{cases}$$

Since $V(\mathscr{G}(u))$ is convex and the unique minimum of $V(\mathscr{G}(u))$ is $(\tilde{C}^{-1}_{\mathscr{A}^+}\mathscr{N}_{\mathscr{A}^+}, 0)^\top$. Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$\Big(\mathscr{G}(\hat{u}^{(n)})\Big)_{\mathscr{A}^+} \xrightarrow{D} \tilde{C}^{-1}_{\mathscr{A}^+}\mathscr{N}_{\mathscr{A}^+} \ , \text{ and } \Big(\mathscr{G}(\hat{u}^{(n)})\Big)_{\mathscr{A}^-} \xrightarrow{D} 0. \tag{4.41}$$

Note that the transformation $\mathscr{G}$ is linear, we have

$$\Big(\mathscr{G}(\hat{u}^{(n)})\Big)_{\mathscr{A}^+} = \sqrt{n}\Big(\mathscr{G}(\hat{\beta}^{*(n)})_{\mathscr{A}^+} - \mathscr{G}(\beta_0)_{\mathscr{A}^+}\Big) \xrightarrow{D} N(0, \sigma^2\tilde{C}^{-1}_{\mathscr{A}^+}),$$

then we complete the proof of the asymptotic normality part.

Next, we show the consistency. Recall the notations in (4.8) and (4.9), for any $i \in \mathscr{A}^+$, the asymptotic normality results indicate that $\hat{\phi}_i^{*(n)} \to \gamma_i^0$ in probability, thus $P(i \in \mathscr{A}_n^+) \to 1$. Then it suffices to show that for any $j \in \mathscr{A}^-$, $P(j \in \mathscr{A}_n^+) \to 0$. Since $j \in \mathscr{A}^-$, there exists some $l \in \{1, \ldots, L\}$ (group) and $s \in \{2, \ldots, p_l\}$ such that $\hat{\phi}_j^{*(n)} = \hat{\gamma}_{ls}^{*(n)}$. By the KKT optimality conditions, we have,

$$2\frac{\tilde{\mathbf{x}}_j^\top(Y - \tilde{X}\mathscr{G}(\hat{\beta}^{*(n)}))}{\sqrt{n}} = \frac{\lambda_n}{2\sqrt{n}}\langle W_{ll} \odot \mathrm{sgn}(\hat{\gamma}_{ls}^{*(n)})U_s\rangle + \frac{\lambda_n}{2\sqrt{n}}\sum_{t=l+1}^{L}\langle W_{lt} \odot p_l^{-1}\mathrm{sgn}(\hat{\Delta}_{lt}^{*(n)})T_{ls}\rangle,$$

where

$$U_s = \begin{pmatrix} & & & s^{\text{th}} & & & \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ -1 & \dots & -1 & 0 & -1 & \dots & -1 \\ 0 & \dots & \dots & 1 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{pmatrix} \begin{matrix} \\ \\ \\ s^{\text{th}} \\ \\ \\ \\ \end{matrix}, T_{ls} = \begin{pmatrix} I_{(s-1)\times p} \\ \text{c} \\ I_{(p-s)\times p} \end{pmatrix}$$

are two $p \times p$ constant matrices with

$$\begin{aligned} \text{c} \;:=\; & \frac{1-p_l}{\operatorname{sgn}(\hat{\Delta}_{lt}^{*(n)})}\Big(\operatorname{sgn}(\hat{\Delta}_{lt}^{*(n)} - \hat{\gamma}_{ls}^{*(n)}), \operatorname{sgn}(\hat{\Delta}_{lt}^{*(n)} - \hat{\gamma}_{ls}^{*(n)} + \hat{\gamma}_{t2}^{*(n)}), \\ & \dots, \operatorname{sgn}(\hat{\Delta}_{lt}^{*(n)} - \hat{\gamma}_{ls}^{*(n)} + \hat{\gamma}_{tp_t}^{*(n)})\Big) \in \mathbb{R}^{1\times p}. \end{aligned}$$

Note that

$$2\frac{\tilde{\mathbf{x}}_j^\top (Y - \tilde{X}\mathscr{G}(\hat{\beta}^{*(n)}))}{\sqrt{n}} = 2\frac{\tilde{\mathbf{x}}_j^\top \tilde{X}\sqrt{n}(\mathscr{G}(\beta_0) - \mathscr{G}(\hat{\beta}^{*(n)}))}{n} + 2\tilde{\mathbf{x}}_j^\top \varepsilon/\sqrt{n}.$$

By (4.41), the first term on the right hand side above tends to some normal distribution and $2\tilde{\mathbf{x}}_j^\top \varepsilon/\sqrt{n} \xrightarrow{D} N(0, 4||\tilde{\mathbf{x}}_j||^2\sigma^2)$. However,

$$\frac{\lambda_n}{2\sqrt{n}}\langle W_{ll} \odot \operatorname{sgn}(\hat{\gamma}_{ls}^{*(n)})U_s\rangle \;=\; \frac{1}{2}\langle \lambda_n n^{(\alpha-1)/2}(W_{ll}/\sqrt{n})^{\ominus\alpha} \odot \operatorname{sgn}(\hat{\gamma}_{ls}^{*(n)})U_s\rangle \to \infty$$

and

$$\frac{\lambda_n}{2\sqrt{n}}\sum_{t=l+1}^{L}\langle W_{lt}\odot p_l^{-1}\mathrm{sgn}(\hat{\Delta}_{lt}^{*(n)})T_{ls}\rangle \;=\; \frac{1}{2}\sum_{t=l+1}^{L}\langle\frac{\lambda_n}{\sqrt{n}}W_{lt}\odot p_l^{-1}\mathrm{sgn}(\hat{\Delta}_{lt}^{*(n)})T_{ls}\rangle\to 0.$$

Therefore,

$$\begin{aligned}
& P(j\in\mathscr{A}_n^+)\\
\leq\;& P\Big(2\frac{\tilde{\mathbf{x}}_j^\top(Y-\tilde{X}\mathscr{G}(\hat{\beta}^{*(n)}))}{\sqrt{n}}\\
&\qquad=\frac{\lambda_n}{2\sqrt{n}}\langle W_{ll}\odot\mathrm{sgn}(\hat{\gamma}_{ls}^{*(n)})U_s\rangle+\frac{\lambda_n}{2\sqrt{n}}\sum_{t=l+1}^{L}\langle W_{lt}\odot p_l^{-1}\mathrm{sgn}(\hat{\Delta}_{lt}^{*(n)})T_{ls}\rangle\Big)\\
\to\;& 0
\end{aligned}$$

We complete the proof. □

Proof of Theorem 4.2.2

The proof of Theorem 4.2.2 is similar to that of Theorem 4.2.1 but with an additional penalty term $\lambda_n\sum_{j=1}^{p}\hat{\omega}_j|\beta_j|$ (to obtain the sparsity) where $\hat{\omega}_j=1/|\hat{\beta}_j|^\alpha$ and now $\mathscr{G}(\beta_0)=(\gamma_1^0,\ldots,\gamma_{L-1}^0,\underbrace{\gamma_L^0,0,\ldots,0}_{p-L+1})^\top=(\gamma_1^0,\ldots,\gamma_{L-1}^0,\underbrace{0,0,\ldots,0}_{p-L+1})^\top$. Thus, we mainly focus on the differences here.

Let

$$\beta=\beta_0+u/\sqrt{n}\Longleftrightarrow\mathscr{G}(\beta)=\mathscr{G}(\beta_0)+\mathscr{G}(u)/\sqrt{n},$$

where $u = (u_1, \ldots, u_p)^\top \in \mathbb{R}^p$ and denote

$$P^\top \mathscr{G}(u) = (v_1, v_{12}, \ldots, v_{1p_1}, v_2, v_{22}, \ldots, v_{2p_2}, \ldots, v_L, v_{L2}, \ldots, v_{Lp_L})^\top. \qquad (4.42)$$

The loss function and penalty function are given respectively as

$$\mathscr{L}(\beta) = \sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2, \;\; \mathscr{P}(\beta) = \lambda_n \sum_{k=1}^{p} \sum_{j=k+1}^{p} \hat{w}_{jk}|\beta_j - \beta_k| + \lambda_n \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|.$$

We first prove the asymptotic normality part. Now let

$$\begin{aligned} \Psi_n(\mathscr{G}(u)) &= \mathscr{L}(\mathscr{G}(\beta)) + \mathscr{P}(\mathscr{G}(\beta)) \\ &= \mathscr{L}\Big(\mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n}\Big) + \mathscr{P}\Big(\mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n}\Big). \end{aligned}$$

Denote $\tilde{u}^{(n)} = \arg\min \Psi_{\mathrm{n}}(\mathrm{u})$ then

$$\tilde{\phi}^{*(n)} = \mathscr{G}(\tilde{\beta}^{*(n)}) = \mathscr{G}(\beta_0) + \mathscr{G}(\tilde{u}^{(n)})/\sqrt{n},$$

that is $\sqrt{n}\left(\mathscr{G}(\tilde{\beta}^{*(n)}) - \mathscr{G}(\beta_0)\right) = \mathscr{G}(\tilde{u}^{(n)})$.

Note that $\Psi_n(\mathscr{G}(u)) - \Psi_n(0) = V_n(\mathscr{G}(u))$, where

$$\begin{aligned} V_n(\mathscr{G}(u)) &= \Big\{\mathscr{L}\Big(\mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n}\Big) - \mathscr{L}\Big(\mathscr{G}(\beta_0)\Big)\Big\} \\ &\qquad + \Big\{\mathscr{P}\Big(\mathscr{G}(\beta_0) + \mathscr{G}(u)/\sqrt{n}\Big) - \mathscr{P}\Big(\mathscr{G}(\beta_0)\Big)\Big\} \\ &:= \mathscr{L}_n(\mathscr{G}(u)) + \mathscr{P}_n(\mathscr{G}(u)). \end{aligned}$$

For the loss function term

$$\begin{aligned}\mathscr{L}_n(\mathscr{G}(u)) &= \mathscr{L}\Big(\mathscr{G}(\beta_0)+\mathscr{G}(u)/\sqrt{n}\Big)-\mathscr{L}\Big(\mathscr{G}(\beta_0)\Big) \\ &= \mathscr{G}(u)^\top\frac{\tilde{X}^\top\tilde{X}}{n}\mathscr{G}(u)-2\mathscr{G}(u)^\top\frac{\tilde{X}^\top\varepsilon}{\sqrt{n}}.\end{aligned} \tag{4.43}$$

For the penalty term $\lambda_n\sum_{k=1}^p\sum_{j=k+1}^p\hat{w}_{jk}|\beta_j-\beta_k|+\lambda_n\sum_{k=1}^p\hat{\omega}_k|\beta_k|$, we defined the same $\tilde{S}(\gamma)$ as in (4.36) and the same $p\times p$ matrix W as in (4.37). Moreover, we denote

$$\hat{\omega}=(\hat{\omega}_1,\ldots,\hat{\omega}_p)^\top:=(\varpi_1,\varpi_{12},\ldots,\varpi_{1p_1},\ldots,\varpi_L,\varpi_{L2},\ldots,\varpi_{Lp_L})^\top,$$

and it follows that

$$\sum_{k=1}^p\hat{\omega}_k|\beta_k| = \sum_{l=1}^L\varpi_l|\gamma_l+\frac{1}{p_l}\sum_{j=2}^{p_l}\gamma_{lj}|+\sum_{l=1}^L\sum_{i=2}^{p_l}\varpi_{li}|\gamma_l+\frac{1}{p_l}\sum_{j=2}^{p_l}\gamma_{lj}-\gamma_{li}| \tag{4.44}$$

Combined with (4.43) and recall the notation (4.42), we have

$$\begin{aligned}&V_n(\mathscr{G}(u))=\mathscr{G}(u)^\top\frac{\tilde{X}^\top\tilde{X}}{n}\mathscr{G}(u)-2\mathscr{G}(u)^\top\frac{\tilde{X}^\top\varepsilon}{\sqrt{n}} \\ &+\frac{\lambda_n}{2\sqrt{n}}\sum_{l=1}^L\langle W_{ll}\odot\sqrt{n}|\tilde{S}_{ll}(P^\top\mathscr{G}(u)/\sqrt{n})|\rangle \\ &+\frac{\lambda_n}{2\sqrt{n}}\sum_{s=1}^L\sum_{t=s+1}^L\langle W_{st}\odot\sqrt{n}\Big(|\Delta^0_{st}I_{p_s\times p_t}+\tilde{S}_{st}(P^\top\mathscr{G}(u)/\sqrt{n})|-|\Delta^0_{st}|I_{p_s\times p_t}\Big)\rangle \\ &+\frac{\lambda_n}{\sqrt{n}}\sum_{l=1}^{L-1}\sqrt{n}\varpi_l(|\gamma^0_l+(v_l+\frac{1}{p_l}\sum_{j=2}^{p_l}v_{lj})/\sqrt{n}|-|\gamma^0_l|)+\frac{\lambda_n}{\sqrt{n}}\varpi_L|v_L+\frac{1}{p_L}\sum_{j=2}^{p_L}v_{Lj}| \\ &+\frac{\lambda_n}{\sqrt{n}}\sum_{l=1}^{L-1}\sum_{i=2}^{p_l}\sqrt{n}\varpi_{li}(|\gamma^0_l+(v_l+\frac{1}{p_l}\sum_{j=2}^{p_l}v_{lj}-v_{li})/\sqrt{n}|-|\gamma^0_l|)\end{aligned} \tag{4.45}$$

$$+\frac{\lambda_n}{\sqrt{n}}\sum_{i=2}^{p_L}\varpi_{Li}|v_L+\frac{1}{p_L}\sum_{j=2}^{p_L}v_{Lj}-v_{Li}|.$$

It is given in (4.12) that $\tilde{X}^\top\tilde{X}/n\to\tilde{C}$ and $\tilde{X}^\top\varepsilon/\sqrt{n}\xrightarrow{D}\mathscr{Z}\sim N(0,\sigma^2\tilde{C})$.

Now we consider the limiting behavior of the penalty terms. We consider the all possible four cases.

(1) When $\beta_{j0}\neq\beta_{k0}$ and $\beta_{j0}\neq 0,\beta_{k0}\neq 0$. In this case, $\hat{w}_{jk}$ is an element of some W_{st} with $s\neq t$. And

$$\sqrt{n}\left(|\Delta_{st}^0 I_{p_s\times p_t}+\tilde{S}_{st}(P^\top\mathscr{G}(u)/\sqrt{n})|-|\Delta_{st}^0|I_{p_s\times p_t}\right)\to\mathrm{sgn}(\Delta_{st}^0)\tilde{S}_{st}(P^\top\mathscr{G}(u)).$$

Since $\lambda_n/\sqrt{n}\to 0$, we have in probability

$$\frac{\lambda_n}{2\sqrt{n}}\sum_{s=1}^{L}\sum_{t=s+1}^{L}\langle W_{st}\odot\sqrt{n}\left(|\Delta_{st}^0 I_{p_s\times p_t}+\tilde{S}_{st}(P^\top\mathscr{G}(u)/\sqrt{n})|-|\Delta_{st}^0|I_{p_s\times p_t}\right)\rangle\to 0 \tag{4.46}$$

Since $\beta_{j0}\neq 0,\beta_{k0}\neq 0$, they should belong to two groups other than group L, say, s and t respectively. Thus for $l=s$ or $l=t$,

$$\frac{\lambda_n}{\sqrt{n}}\sum_{i=2}^{p_l}\sqrt{n}\varpi_{li}(|\gamma_l^0+(v_l+\frac{1}{p_l}\sum_{j=2}^{p_l}v_{lj}-v_{li})/\sqrt{n}|-|\gamma_l^0|)\to_p 0 \tag{4.47}$$

and

$$\frac{\lambda_n}{\sqrt{n}}\sqrt{n}\varpi_l(|\gamma_l^0+(v_l+\frac{1}{p_l}\sum_{j=2}^{p_l}v_{lj})/\sqrt{n}|-|\gamma_l^0|)\to_p 0 \tag{4.48}$$

(2) When $\beta_{j0}=\beta_{k0}\neq 0$, in this case, $\hat{w}_{jk}$ is an element of some W_{ll}. And

$\frac{\lambda_n}{\sqrt{n}}\hat{w}_{jk} = \frac{\lambda_n}{\sqrt{n}} n^{\alpha/2}(\sqrt{n}|\hat{\beta}_j - \hat{\beta}_k|)^{-\alpha} \to \infty$ since $\sqrt{n}\hat{\beta}_n = O_p(1)$. Thus,

$$\begin{aligned} & \frac{\lambda_n}{2\sqrt{n}}\langle W_{ll} \odot \sqrt{n}|\tilde{S}_{ll}(P^\top \mathscr{G}(u)/\sqrt{n})|\rangle \\ = \; & \frac{1}{2}\langle \lambda_n n^{(\alpha-1)/2}(W_{ll}/\sqrt{n})^{\ominus\alpha} \odot |\tilde{S}_{ll}(P^\top \mathscr{G}(u)/\sqrt{n})|\rangle \to \infty \end{aligned} \tag{4.49}$$

However, (4.47) and (4.48) are still satisfied.

(3) When $\beta_{j0} = \beta_{k0} = 0$, in this case, (4.49) is satisfied and moreover

$$\frac{\lambda_n}{\sqrt{n}}\varpi_L|v_L + \frac{1}{p_L}\sum_{j=2}^{p_L} v_{Lj}| \to \infty, \;\; \frac{\lambda_n}{\sqrt{n}}\sum_{i=2}^{p_L}\varpi_{Li}|v_L + \frac{1}{p_L}\sum_{j=2}^{p_L} v_{Lj} - v_{Li}| \to \infty. \tag{4.50}$$

(4) If $\beta_{j0} \neq \beta_{k0} = 0$, then β_{k0} belong to group L. Suppose β_{j0} belong to group $s \neq L$, it is easy to see that (4.46) is satisfied with $t = L$. We also have (4.50). With $l = s$, (4.47) and (4.48) are also satisfied.

Therefore, we have $V_n(\mathscr{G}(u)) \to V(\mathscr{G}(u))$ in probability with

$$V(\mathscr{G}(u)) = \begin{cases} \mathscr{G}(u)_{\tilde{\mathscr{A}}^+}^\top \tilde{C}_{\tilde{\mathscr{A}}^+} \mathscr{G}(u)_{\tilde{\mathscr{A}}^+} - 2\mathscr{G}(u)_{\tilde{\mathscr{A}}^+}^\top \mathscr{Z}_{\tilde{\mathscr{A}}^+}, & \text{if } \mathscr{G}(u)_l = 0, \text{ for } l \in \tilde{\mathscr{A}}^-, \\ \infty, & \text{otherwise.} \end{cases}$$

Since $V(\mathscr{G}(u))$ is convex and the unique minimum of $V(\mathscr{G}(u))$ is $(\tilde{C}_{\tilde{\mathscr{A}}^+}^{-1}\mathscr{Z}_{\tilde{\mathscr{A}}^+}, 0)^\top$. Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$\left(\mathscr{G}(\tilde{u}^{(n)})\right)_{\tilde{\mathscr{A}}^+} \xrightarrow{D} \tilde{C}_{\tilde{\mathscr{A}}^+}^{-1}\mathscr{Z}_{\tilde{\mathscr{A}}^+} \;\text{, and } \left(\mathscr{G}(\tilde{u}^{(n)})\right)_{\tilde{\mathscr{A}}^-} \xrightarrow{D} 0. \tag{4.51}$$

Note that the transformation $\mathscr{G}$ is linear, we have

$$\Big(\mathscr{G}(\tilde{u}^{(n)})\Big)_{\tilde{\mathscr{A}}^+} = \sqrt{n}\Big(\mathscr{G}(\tilde{\beta}^{*(n)})_{\tilde{\mathscr{A}}^+} - \mathscr{G}(\beta_0)_{\tilde{\mathscr{A}}^+}\Big) \xrightarrow{D} N(0, \sigma^2 \tilde{C}^{-1}_{\tilde{\mathscr{A}}^+}),$$

Thus we complete the asymptotic part. The consistency part is similar to that of Theorem 4.2.2. Just notice that for $j \in \tilde{\mathscr{A}}^-$, there are two cases: (1) there exists some $l \in \{1, \ldots, L-1\}$, and $s \in \{2, \ldots, p_l\}$ such that $\tilde{\phi}_j^{*(n)} = \tilde{\gamma}_{ls}^{*(n)}$; (2) $\tilde{\phi}_j^{*(n)} = \tilde{\gamma}_{Ls}^{*(n)}, s \in \{2, \ldots, p_l\}$ or $\tilde{\phi}_j^{*(n)} = \tilde{\gamma}_L^{*(n)}$. Both imply that the KKT condition cannot be satisfied.Thus, $P(j \in \tilde{\mathscr{A}}_n^+) \to 0$. □

Proof of Theorem 4.3.1

We divide the proof into three parts. In part I, we will rewrite the loss term and penalty term of the objective function (4.19) into the function of the transformed variable $\mathscr{T}(\nu)$. In part II, we will show the asymptotic normality result. The asymptotic consistency result will be shown in part III.

- Part I. (Rewrite the objective function)

  We first denote index sets $\mathscr{M} = \{1, \ldots, m\}$ and

  $$\mathscr{M}_l = \{\sum_{i=0}^{l-1} m_i + 1, \sum_{i=0}^{l-1} m_i + 2, ..., \sum_{i=0}^{l-1} m_i + m_l\}$$

  satisfy that $\cup_{l=1}^{L} \mathscr{M}_l = \mathscr{M}$, $\mathscr{M}_l \cap \mathscr{M}_s = \emptyset$, for $l \neq s$, where $l, s = 1, 2, ..., L$ and $m_0 = 0$. Moreover, denote $M_j = m_0 + \ldots + m_{j-1}$, $j = 1, \ldots, L, L+1$,

and

$$\mathscr{M}^+ = \{M_1 + 1, M_2 + 1, \ldots, M_L + 1\}, \;\; \mathscr{M}^- = \mathscr{M} - \mathscr{M}^+. \qquad (4.52)$$

Note that $\mathscr{T}(Y) = P\Gamma QY, \mathscr{T}(\nu) = P\Gamma Q\nu$, we have

$$Y = P^\top \Gamma^{-1} Q^\top \mathscr{T}(Y), \nu = P^\top \Gamma^{-1} Q^\top \mathscr{T}(\nu).$$

Thus, the loss term

$$\begin{aligned} \|\mathscr{Y} - (1_{\mathrm{n}} \otimes \mathrm{I}_{\mathrm{m}})\nu\|^2 &= \sum_{i=1}^{n} \|Y_i - \nu\|^2 \\ &= \sum_{i=1}^{n} (\mathscr{T}(Y_i) - \mathscr{T}(\nu))^\top Q\tilde{\Gamma}Q^\top (\mathscr{T}(Y_i) - \mathscr{T}(\nu)) \end{aligned}$$

where the $m \times m$ matrix

$$\begin{aligned} \tilde{\Gamma} := \left(\Gamma^{-1}\right)^\top \Gamma &= \mathrm{diag}\Big(\left(\Gamma_1^{-1}\right)^\top \Gamma_1^{-1}, \ldots, \left(\Gamma_L^{-1}\right)^\top \Gamma_L^{-1}\Big) \\ &= \mathrm{diag}\Big(m_1, \mathscr{I}_1, m_2, \mathscr{I}_2, \ldots, m_L, \mathscr{I}_L\Big), \end{aligned} \qquad (4.53)$$

where the $(m_l - 1) \times (m_l - 1)$ matrix $\mathscr{I}_l := I_{m_l - 1} - m_l^{-1}(1_{m_l - 1} \otimes 1_{m_l - 1}^\top), l = 1, \ldots, L$ since

$$\Gamma_l^{-1} = \begin{pmatrix} 1 & 1/m_l & 1/m_l & ... & 1/m_l \\ 1 & -1 + 1/m_l & 1/m_1 & ... & 1/m_l \\ 1 & 1/m_l & -1 + 1/m_1 & ... & 1/m_l \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1/m_l & 1/m_l & ... & -1 + 1/m_l \end{pmatrix},$$

and

$$\left(\Gamma_l^{-1}\right)^\top \Gamma_l^{-1} = \begin{pmatrix} m_l & 0 & 0 & ... & 0 \\ 0 & 1-1/m_l & -1/m_l & ... & -1/m_l \\ 0 & -1/m_l & 1-1/m_l & ... & -1/m_l \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & -1/m_l & -1/m_l & ... & 1-1/m_l \end{pmatrix}, \text{ for } l = 1, \ldots, L.$$

Moreover, from (4.53), we can see that the eigenvalues of $\tilde{\Gamma}$ are

$$\{m_l, 1/m_l, \underbrace{1, \ldots, 1}_{m_l - 2}\}$$

which implies that $\tilde{\Gamma}$ is a positive definite matrix and we can decompose it as

$$\tilde{\Gamma} = \tilde{\Gamma}^{1/2}\tilde{\Gamma}^{1/2}. \tag{4.54}$$

Let

$$\tilde{Q} = Q\tilde{\Gamma}^{1/2},$$

the loss term of the objective function can be written as

$$\begin{aligned} \mathscr{L}(\mathscr{T}(\nu)) &:= \|\mathscr{Y} - (1_{\mathrm{n}} \otimes \mathrm{I}_{\mathrm{m}})\nu\|^2 \\ &= \sum_{i=1}^{n} \left(\tilde{Q}^\top(\mathscr{T}(Y_i) - \mathscr{T}(\nu))\right)^\top \left(\tilde{Q}^\top(\mathscr{T}(Y_i) - \mathscr{T}(\nu))\right) \end{aligned} \tag{4.55}$$

To deal with the penalty term $\lambda_n \sum_{j=1}^{m} \sum_{k=j+1}^{m} \hat{w}_{jk}|\nu_j - \nu_k|$, we define

$$S(\nu) = \begin{pmatrix} 0 & \nu_1 - \nu_2 & \nu_1 - \nu_3 & ... & \nu_1 - \nu_m \\ \nu_2 - \nu_1 & 0 & \nu_2 - \nu_3 & ... & \nu_2 - \nu_m \\ \nu_3 - \nu_1 & \nu_3 - \nu_2 & 0 & ... & \nu_3 - \nu_m \\ \vdots & \vdots & \vdots & ... & \vdots \\ \nu_m - \nu_1 & \nu_m - \nu_2 & \nu_m - \nu_{m-1} & ... & 0 \end{pmatrix},$$

and $m \times m$ matrix

$$W = (w_{ij}) \text{ with } w_{jj} = 0, w_{jk} = \hat{w}_{jk} = (|\hat{\nu}_j - \hat{\nu}_k|)^{-\alpha}, j \neq k, \tag{4.56}$$

then

$$\lambda_n \sum_{j=1}^{m} \sum_{k=j+1}^{m} \hat{w}_{jk}|\nu_j - \nu_k| = \lambda_n \frac{1}{2} \mathbf{1}_m^\top \Big(W \odot |S(\nu)| \Big) \mathbf{1}_m \tag{4.57}$$

where each element of the matrix $|S(\nu)|$ is the absolute value of the matrix $S(\nu)$ and $\odot$ denotes the Hadmard product of two matrices.

Using the permutation matrix Q, suppose

$$Q\nu = (\tilde{\nu}_{M_1+1}, \ldots, \tilde{\nu}_{M_2}, \tilde{\nu}_{M_2+1}, \ldots, \tilde{\nu}_{M_3} \ldots, \tilde{\nu}_m)^\top = (\vec{\nu}_1^\top, \ldots, \vec{\nu}_L^\top)^\top,$$

with

$$\vec{\nu}_l = (\tilde{\nu}_{M_l+1}, \ldots, \tilde{\nu}_{M_{l+1}})^\top,$$

and

$$P^\top \mathscr{T}(\nu) = \Gamma Q \nu = (\varphi_1, \varphi_{12}, \ldots, \varphi_{1m_1}, \varphi_2, \varphi_{22}, \ldots, \varphi_{2m_2}, \ldots, \varphi_L, \varphi_{L2}, \ldots, \varphi_{Lm_L})^\top,$$

with

$$\vec{\varphi}_l := \Gamma_l \vec{\nu}_l = (\varphi_l, \varphi_{l2}, \ldots, \varphi_{lm_l})^\top.$$

It follows that

$$\varphi_l = \frac{1}{m_l} \sum_{j=1}^{m_l} \tilde{\nu}_{M_l+j}, \;\; \varphi_{lk} = \tilde{\nu}_{M_l+1} - \tilde{\nu}_{M_l+k}, k = 2, \ldots, m_l$$

and

$$\vec{\nu}_l = \Gamma_l^{-1} \vec{\varphi}_l = \varphi_l + \frac{1}{m_l} \sum_{j=2}^{m_l} \varphi_{lj} - \begin{pmatrix} 0 \\ \varphi_{l2} \\ \vdots \\ \varphi_{lm_l} \end{pmatrix}, \; l = 1, \ldots, L.$$

Now let

$$\tilde{S}(\nu) := QS(\nu)Q^\top = \begin{pmatrix} \tilde{S}_{1,1}(\nu) & \tilde{S}_{1,2}(\nu) & \ldots & \tilde{S}_{1,L}(\nu) \\ \tilde{S}_{2,1}(\nu) & \tilde{S}_{2,2}(\nu) & \ldots & \tilde{S}_{2,L}(\nu) \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{S}_{L,1}(\nu) & \tilde{S}_{L,2}(\nu) & \ldots & \tilde{S}_{L,L}(\nu) \end{pmatrix}, \tag{4.58}$$

then $\tilde{S}_{lc}(\nu)$ is an $m_l \times m_c$ matrix with the (j,k)-th element being

$$\tilde{\nu}_{M_l+j} - \tilde{\nu}_{M_c+k}, j = 1, \ldots, m_l; k = 1, \ldots, m_c; l, c = 1, \ldots, L. \tag{4.59}$$

Notice that

$$\begin{aligned}\sum_{j=1}^{m}\sum_{k=j+1}^{m}\hat{w}_{jk}|\nu_j-\nu_k| &= \sum_{l=1}^{L}\sum_{j=M_l+1}^{M_{l+1}}\sum_{k=j+1}^{M_{l+1}}\hat{w}_{jk}|\tilde{\nu}_j-\tilde{\nu}_k| \\ &\quad+\sum_{l=1}^{L}\sum_{c=l+1}^{L}\sum_{j=M_l+1}^{M_{l+1}}\sum_{k=M_c+1}^{M_{c+1}}\hat{w}_{jk}|\tilde{\nu}_j-\tilde{\nu}_k|.\end{aligned}$$

For $j\in\{M_l+1,M_l+2,\ldots,M_{l+1}\},\ k\in\{j+1,\ldots,M_{l+1}\}$,

$$\begin{aligned}\tilde{\nu}_j-\tilde{\nu}_k &= \begin{cases}\varphi_{lk}, & \text{for } j=M_l+1,\\ -(\varphi_{lj}-\varphi_{lk}), & \text{otherwise.}\end{cases}\\ &= \begin{cases}\varphi_{lk}, & \text{for } j\in\mathscr{M}^+\cap\mathscr{M}_l,\\ -(\varphi_{lj}-\varphi_{lk}), & \text{for } j\in\mathscr{M}^-\cap\mathscr{M}_l;\end{cases}\end{aligned}$$

For $j\in\{M_l+1,M_l+2,\ldots,M_{l+1}\},\ k\in\{M_c+1,\ldots,M_{c+1}\}, l\neq c$,

$$\tilde{\nu}_j-\tilde{\nu}_k = \begin{cases}\Delta_{lc}, & \text{for } j\in\mathscr{M}^+\cap\mathscr{M}_l \ \&\ k\in\mathscr{M}^+\cap\mathscr{M}_c,\\ \Delta_{lc}+\varphi_{ck}, & \text{for } j\in\mathscr{M}^+\cap\mathscr{M}_l \ \&\ k\in\mathscr{M}^-\cap\mathscr{M}_c,\\ \Delta_{lc}-\varphi_{lj}, & \text{for } j\in\mathscr{M}^-\cap\mathscr{M}_l \ \&\ k\in\mathscr{M}^+\cap\mathscr{M}_c;\\ -(\varphi_{lj}-\varphi_{ck})+\Delta_{lc}, & \text{otherwise ;}\end{cases}$$

where $\Delta_{lc}=(\varphi_l-\varphi_c)+\frac{1}{m_l}\sum_{s=2}^{m_l}\varphi_{ls}-\frac{1}{m_c}\sum_{s=2}^{m_c}\varphi_{cs}$.

The penalty term (4.57) of the objective function can be written as

$$\begin{aligned}\mathscr{P}(\mathscr{T}(\nu)) &:= \lambda_n\sum_{j=1}^{m}\sum_{k=j+1}^{m}\hat{w}_{jk}|\nu_j-\nu_k|\\ &= \lambda_n\sum_{l=1}^{L}\Big\{\sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\sum_{k=j+1}^{M_{l+1}}\hat{w}_{jk}|\varphi_{lk}|+\sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\sum_{k=j+1}^{M_{l+1}}\hat{w}_{jk}|\varphi_{lj}-\varphi_{lk}|\Big\}\end{aligned}$$

$$+\lambda_n \sum_{l=1}^{L} \sum_{c=l+1}^{L} \Big\{ \sum_{j \in \mathscr{M}^+ \cap \mathscr{M}_l} \sum_{k \in \mathscr{M}^+ \cap \mathscr{M}_c} \hat{w}_{jk} |\Delta_{lc}| + \sum_{j \in \mathscr{M}^+ \cap \mathscr{M}_l} \sum_{k \in \mathscr{M}^- \cap \mathscr{M}_c} \hat{w}_{jk} |\Delta_{lc} + \varphi_{ck}| + \sum_{j \in \mathscr{M}^- \cap \mathscr{M}_l} \sum_{k \in \mathscr{M}^+ \cap \mathscr{M}_c} \hat{w}_{jk} |\Delta_{lc} - \varphi_{lj}| + \sum_{j \in \mathscr{M}^- \cap \mathscr{M}_l} \sum_{k \in \mathscr{M}^- \cap \mathscr{M}_c} \hat{w}_{jk} |\Delta_{lc} + (\varphi_{ck} - \varphi_{cj})| \Big\}. \quad (4.60)$$

- Part II. (Asymptotic normality)

  Let $\mathscr{T}(\nu) = \mathscr{T}(\nu_0) + \mathscr{T}(u)/\sqrt{n}$, where for vector $u = (u_1, \ldots, u_m)^\top \in \mathbb{R}^m$, the transformation $\mathscr{T}(u)$ satisfies

  $$P^\top \mathscr{T}(u) = (v_1, v_{12}, \ldots, v_{1m_1}, v_2, v_{22}, \ldots, v_{2m_2}, \ldots, v_L, v_{L2}, \ldots, v_{Lm_L})^\top \quad (4.61)$$

  which corresponds to the transformation of ν.

  Let

  $$\begin{aligned} \Psi_n(\mathscr{T}(u)) &= \mathscr{L}(\mathscr{T}(\nu)) + \mathscr{P}(\mathscr{T}(\nu)) \\ &= \mathscr{L}\Big(\mathscr{T}(\nu_0) + \mathscr{T}(u)/\sqrt{n}\Big) + \mathscr{P}\Big(\mathscr{T}(\nu_0) + \mathscr{T}(u)/\sqrt{n}\Big) \end{aligned}$$

  Let $\hat{u}^{(n)} = \arg\min \Psi_n(u)$ then $\hat{\phi}^{*(n)} = \mathscr{T}(\hat{\nu}^{*(n)}) = \mathscr{T}(\nu_0) + \mathscr{T}(\hat{u}^{(n)})/\sqrt{n}$. Note that $\Psi_n(\mathscr{T}(u)) - \Psi_n(0) = V_n(\mathscr{T}(u))$, where

  $$\begin{aligned} V_n(\mathscr{T}(u)) &= \Big\{\mathscr{L}\Big(\mathscr{T}(\nu_0) + \mathscr{T}(u)/\sqrt{n}\Big) - \mathscr{L}\Big(\mathscr{T}(\nu_0)\Big)\Big\} + \Big\{\mathscr{P}\Big(\mathscr{T}(\nu_0) \\ &\qquad + \mathscr{T}(u)/\sqrt{n}\Big) - \mathscr{P}\Big(\mathscr{T}(\nu_0)\Big)\Big\} \\ &:= \mathscr{L}_n(u) + \mathscr{P}_n(u). \end{aligned}$$

First,

$$\begin{aligned}\mathscr{L}_n(\mathscr{T}(u)) &= \mathscr{L}\Big(\mathscr{T}(\nu_0)+\mathscr{T}(u)/\sqrt{n}\Big)-\mathscr{L}\Big(\mathscr{T}(\nu_0)\Big)\\ &= \sum_{i=1}^{n}\Big(\tilde{Q}^\top(\mathscr{T}(Y_i)-(\mathscr{T}(\nu_0)+\mathscr{T}(u)/\sqrt{n}))\Big)^\top\\ &\qquad\qquad\times\Big(\tilde{Q}^\top(\mathscr{T}(Y_i)-(\mathscr{T}(\nu_0)+\mathscr{T}(u)/\sqrt{n}))\Big)\\ &\quad-\sum_{i=1}^{n}\Big(\tilde{Q}^\top(\mathscr{T}(Y_i)-\mathscr{T}(\nu_0))\Big)^\top\Big(\tilde{Q}^\top(\mathscr{T}(Y_i)-\mathscr{T}(\nu_0))\Big)\\ &= \mathscr{T}(u)^\top\tilde{Q}\tilde{Q}^\top\mathscr{T}(u)-2\mathscr{T}(u)^\top\tilde{Q}\tilde{Q}^\top\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathscr{T}(\varepsilon_i).\end{aligned}$$

and also recall from (4.15) and (4.18) that

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\nu}}_n-\nu_0)=\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n}Y_i-\nu_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\nu_0)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\xrightarrow{D}\xi\sim N(0,W),\end{aligned}$$

which implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathscr{T}(\varepsilon_i)=P\Gamma Q\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\xrightarrow{D}\xi\sim N(0,G),$$

where $G=P\Gamma QWQ^\top\Gamma^\top P^\top$.

Second,

$$\begin{aligned}\mathscr{P}\Big(\mathscr{T}(\nu_0)+\mathscr{T}(u)/\sqrt{n}\Big)=\lambda_n\sum_{l=1}^{L}\Big\{&\sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\sum_{k=j+1}^{M_{l+1}}\hat{w}_{jk}|\varphi^0_{lk}+v_{lk}/\sqrt{n}|\\ &+\sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\sum_{k=j+1}^{M_{l+1}}\hat{w}_{jk}|\varphi^0_{lj}-\varphi^0_{lk}+(v_{lj}-v_{lk})/\sqrt{n}|\Big\}\end{aligned}$$

$$
\begin{aligned}
&+\lambda_n \sum_{l=1}^{L}\sum_{c=l+1}^{L}\Big\{ \sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\ \sum_{k\in\mathscr{M}^+\cap\mathscr{M}_c} \hat{w}_{jk}|\Delta_{lc}^0+\delta_{lc}/\sqrt{n}| \\
&+ \sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\ \sum_{k\in\mathscr{M}^-\cap\mathscr{M}_c} \hat{w}_{jk}|\Delta_{lc}^0+\varphi_{ck}^0+(\delta_{lc}+v_{ck})/\sqrt{n}| \\
&+ \sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\ \sum_{k\in\mathscr{M}^+\cap\mathscr{M}_c} \hat{w}_{jk}|\Delta_{lc}^0-\varphi_{lj}^0+(\delta_{lc}-v_{cj})/\sqrt{n}| \\
&+ \sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\ \sum_{k\in\mathscr{M}^-\cap\mathscr{M}_c} \hat{w}_{jk}|\Delta_{lc}^0+(\varphi_{ck}^0-\varphi_{cj}^0)+(\delta_{lc}+(v_{ck}-v_{cj}))/\sqrt{n}|\Big\}
\end{aligned}
$$

where $\delta_{lc}=(v_l-v_c)+{m_l}^{-1}\sum_{s=2}^{m_l}v_{lj}-{m_c}^{-1}\sum_{s=2}^{m_c}v_{cj}$ and the superscript 0 denotes the true value. Moreover, recall the assumption on the true value, we have

$$
\begin{aligned}
&\mathscr{P}_n(\mathscr{T}(u))=\mathscr{P}\Big(\mathscr{T}(\nu_0)+\mathscr{T}(u)/\sqrt{n}\Big)-\mathscr{P}\Big(\mathscr{T}(\nu_0)\Big) \\
=\ & \lambda_n\sum_{l=1}^{L}\sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\sum_{k=j+1}^{M_{l+1}}\hat{w}_{jk}\Big\{|\varphi_{lk}^0+v_{lk}/\sqrt{n}|-|\varphi_{lk}^0|\Big\} && (4.62)\\
&+\lambda_n\sum_{l=1}^{L}\sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\sum_{k=j+1}^{M_{l+1}}\hat{w}_{jk}|(v_{lj}-v_{lk})/\sqrt{n}| && (4.63)\\
&+\lambda_n\sum_{l=1}^{L}\sum_{c=l+1}^{L}\sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\sum_{k\in\mathscr{M}^+\cap\mathscr{M}_c}\hat{w}_{jk}\Big\{|\Delta_{lc}^0+\delta_{lc}/\sqrt{n}|-|\Delta_{lc}^0|\Big\} && (4.64)\\
&+\lambda_n\sum_{l=1}^{L}\sum_{c=l+1}^{L}\sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\sum_{k\in\mathscr{M}^-\cap\mathscr{M}_c}\hat{w}_{jk} && (4.65)\\
&\qquad\cdot\Big\{|\Delta_{lc}^0+\varphi_{ck}^0+(\delta_{lc}+v_{ck})/\sqrt{n}|-|(\delta_{lc}+v_{ck})/\sqrt{n}|\Big\} \\
&+\lambda_n\sum_{l=1}^{L}\sum_{c=l+1}^{L}\sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\sum_{k\in\mathscr{M}^+\cap\mathscr{M}_c}\hat{w}_{jk} && (4.66)\\
&\qquad\cdot\Big\{|\Delta_{lc}^0-\varphi_{lj}^0+(\delta_{lc}-v_{cj})/\sqrt{n}|-|(\delta_{lc}-v_{cj})/\sqrt{n}|\Big\} \\
&+\lambda_n\sum_{l=1}^{L}\sum_{c=l+1}^{L}\sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\sum_{k\in\mathscr{M}^-\cap\mathscr{M}_c}\hat{w}_{jk} && (4.67)
\end{aligned}
$$

$$\cdot\Big\{|\Delta_{lc}^0+\varphi_{ck}^0-\varphi_{cj}^0+\frac{\delta_{lc}+v_{ck}-v_{cj}}{\sqrt{n}}|-|\frac{\delta_{lc}+v_{ck}-v_{cj}}{\sqrt{n}}|\Big\}$$

If we denote the true value $\nu_0=(\nu_{10},\ldots,\nu_{m0})^\top$, then only when $j\in\mathscr{M}_l$ and $k\in\mathscr{M}_l$, $(l=1,\ldots,L)$ can $\nu_{j0}=\nu_{k0}$.

(1) When $\nu_{j0}\neq\nu_{k0}$, $\hat{w}_{jk}\to_p|\nu_{j0}-\nu_{k0}|^{-\alpha}$ and

$$\sqrt{n}\Big(|(\nu_{j0}-\nu_{k0})+\frac{u_j-u_k}{\sqrt{n}}|-|\nu_{j0}-\nu_{k0}|\Big)\to(u_j-u_k)\mathrm{sgn}(\nu_{j0}-\nu_{k0}).$$

Thus

$$\frac{\lambda_n}{\sqrt{n}}\hat{w}_{jk}\sqrt{n}\Big(|(\nu_{j0}-\nu_{k0})+\frac{u_j-u_k}{\sqrt{n}}|-|\nu_{j0}-\nu_{k0}|\Big)\to_p 0.$$

Therefore, the terms from (4.64) to (4.67) goes to 0 in probability as $n\to\infty$.

(2) When $\nu_{j0}=\nu_{k0}$,

$$\sqrt{n}\Big(|(\nu_{j0}-\nu_{k0})+\frac{u_j-u_k}{\sqrt{n}}|-|\nu_{j0}-\nu_{k0}|\Big)=|u_j-u_k|$$

and $\frac{\lambda_n}{\sqrt{n}}\hat{w}_{jk}=\frac{\lambda_n}{\sqrt{n}}n^{\alpha/2}(\sqrt{n}|\hat{\nu}_j-\hat{\nu}_k|)^{-\alpha}\to\infty$ since $\sqrt{n}\hat{\nu}_n=O_p(1)$.

For $l=1,\ldots,L,$, $j\in\mathscr{M}^+\cap\mathscr{M}_l, k=j+1,\ldots,M_{l+1}$,

$$\sqrt{n}(|\varphi_{lk}^0+v_{lk}/\sqrt{n}|-|\varphi_{lk}^0|)\to_p v_{lk}\mathrm{sgn}(\varphi_{lk}^0)$$

and

$$\frac{\lambda_n}{\sqrt{n}}\hat{w}_{jk}=\frac{\lambda_n}{\sqrt{n}}n^{\alpha/2}(\sqrt{n}|\hat{\nu}_j-\hat{\nu}_k|)^{-\alpha}\to\infty.$$

Thus, the term (4.62) goes to infinity except when $v_{lk}=0, k=2,\ldots,M_{l+1}$.

Similarly, for $l = 1, \ldots, L$, $j \in \mathscr{M}^- \cap \mathscr{M}_l, k = j+1, \ldots, M_{l+1}$, if

$$v_{lj} \neq v_{lk} \text{ then } \frac{\lambda_n}{\sqrt{n}} \hat{w}_{jk} \sqrt{n} |(v_{lj} - v_{lk})/\sqrt{n}| \to \infty.$$

In a word, for each group l, $l = 1, \ldots, L$, we require $v_{lj} = v_{lk} = 0, j, k = 2, \ldots, M_{l+1}$ in order to get the finite limit of the penalty term.

Therefore, by Slutsky's theorem, we have $V_n(u) \xrightarrow{D} V(u)$ for every u

$$V(\mathscr{T}(u)) = \begin{cases} \Big(\mathscr{T}(u)\Big)_{\mathscr{A}^+}^\top \big(\tilde{Q}\tilde{Q}^\top\big)_{\mathscr{A}^+} \Big(\mathscr{T}(u)\Big)_{\mathscr{A}^+} - 2\Big(\mathscr{T}(u)\Big)_{\mathscr{A}^+}^\top \big(\tilde{Q}\tilde{Q}^\top\big)_{\mathscr{A}^+} \xi_{\mathscr{A}^+}, \\ \qquad\qquad \text{if } \Big(\mathscr{T}(u)\Big)_s = 0 \text{ for } s \in \mathscr{A}^-, \\ \infty, \qquad\qquad \text{otherwise.} \end{cases}$$

The unique minimum of $V(\mathscr{T}(u))$ is $(\xi_{\mathscr{A}^+}^\top, 0)^\top$. Following the epi-convergence results of Geyer (1994), we have

$$\Big(\mathscr{T}(\hat{u}^{(n)})\Big)_{\mathscr{A}^+} \xrightarrow{D} \xi_{\mathscr{A}^+} \text{ , and } \Big(\mathscr{T}(\hat{u}^{(n)})\Big)_{\mathscr{A}^-} \xrightarrow{D} 0.$$

Note that the transformation $\mathscr{T}$ is linear, we have

$$\Big(\mathscr{T}(\hat{u}^{(n)})\Big)_{\mathscr{A}^+} = \sqrt{n}\Big(\mathscr{T}(\hat{\nu}^{*(n)})_{\mathscr{A}^+} - \mathscr{T}(\nu_0)_{\mathscr{A}^+}\Big) \xrightarrow{D} N(0, G_{\mathscr{A}^+}),$$

then we complete the proof of the asymptotic normality part.

- Part III. (Consistency)

  For $\forall j \in \mathscr{A}^+ = \{1, \ldots, L\}$, based on the asymptotic normality result, we have

  $$\hat{\tau}_j^{*(n)} \to_p \phi_j \text{ thus } P(j \in \mathscr{A}_n^+) \to 1.$$

Then it suffices to show that $\forall j \in \mathscr{A}^-$, $P(j \in \mathscr{A}_n^+) \to 0$. Take the first derivative of the loss function on $\hat{\tau}_j^{*(n)}$ and divide by $\sqrt{n}$, we obtain

$$-2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{(\mathscr{T}(Y_i))_j - \hat{\tau}_j^{*(n)}\right\} \tag{4.68}$$

Note that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{(\mathscr{T}(Y_i))_j - \hat{\tau}_j^{*(n)}\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathscr{T}(Y_i))_j - \sqrt{n}\hat{\tau}_j^{*(n)} = O_p(1), \tag{4.69}$$

from (4.15) and $\sqrt{n}\hat{\tau}_j^{*(n)} \xrightarrow{D} 0$ for $j \in \mathscr{A}^-$.

Now we deal with the penalty function. For the $j \in \mathscr{A}^-$, there exists some l (group) and $s \in \{2, \ldots, m_l\}$ such that

$$\hat{\tau}_j^{*(n)} = \hat{\phi}_{ls}^{*(n)}.$$

The first derivative of the penalty function on $\hat{\tau}_j^{*(n)}$ divided by $\sqrt{n}$ is

$$\begin{aligned}
H_n \;:=\; & \frac{\lambda_n}{\sqrt{n}}\sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\hat{w}_{js}\mathrm{sgn}(\hat{\phi}_{ls}^{*(n)}) \\
& + \frac{\lambda_n}{\sqrt{n}}\sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\hat{w}_{js}\mathrm{sgn}(\hat{\phi}_{lj}^{*(n)} - \hat{\phi}_{ls}^{*(n)})(-1) \\
& + \frac{\lambda_n}{\sqrt{n}}\sum_{c=l+1}^{L}\sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\sum_{k\in\mathscr{M}^+\cap\mathscr{M}_c}\hat{w}_{jk}\mathrm{sgn}(\hat{\Delta}_{lc}^{*(n)})C_1 \\
& + \frac{\lambda_n}{\sqrt{n}}\sum_{c=l+1}^{L}\sum_{j\in\mathscr{M}^+\cap\mathscr{M}_l}\sum_{k\in\mathscr{M}^-\cap\mathscr{M}_c}\hat{w}_{jk}\mathrm{sgn}(\hat{\Delta}_{lc}^{*(n)} + \hat{\phi}_{ck}^{*(n)})C_2 \\
& + \frac{\lambda_n}{\sqrt{n}}\sum_{c=l+1}^{L}\sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\sum_{k\in\mathscr{M}^+\cap\mathscr{M}_c}\hat{w}_{jk}\mathrm{sgn}(\hat{\Delta}_{lc}^{*(n)} - \hat{\phi}_{lj}^{*(n)})C_3
\end{aligned}$$

$$
\begin{aligned}
&+\frac{\lambda_n}{\sqrt{n}}\sum_{c=l+1}^{L}\sum_{j\in\mathscr{M}^-\cap\mathscr{M}_l}\sum_{k\in\mathscr{M}^-\cap\mathscr{M}_c}\hat{w}_{jk}\mathrm{sgn}(\hat{\Delta}_{lc}^{*(n)}+(\hat{\phi}_{cj}^{*(n)}-\hat{\phi}_{cj}^{*(n)}))C_4\\
=\;&\frac{\lambda_n}{\sqrt{n}}\sum_{j=M_l+1}^{M_{l+1}}\hat{w}_{js}C_j'+\frac{\lambda_n}{\sqrt{n}}\sum_{c=l+1}^{L}\sum_{j\in\mathscr{M}_l}\sum_{k\in\mathscr{M}_c}\hat{w}_{jk}D_{jk}\\
:=\;&H_{1n}+H_{2n},
\end{aligned}
$$

where C_1, C_2, C_3, C_4 and C_j', D_{jk} are constants. By the KKT optimality conditions, we know that

$$
-2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{(\mathscr{T}(Y_i))_j-\hat{\tau}_j^{*(n)}\Big\}=H_n.
$$

Therefore

$$
P(j\in\mathscr{A}_n^+)\le P\Big(-2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{(\mathscr{T}(Y_i))_j-\hat{\tau}_j^{*(n)}\Big\}=H_n\Big).
$$

Recall that $j\in\mathscr{A}^-$, similarly as the discussion in the proof of Part II, we have

$$
H_{1n}\to_p\infty \text{ and } \quad H_{2n}\to_p 0
$$

which implies $H_n\to_p\infty$. Therefore,

$$
P(j\in\mathscr{A}_n^+)\le P\Big(-2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{(\mathscr{T}(Y_i))_j-\hat{\tau}_j^{*(n)}\Big\}=H_n\Big)\to 0.
$$

We complete the proof. □

CHAPTER 5

Conclusions and Future Work

In Chapter 2 of the thesis, we proposed to select the threshold variable of the Smooth Threshold Autoregressive (STAR) model by the recently developed L_1 regularization approach. Oracle properties of the adaptive lasso estimator have been obtained. Compared with the threshold variable selection via the hypothesis testing method or classification methods, this method can produce a parsimonious number of nonzero coefficients for the threshold variable, thus leading to a simple way of selecting the threshold variable. In this chapter, a new penalizing approach, Direction Adaptive Lasso (DAL), was also proposed specially for the three models where the shape of the link function cannot be neglected. It was shown from numerical studies that by penalizing the direction of the coefficient vector instead of the coefficients themselves, the threshold variable is more accurately selected. A possible explanation is that the norm of the coefficient vector implies the threshold

shape which should not be penalized. Real data analysis also suggests that the proposed method is able to select the threshold variable more efficiently than the general L_1 regularization method, especially when the sample size is small. Moreover, the experimental result on the popular analyzed real data (Lynx Data Set) is in agreement with the result of the previous studies. This study is very useful in practical application because larger sample size requires more time and money cost and it is not always possible to obtain large sample sets in high-dimensional spaces since an exponentially increasing number of data points are required with increasing dimension. This study has provided a new perspective of threshold variable selection and extended the previous adaptive lasso method to a more efficient one. However, the studies on the new method are restricted to the one specific type of model and the effect of the shape of the link function was only examined numerically. Based on the good numerical performance of the proposed method, further research is needed to examine the theoretical results on the effect of the shape of the link function on the variable selection. In this way, future study could attempt to identify a general class of model where this method can be applied to improve the variable selection efficiency.

In Chapter 3, motivated by the compelling need to improve the numerical stability in high dimension and by practical examples in which different coefficient functions are linearly dependent, we proposed a new varying coefficient model, PVCM, which incorporates the intrinsic patterns in the coefficients. Combined with the kernel smoothing approach, the limiting distributions of the estimators have been obtained under regular conditions. Moreover, incorporating with the L_1 penalty, the estimation can automatically select variables in the linear part and

the nonlinear part. It was shown that the L_1 estimator has the oracle properties. The model possesses superior estimation efficiency over VCM. The advantage of PVCM over VCM increases as p increases. PVCM reduces the actual number of nonparametric functions, and thus has better estimation efficiency. Numerical studies including both simulation study and real data analysis also suggest that the model together with the kernel smoothing estimation method has good estimation performance and is numerically stable even when the number of covariates is large. The gain in estimation efficiency and numerical stability is due to further model identification that only a small number of principal functions need to be estimated non-parametrically, regardless of which smoothing method is used. The key benefit of the proposed model is that the estimation efficiency only depends on a few principal functions. Principal Varying Coefficient Model (PVCM) together with the estimation methods provides a powerful approach towards the analysis of complicated data and results in a considerable improvement for solving the issue "curse of dimensionality". However, this study did not consider the smoothing methods other than the kernel smoothing. Kernel smoothing is popularly used in nonparametric modeling and more theoretically convenient to study. In addition, it is the kernel smoothing that causes the numerical instability in high dimension case. Therefore, this study only focuses on this estimation method. However, the proposed model is a semi-parametric model and thus can be estimated based on other smoothing methods such as spline smoothing. Recent advances indicate that the spline smoothing and the penalized splines enjoy many good properties. See, for example, Wood (2006), and Ruppert *et al* (2009). It would be interesting to incorporate the spline smoothing into the proposed model. The estimation performance based on the splines smoothing needs further investigation.

In Chapter 4, based on the way of dependence in epidemiology, finance study and genetic analysis, where the variates usually function in blocks, we have considered a special L_1 penalty, called cLasso. We have shown that cLasso can achieve the goal of identifying the blocks when the penalty parameters are selected appropriately. On the other hand, the calculation results in all the examples suggest that the sparsity assumption is not appropriate due to its bigger prediction error than the simple regression or ridge regression. Instead, cLasso has much smaller prediction error in all the examples. Moreover, we applied the cLasso to the estimation of covariance matrix. Numerical examples showed that the estimation performance cLasso on the covariance matrix is better than that of Lasso and ridge in some cases.We obtained the oracle properties of cLasso on the covariance matrix. However, it is under the condition that n goes to infinity while p keep fixed. Since the high/ultra-high dimensional problems attract more interest of research with the development of modern technologies, the case $p = p(n) \to \infty$ as $n \to \infty$ would be interesting to be investigated.

Bibliography

[1] Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-722.

[2] An, H. Z. and Huang, F. C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statistica Sinica*, **6**, 943-956.

[3] Belsley, Kuh and Welsch (1980). *Regression diagnostics: identifying influential data and sources of collinearity.* Wiley, New York.

[4] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, **36**, 2577-2604.

[5] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, **24**, 2350-2383.

[6] Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, **95**, 888-902.

[7] Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, **92**, 477-489.

[8] Chan, K. S. and Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, **7**, 179-190.

[9] Chen, R. (1995). Threshold variable selection of open-loop threshold AR models. *Journal of Time Series Analysis*, **16**, 461-481.

[10] Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). Local regression models. In *Statistical Models in S* (Chambers, J. M. and Hastie, T. J., eds), 309 -376. Wadsworth & Brooks, Pacific Grove.

[11] David, H. A. and Gunnink, J. L. (1997). The paired t test under artificial pairing. *The American Statistician* **51**, 9-12.

[12] van Dijk, D. Teräsvirta, T. and Franses, P.H. (2002). Smooth transition autoregressive models - a survey of recent developments. *Econometric Reviews*, **21**, 1-47.

[13] Duan, N. and Li, K.-C. (1991). Slicing regression: A link-free regression method. *Annals of Statistics*, **19(2)**, 505-530.

[14] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407-451.

[15] Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, **33**, 3-56.

[16] Fan, J. and Gijbels, I. (1996). *Local polynomial modeling and its applications.* Chapman and Hall, New York.

[17] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**, 1031-1057.

[18] Fan, J. and Jiang, J. (2005). Nonparametric inferences for additive models. *Journal of the American Statistical Association*, **100**, 890-907.

[19] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

[20] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. (with discussion). *Journal of Royal Statistical Society B*, **70**, 849-911.

[21] Fan, J. and Yao, Q. (2003). *Nonlinear time series. Nonparametric and parametric methods*. Springer-Verlag, New York.

[22] Fan, J. and Zhang, J. T. (2000). Two-step estimation of functional linear model with application to longitudinal data. *Journal of the Royal Statistical Society, Series B*, **62**, 303-322.

[23] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, **27**, 1491-1518.

[24] Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesestesting in varying-coefficient models. *Scandinavian Journal of Statistics*, **27**, 715-731.

[25] Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, **1**, 179-195.

[26] Frank, I.E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

[27] Franses, P. H. and van Dijk, D. (2000). *Nonlinear time series models in empirical finance*. Cambridge, New York.

[28] Fu, W. (1998). Penalized regressions: the bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, **7** 397-416.

[29] Geyer, C. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics*, **22** 1993-2010.

[30] Golub, T. R., Slonim, D. K. , Tamayo, P. , Huard, C., Gaasenbeek, M. , Mesirov, J. P. , Coller, H. , Loh, M. L. , Downing, J. R. ,

Caligiuri, M. A. , Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286** 531-537.

[31] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, **21**, 157-178.

[32] Hastie, T. J. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, **84** 502-516.

[33] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman & Hall.

[34] Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society*, **55**, 757-796.

[35] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.

[36] Huang, J., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85-98.

[37] Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**, 111-128.

[38] Knight, K. (1999). Epi-convergence and stochastic equisemicontinuity. *Technical Report*, University of Toronto, Department of Statistics (http://www.utstat.toronto.edu/keith/papers/).

[39] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics*, **28**, 1356-1378.

[40] Klimko, L. A. and Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes. *Annals of Statistics*, **6**, 629-642.

[41] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics*, **37**, 4254-4278.

[42] Levina, E., Rothman, A. J. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Annals of Applied Statistics*. **2**, 245-263.

[43] Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrasch. Verw. Gebiete*, **61**, 405-415.

[44] Matteson, D. and Tsay, R. (2011). Multivariate volatility modeling: brief review and a new approach. Manuscript, Booth School of Business, University of Chicago.

[45] Osborne, M., B. Presnell, and B. Turlach (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319-337.

[46] Ruppert, D., Wand, M. P. and Carroll, R. J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics*, **3**, 1193-1256.

[47] Siegfried, T. (2010). Odds are, it's wrong: science fails to face the shortcomings of statistics. *Science News* **177**, 26-28.

[48] She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics*, **4**, 1055-1096.

[49] Stock, J. H. and M. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, **97**, 1167-1179.

[50] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, **58**, 267-288.

[51] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society*, **67**, 91-108.

[52] Tong, H. (1990). *Nonlinear time series. A dynamical system approach*. Oxford University Press, New York.

[53] Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society*, **42**, 245-292.

[54] Tsay, R. S. (1989). Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association*, **405**, 231-240.

[55] Wang, H. (2009). Rank reducible varying coefficient model. *Journal of Statistical Planning and Inference*, **139**, 999-1011.

[56] Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553-568.

[57] Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Associate*, **103**, 811-821.

[58] Wood S.N. (2006). *Generalized additive models: An introduction with R.* Chapman & Hall/CRC Press.

[59] Wu, H. and Liang, H. (2004). Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scandinavian Journal of Statistics*, **31**, 3-19.

[60] Wu, S, and Chen, R. (2007). Threshold variable selection and threshold variable driven switching autoregressive models. *Statistica Sinica*, **17**, 241-264.

[61] Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 831-844.

[62] Xia, Y. and Tong, H. (2006). Cumulative effects of air pollution on public health. *Statistics in Medicine*, **25**, 3548-3559.

[63] Yuan, M. and Lin, Y (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, **68**, 49-67.

[64] Zhang, W. H., Collins, A. Maniatis, Tapper, W. and Morton, N. E. (2002). Properties of linkage disequilibrium (LD) maps, *Proceedings of the National Academy of Sciences of the United States*, **99**, 17004-17007.

[65] Zhang, W. Y., Lee, S. Y., and Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis*, **82**, 166-188.

[66] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, **7**, 2541-2563.

[67] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.

[68] Zou, H. and Trevor H. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society B.* **67**, 301-320.