VARIABLE SELECTION AND FUNCTION ESTIMATION USING PENALIZED METHODS

A Dissertation

by

GANGGANG XU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2011

Major Subject: Statistics

VARIABLE SELECTION AND FUNCTION ESTIMATION USING PENALIZED METHODS

A Dissertation

by

GANGGANG XU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Suojin Wang
	Jianhua Huang
Committee Members,	Raymond J. Carroll
	Jianxin Zhou
Head of Department,	Simon J. Sheather

December 2011

Major Subject: Statistics

ABSTRACT

Variable Selection and Function Estimation

Using Penalized Methods. (December 2011)

Ganggang Xu, B.S., Zhejiang University;

M.S., Texas A&M University

Co–Chairs of Advisory Committee: Dr. Suojin Wang
Dr. Jianhua Huang

Penalized methods are becoming more and more popular in statistical research. This dissertation research covers two major aspects of applications of penalized methods: variable selection and nonparametric function estimation. The following two paragraphs give brief introductions to each of the two topics.

Infinite variance autoregressive models are important for modeling heavy-tailed time series. We use a penalty method to conduct model selection for autoregressive models with innovations in the domain of attraction of a stable law indexed by $\alpha \in (0, 2)$. We show that by combining the least absolute deviation loss function and the adaptive lasso penalty, we can consistently identify the true model. At the same time, the resulting coefficient estimator converges at a rate of $n^{-1/\alpha}$. The proposed approach gives a unified variable selection procedure for both the finite and infinite variance autoregressive models.

While automatic smoothing parameter selection for nonparametric function estimation has been extensively researched for independent data, it is much less so for clustered and longitudinal data. Although leave-subject-out cross-validation (CV) has been widely used, its theoretical property is unknown and its minimization is computationally expensive, especially when there are multiple smoothing parameters. By focusing on penalized modeling methods, we show that leave-subject-out CV is

optimal in that its minimization is asymptotically equivalent to the minimization of the true loss function. We develop an efficient Newton-type algorithm to compute the smoothing parameters that minimize the CV criterion. Furthermore, we derive one simplification of the leave-subject-out CV, which leads to a more efficient algorithm for selecting the smoothing parameters. We show that the simplified version of CV criteria is asymptotically equivalent to the unsimplified one and thus enjoys the same optimality property. This CV criterion also provides a completely data driven approach to select working covariance structure using generalized estimating equations in longitudinal data analysis. Our results are applicable to additive, linear varying-coefficient, nonlinear models with data from exponential families.

To my parents and all my grandparents.

ACKNOWLEDGMENTS

First of all, I would like to thank all my committee members: Dr. Suojing Wang, Dr. Jianhua Huang, Dr. Jianxin Zhou and Dr. Raymond J. Carroll. I am very fortunate to have them to guide my work in Texas A&M University.

Special thanks go to Dr.Wang and Dr.Huang, who are my Ph.D. dissertation advisors. During my past five years in Texas A&M University, they have given me numerous helpful advices in my research and daily life. Dr. Wang is a world class researcher in nonparametric statistics, a great teacher and a close friend of mine. Being a graduate student in a foreign country can be extremely difficult with tons of issues other than school work to deal with. Whenever I need help, Dr. Wang was always there and was always supportive for all decisions I made. He would even help fix my old car when it broke down. I respect him for his enthusiasm for work, his attitude to life, his vision of the future and his generosity to his students. I want to thank him for everything he did for me.

Dr. Huang is a very energetic and productive professor who is well known for his work in nonparametric statistics and functional data analysis. He is the one who showed me the door to research in statistics. I was always impressed by his broad and deep knowledge of all areas of statistics and his dedication to first class collaborative research with researchers from outside the Department. He guided me into many exciting areas of statistics and created the best possible research environment for me. I am very grateful to him for all his invaluable help during my years at Texas A&M University.

Dr. Carroll is a world leading statistician and a great teacher. It is a honor to have him on my committee. His insights in research helped me improve my work and made the results more profound than I originally thought. Dr. Zhou is an expert in

numerical computation and optimization. I want to thank him for his input in the computational aspect of the algorithm, which made the algorithm more efficient and stable.

I also want to thank Dr. Michael Longnecker, who is such a great teacher and a good friend. When I was working as a graduate teaching assistant, he has been constantly helpful by giving me useful advice on teaching, encouraging me when I was frustrated and backing me up in front of my classes. He is the reason I fell in love with becoming a good teacher. As the Associate Department Head, he is extremely responsible and efficient, providing timely help to all students and faculty members. I feel very lucky to have him around in my journey to my Ph.D. in Statistics.

I want to thank all my friends and colleagues in the Department. You have made my life in College Station so colorful and exciting. Finally I would like to thank my family, especially my parents. Your support has been essential.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1. Variable selection using penalized methods

Heavy-tailed time series data is often encountered in a variety of fields, such as hydrology (Castillo, 1988), economics and finance (Koedijk et al., 1990) and teletraffic engineering (Duffy et al., 1994). In this situation, the infinite variance autoregressive model is often preferred to the finite variance one, and its statistical theory has been widely studied in the literature. See Resnick (1997) for a comprehensive review and further references.

Model selection is an important aspect of modeling with time series data. An unnecessarily complex model can degrade the efficiency of the resulting parameter estimators and lead to less accurate predictions. For a time series model with finite variance, traditional model selection criteria AIC (Akaike, 1973) and BIC (Schwarz, 1978) can be employed to choose the order of the autoregressive model (McQuarrie and Tsai, 1998). Compared to the case of finite variance autoregressive models, few papers have investigated the model selection for autoregressive models with infinite variance. Bhansali (1988) considered the order determination of the infinite variance autoregressive processes with innovations in the domain of attraction of a stable law, and gave a consistent estimator of the order. Knight (1989) studied the same model and showed that the order selection with AIC is weakly consistent. While most of the literature focuses on the order determination of the time series, Ling (2005) proposed a self-weighted least absolute deviation estimator for the infinite variance autoregressive model under which the coefficient estimates are asymptotically normal

[1]The journal model is *Journal of the American Statistical Association*.

and thus can be used for statistical inference. He also proposed a variable selection procedure with a series of hypothesis tests based on the self-weighted least absolute deviation estimator. However, his method can be unstable and its implementation is complicated.

Using the shrinkage method for variable selection is relatively new in time series literature. Wang et al. (2007a) applied adaptive lasso (Zou, 2006) to the regression model with finite autoregressive errors. They showed that the resulting estimator via adaptive lasso not only has a sparse presentation, but also has the oracle property (Fan and Li, 2001), which means that it can simultaneously select variables and estimate parameters in time series modeling.

One difficulty often encountered in data analysis is that it is generally impossible to know whether a time series of finite length has infinite variance (Granger and Orr, 1972). Many methods have been developed to test for infinite variance of a real time series data; see, for example, Hill (1975). While Wang et al. (2007a)'s method does not apply to infinite variance autoregressive models, using Ling (2005)'s method can cause loss of important information by weighing down large observations, especially in the case of a time series with heavy tails but finite variance.

In Chapter III, we first use the self-weighted least absolute deviation proposed by Ling (2005) as the loss function and the adaptive lasso as the penalty method to do the model selection. Under appropriate conditions, we show that our penalized method can identify the true model consistently and the estimator of the coefficients corresponding to the true model is asymptotically normal, which is important for the statistical inference of infinite variance autoregressive models. After that, we propose a unified variable selection approach that can efficiently deal with heavy-tailed autoregressive models with either finite or infinite variance. By combining the least absolute deviation as the loss function and the adaptive lasso as the penalty

function, we show that under regularity conditions we can identify the true model consistently and obtain a point estimator of the coefficients corresponding to the true model with a convergence rate of $n^{-1/\alpha}$, where $\alpha \in (0, 2)$ is the index of the stable distribution. This convergence rate is faster than that of finite variance time series.

1.2. Nonparametric function estimation using longitudinal data

Longitudinal data analysis has been a subject of intense research in statistics for the past 30 years. Various parametric models (e.g. Vonesh and Chinchilli, 1997; Diggle et al., 2002) and nonparametric or semi-parametric models (e.g. Hart and Wehrly, 1986; Rice and Silverman, 1991; Zeger and Diggle, 1994; Fan and Zhang, 2000; Lin and Carroll, 2000; Wang et al., 2005) have been proposed and studied. In a typical set of longitudinal data, we have observations $(y_{ij}, \boldsymbol{x}_{ij})$, for $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, where y_{ij} is the response variable of jth measurement of the ith subject and the $\boldsymbol{x}_{ij}$ is the corresponding $p \times 1$ vector of covariates. It is reasonable to assume that observations from different subjects are independent and observations within a subject are correlated. For the longitudinal data analysis, there are three main modeling families: marginal models, mixed-effect models, and transition models (Diggle et al., 2002). In Chapter IV, we focus on the marginal approach using generalized estimating equations (GEE, Liang and Zeger, 1986).

By introducing a parameterized working correlation, GEE method has the potential to increase the efficiency of the regression estimates when the marginal distribution of response are from exponential family. More specifically, y_{ij} is from exponential family with mean μ_{ij} and variance v_{ij},

$$f(y_{ij}) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}$$

where $\mu_{ij} = b'(\theta_{ij})$, $v_{ij} = \phi b''(\theta_{ij})$ and with the link function $g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$.

One limitation of the work of Liang and Zeger (1986) is its inflexibility because it assumes fully parametric relationship between the response and covariates. Non-parametric and semi-parametric models are developed to model more complicated relationships in the longitudinal data setup. These work include generalized additive models (GAM, Wild and Yee, 1996; Berhane and Tibshirani, 1998; Lin and Zhang, 1999), varying coefficient models (Hoover et al., 1998; Chiang et al., 2001; Huang et al., 2002), partially linear models (Zeger and Diggle, 1994; He et al., 2002; Wang et al., 2005; Huang et al., 2007), and partial linear varying coefficient model (Ahmad et al., 2005). All above mentioned models can be viewed as special cases with link function defined as:

$$g(\mu_{ij}) = \boldsymbol{x}_{ij0}\boldsymbol{\beta}_0 + \sum_{k=1}^{m} f_k(\boldsymbol{x}_{ijk}),$$

where $\boldsymbol{x}_{ij0}\boldsymbol{\beta}_0$ is the strictly parametric part of the model and f_k, $(k = 1, \ldots, m)$ are unknown smooth functions ($\boldsymbol{x}_{ijk}$ can either be a scaler or a vector).

The flexibility of the GAM is also accompanied by the potential risk of over-fitting the data. Broadly speaking, the estimation of nonparametric terms in (4.2) can be classified into kernel methods (Wand and Jones, 1995) and spline methods (Green and Silverman, 1994). The kernel method avoids over-fitting by selecting appropriate bandwidth for each nonparametric component using cross validation. However, the estimation of kernel coefficient itself can be computationally challenging, the selection of bandwidths can be computationally prohibitive for generalized additive models, if not impossible. In addition, Welsh et al. (2002) pointed out that, by taking into account of the within subject correlation, the spline methods appear to be more efficient than the kernel methods in nonparametric marginal regression model. In Chapter IV, we using spline methods to estimate the nonparametric function. To

avoid over-fitting, Berhane and Tibshirani (1998) proposed to use the "Penalized Quansi-Likelihood" criterion:

$$P(f_1, \cdots, f_k) = Q(\boldsymbol{\eta}; \boldsymbol{y}) - \frac{1}{2}\sum_{k=1}^{m} \lambda_k J(f_k)$$

where $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, $Q(\boldsymbol{\eta}; \boldsymbol{y})$ is some quasi-likelihood score function based on data, $J(\cdot)$ is some penalty functional and $\lambda_1, \cdots, \lambda_k$ are smoothing parameters controlling the tradeoff between model fit and model complexity.

(*Example 1: Partial linear additive model*) The response Y is related to the covariates $X = (X_1, \cdots, X_m)^T \in R^m$ and $Z = (Z_1, \cdots, Z_d)^T \in R^d$ in the way:

$$\mu = E(Y|X = x, Z = z) = g^{-1}(z^T\beta + \sum_{k=1}^{m} f_k(x_k))$$

where x_k is the kth component of x, β is a d-dimensional vector and f_k's are unknown and smooth functions. Then in this case, we can take $Q(\boldsymbol{\eta}; \boldsymbol{y})$ as the log-likelihood function of $\boldsymbol{y}$ and the penalty functional defined as $J(f_k) = \int [f''(x_k)]^2\, dx_k$.

(*Example 2: Varying coefficient model*) Hoover et al. (1998) considered the model

$$y_{ij} = X_{ij}^T\beta(t_{ij}) + \epsilon_i(t_{ij}),$$

where $\beta(t) = (\beta_0(t), \cdots, \beta_m(t))^T$ $(m \geq 0)$ are unknown smooth functions of t, $\epsilon_i(t)$ is a realization of a zero-mean stochastic process $\epsilon(t)$, $(t \in R)$, and X_{ij} and ϵ_i are independent. In this case, $f_k(\cdot)$'s are bivariate functions except for $k = 0$. More specifically, for $k = 0, \cdots, m$, we would have $f_k(\boldsymbol{x}_{k,ij}) = X_{k,ij}\beta_k(t_{ij})$ where $X_{k,ij}$ is the kth component of X_{ij} and $X_{0,ij} = 1$. Here, we take $Q(\boldsymbol{\eta}; \boldsymbol{y}) = -\sum_{i,j}(y_{ij} - X_{ij}^T\beta(t_{ij}))^2$ and the penalty term as $J(f_k) = \int [\beta_k''(t)]^2\, dt$ for $k = 0, \cdots, m$.

The choice of λ_k's is critical for getting a good function estimator. If there is no intra-subject correlation and treat all observations as independent data points, the asymptotic optimality of generalized cross validation (GCV, Craven and Wahba,

1979) in selecting smoothing parameters has been showed by Li (1986) in ridge regression and by Gu and Ma (2005) in the case of mix effect model. However, when there is intra-subject correlation as in longitudinal or cluster data, smoothing parameters selection is still an open problem. One of the popular procedure in this area is called "leave-subject-out cross-validation" (LsoCV), see Rice and Silverman (1991); Hoover et al. (1998); Huang et al. (2002). As popular as it is, there are several issues with this procedure. First of all, the computational cost of doing cross-validation is expensive. Furthermore, in the current practice in longitudinal study, researchers still rely on the grid search to find the optimal λ's using leave one subject out cross-validation. Because of this, current research can only deal with one or two smoothing parameters, searching in a higher dimension is not feasible. This is especially not desirable in varying coefficient model where each nonparametric component is supposed to receive different amount of penalty. The other issue with the LsoCV method is that even though it is widely used, no theoretical properties nor a systematic algorithm for it have yet been developed.

In Chapter IV, we first derive a short cut formulae for the LsoCV score and show that it is asymptotically optimal in selecting smoothing parameters in the sense that under certain conditions, minimizing LsoCV score is equivalent to minimizing the MSE of the function estimator when number of subjects goes to infinity. We then propose a new computationally more efficient criterion for choosing optimal smoothing parameters while maintain the asymptotical optimality. Based on the new criterion, a Newton-Raphson type algorithm is developed for automatically selecting multiple smoothing parameters. In the end, a completely data driven approach of selecting the best working covariance structure is proposed based on the LsoCV method.

CHAPTER II

LITERATURE REVIEW FOR CHAPTER III

2.1. Stable distribution: modeling heavy tailed distribution

Heavy-tailed time series data is often encountered in a variety of fields, such as hydrology (Castillo, 1988), economics and finance (Koedijk et al., 1990) and teletraffic engineering (Duffy et al., 1994). But what is a heavy tail? We use the definition proposed in Resnick (1997). A random variable X is said to have a *light tailed* distribution if it decay exponentially fast as $x \to \infty$,

$$P[|X| > x] \sim \frac{1}{\sqrt{2\pi}} \frac{\exp(-x^2/2)}{x} \to 0.$$

The most famous example in this class is the Normal distribution. A random variable X is said to have a *heavy tailed* distribution $F(x)$ with index $\alpha > 0$ if, for $x > 0$,

$$P(|X| > x) = x^{-\alpha} K(x), \tag{2.1}$$

where $K(x)$ is some slowly varying function, that is, for $x > 0$

$$\lim_{t \to \infty} \frac{K(tx)}{K(t)} = 1.$$

This definition implies that

$$\begin{cases} E(|X|^\beta) < \infty, & \beta < \alpha, \\ E(|X|^\beta) = \infty, & \beta > \alpha. \end{cases} \tag{2.2}$$

As mentioned in Resnick (1997), typical examples of $K(x)$ include:

$$K(x) = \begin{cases} c, & \text{Pareto distribution;} \\ c + o(1), & \text{Stable distribution;} \\ \log(x), & x > 1; \\ 1/\log(x), & x > 1; \end{cases}$$

The difference term $o(1)$ between the tail behavior of the pareto distribution and the stable distribution may look negligible, but it can cause big differences in detecting two types of tails.

One thing worth mentioning is that, when the tail index of $F(x)$ is less than 2, that is $\alpha < 2$, the random variable X would have infinite variance. One consequence is that when a time series or other stochastic processes have error terms of infinite variance, many of the classical methods of analysis based on second moments, for example, regression, autoregressive models and spectral analysis, may not be used properly for such series (Granger and Orr, 1972).

To model distributions with infinite variance, one of the popular choices is the stable distribution law, which can be defined in several different ways. Granger and Orr (1972) gives a detailed summary of stable distribution and we cite some of their results here.

Definition: A distribution function $F(x)$ is called *stable* if for every $a_1 > 0, b_1$ and $a_2 > 0, b_2$, there exists corresponding a and b such that the equation

$$F(a_1 x + b_1) * F(a_2 x + b_2) = F(ax + b)$$

holds, where $*$ denotes the convolution operator.

This definition guarantees the additive property in that if X and Y are indepen-

dent random variables having the same stable distribution function $F(\cdot)$, then the sum $X+Y$ also has the same stable distribution function $F(\cdot)$. This additive definition of stable distribution results in a generalized version of central limit theorem as the following (Granger and Orr, 1972).

Generalized Central Limit Theorem Let $\{X_n\}$ be a sequence of *iid* random variables and $\{a_n\}$ and $\{b_n\}$ be two sequence of numbers, define the sums

$$S_n = \frac{1}{a_n}\sum_{i=1}^{n} X_i - b_n.$$

If weighted sum sequence S_n converges in distribution as $n \to \infty$, then it must converge to a random variable with a stable distribution.

This theorem provides a heuristic justification for the use of stable distribution to model the error terms in time series. For example, if a variable in an economic time series can be considered as sums of a large number of independent terms (like the stock price, which can be viewed as a consequence of numerous independent transactions), the distribution of the series might have infinite variance, when the infinite invariance stable distribution may be a reasonable tool to model this types of data.

A necessary and sufficient condition for the distribution function $F(\cdot)$ to be stable is that its characteristic function $\phi(t)$ admits the following representation:

$$\phi(t) = \exp\{i\gamma t - \delta|t|^{\alpha}[1 + i\beta sgn(t)w(t,\alpha)]\},$$

where $i = \sqrt{-1}$, $0 < \alpha \leq 2$, $-1 \leq \beta \leq 1$, $\delta \geq 0$, γ is any real number and functions $w(t,\alpha)$ and $sgn(\cdot)$ are defined as

$$w(t,\alpha) = \begin{cases} \tan\frac{\pi\alpha}{2}, & \alpha \neq 1; \\ \frac{2}{\pi}\log|t|, & \alpha = 1; \end{cases}, \quad sgn(t) = \begin{cases} 1, & t > 0; \\ 0, & t = 0; \\ -1, & t < 0. \end{cases}$$

This characterization completely describes all members of stable distribution family. Unfortunately, for most values of parameters α, β, γ and δ, $F(\cdot)$ does not have a analytical form. Two special cases are, if $\alpha = 2$, $F(\cdot)$ is the cumulative distribution function of Normal distribution and $\alpha = 1, \beta = 0$ case corresponds to the Cauchy distribution.

Definition For a sequence of *iid* random variables $\{X_n\}$ with distribution function $F(\cdot)$, $F(\cdot)$ is said to belong to the *domain of attraction* of a stable distribution if for some sequence of numbers $\{a_n\}$ and $\{b_n\}$,

$$S_n = \frac{1}{a_n} \sum_{i=1}^{n} X_i - b_n \to F(\cdot) \quad \text{in distribution, as } n \to \infty.$$

A sufficient and necessary condition for the distribution function $F(\cdot)$ to be in *the domain of attraction of the stable law* with index $\alpha \in (0, 2)$ is that

$$\lim_{x \to \infty} \frac{P(X_1 > x)}{P(|X| > x)} \equiv q \in [0, 1]$$

exists and

$$P(|X_1| > x) = x^{-\alpha} K(x),$$

where $K(x)$ is a slowly varying function defined in equation (2.1). This condition together with equation (2.2) implies that distribution functions belong to the domain of attraction of the stable law have heavy tails. Furthermore, since $\alpha \in (0, 2]$, except the Normal case ($\alpha = 2$), all members of stable distributions have infinite variance and even infinite first moment for those $\alpha < 1$.

In Chapter III, we shall use the assumption that the innovations of the autoregressive models with infinite variance belong to the domain of attraction of the stable law with index $\alpha \in (0, 2]$.

2.2. Test for infinite variance: the Hill estimator

As stated in Granger and Orr (1972), having observed a series $\{y_1, \cdots, y_n\}$ with a finite length, it is usually impossible to distinguish whether it has infinite variance or not. Among many others, one important reason why identifying heavy tail distribution is necessary is due to the efficiency of estimation. For a parametric model, the most efficient estimator for parameters in the model is the maximum likelihood estimate. Misspecification of distributions of the observations would lead to loss in efficiency of estimators. For example, it is shown in Davis et al. (1992) that, for autoregressive models with innovations belong to the domain of attraction of the stable law with index $\alpha \in [1, 2)$, the least absolute deviation (LAD) estimator is asymptotically much more efficient than the least square (LS) estimator. However, if the innovations are from Gaussian process, then the LS estimator is the most efficient estimator that outperforms the LAD estimator. So in this sense, it is important to identify whether the heavy tail exists in a observed process in order to choose most efficient estimation tools.

Many numerical and graphical testing procedures have been proposed for testing the existence of heavy tail distributions. One of the widely used procedure is the Hill estimator (Hill, 1975). Suppose $X_1, \cdots, X_n$ are *iid* from a distribution $F(\cdot)$. The left-hand and right-hand Hill index are defined as:

$$H_{L,k} = \left\{ \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(i)}}{X_{(k+1)}} \right\}^{-1}, \quad H_{R,k} = \left\{ \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{(n-i+1)}}{X_{(n-k)}} \right\}^{-1} \tag{2.3}$$

where $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are order statistics and $k < n$. It has been shown that if $\{X_n\}$ is a stationary $MA(\infty)$ process and the marginal distribution satisfies

$$P(|X_1| > x) = x^{-\alpha} K(x), \text{as} \quad x \to \infty,$$

then if $k/n \to 0$ as $k \to \infty$ and $n \to \infty$, we have

$$H_{L,k} \xrightarrow{p} \alpha, \text{and} \quad H_{R,k} \xrightarrow{p} \alpha,$$

where the notation $\xrightarrow{p}$ stands for converge in probability. More details about the Hill estimator can be found in, for example, Resnick (1997). Applying to our setting of stable distribution, this result asserts that the Hill estimator is an consistent estimator of the index α if the marginal distribution of a process is from the domain of attraction of the stable law. If the estimated $\hat{\alpha} < 2$, then we will have strong evidence to believe that this process has an infinite variance.

We are interested in the autoregressive model with innovation having infinite variance, Resnick (1997) proposed following two ways to estimate α. Suppose now we have observations $y_1, \cdots, y_n$ from an autoregressive model

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t,$$

where $P(|\epsilon_1| > x) = x^{-\alpha} K(x)$. Then

1. we can apply the Hill estimator defined in (2.3) directly to observed $y_1, \cdots, y_n$. The reason is that, by a result of Cline (1983), one has

$$P(|y_1| > x) \sim (const) P(|\epsilon_1| > x),$$

   which implies that the tail of y_1 contains the same information as the tail of ϵ_1.

2. Find consistent estimates for parameters $\phi_1, \cdots, \phi_p$ first and then apply the Hill estimator to the estimated residuals as in (2.3).

Based on the existing empirical results in literature, the second one is usually considered to be a better procedure.

One drawback of the Hill estimator is that the choice of k is very subjective. So

in practice, the Hill estimator is used by plotting graphs $\{(k, H_{L,k}), 1 \leq k \leq n\}$ and $\{(k, H_{R,k}), 1 \leq k \leq n\}$, hoping both graphs look stable so that we can pick out a value of α. These graphs are useful even when a good value of α cannot be observed but a rough range of α is observable from these graphs, which is sufficient for us to determine whether the distribution has a heavy tail.

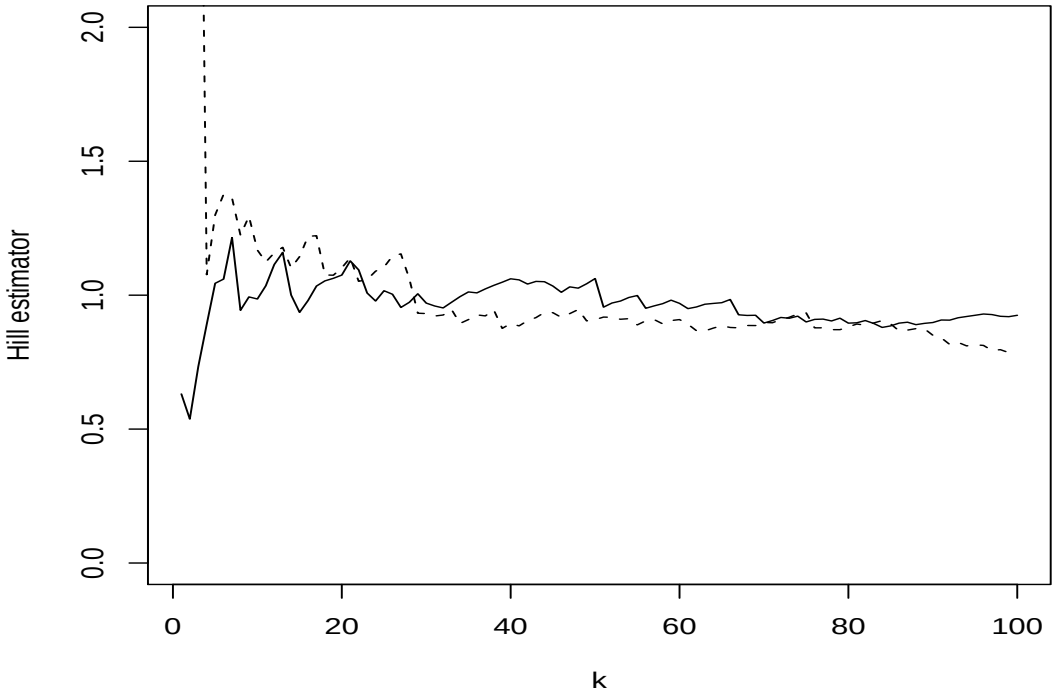

Figure 1. Hill estimators of the left-handed tail index $H_{L,k}$ (dashed line) and right-handed tail index $H_{R,k}$ (solid line) using *iid* sample

To illustrate the use of the Hill estimator, we simulate $n = 400$ independent random numbers from standard Cauchy distribution ($\alpha = 1, \beta = 0$). Figure 1 graphs $\{(k, H_{L,k}) : 1 \leq k \leq 100\}$ and $\{(k, H_{R,k}) : 1 \leq k \leq 100\}$, where we can clearly see that both curves stabilize around true value $\alpha = 1$ when k increases. From Figure 1, we

can easily conclude that this distribution has a very heavy tail. To further illustrate the application of the Hill estimator to the autoregressive model, we simulate the process $y_1, \cdots, y_{400}$ from the model

$$y_t = 0.5y_{t-1} + \epsilon_t,$$

where ϵ_t is generated independently from standard cauchy distribution. Figure 2 graphs $\{(k, H_{L,k}) : 1 \leq k \leq 100\}$ and $\{(k, H_{R,k}) : 1 \leq k \leq 100\}$ using the observed data $y_1, \cdots, y_{400}$ and the estimated residuals by plugging in the least square estimator of ϕ whose true value is 0.5, repectively. As proposed in Resnick (1997), applying the Hill estimator to the estimated residuals appears to be much better in turns of producing a stable value of α in the graph. However, the Hill estimator applying to the observed $AR(1)$ process also provides sufficient evidence to reveal the heavy tailed nature of the innovation process $\{\epsilon_t\}$.

2.3. Estimation of infinite variance autoregressive model

Consider a stationary autoregressive time series $\{y_t\}$ which is generated by

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t, \tag{2.4}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$ is an unknown parameter vector with its true value $\boldsymbol{\phi}_0 = (\phi_1^0, \ldots, \phi_p^0)^T$ and $\{\epsilon_t\}$ is a sequence of independent and identically distributed errors whose common distribution belongs to the domain of attraction of a stable distribution with index $0 < \alpha < 2$. In other words,

$$P(|\epsilon_t|) = x^{-\alpha} K(x)\{1 + o(1)\}, \tag{2.5}$$

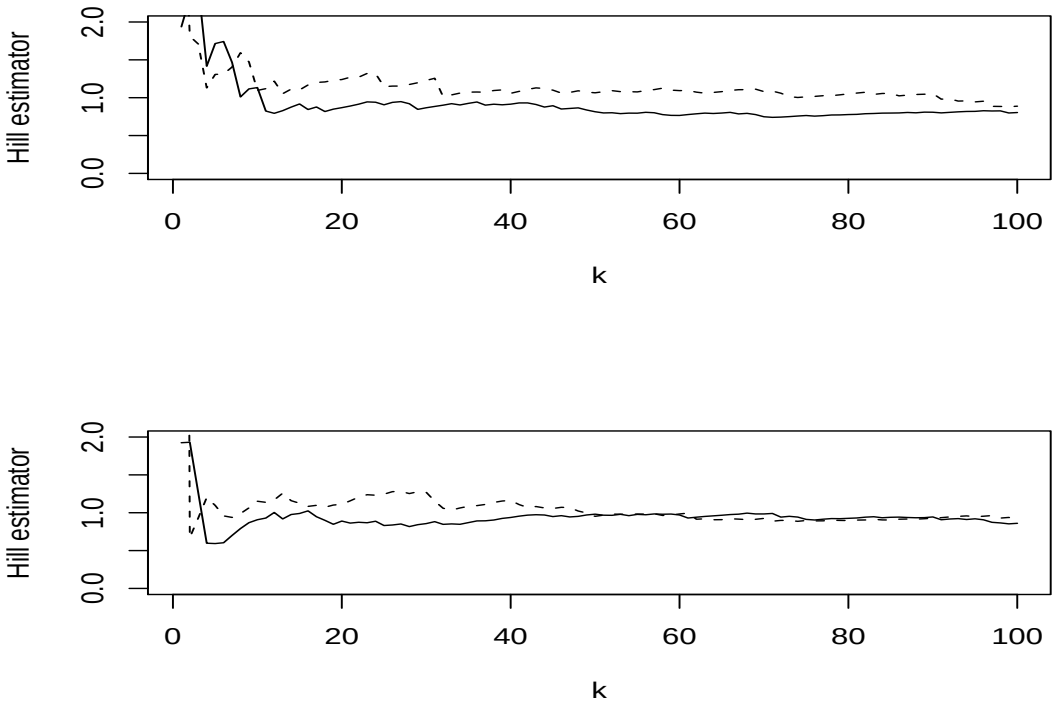

Figure 2. Hill estimators of the left-handed tail index $H_{L,k}$ (dashed line) and right-handed tail index $H_{R,k}$ (solid line) using $AR(1)$ sample (above) and estimated residuals (below)

where $K(x)$ is a slowly varying function at ∞ and

$$\lim_{x\to\infty} P(\epsilon_t > x)/p(|\epsilon_t| > x) = q, \qquad 0 \le q \le 1. \tag{2.6}$$

This assumption on the innovation process is wildly used in the literatures (Knight, 1989; Davis et al., 1992) and it appears that many financial data series are heavy tailed in this sense. Notice that if $K(x)$ is a constant, then the corresponding distribution is a Pareto-like distribution, which contains the Cauchy distribution and general stable distributions as its special cases.

Furthermore, we assume that the characteristic polynomial $\phi(z) = 1-\phi_1^0 z-\cdots-\phi_p^0 z^p$ of model (2.4) has all roots outside the unit circle, which makes $\{y_t\}$ strictly stationary and ergodic. Thus we can represent the infinite variance autoregressive model (2.4) as a linear process

$$y_t = \sum_{j=0}^{\infty} \psi_j^0 \epsilon_{t-j}, \tag{2.7}$$

where ψ_j^0's are the coefficients of z^j in the power series expansion of $1/\phi(z)$.

2.3.1. Least square and least absolute deviation estimator

The least square (LS) estimator $\hat{\boldsymbol{\phi}}_{LS}$ of $\boldsymbol{\phi}$ is defined as the minimizer of

$$V_{LS}(\boldsymbol{\phi}) = \sum_{t=p+1}^{n} (y_t - \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p})^2, \tag{2.8}$$

and the least absolute deviation (LAD) estimator $\hat{\boldsymbol{\phi}}_{LAD}$ is obtained by minimizing

$$V_{LAD}(\boldsymbol{\phi}) = \sum_{t=p+1}^{n} |y_t - \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}|. \tag{2.9}$$

Although intuitively, $\hat{\boldsymbol{\phi}}_{LS}$ and $\hat{\boldsymbol{\phi}}_{LAD}$ may not work since under assumptions (2.5) and (2.6), the autoregressive model (2.4) has infinite variance when $\alpha < 2$, and even infinite mean when $\alpha < 1$. However, both of them perform surprisingly well in practice. Davis et al. (1992) provides a heuristic explanation of this phenomenon. They argued that it is true that large positive or negative values of ϵ_t produce points appearing to be *outliers*. However, each one of these *outliers* will produce a sequence of *leverage points*, which would compensate for the negative effect of the outliers and lead to faster convergence rates of both $\hat{\boldsymbol{\phi}}_{LS}$ and $\hat{\boldsymbol{\phi}}_{LAD}$ than in the finite variance setting. Furthermore, since $V_{LAD}(\boldsymbol{\phi})$ gives less weight to the outliers while giving similar weight to the leverage points, $\hat{\boldsymbol{\phi}}_{LAD}$ is reasonably expected to be more efficient

than $\hat{\boldsymbol{\phi}}_{LS}$, which is later confirmed by their theoretical results.

For the LS estimator $\hat{\boldsymbol{\phi}}_{LS}$, Davis and Resnick (1985) and Davis and Resnick (1986) show that, under assumptions (2.5) and (2.6), there exists a slowly varying function $K_0(n)$ such that:

$$n^{1/\alpha}K_0(n)(\hat{\boldsymbol{\phi}}_{LS} - \boldsymbol{\phi}_0) \to \xi_0 \quad \text{in distribution, as } n \to \infty, \tag{2.10}$$

where ξ_0 is the ratio of two stable random variables. If $\{\epsilon_t\}$ is generated from a stable distribution, then $K_0(n) = (\log n)^{-1/\alpha}$.

Theorem 4.1 in Davis et al. (1992) establishes the asymptotic property of $\hat{\boldsymbol{\phi}}_{LAD}$, which asserts that under conditions of section 2.3 together with several mild technical conditions, one has

$$n^{1/\alpha}K_1(n)(\hat{\boldsymbol{\phi}}_{LAD} - \boldsymbol{\phi}_0) \to \xi \quad \text{in distribution, as } n \to \infty, \tag{2.11}$$

where $K_1(x)$ is some slowly varying function such that $n^{1/\alpha}K_1(n) = b_n$ with $b_n = \{\inf x : P(|\epsilon_1| > x) \leq n^{-1}\}$ and ξ is some unknown random vector. For more details, please refer to Davis et al. (1992).

Now compare equations (2.10) and (2.11), since for Pareto-like and stable distributions, $K_1(x)$ is constant and $K_0(n) = (\log n)^{-1/\alpha}$ (Davis et al., 1992), one can immediately get that, as $n \to \infty$

$$\frac{||\hat{\boldsymbol{\phi}}_{LAD} - \boldsymbol{\phi}_0||}{||\hat{\boldsymbol{\phi}}_{LS} - \boldsymbol{\phi}_0||} \xrightarrow{p} 0,$$

which proves the conjecture that $\hat{\boldsymbol{\phi}}_{LAD}$ is more efficient than $\hat{\boldsymbol{\phi}}_{LS}$, at least for Pareto-like and stable distributions.

2.3.2. Self-weighted least absolute deviation estimator

One of the major problem with the LS and LAD estimators is that their limiting distributions do not have closed forms. This can be seen from the fact that ξ_0 and ξ in equations (2.10) and (2.11) generally do not have closed form distributions. The immediate consequence is that we cannot perform statistical inference based on $\hat{\boldsymbol{\phi}}_{LS}$ and $\hat{\boldsymbol{\phi}}_{LAD}$. To overcome this difficulty, Ling (2005) proposed a new estimation method named self-weighted least absolute deviation (SLAD) estimation for infinite variance autoregressive models, where the estimator $\hat{\boldsymbol{\phi}}_{SLAD}$ is obtained by minimizing

$$V_{SLAD}(\boldsymbol{\phi}) = \sum_{t=p+1}^{n} w_t |y_t - \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}|, \tag{2.12}$$

with w_t as a pre-given function of $\{y_{t-1}, \cdots, y_{t-p}\}$. By imposing some conditions on the choice of w_t and the distribution of ϵ_t, Ling (2005) shows that the limiting distribution of $\hat{\boldsymbol{\phi}}_{SLAD}$ is normal distribution. Denote $X_t = (y_{t-1}, \cdots, y_{t-p})^T$. Following are two additional conditions to those in section 2.3 used in Ling (2005):

Condition 1: $E\{(w_t + w_t^2)(||X_t||^2 + ||X_t||^3)\} < \infty$.

Condition 2: The error process $\{\epsilon_t\}$ has a marginal distribution with 0 median and a differentiable density $f(\cdot)$ such that $f(0) > 0$ and $\sup_{x \in R} |f'(x)| < \infty$.

The choice of the weight function is the critical step to ensure the asymptotic normality of $\hat{\boldsymbol{\phi}}_{SLAD}$. Ling (2005) proposed to use the following weight function

$$w_t = \begin{cases} 1 & \text{if} \quad c_t = 0, \\ C^3/c_t^3, & \text{if} \quad c_t \neq 0, \end{cases}$$

where $c_t = \sum_{k=1}^{p} |y_{t-k}|(|y_{t-k}| \geq C)$, and C can be chosen as the 90% or 95% quantile of data points $\{y_1, \cdots, y_n\}$. Under conditions of section 2.3 and conditions 1-2, Ling

(2005) shows that

$$n^{1/2}(\hat{\boldsymbol{\phi}}_{SLAD} - \boldsymbol{\phi}_0) \to N\left(\mathbf{0}, \frac{1}{4f^2(0)\Sigma^{-1}\Gamma\Sigma^{-1}}\right) \quad \text{in distribution,} \quad \text{as} \quad n \to \infty,$$

where $\Gamma = E(w_t^2 X_t X_t^T)$ and $\Sigma = E(w_t X_t X_t^T)$.

The normality of the estimator $\hat{\boldsymbol{\phi}}_{SLAD}$ enable us to do statistical inferences such as hypothesis tests as in the finite variance case, which is a break through in the research of infinite variance autoregressive models. By conducting a series of Wald tests, one should be able to do the forward, backward or stepwise model selection. However, as in the linear regression case, these model selection methods can be unstable and it is difficult to control the overall type I error of conducting multiple hypothesis tests. To overcome this difficulty, we propose to conduct the model selection of infinite variance autoregressive model using penalized methods as will been shown later.

2.4. Order determination

Order determination is an important aspect of using an autoregressive model. Given a time series $\{y_t\}$, if the true underlying structure of this process is autoregressive, what is the true value of p in model (2.4)? If the true underlying structure is not autoregressive, for example, the moving average process, what is the smallest p that will give a reasonable fit to the observed series? These problems have been studied extensively for finite variance autoregressive models, but much less for the case when the error process $\{\epsilon_t\}$ has infinite variance.

Bhansali (1988) considered the order determination for autoregressive processes under the same assumptions as in section 2.3, and gave a consistent estimator of the

order p. Suppose we have observed a series $y_1, \cdots, y_n$, define quantities

$$\gamma(k,n) = \sum_{t=1}^{n-k} y_t y_{t+k}, \quad \text{and} \quad \rho(k,n) = \gamma(k,n)/\gamma(0,n),$$

where $k = 0, \pm 1, \cdots, \pm n-1$. And the estimated normalized variance is given by

$$\hat{\sigma}^2(p) = \sum_{j=0}^{p} \hat{\phi}_j \rho(j,n), \quad p = 0, \cdots, P, \tag{2.13}$$

where $\hat{\boldsymbol{\phi}}$ is some estimator of $\boldsymbol{\phi}$ and P is some given integer. To obtain the optimal order $\hat{p}_{opt}$, Bhansali (1988) proposed to choose the best p from $0, \cdots, P$ by minimizing following two criterions

$$FPEY_\alpha(p) = \hat{\sigma}_Y^2(p)(1 + \alpha p/n),$$

$$FPEL_\alpha(p) = \hat{\sigma}_L^2(p)(1 + \alpha p/n),$$

where $\alpha \in (0, 2]$ is the index of the stable law distribution of $\{\epsilon_t\}$ and $\hat{\sigma}_Y^2(p)$ and $\hat{\sigma}_L^2(p)$ are obtained by plugging Yule-Walker and least square estimates of $\boldsymbol{\phi}$ into equation (2.13), respectively. Bhansali (1988) later proved that under conditions of section 2.3, minimizing either $FPEY_\alpha(p)$ or $FPEY_\alpha(p)$ would consistently choose the true value of p with probability 1, as $n \to \infty$.

Knight (1989) also studied the order determination of the autoregressive models under the same conditions as in Bhansali (1988). Knight (1989) proposed to minimizing the following AIC type criterion

$$\text{AIC}(p) = n \log \hat{\sigma}_Y^2(p) + 2p, \quad p = 0, \cdots, P,$$

where $\hat{\sigma}_Y^2(p)$ is the same as in $FPEY_\alpha(p)$. The conclusion of Knight (1989) is that, under conditions of section 2.3, if $\hat{p} = arg\min_{0 \le p \le P} \text{AIC}(p)$, then we have

$$\hat{p} \xrightarrow{p} p_{true},$$

as $n \to \infty$, where $\xrightarrow{p}$ stands for convegence in probability. There have been a few of works studying the order determination of time series other than autoregressive models, for example, GARCH model with infinite variance, but our focus here is on the stationary autoregressive models with infinite variance.

2.5. Variable selection using penalized methods

2.5.1. Variable selection of linear regression model

The consistent order estimators in section 2.4 can significantly reduce the model complexity of the autoregressive model and thus lead to more efficient estimation of the model coefficients. However, even when the order of a time series is correctly identified, there is still a possibility that some of the coefficients ϕ_j^0's are zeros and including those zero coefficients will also result in an unnecessarily complex model which degrade the efficiency of the coefficient estimators and leads to less accurate predictions. This is especially true for long-memory autoregressive models whose order can increase as n increases. In addition, a model with a sparse representation reveals the underlying structure of the observed process. Therefore, variable selection can be a very important aspect of autoregressive models.

The idea of using penalized methods to do variable selection is pioneered by the revolutionary paper Tibshirani (1996) in the linear regression setting. Consider the linear regression model:

$$y_i = x_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \cdots, n, \tag{2.14}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector and ϵ_i's are iid random errors with variance σ^2.

To obtain the estimate of β, the *Lasso* method aims at minimizing

$$Lasso(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p}|\beta_j|, \tag{2.15}$$

where $\lambda_n > 0$ is a tuning parameter used to obtain a balance between model fit and model complexity. By shrinking the value of λ towards 0, some components of $\boldsymbol{\beta}$ will be shrunk to exact 0, which means those corresponding covariates are excluded from the model. The primary advantage of the *Lasso* method is that it can simultaneously do variable selection and model estimation, which is more stable than subsets selection in the sense that small changes in the data will not result in big change of the model selection result. Another advantage is that, as in ridge regression, the shrinkage in coefficients will help improve the prediction accuracy of the fitted model.

As appealing as the *Lasso* method is, Zou (2006) along with several other researchers pointed out that the *Lasso* variable selection result is not consistent under certain conditions. Denote $\boldsymbol{\beta}_0 = \{\beta_1^0, \cdots, \beta_p^0\}$ as the true value of $\boldsymbol{\beta}$ and $\mathcal{S} = \{j : \beta_j^0 \neq 0, j = 1, \ldots, p\}$ and $\mathcal{S}_n^{lasso} = \{j : \hat{\beta}_j^{lasso} \neq 0, j = 1, \ldots, p\}$ as the nonzero coefficients estimated via the *Lasso* method. By inconsistency, we mean that

$$\lim_{n\to\infty} P(\mathcal{S}_n^{lasso} = \mathcal{S}) < 1.$$

In other words, under certain conditions, no matter how large your sample sizes is, there is a positive possibility that we will end up with an incorrect model using the *Lasso* method. To solve this problem, Zou (2006) proposed to use a modification of the *Lasso* method, named as the *Adaptive Lasso* method, which estimates $\boldsymbol{\beta}$ by minimizing

$$aLasso(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} w_j|\beta_j|, \tag{2.16}$$

where $\boldsymbol{w}$ is a known weights vector. Zou (2006) suggested using $w_j = 1/|\tilde{\beta}_j|^\gamma$ with

$\gamma > 0$ and $\hat{\beta}$ being a $\sqrt{n}$-consistent estimator to β_0. Again, define $\mathcal{S}_n^{alasso} = \{j : \hat{\beta}_j^{alasso} \neq 0, j = 1, \ldots, p\}$ as the nonzero coefficients estimated via the *Adaptive Lasso* method, Zou (2006) showed that if $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$, then the *Adaptive Lasso* estimator enjoys a so-called "Oracle property" (Fan and Li, 2001), which includes:

1. Consistency in variable selection: $\lim_{n\to\infty} P(\mathcal{S}_n^{alasso} = \mathcal{S}) = 1$,
2. Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{alasso} - \boldsymbol{\beta}_{0\mathcal{S}}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 C_{\mathcal{S}}^{-1})$,

where $C_{\mathcal{S}} = \lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}_{\mathcal{S}}^T \boldsymbol{X}_{\mathcal{S}}$ with $\boldsymbol{X}_{\mathcal{S}}$ being the design matrix only using covariates with nonzero estimated coefficients. "Oracle property" means that we can simultaneously do variable selection and model estimation as if the true model is known.

The *Adaptive Lasso* method is not the only penalized method that enjoys this "Oracle property". Another famous example would be the smoothly clipped absolute deviation (SCAD) penalty function proposed in Fan and Li (2001). Zou and Li (2008) further proposed to modify the penalty term in (2.16) by replacing each $\lambda_n w_j$ term with $p'_{\lambda_n}(|\tilde{\phi}_{1j}|)$ for some general penalty function $p_\lambda(\cdot)$, for example, the SCAD penalty function, which maintains the "Oracle property".

Wang et al. (2007b) considered the model (2.14) with the error term ϵ_i from some heavy tailed distribution, where they proposed to do model estimation and variable selection using the *Lad-Lasso* method by minimizing

$$LadLasso(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - x_i^T \boldsymbol{\beta}| + \sum_{j=1}^{p} \lambda_j |\beta_j|, \tag{2.17}$$

where the tuning parameters can be chosen as

$$\lambda_j = \lambda_n \frac{\log n}{n|\tilde{\beta}_j|}, \quad j = 1, \cdots, p,$$

with $\tilde{\boldsymbol{\beta}}$ being the unpenalized least square estimator or other $\sqrt{n}-$consistent estima-

tors of $\boldsymbol{\beta}$. The use of least absolute deviation loss function in (2.17) instead of the least square loss function handles the problem of having residuals from heavy tailed distributions including those with infinite variances by assigning smaller weights to large values of deviations. Assuming that the error ϵ_i has a continuous density function $f(\cdot)$ such that $f(0) > 0$, then under certain conditions, Wang et al. (2007b) showed that as $n \to \infty$,

$$P(\hat{\boldsymbol{\beta}}_{\mathcal{S}^c} = 0) \to 1, \quad \text{and} \quad \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{0\mathcal{S}}) \xrightarrow{d} N(\mathbf{0}, \frac{1}{4f^2(0)} C_{\mathcal{S}}^{-1}),$$

which implies that the *Lad-Lasso* method also enjoys the "Oracle property". This actually motivates us to consider apply the *Lad-Lasso* method to model infinite variance autoregressive model.

2.5.2. Variable selection of autoregressive model

Using the shrinkage method for variable selection is relatively new in time series literature. Wang et al. (2007a) applied adaptive lasso (Zou, 2006) to the regression model with finite autoregressive errors. They considered the model

$$y_t = x_t^T \boldsymbol{\beta} + \epsilon_t, \quad t = 1, \cdots, n$$

with the error term ϵ_t having a finite fourth moment and following a $AR(q)$ process

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q} + e_t$$

where $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_q)^T$ is the coefficient vector. The estimation of this model involves the regression parameter $\boldsymbol{\beta}$ and the autoregressive parameter $\boldsymbol{\phi}$, which is achieved by minimizing

$$\sum_{t=q+1}^{n} \left[y_t - x_t^T \boldsymbol{\beta} - \sum_{l=1}^{q} \phi_l (y_{t-l} - x_{t-l}^T \boldsymbol{\beta}) \right]^2 + \sum_{j=1}^{p} \lambda_j |\beta_j| + \sum_{l=1}^{q} \gamma_l |\phi_l|,$$

where the tuning parameters can be chosen in the following manner

$$\lambda_j = \lambda_n \frac{\log n}{n|\tilde{\beta}_j|} \quad \text{and} \quad \gamma_l = \gamma_n \frac{\log n}{n|\tilde{\phi}_l|},$$

with $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\phi}}$ being the unpenalized least square estimator or other $\sqrt{n}-$consistent estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$. Define the index sets $\mathcal{S}_1 = \{1 \leq j \leq p : \beta_j \neq 0\}$ and $\mathcal{S}_2 = \{1 \leq l \leq q : \phi_l \neq 0\}$, Wang et al. (2007a) showed that under certain conditions, as $n \to \infty$, the resulting estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\phi}}$ have the following property

$$P(\hat{\boldsymbol{\beta}}_{\mathcal{S}_1^c} = 0) \to 1 \quad \text{and} \quad P(\hat{\boldsymbol{\phi}}_{\mathcal{S}_2^c} = 0) \to 1,$$

which means that all those insignificant components of regression and autoregressive coefficients can be consistently excluded from the estimated model. This is an appealing property that for the autoregressive part, one would not only be able to do the order determination but also variable selection. We would apply a similar idea to the infinite variance autoregressive model.

2.6. Autoregressive approximation for a stationary process

Let (Ω, F_Y, P) be a probability space. A zero mean stochastic process $\{y_t\}$ is said to be strictly stationary if the finite dimension joint cumulative distribution function of $\{y_t\}$ at times $t_1 + s, \cdots, t_k + s$ satisfies

$$F_Y(y_{t_1+s}, \cdots, y_{t_k+s}) = F_Y(y_{t_1}, \cdots, y_{t_k})$$

for all k and $s > 0$. A simple example would be the white noise process with identical distribution. Following similar notations of Cheng et al. (2000), for each process $\{y_t\}$

with $y_t \in L^p(\Omega)$, that is, $\int_\Omega |y|^p dF_Y(y) < \infty$, we define the following subspaces:

$$H_t(Y) = \bar{sp}\{y_s, s \leq t\}, \text{and} \quad H_{-\infty}(Y) = \bigcap_{t \leq 0} H_t(Y)$$

where $\bar{sp}\{\cdots\}$ represents the closed linear space spanned by the elements in the bracket under the L^p norm.

The process $\{y_t\}$ is said to be *deterministic* if

$$H_{t-1}(Y) = H_t(Y)$$

for all t, and is called *nondeterministic* otherwise.

If a nondeterministic process satisfies

$$H_{-\infty} = \{0\},$$

then it is said to be a *pure nondeterministic* process.

2.6.1. Weakly stationary process

In most situations, strict stationarity is too strong of an assumption in prediction theory of stationary process. A zero mean stochastic process $\{y_t\}$ is called a weakly stationary process if

$$E|y_t|^2 < \infty, \text{and} \quad cov(y_s, y_t) = \gamma(s-t),$$

for all s, t, where $\gamma(\cdot)$ is referred to as the covariance function.

The weakly stationary process has been extensively studied and it can be shown that, any weakly stationary process with a continuous spectral density can be approximated by a weakly stationary autoregressive model with a large order (Brockwell and Davis, 1991). In fact, it is shown in Pourahmadi (1988) that for a purely nondeterministic weakly stationary process $\{y_t\}$, there exists a unique series $\{a_k\}$ such that

for all t, one has

$$y_t = \sum_{k=1}^{\infty} a_k y_{t-k} + \epsilon_t,$$

provided that $\sum_{k=1}^{\infty} a_k y_{t-k}$ is convergent in the L^2 norm. A sufficient condition for the convergence of $\sum_{k=1}^{\infty} a_k y_{t-k}$ is that $\sum_{k=1}^{\infty} |a_k| < \infty$.

Another nice property of this decomposition is that variables in the innovation process $\{\epsilon_t\}$ are orthogonal under the inner product induced by the L^2 norm, i.e., they are uncorrelated. This is a very useful result which indicates that, for a general weakly stationary process, we can use an autoregressive model with a sufficiently large order to do one-step or multi-step predictions, without knowing the true probability structure of the process.

2.6.2. p-stationary process

The popularity of the autoregressive model in time series studies is largely due to the fact that any second order stationary process with symmetric continuous spectral density can be approximated by an autoregressive process (Brockwell and Davis, 1991). It would be very appealing if this type of approximation still holds for the infinite variance process, which can justify the use of autoregressive model to do predictions. However, even for the strictly stationary process with infinite variance, this is difficult to show.

Miamee and Pourahmadi (1988) established such a relationship for the p-stationary process. A discrete time stochastic process $\{y_t\}$ is said to be a p-stationary process if

$$E|y_t|^p < \infty, \text{and} \quad E\left|\sum_{k=1}^{n} c_k y_{t_k+h}\right|^p = E\left|\sum_{k=1}^{n} c_k y_{t_k}\right|^p,$$

$(1 < p \leq 2)$ for all integers $n \geq 1$, $t_1, \ldots, t_n$, h, and scalars $c_1, \ldots, c_n$. Note that, when $p = 2$, it is a weakly stationary process and it is the only case in this class with finite

variance. This class of processes includes the harmonizable stable processes of order α with $\alpha \in (1, 2]$ and strictly stationary processes with finite p-th moment. Miamee and Pourahmadi (1988) showed that for a purely nondeterministic p-stationary process $\{y_t\}$ with innovation $\{\epsilon_t\}$, there exists a unique series $\{a_k\}$ such that for all t, one has

$$y_t = \sum_{k=1}^{\infty} a_k y_{t-k} + \epsilon_t,$$

provided that $\sum_{k=1}^{\infty} a_k y_{t-k}$ is convergent in the mean of order p. A sufficient condition for the convergence of $\sum_{k=1}^{\infty} a_k y_{t-k}$ is that $\sum_{k=1}^{\infty} |a_k| < \infty$. For regularity conditions and more recent advances in this area, see Cheng et al. (2000).

Compare to the weakly stationary process, the autoregressive representation above does not have the property that variables in the innovation process $\{\epsilon_t\}$ are not uncorrelated for the case of $0 < p < 2$. So the above representation does provide some insights for using an autoregressive model for predicting a general stationary infinite variance time series in that even though the underlying structure of the time series is not autoregressive, it can be approximated by an autoregressive model under certain conditions. However things are not as nicely done as in $p = 2$ case.

CHAPTER III

VARIABLE SELECTION FOR INFINITE VARIANCE AUTOREGRESSIVE MODELS

3.1. Introduction

Heavy-tailed time series data is often encountered in a variety of fields, such as hydrology (Castillo, 1988), economics and finance (Koedijk et al., 1990) and teletraffic engineering (Duffy et al., 1994). In this situation, the infinite variance autoregressive model is often preferred to the finite variance one, and its statistical theory has been widely studied in the literature. See Resnick (1997) for a comprehensive review and further references.

Model selection is an important aspect of modeling with time series data. An unnecessarily complex model can degrade the efficiency of the resulting parameter estimators and lead to less accurate predictions. For a time series model with finite variance, traditional model selection criteria AIC (Akaike, 1973) and BIC (Schwarz, 1978) can be employed to choose the order of the autoregressive model (McQuarrie and Tsai, 1998). Compared to the case of finite variance autoregressive models, few papers have investigated the model selection for autoregressive models with infinite variance. Bhansali (1988) considered the order determination of the infinite variance autoregressive processes with innovations in the domain of attraction of a stable law, and gave a consistent estimator of the order. Knight (1989) studied the same model and showed that the order selection with AIC is weakly consistent. While most of the literature focuses on the order determination of the time series, Ling (2005) proposed a self-weighted least absolute deviation estimator for the infinite variance autoregressive model under which the coefficient estimates are asymptotically normal

and thus can be used for statistical inference. He also proposed a variable selection procedure with a series of hypothesis tests based on the self-weighted least absolute deviation estimator. However, his method can be unstable and its implementation is complicated.

Using the shrinkage method for variable selection is relatively new in time series literature. Wang et al. (2007a) applied adaptive lasso (Zou, 2006) to the regression model with finite autoregressive errors. They showed that the resulting estimator via adaptive lasso not only has a sparse presentation, but also has the oracle property (Fan and Li, 2001), which means that it can simultaneously select variables and estimate parameters in time series modeling.

One difficulty often encountered in data analysis is that it is generally impossible to know whether a time series of finite length has infinite variance (Granger and Orr, 1972). Many methods have been developed to test for infinite variance of a real time series data; see, for example, Hill (1975). While Wang et al. (2007a)'s method does not apply to infinite variance autoregressive models, using Ling (2005)'s method can cause loss of important information by weighing down large observations, especially in the case of a time series with heavy tails but finite variance.

In this chapter, we first use the self-weighted least absolute deviation proposed by Ling (2005) as the loss function and the adaptive lasso as the penalty method to do the model selection. Under appropriate conditions, we show that our penalized method can identify the true model consistently and the estimator of the coefficients corresponding to the true model is asymptotically normal, which is important for the statistical inference of infinite variance autoregressive models. After that, we propose a unified variable selection approach that can efficiently deal with heavy-tailed autoregressive models with either finite or infinite variance. By combining the least absolute deviation as the loss function and the adaptive lasso as the penalty

function, we show that under regularity conditions we can identify the true model consistently and obtain a point estimator of the coefficients corresponding to the true model with a convergence rate of $n^{-1/\alpha}$, where $\alpha \in (0, 2)$ is the index of the stable distribution. This convergence rate is faster than that of finite variance time series.

Computationally, the algorithm of our methods can be formulated as an estimation problem of ordinary least absolute deviation, and consequently, any standard unpenalized least absolute deviation program can be used to find the final estimator without much programming effort. A simulation study is carried out that confirms our theoretical findings. Finally, We apply the proposed penalty method to the Hang Seng Index data set, which has been examined by Ling (2005) using a series of hypothesis tests.

3.2. Adaptive lasso for infinite variance autoregressive models

3.2.1. Notations and Preliminaries

Consider a stationary autoregressive time series $\{y_t\}$ which is generated by

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t, \tag{3.1}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$ is an unknown parameter vector with true value $\boldsymbol{\phi}_0 = (\phi_1^0, \ldots, \phi_p^0)^T$. We assume that there are a total of $p_0 \leq p$ non-zero coefficients within ϕ_0. Denote $\mathcal{S} = \{j : \phi_j^0 \neq 0, j = 1, \ldots, p\}$ and $\mathcal{S}^c = \{j : \phi_j^0 = 0, j = 1, \ldots, p\}$. Assume that $\{\epsilon_t\}$'s are independent and identically distributed in the domain of attraction of a stable law with index $\alpha \in (0, 2)$. More specifically,

$$P(|\epsilon_t| > x) = x^{-\alpha} K(x)(1 + o(1)), \tag{3.2}$$

where $K(x)$ is a slowly varying function such that $lim_{x\to\infty}\frac{K(tx)}{K(x)} = 1$ for any $t > 0$ and

$$\lim_{x\to\infty}\frac{P(\epsilon_t > x)}{P(|\epsilon_t| > x)} = q, \quad 0 \le q \le 1. \tag{3.3}$$

This type of innovation is popular in modeling infinite variance autoregressive models; see Knight (1989) and Davis et al. (1992). It appears appears that some financial data are heavy tailed in this sense. Here $K(x)$ is a constant for the class of Pareto-like distributions, which includes the Cauchy and stable distributions. We also assume that

$$\phi(z) = 1 - \phi_1^0 z - \cdots - \phi_p^0 z^p \neq 0$$

for all complex z with $|z| \le 1$, which makes $\{y_t\}$ strictly stationary and ergodic. Thus Model (3.1) can be represented as

$$y_t = \sum_{j=0}^{\infty} \psi_j^0 \epsilon_{t-j},$$

where ψ_j^0's are the coefficients of z^j in the power series expansion of $1/\phi(z)$.

3.2.2. Adaptive lasso with self-weighted least absolute deviation

In practice, even when the order of a time series is correctly identified, an unnecessarily complex model can still degrade the efficiency of the coefficient estimators and lead to less accurate predictions. In addition, a model with a sparse representation reveals the underlying structure of the observed process. We propose the following procedure for simultaneous order determination and variable selection of a time series.

We first choose the self-weighted least absolute deviation (SLAD)proposed by Ling (2005) as the loss function, which is defined as

$$L_{1n}(\phi) = \sum_{t=p+1}^{n} h_t |y_t - X_t^T \phi|, \tag{3.4}$$

where $X_t = (y_{t-1}, \ldots, y_{t-p})^T$ and h_t is a given function of $\{y_{t-1}, \ldots, y_{t-p}\}$. Then the SLAD estimator is defined as $\tilde{\phi}_{1n} = \arg\min_{\phi}\{L_{1n}(\phi)\}$. Ling (2005) showed that, unlike other estimators of model (3.1), the SLAD estimator has an asymptotic normal distribution under the following two conditions:

Condition 1 A appropriate weight function in (3.4), h_t, is chosen such that $E\{(h_t + h_t^2)(\| X_t \|^2 + \| X_t \|^3)\} < \infty$;

Condition 2 The errors ϵ_t have zero median and a differentiable density $f(x)$ everywhere in R such that $f(0) > 0$ and $\sup_{x \in R} \mid f^{'}(x) \mid < \infty$.

The following Lemma 3.2.1 is the Theorem 1 of Ling (2005). It states that the SLAD estimator is root-n consistent and asymptotically normally distributed.

Lemma 3.2.1. *If Conditions* $1 - 2$ *hold, then it follows that*

$$n^{\frac{1}{2}}(\tilde{\phi}_{1n} - \phi_0) \rightarrow N\left\{0, \frac{1}{4f^2(0)}\Sigma^{-1}\Omega\Sigma^{-1}\right\} \tag{3.5}$$

in distribution, where $\Sigma = E(h_t X_t X_t^T)$ *and* $\Omega = E(h_t^2 X_t X_t^T)$.

Abbreviating the adaptive lasso method with SLAD function as SLAD-alasso. The SLAD-alasso estimator $\hat{\phi}_{1n}$ is obtained by minimizing the following objective function

$$V_{1n}(\phi) = L_{1n}(\phi) + \lambda_n \Sigma_{j=1}^{p} r_{1j} \mid \phi_j \mid, \tag{3.6}$$

where the weight $r_{1j} = |\tilde{\phi}_{1j}|^{-\gamma}$ with $\gamma > 0$ and $\tilde{\phi}_{1j}$ is the jth element of $\tilde{\phi}_{1n}$. By Lemma 3.2.1, as the sample size grows, the weights for zero coefficients go to infinity, whereas the weights for nonzero coefficients converge to finite constants which enables us to use SLAD-alasso as a tool to simultaneously select variables and estimate coefficients.

Now we give the following main theorem about the property of the SLAD-alasso estimator.

Theorem 3.2.1. *Denote* $\mathcal{S}_1^* = \{1 \le j \le p : \hat{\phi}_{1j} \neq 0\}$, *where* $\hat{\phi}_{1j}$ *is the* j*th element of* $\hat{\phi}_{1n}$. *Under Conditions* 1 *and* 2, *suppose that* $\lambda_n n^{-\frac{1}{2}} \to 0$ *and* $\lambda_n n^{(\frac{\gamma}{2}-1)} \to \infty$. *Then the minimizer of (3.6)* $\hat{\phi}_{1n}$ *satisfies the following properties:*

(1) Consistency in variable selection:

$$\lim_{n\to\infty} P(\mathcal{S}_1^* = \mathcal{S}) = 1;$$

(2) Asymptotic normality: as $n \to \infty$,

$$n^{\frac{1}{2}}(\hat{\phi}_{1\mathcal{S}} - \phi_{\mathcal{S}}^0) \to N\left\{0, \frac{1}{4f^2(0)}\Sigma_{\mathcal{S}}^{-1}\Omega_{\mathcal{S}}\Sigma_{\mathcal{S}}^{-1}\right\} \quad \textit{in distribution,}$$

where $\phi_{\mathcal{S}}^0$ *and* $\hat{\phi}_{1\mathcal{S}}$ *are the subvector of* ϕ_0 *and* $\hat{\phi}_{1n}$ *corresponding to the nonzero coefficients, and* $\Sigma_{\mathcal{S}}$ *and* $\Omega_{\mathcal{S}}$ *are the submatrix of* Σ *and* Ω *corresponding to* $\phi_{\mathcal{S}}^0$, *respectively.*

The proof of Theorem 3.2.1 is given in the Appendix.

Remark 3.2.2. *At the beginning of this chapter, we assume that the distribution of* $\{\epsilon_t\}$ *belongs to the domain of attraction of a stable distribution with index* $\alpha \in (0, 2)$. *In fact, this assumption is only necessary for proving the asymptotic property of LAD-alasso in the next section. For SLAD-alasso here,* $E(|\epsilon|^\delta) < \infty$ *for some* $0 < \delta < 2$ *is sufficient to prove Theorem 3.2.1.*

Remark 3.2.3. *The choice of weights* r_{1j}*'s can incorporate prior information in practice. For example, if previous experience suggests that some variables must be selected, we can simply set* $r_{1j} = 0$ *for these variables. The choice of penalty term can be made more general by replacing each* $\lambda_n r_j$ *term in (3.6) with* $p'_{\lambda_n}(|\tilde{\phi}_{1j}|)$ *for some penalty function* $p_\lambda(\cdot)$*; see Zou and Li (2008). A special choice would be the famous smoothly clipped absolute deviation (Fan and Li, 2001) penalty function.*

Theorem 3.2.1 states that by choosing a suitable pair of (λ_n, γ), the SLAD-alasso method can consistently select the true model and the estimator of the coefficients corresponding to the true model is asymptotically normal. As an example of the choice of (λ_n, γ), one can take $\gamma = 2$ and $\lambda_n = \log n$. Because of its asymptotical normality, we can use the SLAD-alasso estimator to make statistical inferences, which is the main reason why we choose self-weighted least absolute deviation as the loss function.

In practice, we need to select a suitable weight h_t for the loss function part. Ling (2005) suggested using the following weight function:

$$h_t = \begin{cases} 1 & (c_t = 0), \\ C^3/c_t^3 & (c_t \neq 0), \end{cases} \tag{3.7}$$

where $c_t = \sum_{j=1}^{p} |y_{t-j}| \{I(|y_{t-j}| \geq C)\}$ and $C > 0$ is a constant. It is easy to see that this weight function satisfies Condition 1. Similar to Ling (2005), we take C as the ρth quantile of data $\{y_1, \ldots, y_n\}$.

As stated in Ling (2005), with random errors from distributions satisfying (3.2) and (3.3), it can be shown theoretically that larger C would result in smaller asymptotic variance of the SLAD estimator. However, a overly large C would make the distribution of the SLAD estimator asymptotically non-normal even with a large sample size. Our simulation results show that with Cauchy errors, the empirical standard errors matched well with the asymptotical standard errors in the case of $\rho = 90\%$ but matched much worse in the case of $\rho = 95\%$. However, when ϵ_i's are from the $S(1{\cdot}5,0;1)$ distribution, both $\rho = 90\%$ and $\rho = 95\%$ matched well. This indicates that the optimal choice of C varies for different models and error distributions to ensure that the conclusion of Theorem 3.2.1 still holds for the SLAD-alasso estimator.

A realistic question is, if the limiting distribution of SLAD-alasso estimator is

not normal because of a poor choice of C, is it still possible for us to do the model selection using adaptive lasso? Fortunately, the answer is yes. Our simulation results indicate that the model selection result of SLAD-alasso becomes better as C increases. Particularly, if we take C to be the 100% quantile of y_i's, in which case we have $h_t = 1$, the model selection results are the best. It motivates us to consider the ordinary least absolute deviation (LAD) as the loss function combining with the adaptive loss penalty function, which we name LAD-alasso. In the following subsection, we study the asymptotic property of LAD-alasso and explain why LAD-alasso performs better than SLAD-alasso in model selection in spite of the fact that the limiting distribution of the LAD-alasso estimator does not have a closed form.

3.2.3. Adaptive lasso with least absolute deviation

Denote $L_{2n}(\boldsymbol{\phi}) = \sum_{t=p+1}^{n} |y_t - X_t^T \boldsymbol{\phi}|$, where $X_t = (y_{t-1}, \ldots, y_{t-p})^T$. Define the LAD estimator of Model (3.1) as $\tilde{\boldsymbol{\phi}}_{2n} = \arg\min_{\phi}\{L_{2n}(\boldsymbol{\phi})\}$. And then the LAD-alasso estimator $\hat{\phi}_{1n}$ is defined as the minimizer of

$$V_{2n}(\boldsymbol{\phi}) = L_{2n}(\boldsymbol{\phi}) + \lambda_n \Sigma_{j=1}^{p} r_{2j} |\phi_j|, \tag{3.8}$$

where the weight $r_{2j} = |\tilde{\phi}_{2j}|^{-\gamma}$ with $\gamma > 1$ and $\tilde{\phi}_{2j}$ being the jth element of $\tilde{\boldsymbol{\phi}}_{2n}$. Note that $\tilde{\phi}_{1n}$ can be obtained by setting $\lambda_n = 0$ when minimizing (3.8). As stated in Davis et al. (1992), although Model (3.1) has an infinite variance and even infinite mean if $\alpha < 1$, the LAD estimator performs surprisingly well. In fact, $\tilde{\boldsymbol{\phi}}_{2n}$ usually converges in a rate faster than $n^{-1/2}$. In this sense, we obtain a better choice of weights r_j's, and hence for a given sample size n, minimizing (3.8) would yield better variable selection results than that in the finite variance case.

The asymptotic theory for $\tilde{\phi}_{2n}$ was established by Davis et al. (1992). Denote

$$W_n(u) = \sum_{t=p+1}^{n} \left(|\epsilon_t - b_n^{-1} X_t^T u| - |\epsilon_t| \right), \tag{3.9}$$

where $b_n = \inf\{x : P(|\epsilon_t| > x) \leq n^{-1}\}$. As stated in Davis et al. (1992), for Pareto-like distributions we may take $b_n = n^{1/\alpha}$, and in general $b_n = n^{1/\alpha} K_1(x)$ for some slowly varying function $K_1(\cdot)$. Recall that $y_t = \sum_{j=0}^{\infty} \psi_j^0 \epsilon_{t-j}$, and define quantity

$$W(u) = \sum_{i=1}^{\infty} \sum_{k=1}^{\infty} \{ |\epsilon_{k,i} - (\psi_{i-1}^0 u_1 + \cdots + \psi_{i-p}^0 u_p) \varrho_k \Theta_k^{-1/\alpha}| - |\epsilon_{k,i}| \}, \tag{3.10}$$

where $\{\epsilon_{k,i}\}$, $\{\varrho_k\}$, and $\{\Theta_k\}$ are three independent sequences defined as:

1. $\{\epsilon_{k,i}\}$ is independent and identically distributed as ϵ_t;

2. $\{\varrho_k\}$ is independent and identically distributed with $P(\varrho_k = 1) = q$ and $P(\varrho_k = -1) = 1 - q$, with q given in (3.3);

3. $\Theta_k = \sum_{j=1}^{k} \Gamma_j$, where $\{\Gamma_j\}$ is a sequence of independent and identically distributed unit exponential random variables.

The following lemma is Theorem 4.1 in Davis et al. (1992), which establishes the asymptotic property of $\tilde{\phi}_{2n}$.

Lemma 3.2.2. *Suppose that $\{\epsilon_t\}$ satisfies (3.2) and (3.3) with $\alpha \in (0, 2)$, and has median 0 if $\alpha \geq 1$. If either (a) $\alpha < 1$, or (b) $\alpha > 1$ and $E(|\epsilon_t|^\beta) < \infty$ for some $\beta < 1 - \alpha$, or (c) $\alpha = 1$ and $E(\log |\epsilon_t|) > -\infty$, then $W_n(\cdot) \to W(\cdot)$ in distribution. Moreover, if $W(\cdot)$ has a unique minimum almost surely, then*

$$b_n(\tilde{\phi}_{1n} - \phi_0) \to \xi \quad \textit{in distribution, as } n \to \infty, \tag{3.11}$$

where $b_n = n^{1/\alpha} K_1(x)$ for some slowly varying function $K_1(x)$ and ξ is the minimum of $W(\cdot)$.

Remark 3.2.4. *The conditions in Lemma 3.2.2 guarantee that $W(\cdot)$ is well-defined. To guarantee that $W(\cdot)$ has a unique minimum almost surely, Davis et al. (1992) showed that the following condition is sufficient: for all $\varepsilon > 0$ there exists a constant $d > 0$ such that*

$$P(x < \epsilon_t < y) \geq \begin{cases} d(y-x)^{1/\alpha}, & \alpha < 1, \\ d(y-x), & \alpha \geq 1, \end{cases}$$

whenever $-\varepsilon < x < y < \varepsilon$. For the Cauchy distribution and most stable distributions with index $\alpha \in (0, 2)$, this condition is obviously satisfied.

Let $\hat{\boldsymbol{\phi}}_{2n}$ be the minimizer of (3.8) and denote $\mathcal{S}^* = \{j : \hat{\phi}_{2j} \neq 0, j = 1, \ldots, p\}$, where $\hat{\phi}_{2j}$ is the jth element of $\hat{\boldsymbol{\phi}}_{2n}$. As our main theoretical result, the next theorem states the variable selection consistency of adaptive lasso method as well as the weak convergence of coefficient estimators to their true values.

Theorem 3.2.5. *Suppose that $\{\epsilon_t\}$ satisfies the conditions stated in Lemma 3.2.1 and Remark 3.2.4. If $\lambda_n b_n^{-1} \to 0$ and $\lambda_n b_n^{\gamma-2} \to \infty$, with $b_n = n^{1/\alpha} K_1(x)$ for some slowly varying function $K_1(x)$, then we have $\lim_{n\to\infty} P(\mathcal{S}^* = \mathcal{S}) = 1$ and $b_n(\hat{\phi}_{1\mathcal{S}} - \phi^0_{\mathcal{S}}) = O_p(1)$, where $\phi^0_{\mathcal{S}}$ and $\hat{\phi}_{1\mathcal{S}}$ are the subvectors of ϕ_0 and $\hat{\phi}_{1n}$ corresponding to the non-zero coefficients, respectively.*

The proof of Theorem 3.2.5 is given in the Appendix. We would like to point out that it is generally not possible to obtain an explicit representation of the limiting distribution of $b_n(\hat{\phi}_{1\mathcal{S}} - \phi^0_{\mathcal{S}})$.

Remark 3.2.6. *When $\{\epsilon_t\}$'s in Model (3.1) have finite variance, the result in Theorem 3.2.5 still holds, see Wang et al. (2007b). In this case, we have $b_n = n^{1/2}$ and $\hat{\phi}_{1\mathcal{S}}$ is asymptotically normal.*

Remark 3.2.7. *The conditions on λ_n seem to be complicated. However, simply taking $\lambda_n = \log n$ would satisfy all conditions for any $\gamma > 1$ and $\alpha \in (0, 2)$. In addition, this choice of λ_n and γ also works when $\{\epsilon_t\}$ has finite variance, see Wang et al. (2007b).*

Theorem 3.2.5 states that adaptive lasso method is an estimation and variable selection procedure that takes genuine advantage of the LAD estimators in the infinite variance autoregressive model. It can perform variable selection consistently with the resulting estimator corresponding to the nonzero coefficient part weakly converges more quickly than that of SLAD-alasso. However, compared to SLAD-alasso, the limiting distribution of LAD-alasso estimator does not have a closed form.

Remarks 3.2.6 and 3.2.7 state that the LAD loss function combined with the adaptive lasso penalty function works for both finite and infinite variance situations. There is no need to distinguish between these two cases when the primary concern is to obtain a sparse model. When inference is needed, one can apply tools such as the self-weighted least absolute deviation method proposed by Ling (2005) to the selected model.

3.2.4. Comparison with self-weighted least absolute deviation method

We have seen that each method, SLAD-alasso and LAD-alasso, has its own merit. The LAD-alasso method gives a better variable selection results while the SLAD-alasso estimator is asymptotically normally distributed and hence can be used to perform statistical inference.

Since $\tilde{\phi}_{1n}$ is asymptotically normal, Ling (2005) proposed an variable selection procedure based on a series of hypothesis tests using Chi-square test statistics. However, there are a few drawbacks of this approach:

1. When using this method, if the data do not have infinite variance, weighing

down large observations would lead to unnecessary loss of important information;

2. The choice of C is subjective. There is no theoretical justification for a best choice of C. Our simulation results show that different choices of C would lead to different model selection results. In addition, there is not a universal choice of C that can guarantee the asymptotical normality of SLAD-alasso estimator for all distributions;

3. It is difficult to manage the overall type I error when conducting a series of hypothesis tests;

4. The unknown term $f(0)$ in (3.5) needs to be estimated. The most commonly used kernel density estimator is effective but would also make the conclusions of the hypothesis tests less reliable.

The least absolute deviation adaptive lasso method suffers from none of the above problems. One limitation is that there is no closed form for the limit distribution of the resulting estimator. Recall that the motivation for our method is to develop a better variable selection strategy and to produce a faster convergent point estimator in the infinite variance case. To make inference, we can apply existing methods such as Ling (2005)'s to the model selected by our method.

3.2.5. p-Stationary process

The popularity of the autoregressive model in time series studies is largely due to the fact that any second order stationary process with symmetric continuous spectral density can be approximated by an autoregressive process (Brockwell and Davis, 1991). However, when it comes to stationary process with infinite variance, this type

of relationship is difficult to establish.

Miamee and Pourahmadi (1988) established such a relationship for the p-stationary process. A discrete time stochastic process $\{y_t\}$ is said to be a p-stationary process if $E|y_t|^p < \infty$, and $E\left|\sum_{k=1}^{n} c_k y_{t_k+h}\right|^p = E\left|\sum_{k=1}^{n} c_k y_{t_k}\right|^p$, $(1 < p \leq 2)$ for all integers $n \geq 1$, $t_1, \ldots, t_n$, h, and scalars $c_1, \ldots, c_n$. Note that, when $p = 2$, it is a second order weakly stationary process and it is the only case in this class with finite variance. This class of processes includes the harmonizable stable processes of order α with $\alpha \in (1, 2]$ and strictly stationary processes with finite p-th moment. Miamee and Pourahmadi (1988) showed that for a purely nondeterministic p-stationary process $\{y_t\}$ with innovation $\{\epsilon_t\}$, there exists a unique series $\{a_k\}$ such that for all t, one has

$$y_t = \sum_{k=1}^{\infty} a_k y_{t-k} + \epsilon_t,$$

provided that $\sum_{k=1}^{\infty} a_k y_{t-k}$ is convergent in the mean of order p. Note that a sufficient condition for the convergence of $\sum_{k=1}^{\infty} a_k y_{t-k}$ is that $\sum_{k=1}^{\infty} |a_k| < \infty$. For regularity conditions and more recent advances in this area, see Cheng et al. (2000). The autoregressive representation of $\{y_t\}$ above provides justifications for using an autoregressive model for some stationary infinite variance time series in that even though the underlying structure of a time series is not autoregressive, it can be approximated by an autoregressive model under certain conditions.

3.3. A simulation study

3.3.1. Computational formulation

In this section we run a simulation study to support our theoretical results. First of all, we discuss the computational issues of SLAD-alasso. Computationally, LAD-alasso is a special case of SLAD-alasso with $h_t \equiv 1$. The algorithm of SLAD-alasso

can be converted to an estimation problem of least absolute deviation. Consider a data set $\{(y_t^*, X_t^{*T})\}$ $(t = 1, \ldots, n)$, where $(y_t^*, X_t^{*T}) = (h_t y_t, h_t X_t^T)$ $(t = p+1, \ldots, n)$, $(y_t^*, X_t^{*T}) = (0, \lambda_n r_{1t} e_t^T)$ $(t = 1, \ldots, p)$ and e_j is a p-dimensional vector with the jth component equal to 1 and all others equal to 0. Then (3.6) is equal to

$$V_{1n}(\boldsymbol{\phi}) = \Sigma_{t=1}^{n} |y_t^* - X_t^{*T}\boldsymbol{\phi}|.$$

Consequently, any standard unpenalized least absolute deviation program can be used to find the SLAD-alasso estimator without much programming effort. In our simulation study, we used the existing function rq in the R package QUANTREG to solve the LAD problem.

3.3.2. Tuning parameter selection

Tuning parameter selection is another key issue in implementing our penalty methods. First, we chose the weight h_t presented in (3.7) for SLAD-alasso. In order to support our theoretical findings, we took C to be the 90%, 95% and 100% quantile of data $\{y_1, \ldots, y_n\}$, where the third choice of C is the ordinary LAD-alasso method. To select the optimal pair of (λ_n, γ) that meets the conditions in Theorem 3.2.1 and 3.2.5, we perform a two dimensional grid search. The tuning parameter γ is selected from the set $\{2, 3, 4, 5, 6\}$. As stated in Remark 3.2.7, for any $\gamma > 1$ and $\alpha \in (0, 2)$, taking $\lambda_n = \log n$ would always satisfy conditions of Theorem 3.2.1 and 3.2.5. Based on this observation, we took $\lambda_n = \lambda^* \log n$, where λ^* is selected from 10 equally spaced grid points from 0 to 1. Finally, to select the optimal (γ, λ^*), one possibility is to do cross-validation. We conduct 5-fold cross-validation where the optimal (γ, λ^*) is selected by minimizing the least absolute prediction error of the validation data. We

also consider the following Schwartz-type information criterion.

$$\text{SIC}_{\gamma,\lambda^*} = \log\left(\frac{1}{n}\sum_{t=p+1}^{n}|y_t - X_t^T\hat{\phi}_{\gamma,\lambda^*}|\right) + \hat{df}_{\gamma,\lambda^*} \times \frac{\log n}{2n}, \tag{3.12}$$

where $\hat{df}_{\gamma,\lambda^*}$ is the number of non-zero coefficients of $\hat{\phi}_{\gamma,\lambda^*}$. This criterion is first suggested by Koenker et al. (1994) and He and Ng (1999) for choosing the regularization parameter in quantile smoothing splines and has been widely used in quantile regression literature. Since least absolute deviation regression is a special case of quantile regression, we can expect SIC to yield reasonably good results. A similar BIC type criterion has also been used in Wang et al. (2007a), where they showed that such a BIC type criterion performs much better than cross-validation in model selection. Our simulation results also support this conclusion.

3.3.3. Simulation results

We generated the data from the autoregressive model $y_t = 0.5y_{t-1} - 0.7y_{t-3} + \epsilon_t$, which was also used by Wang et al. (2007a) as the autoregressive errors for the regression model. Three error distributions, Cauchy, $S(1.5, 0; 1)$ and $N(0, 1)$, were considered, where $S(1.5, 0; 1)$ is the symmetric α-stable distribution with unit scale factor and $\alpha = 1.5$. For the LAD-alasso and SLAD-alasso methods, we start with a full model of order $p = 5$. Other choices may also be used. Our observation is that as long as p is not overly large, the model selection results will not change much when p changes.

In each case, we used (3.7) as the weights for SLAD-alasso, and took C to be the ρ quantile of data $\{y_1, \ldots, y_n\}$. Specifically, we took $\rho = 90\%$, $\rho = 95\%$ and $\rho = 100\%$, where the third choice led to LAD-alasso. For comparison, we also used the hypothesis test method proposed by Ling (2005) to do the variable selection. A series of hypothesis were conducted in the following way: start from $p = 5$ and run

a Chi-square test for each coefficient at significant level 0.05. If any test statistic is insignificant, we delete the coefficient with the smallest p-value and run the procedure again until all remaining coefficients are significant. The sample sizes were chosen as 50, 100, 200 and 400, and the summary statistics were based on 500 replications. To measure variable selection performance, we summarized the average number of correctly identified zero coefficients (CT) and the percent of times when the true model is correctly identified (PCM) using each method. We also present the average number of coefficients erroneously set to zero (ICT). To measure the estimation accuracy of each method, we calculated the empirical means and standard errors (SE) of the resulting estimator over 500 replications. We also gave the asymptotic standard errors (AE) of the SLAD-alasso estimator as in Lemma 3.2.1 for $\rho = 90\%$ and $\rho = 95\%$ cases. All simulation results are presented in Tables 1–9.

1. In all three cases, we can see that the model selection results get better when ρ grows from 90% to 100% with the LAD-alasso be the best, which is consistent with our theoretical findings.

2. On the other hand, in the Cauchy error case, the discrepancy between SE and AE becomes larger when ρ move from 90% to 95%, which might be an indication of the asymptotical distribution of the SLAD estimator becomes nonnormal even when the sample size is as large as $n = 400$.

3. In most cases, our method appears to outperform the hypothesis testing method of Ling (2005) on both model selection consistency and the accuracy of coefficient estimators, given sufficiently large sample size.

4. For SLAD-alasso, we can see that $\rho = 95\%$ is a good choice and the model selection results are pretty good, even competitive to the LAD-alasso method.

5. From Table 9, we can see that the results by LAD-alasso with SIC are also acceptable and competitive to those by the hypothesis test method with normal error terms. This suggests that LAD-alasso can do well even when the variance of error is finite.

6. For the hypothesis testing method of Ling (2005), choosing $\rho = 95\%$ has better performance than using $\rho = 90\%$ in the case of Cauchy innovations, which supports our point that the choice of C in the method of Ling (2005) has an impact on the model selection results.

7. From the three tables we can see that the simulation results by SIC outperform that of cross-validation in almost every case. Although the statistical property of the SIC criterion has not been established in our scenario, but the good empirical performance suggests studying the limiting behavior of the SIC might be a promising future research direction.

To summarize, both of two proposed automatic variable selection procedures, SLAD-alasso and LAD-alasso methods, can simultaneously perform consistent variable selection and model estimation. SLAD-alasso methods enjoys the the advantage that the resulting estimator has asymptotically normal distribution, which make the post-model selection statistical inference possible. LAD-alasso method can serve as a unified variable selection approach for heavy-tailed autoregressive models with either finite or infinite variance. In the infinite variance case, LAD-alasso method tends to provide better variable selection results than both SLAD-alasso method and Ling (2005)'s hypothesis testing procedure.

Table 1. Simulation results with Cauchy errors using SLAD-alasso with $\rho = 90\%$

		Variable Selection			Estimation accuracy					
					$\hat{\phi}_1$			$\hat{\phi}_3$		
n	Method	ICT	CT	PCM	Mean	SE	AE	Mean	SE	AE
50	CV	0.01	2.48	73.2	0.499	2.5	1.4	-0.697	3.5	1.3
	SIC	0.01	2.85	91.9	0.498	2.7	1.4	-0.697	3.7	1.3
	HTM	0.01	2.27	58.5	0.501	2.9	1.4	-0.698	3.9	1.3
100	CV	0	2.63	84.0	0.495	2.3	1.0	-0.699	2.0	0.9
	SIC	0	2.93	96.1	0.495	2.3	1.0	-0.701	2.2	0.9
	HTM	0	2.32	56.0	0.496	2.6	1.0	-0.700	2.4	0.9
200	CV	0	2.58	84.7	0.499	0.7	0.7	-0.700	0.7	0.6
	SIC	0	3.00	100.0	0.500	0.8	0.7	-0.700	0.9	0.6
	HTM	0	2.55	70.1	0.499	0.8	0.7	-0.700	0.8	0.6
400	CV	0	2.58	86.9	0.500	0.6	0.5	-0.700	0.6	0.4
	SIC	0	3.00	100.0	0.501	0.6	0.5	-0.700	0.6	0.4
	HTM	0	2.77	83.4	0.500	0.6	0.5	-0.701	0.6	0.4

ICT, the average number of coefficients erroneously set to zero; CT, the average number of zero coefficients corresponding to true zero coefficients; PCM (%), the percentage of times correct model selected; SE($\times 10^{-2}$), the empirical standard deviation; AE($\times 10^{-2}$), the asymptotic standard deviation; CV, cross-validation; SIC, Schwartz-type information criterion; HTM, the hypothesis test method; The true value of nonzero coefficients $\phi_1^0 = 0.500$ and $\phi_3^0 = -0.700$.

Table 2. Simulation results with Cauchy errors using SLAD-alasso with $\rho = 95\%$

		Variable Selection			Estimation accuracy					
					$\hat{\phi}_1$			$\hat{\phi}_3$		
n	Method	ICT	CT	PCM	Mean	SE	AE	Mean	SE	AE
50	CV	0	2.64	83.5	0.498	2.2	0.6	-0.697	1.9	0.6
	SIC	0.01	2.91	91.7	0.499	2.1	0.6	-0.698	2.1	0.6
	HTM	0	2.34	58.3	0.499	2.5	0.6	-0.700	2.4	0.6
100	CV	0	2.62	85.7	0.499	1.3	0.5	-0.699	0.9	0.4
	SIC	0	2.99	99.9	0.500	1.2	0.5	-0.699	1.0	0.4
	HTM	0	2.41	64.3	0.499	1.3	0.5	-0.699	1.1	0.4
200	CV	0	2.59	85.2	0.500	0.8	0.3	-0.700	0.6	0.3
	SIC	0	3.00	100.0	0.500	0.7	0.3	-0.700	0.6	0.3
	HTM	0	2.62	76.8	0.501	0.8	0.3	-0.700	0.7	0.3
400	CV	0	2.60	86.1	0.501	0.5	0.2	-0.700	0.5	0.2
	SIC	0	3.00	100.0	0.501	0.5	0.2	-0.700	0.5	0.2
	HTM	0	2.61	73.3	0.501	0.5	0.2	-0.700	0.5	0.2

ICT, the average number of coefficients erroneously set to zero; CT, the average number of zero coefficients corresponding to true zero coefficients; PCM (%), the percentage of times correct model selected; SE($\times 10^{-2}$), the empirical standard deviation; AE($\times 10^{-2}$), the asymptotic standard deviation; CV, cross-validation; SIC, Schwartz-type information criterion; HTM, the hypothesis test method; The true value of nonzero coefficients $\phi_1^0 = 0.500$ and $\phi_3^0 = -0.700$.

Table 3. Simulation results with Cauchy errors using LAD-alasso

		Variable Selection			Estimation accuracy			
					$\hat{\phi}_1$		$\hat{\phi}_3$	
n	Method	ICT	CT	PCM	Mean	SE	Mean	SE
50	CV	0	2.69	85.0	0.497	1.4	-0.696	1.7
	SIC	0	2.93	95.6	0.499	1.3	-0.695	1.9
100	CV	0	2.67	86.3	0.499	1.0	-0.698	0.9
	SIC	0	2.99	99.9	0.499	0.9	-0.699	0.8
200	CV	0	2.69	89.1	0.500	0.5	-0.700	0.4
	SIC	0	3.00	100.0	0.500	0.5	-0.700	0.4
400	CV	0	2.68	89.6	0.500	0.2	-0.700	0.2
	SIC	0	3.00	100.0	0.500	0.2	-0.700	0.2

ICT, the average number of coefficients erroneously set to zero; CT, the average number of zero coefficients corresponding to true zero coefficients; PCM (%), the percentage of times correct model selected; SE($\times 10^{-2}$), the empirical standard deviation; AE($\times 10^{-2}$), the asymptotic standard deviation; CV, cross-validation; SIC, Schwartz-type information criterion; The true value of nonzero coefficients $\phi_1^0 = 0.500$ and $\phi_3^0 = -0.700$.

Table 4. Simulation results with $S(1.5, 0; 1)$ errors using SLAD-alasso with $\rho = 90\%$

		Variable Selection			Estimation accuracy					
					$\hat{\phi}_1$			$\hat{\phi}_3$		
n	Method	ICT	CT	PCM	Mean	SE	AE	Mean	SE	AE
50	CV	0.07	2.06	50.3	0.472	10.4	6.2	-0.691	9.4	5.7
	SIC	0.12	2.50	62.7	0.488	9.1	6.2	-0.702	8.2	5.7
	HTM	0.16	2.61	71.9	0.502	8.2	6.2	-0.702	8.3	5.7
100	CV	0	2.50	74.3	0.489	6.0	4.4	-0.699	5.5	4.1
	SIC	0.01	2.84	86.1	0.496	6.0	4.4	-0.702	5.1	4.1
	HTM	0	2.75	83.5	0.498	5.6	4.4	-0.703	5.2	4.1
200	CV	0	2.61	80.7	0.492	4.0	3.1	-0.699	3.6	2.9
	SIC	0	2.89	92.2	0.497	3.9	3.1	-0.702	3.5	2.9
	HTM	0	2.83	86.8	0.498	4.0	3.1	-0.702	3.5	2.9
400	CV	0	2.57	82.6	0.494	2.6	2.2	-0.699	2.4	2.0
	SIC	0	2.97	97.3	0.497	2.5	2.2	-0.700	2.3	2.0
	HTM	0	2.85	86.1	0.497	2.6	2.2	-0.700	2.3	2.0

ICT, the average number of coefficients erroneously set to zero; CT, the average number of zero coefficients corresponding to true zero coefficients; PCM (%), the percentage of times correct model selected; SE($\times 10^{-2}$), the empirical standard deviation; AE($\times 10^{-2}$), the asymptotic standard deviation; CV, cross-validation; SIC, Schwartz-type information criterion; HTM, the hypothesis test method; The true value of nonzero coefficients $\phi_1^0 = 0.500$ and $\phi_3^0 = -0.700$.

Table 5. Simulation results with $S(1.5, 0; 1)$ errors using SLAD-alasso with $\rho = 95\%$

		Variable Selection			Estimation accuracy					
					$\hat{\phi}_1$			$\hat{\phi}_3$		
n	Method	ICT	CT	PCM	Mean	SE	AE	Mean	SE	AE
50	CV	0.22	1.89	35.8	0.404	13.4	13.6	-0.617	16.6	12.7
	SIC	0.23	2.19	39.2	0.436	13.7	13.6	-0.635	12.8	12.7
	HTM	0.30	2.75	65.1	0.542	13.1	13.6	-0.739	10.5	12.7
100	CV	0.03	2.24	57.6	0.459	13.1	9.6	-0.670	9.6	9.0
	SIC	0.03	2.56	65.6	0.482	10.4	9.6	-0.683	8.3	9.0
	HTM	0.05	2.86	87.1	0.487	8.8	9.6	-0.694	7.6	9.0
200	CV	0	2.35	69.3	0.491	6.8	6.8	-0.688	5.8	6.3
	SIC	0	2.67	74.2	0.506	6.3	6.8	-0.692	5.7	6.3
	HTM	0.01	2.84	86.8	0.505	6.2	6.8	-0.706	5.7	6.3
400	CV	0	2.51	78.6	0.494	5.3	4.8	-0.698	4.3	4.5
	SIC	0	2.80	83.5	0.501	4.6	4.8	-0.702	3.9	4.5
	HTM	0	2.84	88.4	0.498	4.5	4.8	-0.699	3.9	4.5

ICT, the average number of coefficients erroneously set to zero; CT, the average number of zero coefficients corresponding to true zero coefficients; PCM (%), the percentage of times correct model selected; SE($\times 10^{-2}$), the empirical standard deviation; AE($\times 10^{-2}$), the asymptotic standard deviation; CV, cross-validation; SIC, Schwartz-type information criterion; HTM, the hypothesis test method; The true value of nonzero coefficients $\phi_1^0 = 0.500$ and $\phi_3^0 = -0.700$.

Table 6. Simulation results with $S(1.5, 0; 1)$ errors using LAD-alasso

		Variable Selection			Estimation accuracy			
					$\hat{\phi}_1$		$\hat{\phi}_3$	
n	Method	ICT	CT	PCM	Mean	SE	Mean	SE
50	CV	0	2.43	72.5	0.488	6.1	-0.691	5.9
	SIC	0	2.82	85.5	0.493	5.6	-0.694	5.8
100	CV	0	2.46	75.3	0.496	3.3	-0.691	3.5
	SIC	0	2.92	93.7	0.498	3.2	-0.693	3.4
200	CV	0	2.55	83.1	0.497	2.0	-0.697	2.0
	SIC	0	2.98	98.8	0.499	1.9	-0.698	2.0
400	CV	0	2.62	85.1	0.499	1.2	-0.698	1.2
	SIC	0	2.99	99.9	0.499	1.2	-0.698	1.3

ICT, the average number of coefficients erroneously set to zero; CT, the average number of zero coefficients corresponding to true zero coefficients; PCM (%), the percentage of times correct model selected; SE($\times 10^{-2}$), the empirical standard deviation; AE($\times 10^{-2}$), the asymptotic standard deviation; CV, cross-validation; SIC, Schwartz-type information criterion; The true value of nonzero coefficients $\phi_1^0 = 0.500$ and $\phi_3^0 = -0.700$.

Table 7. Simulation results with $N(0,1)$ errors using SLAD-alasso with $\rho = 90\%$

		Variable Selection			Estimation accuracy					
					$\hat{\phi}_1$			$\hat{\phi}_3$		
n	Method	ICT	CT	PCM	Mean	SE	AE	Mean	SE	AE
50	CV	0.22	1.89	35.8	0.404	13.4	13.6	-0.617	16.6	12.7
	SIC	0.23	2.19	39.2	0.436	13.7	13.6	-0.635	12.8	12.7
	HTM	0.30	2.75	65.1	0.542	13.1	13.6	-0.739	10.5	12.7
100	CV	0.03	2.24	57.6	0.459	13.1	9.6	-0.670	9.6	9.0
	SIC	0.03	2.56	65.6	0.482	10.4	9.6	-0.683	8.3	9.0
	HTM	0.05	2.86	87.1	0.487	8.8	9.6	-0.694	7.6	9.0
200	CV	0	2.35	69.3	0.491	6.8	6.8	-0.688	5.8	6.3
	SIC	0	2.67	74.2	0.506	6.3	6.8	-0.692	5.7	6.3
	HTM	0.01	2.84	86.8	0.505	6.2	6.8	-0.706	5.7	6.3
400	CV	0	2.51	78.6	0.494	5.3	4.8	-0.698	4.3	4.5
	SIC	0	2.80	83.5	0.501	4.6	4.8	-0.702	3.9	4.5
	HTM	0	2.84	88.4	0.498	4.5	4.8	-0.699	3.9	4.5

ICT, the average number of coefficients erroneously set to zero; CT, the average number of zero coefficients corresponding to true zero coefficients; PCM (%), the percentage of times correct model selected; SE($\times 10^{-2}$), the empirical standard deviation; AE($\times 10^{-2}$), the asymptotic standard deviation; CV, cross-validation; SIC, Schwartz-type information criterion; HTM, the hypothesis test method; The true value of nonzero coefficients $\phi_1^0 = 0.500$ and $\phi_3^0 = -0.700$.

Table 8. Simulation results with $N(0,1)$ errors using SLAD-alasso with $\rho = 95\%$

		Variable Selection			Estimation accuracy					
					$\hat{\phi}_1$			$\hat{\phi}_3$		
n	Method	ICT	CT	PCM	Mean	SE	AE	Mean	SE	AE
50	CV	0.11	1.93	45.1	0.427	10.9	9.0	-0.641	13.0	8.5
	SIC	0.11	2.42	55.7	0.453	9.6	9.0	-0.664	9.7	8.5
	HTM	0.20	2.72	73.0	0.494	8.8	9.0	-0.704	9.9	8.5
100	CV	0	2.30	63.7	0.470	8.2	6.3	-0.671	8.4	6.0
	SIC	0	2.64	71.3	0.486	7.9	6.3	-0.681	7.7	6.0
	HTM	0.04	2.84	88.9	0.491	8.4	6.3	-0.688	7.3	6.0
200	CV	0	2.42	71.5	0.487	4.9	4.5	-0.690	4.4	4.3
	SIC	0	2.74	76.6	0.496	4.8	4.5	-0.696	4.1	4.3
	HTM	0	2.81	85.4	0.488	5.0	4.5	-0.700	4.2	4.3
400	CV	0	2.60	80.6	0.493	3.9	3.2	-0.693	3.4	3.0
	SIC	0	2.90	91.1	0.498	3.6	3.2	-0.695	3.2	3.0
	HTM	0	2.87	88.8	0.509	3.5	3.2	-0.695	3.2	3.0

ICT, the average number of coefficients erroneously set to zero; CT, the average number of zero coefficients corresponding to true zero coefficients; PCM (%), the percentage of times correct model selected; SE($\times 10^{-2}$), the empirical standard deviation; AE($\times 10^{-2}$), the asymptotic standard deviation; CV, cross-validation; SIC, Schwartz-type information criterion; HTM, the hypothesis test method; The true value of nonzero coefficients $\phi_1^0 = 0.500$ and $\phi_3^0 = -0.700$.

Table 9. Simulation results with $N(0,1)$ errors using LAD-alasso

		Variable Selection			Estimation accuracy			
					$\hat{\phi}_1$		$\hat{\phi}_3$	
n	Method	ICT	CT	PCM	Mean	SE	Mean	SE
50	CV	0.05	1.93	47.4	0.449	8.8	-0.625	8.7
	SIC	0.07	2.46	59.3	0.452	8.4	-0.631	9.0
100	CV	0	2.41	69.8	0.477	6.7	-0.680	6.6
	SIC	0	2.74	79.2	0.482	6.4	-0.685	6.5
200	CV	0	2.56	78.0	0.490	4.6	-0.686	3.8
	SIC	0	2.86	89.8	0.500	4.5	-0.690	3.6
400	CV	0	2.61	81.7	0.494	2.9	-0.694	2.9
	SIC	0	2.93	93.3	0.497	2.8	-0.696	2.8

ICT, the average number of coefficients erroneously set to zero; CT, the average number of zero coefficients corresponding to true zero coefficients; PCM (%), the percentage of times correct model selected; SE($\times 10^{-2}$), the empirical standard deviation; AE($\times 10^{-2}$), the asymptotic standard deviation; CV, cross-validation; SIC, Schwartz-type information criterion; The true value of nonzero coefficients $\phi_1^0 = 0.500$ and $\phi_3^0 = -0.700$.

3.4. A real data example

In this section, we employ our new method to analyze the Hang Seng Index data, which has been examined by Ling (2005). The data consists of 497 Hang Seng Index daily closing indices from June 3rd, 1996 to May 31st, 1998. Let x_t be the original data and $y_t = \log(x_t/x_{t-1})$. The original data and transformed data are displayed in Figure 3, where we can clearly see some outliers in the $\{y_t\}$ sequence, which indicates that this process may have infinite variance. Ling (2005) adopted the Hill estimator to test the tail index of y_t and showed that the data $\{y_t\}$ has an infinite variance.

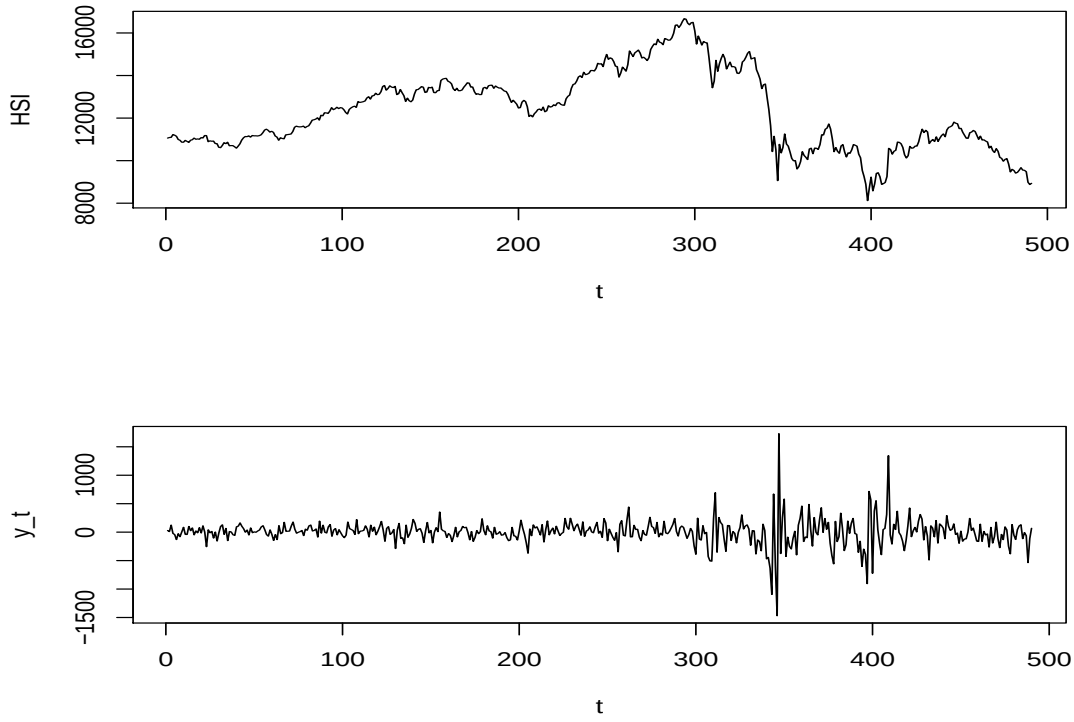

Figure 3. Original HSI data x_t (above) and the transformed data y_t(below).

To fit the data $\{y_t\}$ with an appropriate infinite variance autoregressive model,

Ling (2005) selected the best model by a series of hypothesis tests based on the self-weighted least absolute deviation estimator. The final model used by Ling (2005) is $y_t = \phi_3 y_{t-3} + \epsilon_t$, where the estimator $\tilde{\phi}_3 = 0.123$.

Table 10. The final model for the Hang Seng Index data

Method	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	y_{t-5}	y_{t-6}	y_{t-7}
LAD-alasso	0	0	0.117	0	0	0	0
Lin-95%	0	0	0.123	0	0	0	0

We employed the least absolute deviation adaptive lasso method to fit the data $\{y_t\}$. We used criterion (3.12) to select the optimal pair (γ, λ^*) in the same way as described in Section 3.3.1. The maximum autoregressive order was taken to be 7, which was the same as in Ling (2005). The estimation results are presented in Table 10. The least absolute deviation adaptive lasso method chose the model with y_{t-3} as the only relevant variable. This result coincides with the variable selection result of Ling (2005), but with a slightly different estimate $\hat{\phi}_3 = 0.117$.

CHAPTER IV

NONPARAMETRIC FUNCTION ESTIMATION USING LONGITUDINAL DATA

4.1. Introduction

Recent years have seen growing interests in developing flexible statistical models for analyzing longitudinal data or the more general cluster data. Various semiparametric (Zeger and Diggle, 1994; Zhang et al., 1998; Lin and Ying, 2001; Wang et al., 2005) and nonparametric (Rice and Silverman, 1991; Wang, 1998; Fan and Zhang, 2000; Lin and Carroll, 2000; Welsh et al., 2002; Wang, 2003; Zhu et al., 2008) models have been proposed and studied in the literature. All of these flexible, semiparametric or nonparametric, methods require specification of tuning parameters, such as the bandwidth for the local polynomial kernel method, the number of knots for regression splines, and the penalty parameter for penalized splines and smoothing splines.

The "leave-subject-out cross-validation" (LsoCV) or more generally called "leave-cluster-out cross-validation", introduced by Rice and Silverman (1991), has been widely used as the method for selecting tuning parameters in analyzing longitudinal data and clustered data. See, for example, Hoover et al. (1998); Huang et al. (2002); Wu and Zhang (2006); Wang et al. (2008). The LsoCV is intuitively appealing since the within-subject dependence is preserved by leaving out all observations from the same subject together in the cross-validation. In spite of its broad acceptance in practice, the use of LsoCV still lacks a theoretical justification to date. Moreover, the existing literature has focused on the grid search method for finding the minimizer of the LsoCV score (Chiang et al., 2001; Huang et al., 2002; Wang et al., 2008), which is computationally rather inefficient and is computationally prohibitive when there are multiple smoothing parameters. The goal of this project is twofold: First, we

develop a theoretical justification of the LsoCV by showing that the LsoCV score is asymptotically equivalent to an appropriately defined loss function; second, we develop a computationally efficient algorithm to optimize the LsoCV score for selecting multiple smoothing parameters.

Now we introduce the modeling framework to facilitate our discussion. Although all discussions in this project apply to cluster data analysis, we shall focus our presentation on longitudinal data. For a typical longitudinal data set, we have observations $(y_{ij}, \boldsymbol{x}_{ij})$, for $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, with y_{ij} being the jth response from the ith subject and $\boldsymbol{x}_{ij}$ being the corresponding vector of covariates. It is assumed that observations within a subject are correlated while observations between subjects are independent. We further assume that y_{ij} is from an exponential family with mean μ_{ij}, variance v_{ij}, the density function

$$f(y_{ij}) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \tag{4.1}$$

and the mean is related to the covariates $\boldsymbol{x}_{ijk}$, $k = 0, 1, \ldots, m$, through

$$g(\mu_{ij}) = \boldsymbol{x}_{ij0}\boldsymbol{\beta}_0 + \sum_{k=1}^{m} f_k(\boldsymbol{x}_{ijk}), \tag{4.2}$$

where g is a known monotone increasing link function, β_0 is a vector of linear regression coefficients, and f_k, $(k = 1, \ldots, m)$ are unknown smooth functions (possibly multidimensional). This is a very general framework including as special cases the generalized additive models (Berhane and Tibshirani, 1998; Lin and Zhang, 1999), the varying coefficient models (Hoover et al., 1998; Chiang et al., 2001; Huang et al., 2002), the partially linear models (He et al., 2002; Wang et al., 2005; Huang et al., 2007), and the partial linear varying coefficient models (Ahmad et al., 2005). As in the generalized linear models, the exponential distribution assumption can be relaxed; it is sufficient to specify the mean-variance relationship.

Consider first the identity link function in (4.2). Denote $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})^T$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{in_i})^T$. By using a polynomial spline basis expansion to approximate each f_k, the mean vector $\boldsymbol{\mu}_i$ can be approximated by $\boldsymbol{\mu}_i \approx \mathbf{X}_i\boldsymbol{\beta}$ for some matrix $\mathbf{X}_i$ and unknown parameter vector $\boldsymbol{\beta}$. By extending of the generalized estimating equations (GEE) of Liang and Zeger (1986), we estimate $\boldsymbol{\beta}$ by minimizing

$$pl(\boldsymbol{\beta}) = \sum_{i=1}^{n}(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{W}_i^{-1}(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta}) + \sum_{k=1}^{m}\lambda_k\boldsymbol{\beta}^T\mathbf{S}_k\boldsymbol{\beta}, \tag{4.3}$$

where $\mathbf{W}_i = \mathbf{J}_i^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{J}_i^{1/2}$ with $\mathbf{J}_i = diag\{v_{i1}, \ldots, v_{in_i}\}$, $\mathbf{R}(\boldsymbol{\alpha})$ is the possibly misspecified working correlation parameterized with $\boldsymbol{\alpha}$, $\mathbf{S}_k$ is a quadratic penalty matrix such that $\boldsymbol{\beta}^T\mathbf{S}_k\boldsymbol{\beta}$ is a roughness penalty for f_k, and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_m)$ is a vector of penalty parameters controlling the tradeoff between the model fitting and the model complexity.

For a general link function in (4.2), we estimate the parameters using an iterative reweighted penalized least square algorithm. Following the theory of Fisher's scoring method, define $\boldsymbol{z}_i^{[l]} = \mathbf{X}_i\boldsymbol{\beta}^{[l]} + (\boldsymbol{\Delta}_i^{[l]})^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i^{[l]})$, where $\boldsymbol{\beta}^{[l]}$ is the estimate at the lth step, $\boldsymbol{\Delta}_i^{[l]}$ is a diagonal matrix with diagonal elements as the first derivative of $g^{-1}(\cdot)$ evaluated at $\hat{\mu}_{ij}^{[l]}$, $(\mathbf{W}_i^{[l]})^{-1} = \boldsymbol{\Delta}_i^{[l]}(\mathbf{W}_i^{[l-1]})^{-1}\boldsymbol{\Delta}_i^{[l]}$, then at the $(l+1)$th step, $\boldsymbol{\beta}^{[l+1]}$ can be obtained by minimizing

$$pl^{[l]}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(\boldsymbol{z}_i^{[l]} - \mathbf{X}_i\boldsymbol{\beta})^T(\mathbf{W}_i^{[l]})^{-1}(\boldsymbol{z}_i^{[l]} - \mathbf{X}_i\boldsymbol{\beta}) + \sum_{k=1}^{m}\lambda_k\boldsymbol{\beta}^T\mathbf{S}_k\boldsymbol{\beta},$$

which has the same form as (4.3). Thus we will focus on (4.3) for our discussion.

Let $\hat{\mu}(\cdot)$ denote the estimate of the mean function obtained by using basis expansion of unknown functions and solving the minimization problem (4.3) for estimating the coefficients of the basis expansion. Let $\hat{\mu}^{[-i]}(\cdot)$ be the estimate of the mean function $\mu(\cdot)$ by the same method but using all the data except observations from subject

i, $1 \leq i \leq n$. The LsoCV criterion is defined as

$$\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \{\boldsymbol{y}_i - \hat{\mu}^{[-i]}(\mathbf{X}_i)\}^T \{\boldsymbol{y}_i - \hat{\mu}^{[-i]}(\mathbf{X}_i)\}. \tag{4.4}$$

The main theoretical contribution of this project is that we show, under reasonable regularity conditions, minimization of LsoCV is equivalent to minimization of the squared error loss function. In the case of penalized regression for independent data, Li (1986) established the asymptotic optimality of the generalized cross validation (GCV) (Craven and Wahba, 1979) in choosing penalty parameters by showing that the GCV score is asymptotically equivalent to the squared error loss. Our result can be viewed as an extension of the result by Li (1986) to the longitudinal data setting.

Gu and Ma (2005) and Gu and Han (2008) have extended the GCV to handle dependent data and established its asymptotic optimality by showing that their modified GCV scores are asymptotically equivalent to some tailor-made loss functions. The dependence of their GCV scores and the corresponding loss functions on the assumed correlation structure is a shortcoming, as commented in Gu and Ma (2005): " While many correlated errors can be cast as variance components with low-rank random effects, some others do not conform, which spells the limitation of the techniques developed here." Contrasting to these work, our LsoCV and the asymptotic equivalent loss function are not attached to any specific correlation structure. As an important by-product of this observation, the LsoCV can be used to select not only the penalty parameters but also the correlation structure.

Another contribution of this project is development of a fast algorithm for optimizing the LsoCV score. To avoid computation of a large number of matrix inversions, we first derive an asymptotically equivalent approximation of the LsoCV score and then derive a Newton–Raphson type algorithm. Such an algorithm is very useful when we need to select multiple penalty parameters.

4.2. Leave-subject-out cross validation

4.2.1. Heuristic justification

The initial, heuristic justification of LsoCV by Rice and Silverman (1991) is that it mimics the mean squared prediction error (MSPE). Consider some new observations $(\mathbf{X}_i, \boldsymbol{y}_i^*)$, $i = 1, \ldots, n$, taken at the same design points as the observed data. For a given estimator of the mean function $\mu(\cdot)$, denoted as $\hat{\mu}(\cdot)$, the MSPE is defined as

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n} E\|\boldsymbol{y}_i^* - \hat{\mu}(\mathbf{X}_i)\|^2 = \frac{1}{n}tr(\boldsymbol{\Sigma}) + \frac{1}{n}\sum_{i=1}^{n} E\|\mu(\mathbf{X}_i) - \hat{\mu}(\mathbf{X}_i)\|^2.$$

The independence between $\hat{\mu}^{[-i]}(\cdot)$ and $\boldsymbol{y}_i$ implies that

$$E\{\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda})\} = \frac{1}{n}tr(\boldsymbol{\Sigma}) + \frac{1}{n}\sum_{i=1}^{n} E\|\mu(\mathbf{X}_i) - \hat{\mu}^{[-i]}(\mathbf{X}_i)\|^2.$$

When n is large, $\hat{\mu}^{[-i]}(\cdot)$ should be close to $\hat{\mu}(\cdot)$, the estimate that uses observations from all subjects. Thus, we would expect that $E\{\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda})\}$ would be close to the MSPE. By leaving out together all observations from the same subject, the within-subject correlation is preserved in LsoCV and without having to model and estimate this correlation.

4.2.2. Loss function

We shall provide a formal justification of LsoCV by showing that the LsoCV is asymptotically equivalent to an appropriately defined loss function. To define the loss function, we need some notations. Denote $\boldsymbol{Y} = (\boldsymbol{y}_1^T, \cdots, \boldsymbol{y}_n^T)^T$, $\mathbf{X} = (\mathbf{X}_1^T, \cdots, \mathbf{X}_n^T)^T$, and $\mathbf{W} = diag\{\mathbf{W}_1, \cdots, \mathbf{W}_n\}$. Then, for fixed $\boldsymbol{\lambda}$ and $\mathbf{W}$, the minimizer of (4.3) has the closed-form expression

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X} + \sum_{k=1}^{m}\lambda_k\mathbf{S}_k\right)^{-1}\mathbf{X}^T\mathbf{W}^{-1}\boldsymbol{Y}. \tag{4.5}$$

The fitted mean function evaluated at the design points of the data is given by

$$\hat{\mu}(\mathbf{X}|\boldsymbol{Y}, \mathbf{W}, \boldsymbol{\lambda}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{A}(\mathbf{W}, \boldsymbol{\lambda})\boldsymbol{Y}, \tag{4.6}$$

where $\mathbf{A}(\mathbf{W}, \boldsymbol{\lambda})$ is the hat matrix defined as

$$\mathbf{A}(\mathbf{W}, \boldsymbol{\lambda}) = \mathbf{X}\big(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X} + \sum_{k=1}^{m}\lambda_k\mathbf{S}_k\big)^{-1}\mathbf{X}^T\mathbf{W}^{-1}. \tag{4.7}$$

From now on, we use $\mathbf{A}$ to represent $\mathbf{A}(\mathbf{W}, \boldsymbol{\lambda})$ without causing any confusion.

For a given estimator $\hat{\mu}(\cdot)$ of $\mu(\cdot)$, define the mean square error (MSE) loss as the true loss function

$$L(\hat{\boldsymbol{\mu}}) = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}(\mathbf{X}_i) - \mu(\mathbf{X}_i)\}^T\{\hat{\mu}(\mathbf{X}_i) - \mu(\mathbf{X}_i)\}. \tag{4.8}$$

Using (4.6), we obtain that, for the estimator obtained by minimizing (4.3), the true loss function (4.8) is

$$\begin{aligned} L(\mathbf{W}, \boldsymbol{\lambda}) &= \frac{1}{n}(\mathbf{A}\boldsymbol{Y} - \boldsymbol{\mu})^T(\mathbf{A}\boldsymbol{Y} - \boldsymbol{\mu}) \\ &= \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I} - \mathbf{A})^T(\mathbf{I} - \mathbf{A})\boldsymbol{\mu} + \frac{1}{n}\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon} - \frac{2}{n}\boldsymbol{\mu}^T(\mathbf{I} - \mathbf{A}^T)\mathbf{A}\boldsymbol{\epsilon}, \end{aligned} \tag{4.9}$$

where $\boldsymbol{\mu} = (\mu(\mathbf{X}_1)^T, \cdots, \mu(\mathbf{X}_n)^T)^T$, $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$. Since $E(\boldsymbol{\epsilon}) = 0$ and $\boldsymbol{\Sigma} = Var(\boldsymbol{\epsilon}) = diag\{\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_n\}$, the risk function is

$$R(\mathbf{W}, \boldsymbol{\lambda}) = E\{L(\mathbf{W}, \boldsymbol{\lambda})\} = \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I} - \mathbf{A})^T(\mathbf{I} - \mathbf{A})\boldsymbol{\mu} + \frac{1}{n}tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T). \tag{4.10}$$

4.2.3. Regularity conditions

This section states some necessary regularity conditions needed for our theoretical results. Notice that unless $\mathbf{W} = \mathbf{I}$, $\mathbf{A}$ is not symmetric. Define a symmetric version of $\mathbf{A}$ as $\tilde{\mathbf{A}} = \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{1/2}$. Let $\mathbf{C}_{ii}$ be the diagonal block of $\tilde{\mathbf{A}}^2$ corresponding to the ith subject. Now we state the regularity conditions. With a slight abuse of notations,

denote by $\lambda_{max}(\cdot)$ and $\lambda_{min}(\cdot)$ the largest and the smallest eigenvalues of a matrix, respectively. Let $\xi(\mathbf{\Sigma}, \mathbf{W}) = \lambda_{max}(\mathbf{\Sigma}\mathbf{W}^{-1})\lambda_{max}(\mathbf{W})$, $\boldsymbol{e}_i = \mathbf{\Sigma}_i^{-1/2}\boldsymbol{\epsilon}_i$ and $\boldsymbol{u}_i$, $\boldsymbol{v}_i$ be $n_i \times 1$ vectors such that $\boldsymbol{u}_i^T\boldsymbol{u}_i = \boldsymbol{v}_i^T\boldsymbol{v}_i = 1$, and $i = 1, \cdots, n$.

Condition 1. For some $K > 0$, $E\{(\boldsymbol{u}_i^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{v}_i)^2\} \leq K$, $i = 1, \ldots, n$.

Condition 2.

(i) $\max_{1\leq i\leq n}\{tr(\mathbf{A}_{ii})\} = O(tr(\mathbf{A})/n)$;

(ii) $\max_{1\leq i\leq n}\{tr(\mathbf{C}_{ii})\} = o(1)$.

Condition 3. $\xi(\mathbf{\Sigma}, \mathbf{W})/n = o(R(\mathbf{W}, \boldsymbol{\lambda}))$.

Condition 4. $\xi(\mathbf{\Sigma}, \mathbf{W})\{n^{-1}tr(\mathbf{A})\}^2/\{n^{-1}tr(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma})\} = o(1)$.

Condition 5. $\lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})O(n^{-2}tr(\mathbf{A})^2) = o(1)$.

Condition 1 is a mild moment condition that requires that each component of the standardized residual $\boldsymbol{e}_i = \Sigma_i^{-1/2}\boldsymbol{\epsilon}_i$ has uniformly bounded fourth moment. In particular, when $\boldsymbol{\epsilon}_i$'s are from the Gaussian distribution, the condition holds with $K = 3$.

Condition 2 extends the usually condition on leverage points used in theoretical analysis of the standard linear regression. It says that the number of dominant or extremely influential subjects is negligible compared to the total number of subjects. In the special case that all subjects have the same design points, the condition holds since $tr(\mathbf{A}_{ii}) = tr(\mathbf{A})/n$ for all $i = 1, \cdots, n$.

In this paper we assume that n_i's are bounded. Then any reasonable choice of $\mathbf{W}$ would generally yield a finite value of the quantity $\lambda_{max}(\mathbf{\Sigma}\mathbf{W}^{-1})\lambda_{max}(\mathbf{W})$, and thus Condition 3 becomes a very mild condition, because one would not expect

nonparametric estimation to deliver a parametric convergence rate of $O(n^{-1})$ (Gu and Ma, 2005).

Condition 4 is also a mild condition that extends similar conditions in the smoothing spline literature. In fact, it typically holds that, at least for univariate smoothing splines and $\mathbf{W} = \mathbf{I}$, $tr(\mathbf{A}(\lambda)) \sim O((\lambda/n)^{-1/r})$ and $tr(\mathbf{A}^2(\lambda)) \sim O((\lambda/n)^{-1/r})$ for some $r > 1$, see Gu and Ma (2005), and thus in this case, Condition 4 follows if $n(\lambda/n)^{1/r}\lambda_{min}(\mathbf{\Sigma})/\lambda_{max}(\mathbf{\Sigma}) \to \infty$ as $\lambda \to 0$.

Conditions 3–5 all indicate that a bad choice of the working covariance matrix $\mathbf{W}$ may also deteriorate the performance of the LsoCV method. For example, Condition 3–5 may be violated when $\mathbf{\Sigma}^{-1}\mathbf{W}$ or $\mathbf{W}$ is nearly singular.

4.2.4. Optimality of leave-subject-out CV

Now we provide a theoretical justification of using the minimizer of LosCV$(\mathbf{W}, \boldsymbol{\lambda})$ to select the optimal value of the penalty parameters $\boldsymbol{\lambda}$ for a fixed working covariance matrix $\mathbf{W}$. Naturally, it is reasonable to consider the value of $\boldsymbol{\lambda}$ that minimizes the true loss function $L(\mathbf{W}, \boldsymbol{\lambda})$ as the optimal value of the penalty parameters for a fixed $\mathbf{W}$. However, $L(\mathbf{W}, \boldsymbol{\lambda})$ can not be evaluated using data alone since the true mean function in the definition of $L(\mathbf{W}, \boldsymbol{\lambda})$ is unknown. One idea is to use an unbiased estimate of the risk function $R(\mathbf{W}, \boldsymbol{\lambda})$ as a proxy of $L(\mathbf{W}, \boldsymbol{\lambda})$. Define

$$U(\mathbf{W}, \boldsymbol{\lambda}) = \frac{1}{n}\mathbf{Y}^T(\mathbf{I} - \mathbf{A})^T(\mathbf{I} - \mathbf{A})\mathbf{Y} + \frac{2}{n}tr(\mathbf{A}\mathbf{\Sigma}). \tag{4.11}$$

It is easy to show that

$$U(\mathbf{W}, \boldsymbol{\lambda}) - L(\mathbf{W}, \boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \frac{2}{n}\boldsymbol{\mu}^T(\mathbf{I} - \mathbf{A})^T\boldsymbol{\epsilon} - \frac{2}{n}\{\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon} - tr(\mathbf{A}\mathbf{\Sigma})\}, \tag{4.12}$$

which has expectation zero. Thus, if Σ is known, $U(\mathbf{W}, \boldsymbol{\lambda}) - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n$ is an unbiased estimate of the risk $R(\mathbf{W}, \boldsymbol{\lambda})$.

Theorem 4.2.1. *Under Conditions 1–4, for fixed* $\mathbf{W}$ *and* $\boldsymbol{\lambda}$*, as* $n \to \infty$*,*

$$L(\mathbf{W}, \boldsymbol{\lambda}) - R(\mathbf{W}, \boldsymbol{\lambda}) = o_p(R(\mathbf{W}, \boldsymbol{\lambda})) \tag{4.13}$$

and

$$U(\mathbf{W}, \boldsymbol{\lambda}) - L(\mathbf{W}, \boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\mathbf{W}, \boldsymbol{\lambda})).$$

This theorem shows that, the function $U(\mathbf{W}, \boldsymbol{\lambda}) - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n$, the loss function $L(\mathbf{W}, \boldsymbol{\lambda})$, and the risk function $R(\mathbf{W}, \boldsymbol{\lambda})$ are asymptotically equivalent. Thus, if $\boldsymbol{\Sigma}$ is known, $U(\mathbf{W}, \boldsymbol{\lambda}) - \boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n$ is a consistent estimator of the risk function and moreover, $U(\mathbf{W}, \boldsymbol{\lambda})$ can be used as a reasonable surrogate of $L(\mathbf{W}, \boldsymbol{\lambda})$ for selecting the penalty parameters, since the $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}/n$ term does not depend on $\boldsymbol{\lambda}$.

However, $U(\mathbf{W}, \boldsymbol{\lambda})$ depends on knowledge of the true covariance matrix $\boldsymbol{\Sigma}$, which is usually not available. The following result states that the LsoCV score provides a good approximation of $U(\mathbf{W}, \boldsymbol{\lambda})$.

Theorem 4.2.2. *Under Conditions 1–5, for fixed* $\mathbf{W}$ *and* $\boldsymbol{\lambda}$*, as* $n \to \infty$*,*

$$\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) - U(\mathbf{W}, \boldsymbol{\lambda}) = o_p(L(\mathbf{W}, \boldsymbol{\lambda})),$$

and therefore

$$\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) - L(\mathbf{W}, \boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\mathbf{W}, \boldsymbol{\lambda})). \tag{4.14}$$

This theorem suggests that minimizing $\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ is asymptotically equivalent to minimizing $U(\mathbf{W}, \boldsymbol{\lambda})$ and is also equivalent to minimizing the true loss function $L(\mathbf{W}, \boldsymbol{\lambda})$. Unlike $U(\mathbf{W}, \boldsymbol{\lambda})$, $\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda})$ can be evaluated simply using the data. The theorem provides the justification of using LsoCV, as a consistent estimator of the loss or risk function, for selecting the penalty parameters.

Remark. Since our definition of the true loss function does not depend on a

specific model of the true covariance structure, we can use the loss function as a benchmark for selecting the working covariance structure. Thus the result in Theorem 4.2.2 suggests and provides a justification to use the LsoCV for selecting the working covariance matrix. This suggestion is evaluated using a simulation study in Section 4.4.3. When using the LsoCV to select the working covariance matrix, we recommend to use $\boldsymbol{\lambda} = \mathbf{0}$ so that no shrinkage bias is introduced to the parameter estimation. When the number of knots is relatively large, the bias of parameter estimation is negligible compared with the variance, and consequently minimization of the risk is equivalent to minimization of the variance.

4.2.5. Selection of working covariance structure

The major purpose of the introduction of GEE method is to improve the efficiency of resulting estimator, which makes the choice of $\mathbf{W}$ in (4.3) rather important. For a special case where all n_i's are equal and $\boldsymbol{\lambda} = \mathbf{0}$ with appropriate chosen knots, Zhu et al. (2008) shows that the function estimator using GEE based regression splines is most efficient when the true covariance structure is specified. In our setting, if we ignore the estimation bias using splines approximation, when $\boldsymbol{\lambda} = \mathbf{0}$, the variance of estimator $\hat{\boldsymbol{\beta}}(\mathbf{W})$ minimizing (4.3) is minimized when $\mathbf{W} = \boldsymbol{\Sigma}$ in the sense that

$$E\{\hat{\boldsymbol{\beta}}(\mathbf{W}) - \boldsymbol{\beta}(\mathbf{W})\}\{\hat{\boldsymbol{\beta}}(\mathbf{W}) - \boldsymbol{\beta}(\mathbf{W})\}^T - E\{\hat{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) - \boldsymbol{\beta}(\boldsymbol{\Sigma})\}\{\hat{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) - \boldsymbol{\beta}(\boldsymbol{\Sigma})\}^T$$

is positive semi-definite for any $\mathbf{W}$. Denote $\boldsymbol{\Omega}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i$, we can rewrite the risk function (4.10) as

$$\begin{aligned} R(\mathbf{W}, \mathbf{0}) &= \frac{1}{n} E \sum_{i=1}^{n} ||\mathbf{X}_i\{\hat{\boldsymbol{\beta}}(\mathbf{W}) - \boldsymbol{\beta}(\mathbf{W})\}||^2 \\ &= tr[\boldsymbol{\Omega} E\{\hat{\boldsymbol{\beta}}(\mathbf{W}) - \boldsymbol{\beta}(\mathbf{W})\}\{\hat{\boldsymbol{\beta}}(\mathbf{W}) - \boldsymbol{\beta}(\mathbf{W})\}^T], \end{aligned}$$

which implies that $R(\mathbf{W}, \mathbf{0})$ is minimized when $\mathbf{W} = \boldsymbol{\Sigma}$, given that $\boldsymbol{\Omega}_n$ is positive definite. Under certain conditions, Zhu et al. (2008) shows that the estimation bias using regression spline does not depend on $\mathbf{W}$, but it remains unclear whether this property holds in a more general case.

Nevertheless, in practice $R(\mathbf{W}, \mathbf{0})$ can still serve as a good criterion for selection of $\mathbf{W}$ if it can be consistently estimated. Equations (4.13), (4.14) and (4.16) indicate that LsoCV$(\mathbf{W}, \mathbf{0})$ and LsoCV*$(\mathbf{W}, \mathbf{0})$ can be used to select the best $\mathbf{W}$ that will yield efficient estimator.

4.3. Efficient computation

In this section, we develop a computationally efficient Newton–Raphson-type algorithm to minimize the LsoCV score.

4.3.1. Shortcut formula

The definition of LsoCV would indicate that it is necessary to solve n separate minimization problems in order to find the LsoCV score. However, a computational shortcut is available that requires solving only one minimization problem that involves all data. Recall that $\mathbf{A}$ is the hat matrix. Let $\mathbf{A}_{ii}$ denote the diagonal block of $\mathbf{A}$ corresponding to the observations of subject i.

Theorem 4.3.1. *(Shortcut Formula) The LsoCV score satisfies*

$$LsoCV(\mathbf{W}, \boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T (\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-T} (\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) \tag{4.15}$$

where $\mathbf{I}_{ii}$ *is a* $n_i \times n_i$ *identity matrix, and* $\hat{\boldsymbol{y}}_i = \hat{\mu}(\mathbf{X}_i)$.

This result, whose proof is given in the Appendix, extends a similar result for independent data (Green and Silverman, 1994, page 31). Indeed, if each subject has

only one observation and $\mathbf{W}$ is the identity matrix, then (4.15) reduces to LsoCV $= (1/n)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(1-a_{ii})^2$, which is exactly the shortcut formula for the ordinary cross-validation score.

4.3.2. An approximation of leave-subject-out CV

A close inspection of the short-cut formula of LsoCV$(\mathbf{W}, \boldsymbol{\lambda})$ given in (4.15) suggests that, the evaluation of LsoCV$(\mathbf{W}, \boldsymbol{\lambda})$ can still be computationally expensive because of the requirement of matrix inversion and formulation of the hat matrix $\mathbf{A}$. To further reduce the computational cost, using the Taylor's expansion $(\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-1} \approx \mathbf{I}_{ii} + \mathbf{A}_{ii}$, we obtain the following approximation of LsoCV$(\mathbf{W}, \boldsymbol{\lambda})$:

$$\text{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda}) = \frac{1}{n}\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\mathbf{Y} + \frac{2}{n}\sum_{i=1}^{n}\hat{\boldsymbol{e}}_i^T\mathbf{A}_{ii}\hat{\boldsymbol{e}}_i,$$

where $\hat{\boldsymbol{e}} = (\mathbf{I}-\mathbf{A})\mathbf{Y}$. The next theorem shows that this approximation is a good one in the sense that its minimization is also asymptotically equivalent to the minimization of the true loss function.

Theorem 4.3.2. *Under Conditions 1–5, for fixed* $\boldsymbol{\lambda}$*, as* $n \to \infty$*, we have*

$$\text{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda}) - L(\mathbf{W}, \boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\mathbf{W}, \boldsymbol{\lambda})). \tag{4.16}$$

This result and Theorem 4.2.2 imply that LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda})$ and LsoCV$(\mathbf{W}, \boldsymbol{\lambda})$ are asymptotically equivalent, that is, for nonrandom $\mathbf{W}$ and $\boldsymbol{\lambda}$, LsoCV$(\mathbf{W}, \boldsymbol{\lambda})$ − LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda}) = o_p(L(\mathbf{W}, \boldsymbol{\lambda}))$. The proof Theorem 4.3.2 is given in the Appendix.

4.3.3. Algorithm

We develop an efficient algorithm based on the works of Gu and Wahba (1991) and Wood (2004). The idea is to optimize the log transform of $\boldsymbol{\lambda}$ using the Newton–

Raphson method. Our algorithm can be viewed as an extension of the stable and fast algorithm of minimizing the GCV score in Wood (2004) to the longitudinal data case. Define the transformed data using the working covariance structure as $\tilde{\boldsymbol{Y}} = \mathbf{W}^{-1/2}\boldsymbol{Y}$, $\tilde{\mathbf{X}} = \mathbf{W}^{-1/2}\mathbf{X}$, and the corresponding hat matrix as

$$\tilde{\mathbf{A}} = \tilde{\mathbf{X}}\big(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \boldsymbol{S}\big)^{-1}\tilde{\mathbf{X}}^T,$$

where $\mathbf{S} = \sum_{k=1}^{m} \lambda_k \mathbf{S}_k$. Since $\mathbf{S}$ is positive semi-definite, we can find a matrix $\mathbf{B}$ with full column rank such that $\mathbf{S} = \mathbf{B}^T\mathbf{B}$ using, for example, the Cholesky decomposition. Then, form the QR decomposition $\tilde{\mathbf{X}} = \mathbf{Q}^T\mathbf{R}$, where $\mathbf{Q}$ is a $N \times p$ column orthonormal matrix and $\mathbf{R}$ is a $p \times p$ upper triangular matrix, N is the total number of observations in all subjects and p is the number of columns in the design matrix $\mathbf{X}$. The identity $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \mathbf{S} = \mathbf{R}^T\mathbf{R} + \mathbf{B}^T\mathbf{B}$ motivates us to form the singular value decomposition

$$\begin{pmatrix} \mathbf{R} \\ \mathbf{B} \end{pmatrix} = \mathbf{U}\mathbf{D}\mathbf{V}^T \approx \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*T}, \tag{4.17}$$

where $\mathbf{D}$ is the diagonal matrix of singular values, $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices. Some of the diagonal elements of $\mathbf{D}$ can be very small and thus can be removed without causing appreciable errors. The matrices $\mathbf{U}^*, \mathbf{D}^*, \mathbf{V}^*$ in (4.17) are obtained by removing small singular values from $\mathbf{D}$ along with the corresponding columns of $\mathbf{U}$ and $\mathbf{V}$. Define the sub matrix $\mathbf{U}_1^*$ of $\mathbf{U}^*$ such that $\mathbf{R} = \mathbf{U}_1^*\mathbf{D}^*\mathbf{V}^{*T}$. Then we can rewrite the matrix $\tilde{\mathbf{A}}$ as

$$\tilde{\mathbf{A}} = \mathbf{Q}^T\mathbf{R}(\mathbf{R}^T\mathbf{R} + \mathbf{B}^T\mathbf{B})^{-1}\mathbf{R}^{-1}\mathbf{Q} = \mathbf{Q}\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}^T.$$

Note that $\mathbf{Q}$ is a $N \times p$ matrix, $\mathbf{U}_1^*$ is a $p \times p$ matrix. The fast algorithm for GCV optimization in Wood (2004) takes advantage of the fact that $tr(\tilde{\mathbf{A}}) = tr(\mathbf{U}_1^*\mathbf{U}_1^{*T})$, which only takes $O(p^3)$ floating operations to evaluate. However, this appealing

property does not hold for the evaluation of LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda})$. Define

$$\alpha = \frac{1}{n}\tilde{\boldsymbol{Y}}^T(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{W}(\mathbf{I}-\tilde{\mathbf{A}})\tilde{\boldsymbol{Y}},$$

$$\beta = \frac{2}{n}\tilde{\boldsymbol{Y}}^T(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{W}\left(\sum_{i=1}^{n}\mathbf{L}_i^T\mathbf{L}_i\tilde{\mathbf{A}}\mathbf{L}_i^T\mathbf{L}_i\right)(\mathbf{I}-\tilde{\mathbf{A}})\tilde{\boldsymbol{Y}},$$

where $\mathbf{L}_i = [\mathbf{0}, \cdots, \mathbf{I}_{ii}, \cdots, \mathbf{0}]_{n_i \times N}$. It is easy to see that LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda}) = \alpha + \beta$. To make good use of the QR decomposition given above, we define the $p \times 1$ vectors $\tilde{\boldsymbol{Y}}_{\mathbf{Q}} = \mathbf{Q}^T\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{Y}}_W = \mathbf{Q}^T\mathbf{W}\tilde{\boldsymbol{Y}}$, the $p \times p$ matrix $\mathbf{Q}_W = \mathbf{Q}^T\mathbf{W}\mathbf{Q}$ and the $n_i \times p$ matrices $\mathbf{Q}_i = \mathbf{L}_i\mathbf{Q}$, $(i = 1, \cdots, n)$. Then, α and β can be computed using

$$\alpha = \frac{1}{n}(\tilde{\boldsymbol{Y}}^T\mathbf{W}\tilde{\boldsymbol{Y}} - 2\tilde{\boldsymbol{Y}}_Q^T\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_W + \tilde{\boldsymbol{Y}}_Q^T\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_W\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q),$$

$$\beta = \frac{2}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q)^T\mathbf{W}_i\mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q).$$

Following Gu and Wahba (1991), we define $\eta_j = log(\lambda_j)$, $j = 1, \cdots, m$, and compute the gradients and Hessian matrix of LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda})$ with respect to η_j's. Define $\mathbf{M}_k = \mathbf{D}^{*-1}\mathbf{V}^{*T}\mathbf{S}_k\mathbf{V}^*\mathbf{D}^{*-1}$, $\mathbf{M}_k^* = \mathbf{U}_1^*\mathbf{M}_k\mathbf{U}_1^{*T}$ and $\mathbf{K}_k = \mathbf{M}_k\mathbf{U}_1^{*T}\mathbf{Q}_W\mathbf{U}_1^*$, then

$$\frac{\partial\alpha}{\partial\eta_k} = \frac{2\lambda_k}{n}(\tilde{\boldsymbol{Y}}_Q^T\mathbf{M}_k^*\tilde{\boldsymbol{Y}}_W - \tilde{\boldsymbol{Y}}_Q^T\mathbf{U}_1^*\mathbf{K}_k\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q),$$

$$\begin{aligned}\frac{\partial\beta}{\partial\eta_k} = {} & \frac{2\lambda_k}{n}\tilde{\boldsymbol{Y}}_Q^T\mathbf{M}_k^*\sum_{i=1}^{n}\mathbf{Q}_i^T(\mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T\mathbf{W}_i)^\dagger(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q) \\ & - \frac{2\lambda_k}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q)^T\mathbf{W}_i\mathbf{Q}_i\mathbf{M}_k^*\mathbf{Q}_i^T(\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q),\end{aligned}$$

where $(\mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T\mathbf{W}_i)^\dagger = \mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T\mathbf{W}_i + \mathbf{W}_i\mathbf{Q}_i\mathbf{U}_1^*\mathbf{U}_1^{*T}\mathbf{Q}_i^T$. To derive the second derivatives of LsoCV$^*(\mathbf{W}, \boldsymbol{\lambda})$, define $\mathbf{H}_{jk} = \mathbf{U}_1^*(\mathbf{M}_k\mathbf{M}_j + \mathbf{M}_j\mathbf{M}_k)\mathbf{U}_1^{*T}$, and $\mathbf{G}_{jk} = \mathbf{M}_k\mathbf{K}_j + \mathbf{M}_j\mathbf{K}_k + \mathbf{M}_k\mathbf{Q}_W\mathbf{M}_j$. Then

$$\frac{\partial^2\alpha}{\partial\eta_k\partial\eta_j} = \frac{2\lambda_k\lambda_j}{n}\{\tilde{\boldsymbol{Y}}_Q^T\mathbf{U}_1^*\mathbf{G}_{jk}\mathbf{U}_1^{*T}\tilde{\boldsymbol{Y}}_Q - \tilde{\boldsymbol{Y}}_Q^T\mathbf{H}_{jk}\tilde{\boldsymbol{Y}}_W\} + \delta_k^j\frac{\partial\alpha}{\partial\eta_k},$$

$$\frac{\partial^2 \beta}{\partial \eta_k \partial \eta_j} = \mathbf{T}_{1,kj} + \mathbf{T}_{2,kj} + (\mathbf{T}_{3,kj} + \mathbf{T}_{3,jk}) + \mathbf{T}_{4,kj} + \delta_k^j \frac{\partial \beta}{\partial \eta_k},$$

where $\delta_k^j = 1$ if $k = j$ and 0 otherwise, and

$$\mathbf{T}_{1,kj} = -\frac{2\lambda_k \lambda_j}{n} \tilde{\boldsymbol{Y}}_Q^T \mathbf{H}_{kj} \sum_{i=1}^{n} \mathbf{Q}_i^T (\mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \mathbf{Q}_i^T \mathbf{W}_i)^\dagger (\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \tilde{\boldsymbol{Y}}_Q),$$

$$\mathbf{T}_{2,kj} = \frac{2\lambda_k \lambda_j}{n} \sum_{i=1}^{n} (\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \tilde{\boldsymbol{Y}}_Q)^T \mathbf{W}_i \mathbf{Q}_i \mathbf{H}_{kj} \mathbf{Q}_i^T (\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \tilde{\boldsymbol{Y}}_Q),$$

$$\mathbf{T}_{3,kj} = -\frac{2\lambda_k \lambda_j}{n} \tilde{\boldsymbol{Y}}_Q^T \mathbf{M}_k^* \sum_{i=1}^{n} \mathbf{Q}_i^T (\mathbf{W}_i \mathbf{Q}_i \mathbf{M}_j^* \mathbf{Q}_i^T)^\dagger (\tilde{\boldsymbol{y}}_i - \mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \tilde{\boldsymbol{Y}}_Q),$$

$$\mathbf{T}_{4,kj} = \frac{2\lambda_k \lambda_j}{n} \tilde{\boldsymbol{Y}}_Q^T \mathbf{M}_k^* \sum_{i=1}^{n} \mathbf{Q}_i^T (\mathbf{Q}_i \mathbf{U}_1^* \mathbf{U}_1^{*T} \mathbf{Q}_i^T \mathbf{W}_i)^\dagger \mathbf{Q}_i \mathbf{M}_j^* \tilde{\boldsymbol{Y}}_Q.$$

Using the formulas of the gradients and the Hessian matrix, the minimization of $\text{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ can be done using the iterative Newton–Raphson method. The key of the algorithm is the QR decomposition of $\tilde{\mathbf{X}}$ used in (4.17), which is the computationally most expensive step of the algorithm with the cost of Np^2 floating point operations. However, this QR decomposition needs only to be carried out once for all iterations of the Newton–Raphson algorithm since $\tilde{\mathbf{X}}$ does not depend on $\boldsymbol{\lambda}$. After the $\tilde{\boldsymbol{Y}}_Q$ and $\mathbf{Q}_i$'s are obtained, the evaluations of α and β cost $O(p^2)$ and $O(p^2 + Np)$ floating point operations, respectively. The computation of gradients and the Hessian matrix of $\text{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda})$ can be efficiently computed in a similar manner as α and β by using the formulas given above. As a comparison, using the Newton–Raphson method to find the minimizer of $\text{LsoCV}(\boldsymbol{\lambda})$ given in (4.15) is much more expensive. For each iteration, it involves formation of the hat matrix $\mathbf{A}$ ($O(Np^2)$ operations), the inversion of $\mathbf{A}_{ii}$'s ($O(\sum_{i=1}^{n} n_i^3)$ operations), and the summation ($O(\sum_{i=1}^{n} n_i^2)$ operations). The overall computational cost for each iteration is $O(Np^2)$, which is much more than the cost of minimizing $\text{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda})$ ($O(Np)$ operations), especially when p is large.

In our implementation of the Newton–Raphson algorithm, we followed suggestions of Wood (2004) on convergence criteria and choosing searching directions in each iteration.

4.4. Simulation studies

4.4.1. Function estimation

In this section, we illustrate the finite-sample performance of LsoCV* in selecting the penalty parameters for function estimation. In each simulation run, we set $n = 100$ and $n_i = 5$, $(i = 1, \cdots, n)$. A random sample is generated from the model

$$y_{ij} = f_1(x_{1,i}) + f_2(x_{2,ij}) + \epsilon_{ij}, \qquad j = 1, \cdots, 5, i = 1, \cdots, 100, \tag{4.18}$$

where x_1 is a subject level covariate and x_2 is an observational level covariate, both of which are drawn from $Uniform(-2, 2)$. Functions used here are from Welsh et al. (2002) with slight modifications:

$$f_1(x) = 2\sqrt{z(1-z)}\sin(2\pi\frac{1+2^{-3/5}}{1+z^{-3/5}}),$$

$$f_2(x) = \sin(8z-4) + 2\exp(-256(z-0.5)^2),$$

where $z = (x+2)/4$. The error term ϵ_{ij}'s are generated from a Gaussian distribution with zero mean, variance σ^2, and the compound symmetry correlation structure within a subject, that is

$$Cov(\epsilon_{ij}, \epsilon_{kl}) = \begin{cases} \sigma^2, & \text{if } i = j = k = l; \\ \rho\sigma^2, & \text{if } i = k,\ j \neq l, \\ 0, & otherwise; \end{cases} \tag{4.19}$$

$j, l = 1, \cdots, 5$, $i, k = 1, \cdots, 100$. In this subsection, we take $\sigma = 1$ and $\rho = 0.8$. A cubic splines with 10 equally spaced knots in $(-2, 2)$ was used for estimating each function components. Functions were estimated by minimizing (4.3) with two working correlations: the working independence (denoted as $\mathbf{W}_1 = \mathbf{I}$) and the compound symmetry with $\rho = 0.8$ (denoted as $\mathbf{W}_2$). Penalty parameters were selected by minimizing LsoCV*. Figure 4 shows the bias and the variance of estimating each component function based on 200 Monte Carlo runs, calculated over 100 equally spaced grid points in [-2,2]. The top two panels of Figure 4 show that the biases using $\mathbf{W}_1$ and $\mathbf{W}_2$ are almost the same, which is consistent with the conclusion in Zhu et al. (2008) that the bias of function estimation using regression splines does not depend on the choice of the working correlation. The bottom two panels indicate that using the true correlation structure $\mathbf{W}_2$ yields more efficient function estimation; the message is more clear in the estimation of $f_2(x)$.

4.4.2. Comparison with GCV

In this section, we compare the penalty parameter selection using the LsoCV* and the GCV (Craven and Wahba, 1979). Since the GCV is designed for independent data, we use working independence when applying LsoCV*. This means that we do not take into account the dependence in the fitting procedure for a fair comparison. Thus the difference of the results by two methods are mainly caused by the ability to take into account of dependence in the delete-subject-out CV. The data were generated using (4.18) and (4.19) in the same way as in the previous subsection. For each simulation run, to compare efficiencies of the estimated mean functions using different penalty parameter selection approaches, we calculated the ratio of true losses at different choices of penalty parameters: $L(\mathbf{I}, \boldsymbol{\lambda}_{\text{LsoCV}^*})/L(\mathbf{I}, \boldsymbol{\lambda}_{\text{GCV}})$ and $L(\mathbf{I}, \boldsymbol{\lambda}_{\text{LsoCV}^*})/L(\mathbf{I}, \boldsymbol{\lambda}_{\text{Opt}})$, where $\boldsymbol{\lambda}_{\text{GCV}}$ and $\boldsymbol{\lambda}_{\text{LsoCV}^*}$ are penalty parameters selected by using GCV and LsoCV*,

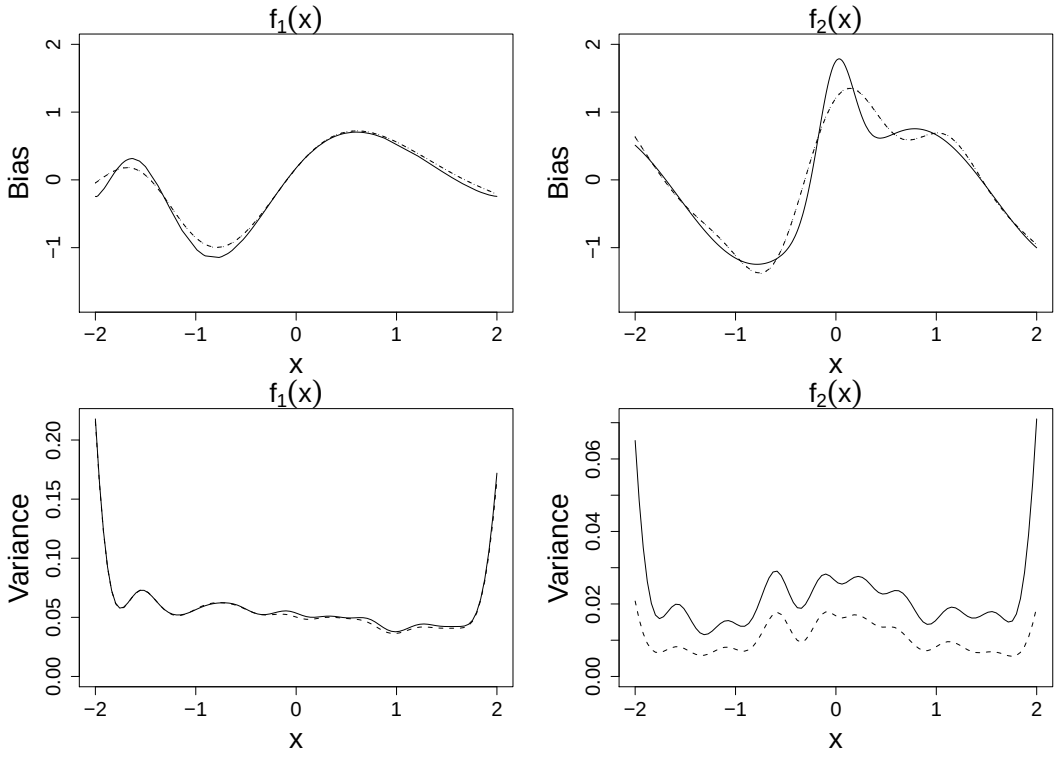

Figure 4. Simulation results for function estimation. Top panels: bias of estimated functions. Bottom panels: variance of estimated functions. In all panels, solid curves correspond to $\mathbf{W}_1$, and dashed curves $\mathbf{W}_2$.

respectively, and $\boldsymbol{\lambda}_{\text{Opt}}$ is obtained by minimizing the true loss function defined in (4.8) as if the mean function $\mu(\cdot)$ is known with $\mathbf{W} = \mathbf{I}$. A cubic spline with 10 equally spaced knots was used for estimating each function component. For the first experiment, we fixed $\rho = 0.8$ and increased the noise standard deviation σ from 0.5 to 1. For the second experiment, we fixed $\sigma = 1$ and varied ρ from -0.2 to 0.9. Results are presented in Figure 5. We see that, when σ or ρ increases, LsoCV* becomes more efficient than GCV in terms of minimizing the true loss of the estimated mean function $\hat{\mu}(\cdot)$. In addition, from the right two panels of Figure 5, we see that the minimizers of LsoCV* and the true loss function using the information of the true function are reasonably close, which supports the conclusion of Theorem 4.3.2.

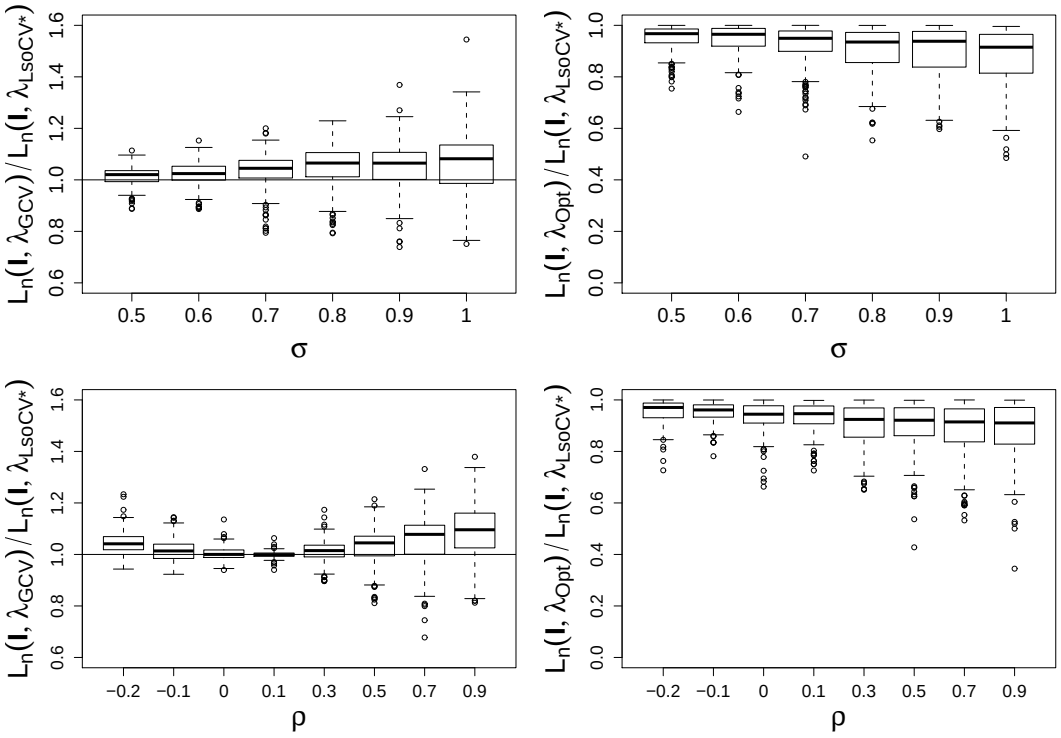

Figure 5. Relative efficiency of LsoCV* to GCV and the true loss using working independence.

4.4.3. Covariance structure selection

In this subsection, we study the performance of LsoCV* in selecting the covariance structure. The data was generated using the model (4.18) with $\sigma = 1$, $n_i = 5$ for all $i = 1, \cdots, n$. The only difference of the setup from that in Section 4.4.1 is that in this experiment, both x_1 and x_2 are set to be observational level covariates drawn from $Uniform(-2, 2)$. Four types of within-subject correlation structures were considered: independence (IND), compound symmetry with correlation coefficient ρ (CS), AR(1) with lag-one correlation ρ (AR), and unstructured correlation matrix with $\rho_{12} = \rho_{23} = 0.8$, $\rho_{13} = 0.3$ and 0 otherwise (UN). Data were generated using each one of these correlation structures as the true structure and then the LsoCV* was used to select the best working correlation from the four possible candidates.

Table 11. Simulation results for working covariance structure selection.

			Selected Structure			
n	ρ	True Structure	IND	CS	AR	UN
50	0.3	IND	89.5	4.0	6.5	0
		CS	8.0	64.5	27.5	0
		AR	11.5	11.7	77	0
		UN	0.5	0.5	12.5	86.5
	0.5	IND	98.5	0.5	1.0	0
		CS	6.0	71	23	0
		AR	3.5	13.5	83	0
		UN	3.0	3.0	11.5	82.5
	0.8	IND	99.5	0.5	0	0
		CS	3.5	69	27.5	0
		AR	2.5	21	73	3.5
		UN	6.0	3.0	5.0	86

A cubic spline with the 10 equally spaced knots in $(-2, 2)$ was used to model each unknown function and we set the penalty parameter vector $\boldsymbol{\lambda} = \mathbf{0}$. Simulation results based on 200 runs were summarized in Tables 11–13, which show very good selection results, that is, the true correlation structure was selected in majority of times.

4.5. A real data example

As a subset from the Multi-center AIDS Cohort Study, the data include the repeated measurements of CD4 cell counts and percentages on 283 homosexual men who became HIV-positive between 1984 and 1991. All subjects were scheduled to take their measurements at semi-annual visits. However, since many subjects missed some of

Table 12. Simulation results for working covariance structure selection.

			Selected Structure			
n	ρ	True Structure	IND	CS	AR	UN
100	0.3	IND	97.5	1.5	1.0	0
		CS	1.0	82.5	16.5	0
		AR	5.5	6.0	88.5	0
		UN	0.0	1.5	11	87.5
	0.5	IND	99.5	0.5	0	0
		CS	3.0	81.5	15	0.5
		AR	2.5	7.5	90	0
		UN	1.5	1.5	13.5	83.5
	0.8	IND	100	0	0	0
		CS	1.0	77.5	20	1.5
		AR	0.5	16	81	2.5
		UN	3.5	3.5	8.5	84.5

Table 13. Simulation results for working covariance structure selection.

			Selected Structure			
n	ρ	True Structure	IND	CS	AR	UN
150	0.3	IND	99	1.0	0	0
		CS	1.5	87	11.5	0
		AR	2.0	4.5	93.5	0
		UN	0	0	16	84
	0.5	IND	100	0	0	0
		CS	2.0	89	9.0	0
		AR	1.0	11.0	87	1.0
		UN	0.5	1.0	11.5	87
	0.8	IND	100	0	0	0
		CS	3.0	76	21	0
		AR	2.5	17	77	3.0
		UN	2.0	4.0	6.5	87.5

their scheduled visits, there are unequal numbers of repeated measurements and different measurement times per subject. Further details of the study can be found in Kaslow et al. (1987).

Our goal is to do statistical analysis of the trend of mean CD4 percentage depletion over time. Denote by t_{ij} the time in years of the jth measurement of the ith individual after HIV infection, by y_{ij} the ith individual's CD4 percentage at time t_{ij} and by $X_i^{(1)}$ the ith individual's smoking status with values 1 or 0 for the ith individual ever or never smoked cigarettes, respectively, after the HIV infection. To obtain a clear biological interpretation, we define $X_i^{(2)}$ to be the ith individual's centered age at HIV infection, which is obtained by the ith individual's age at infection subtract the sample average age at infection. Similarly, the ith individual's centered pre-infection CD4 percentage, denoted by $X_i^{(3)}$, is computed by subtracting the average pre-infection CD4 percentage of the sample from the ith individual's actual pre-infection CD4 percentage. These covariates, except the time, are time-invariant. Consider the varying-coefficient model

$$y_{ij} = \beta_0(t_{ij}) + X_i^{(1)}\beta_1(t_{ij}) + X_i^{(2)}\beta_2(t_{ij}) + X_i^{(2)}\beta_2(t_{ij}) + \epsilon_{ij}, \tag{4.20}$$

where $\beta_0(t)$ represents the trend of mean CD4 percentage changing over time after the infection for a non-smoker with average pre-infection CD4 percentage and average age at HIV infection, and $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$ describe the time-varying effects for cigarette smoking, age at HIV infection, and pre-infection CD4 percentage, respectively, on the post-infection CD4 percentage. Since the number observations are very uneven among subjects, we only used subjects with at least 4 observations. A cubic spline with $k = 10$ equally spaced knots was used for modeling each function. We first used the working independence $\mathbf{W}_1 = \mathbf{I}$ covariance structure to fit the data and then

use the residuals form this model to estimate parameters in the correlation function

$$\gamma(u, \alpha, \theta) = \alpha + (1 - \alpha) \exp(-\theta u),$$

where u is the lag in time and $0 < \alpha < 1$, $\theta > 0$. This correlation function was considered previously in Zeger and Diggle (1994). The estimated parameter values are $(\hat{\alpha}, \hat{\theta}) = (0.40, 0.75)$. The second working correlation matrix $\mathbf{W}_2$ considered was formed using $\gamma(u, \hat{\alpha}, \hat{\theta})$. We computed that LsoCV$(\mathbf{W}_1, \mathbf{0}) = 881.88$ and LsoCV$(\mathbf{W}_2, \mathbf{0}) = 880.33$, which implies that using $\mathbf{W}_2$ may be more desirable. This conclusion remains unchanged when the number of knots varies. To visualize the gain in estimation efficiency by using $\mathbf{W}_2$ instead of the working independence, we calculated the width of the 95% pointwise bootstrap confidence intervals based on 1000 bootstrap samples, which is displayed in Figure 6. We see that the bootstrap intervals using $\mathbf{W}_2$ is almost uniformly narrower than those using working independence, indicating more estimation efficiency.

In Figure 7, we present the fitted coefficient functions using $\mathbf{W}_2$ with the penalty parameters $\boldsymbol{\lambda}$ selected by minimizing LsoCV*$(\mathbf{W}_2, \boldsymbol{\lambda})$. The findings are consistent with previous studies conducted on the same data set; see for example, Wu and Chiang (2000), Fan and Zhang (2000), and Huang et al. (2002).

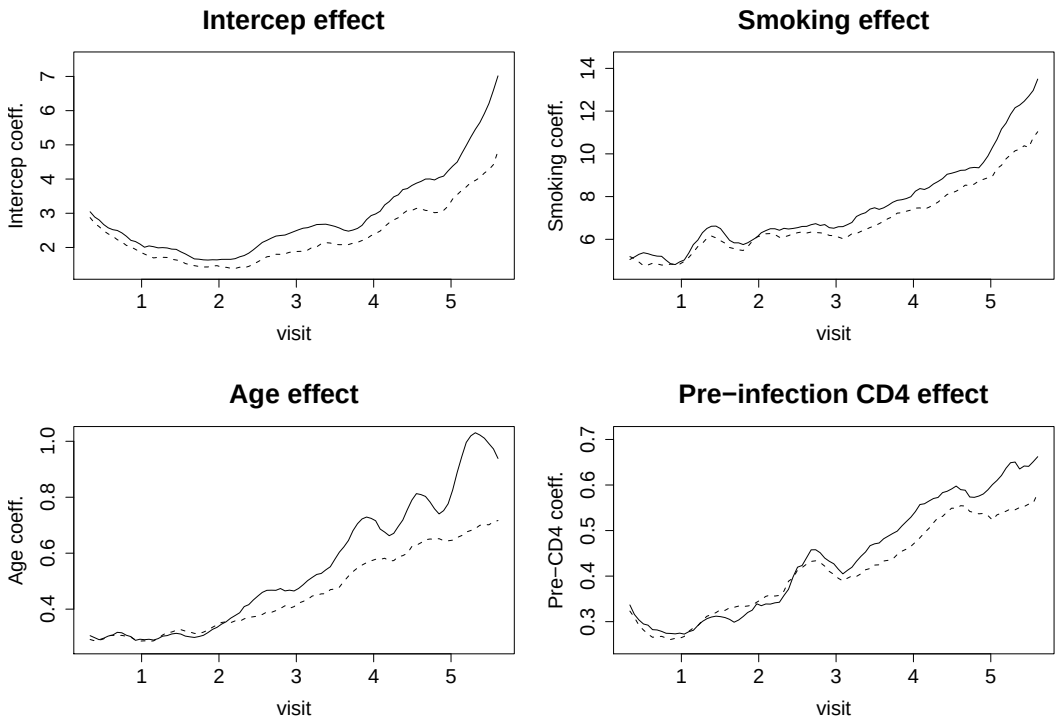

Figure 6. Width of the 95% pointwise bootstrap confidence intervals based on 1000 bootstrap samples, using the working independence (solid line) and the covariance matrix $\mathbf{W}_2$ (dashed line).

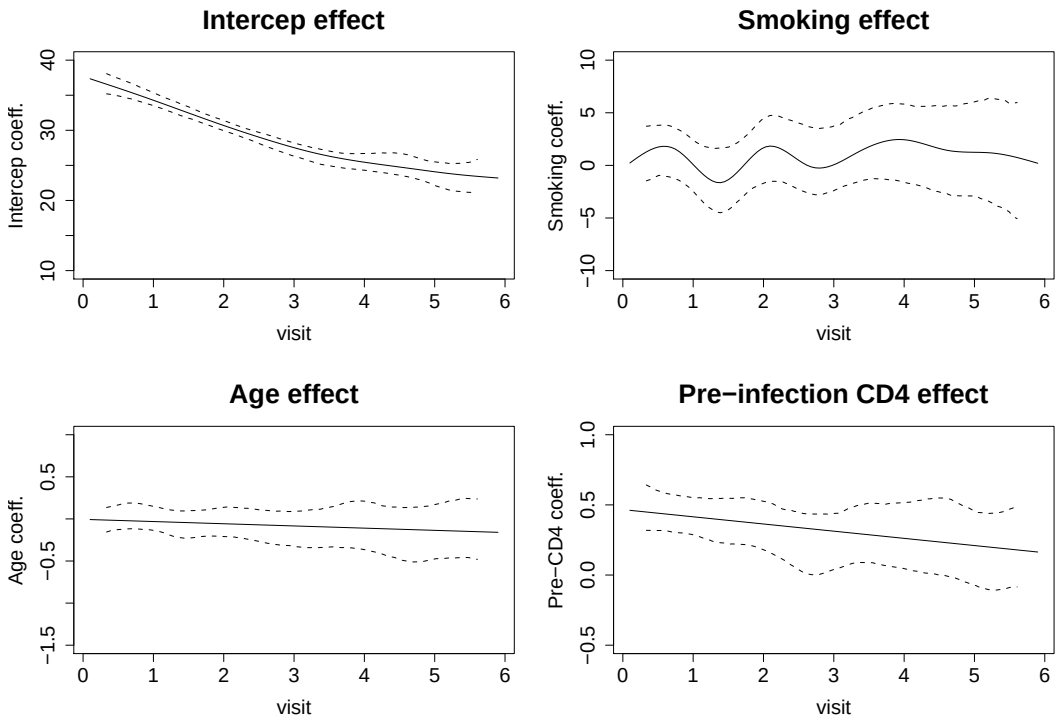

Figure 7. Fitted varying coefficient model of the CD4 data using the working covariance matrix $\mathbf{W}_2$. Solid curves are fitted coefficient functions; dotted curves show the 95% bootstrap pointwise confidence intervals.

REFERENCES

Ahmad, I., Leelahanon, S., and Li, Q. (2005), "Efficient Estimation of a Semiparametric Partially Linear Varying Coefficient Model," *The Annals of Statistics*, 33, 258–283.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," *Proceedings of the 2nd International Symposium Information Theory*, pp. 267–281.

Anderson, T. W., and Gupta, S. D. (1963), "Some Inequalities on Characteristic Roots of Matrices," *Biometrika*, 50, 522–524.

Benasseni, J. (2002), "A Complementary Proof of an Eigenvalue Property in Correspondence Analysis," *Linear Algebra and Its Applications*, 354, 49–51.

Berhane, K., and Tibshirani, R. J. (1998), "Generalized Additive Models for Longitudinal Data," *The Canadian Journal of Statistics*, 26, 517–535.

Bhansali, R. J. (1988), "g," *Journal of the Royal Statistical Society, Ser. B*, 50, 46–60.

Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Method. 2nd Ed*, New York: Springer.

Castillo, E. (1988), *Extreme Value Theory in Engineering*, Cambridge: Cambridge University Press.

Cheng, R., Miamee, A. G., and Pourahmadi, M. (2000), "Regularity and Minimality of Infinite Variance Processes," *Journal of Theoretical Probability*, 13, 1115–1122.

Chiang, C. T., Rice, J. A., and Wu, C. O. (2001), "Smoothing Spline Estimation for Varying Coefficient Models with Repeatedly Measured Dependent Variables," *Journal of the American Statistical Association*, 96, 605–619.

Cline, D. (1983), "Estimation and Linear Prediction for Regression, Autoregression and ARMA with Infinite Variance Data," PhD Thesis, Department of Statistics, Colorado State University.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-validation," *Numerische Mathematik*, 31, 377–403.

Davis, R. A., Knight, K., and Liu, J. (1992), "M-estimation for Autoregressions with Infinite Variance," *Stochastic Processes and their Applications*, 40, 145–180.

Davis, R. A., and Resnick, S. (1985), "More Limit Theory for the Sample Correlation Function of Moving Averages," *Stochastic Processes and their Applications*, 20, 257–279.

Davis, R. A., and Resnick, S. (1986), "Limit Theory for the Sample Covariance and Correlation Functions of Moving Averages," *Annals of Statistics*, 14, 533–558.

Diggle, P. J., Liang, K., and Zeger, S. L. (2002), *Analysis of Longitudinal Data, 2nd Edition*, Oxford: Oxford University Press.

Duffy, D., Mcintosh, A., Rosenstein, M., and Willinger, W. (1994), "Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks," *IEEE Journal on Selctd Areas in Communications*, 12, 544–551.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalised Likeli-

hood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., and Zhang, J. T. (2000), "Two-Step Estimation of Functional Linear Models with Applications to Longitudinal Data," *Journal of the Royal Statistical Society, Ser. B*, 62, 303–322.

Geyer, C. (1994), "On the Asymptotics of Constrained M-estimation," *The Annals of Statistics*, 22, 1993–2010.

Granger, C., and Orr, D. (1972), "Infinite Variance and Research Strategy in Time Series Analysis," *Journal of the American Statistical Association*, 67, 275–285.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall.

Gu, C., and Han, C. (2008), "Optimal smoothing with correlated data," *Sankhya: The Indian Journal of Statistics*, 70, 38–72.

Gu, C., and Ma, P. (2005), "Optimal Smoothing in Nonparametric Mixed-Effect Models," *The Annals of Statistics*, 33, 1357–1379.

Gu, C., and Wahba, G. (1991), "Minimizing GCV/GML Scores with Multiple Smoothing Parameters via the Newton Method," *SIAM Journal on Scientific and Statistical Computation*, 12, 383–398.

Hart, J., and Wehrly, T. (1986), "Kernel Regression Estimation Using Repeated Measurements Data," *Journal of the American Statistical Association*, 81, 1080–1088.

He, X., and Ng, P. (1999), "COBS: Qualitatively Constrained Smoothing via Linear Programming," *Computational Statistics*, 14, 315–337.

He, X., Zhu, Z. Y., and Fung, W. K. (2002), "Estimation in a Semiparametric Model for Longitudinal Data with Unspecified Dependence Structure," *Biometrika*, 89, 579–590.

Hill, B. (1975), "A Simple General Approach to Inference about the Tail of a Distribution," *The Annals of Statistics*, 3, 1162–1174.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models with Longitudinal Data," *Biometrika*, 85, 809–822.

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.

Huang, J. Z., Zhang, L., and Zhou, L. (2007), "Efficient Estimation in Marginal Partially Linear Models for Longitudinal/Clustered Data Using Splines," *Scandinavian Journal of Statistics*, 126, 310–318.

Knight, K. (1989), "Consistency of Akaike's Information Criterion for Infinite Variance Autoregressive Processes," *The Annals of Statistics*, 17, 824–840.

Knight, K., and Fu, W. (2000), "Consistency of Akaike's Information Criterion for Infinite Variance Autoregressive Processes," *The Annals of Statistics*, 28, 1356–1378.

Koedijk, K., Schafgans, M., and De vries, C. (1990), "The Tail Index of Exchange Rate Returns," *Journal of International Economics*, 29, 93–108.

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile Smoothing Splines," *Biometrika*, 81, 673–680.

Li, K. C. (1986), "Asymptotic Optimality of CL and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing," *The Annals of Statistics*, 14, 1101–1112.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, D. Y., and Ying, Z. (2001), "Semiparametric and Nonparametric Regression Analysis of Longitudinal Data (With Discussion)," *Journal of the American Statistical Association*, 96, 103–126.

Lin, X., and Carroll, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor is Measured without/with Error," *Journal of the American Statistical Association*, 95, 520–534.

Lin, X., and Zhang, D. (1999), "Inference in Generalized Additive Mixed Models by Using Smoothing Splines," *Journal of the Royal Statistical Society, Ser. B*, 61, 381–400.

Ling, S. (2005), "Self-weighted Least Absolute Deviation Estimation for Infinite Variance Autoregressive Models," *Journal of the Royal Statistical Society, Ser. B*, 67, 381–393.

McQuarrie, D. R., and Tsai, C. L. (1998), *Regression and Time Series Model Selection*, Singapore: World Scientific.

Miamee, A. G., and Pourahmadi, M. (1988), "Wold Decomposition, Prediction and Parameterization of Stationary Processes with Infinite Variance," *Probability Theory and Related Fields*, 79, 145–164.

Pourahmadi, M. (1988), "Autoregressive Representations of Multivariate Stationary Stochastic Processes," *Probability Theory and Related Fields*, 80, 315–322.

Resnick, S. I. (1997), "Heavy Tail Modeling and Teletraffic Data (with Discussion)," *The Annals of Statistics*, 25, 1805–1869.

Rice, J. A., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically when the Data are Curves," *Journal of the Royal Statistical Society, Ser. B*, 53, 233–243.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.

Vonesh, E. F., and Chinchilli, V. M. (1997), *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, New York: Marcel Dekker.

Wand, M., and Jones, C. (1995), *Kernel Smoothing*, London: Chapman and Hall.

Wang, H., Li, G., and Jiang, G. (2007b), "Robust Regression Shrinkage and Consistent Variable Selection via the Lad-lasso," *Journal of Business & Economic Statistics*, 25, 347–355.

Wang, H., Li, G., and Tsai, C. L. (2007a), "Regression Coefficients and Autoregressive Order Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 69, 63–78.

Wang, L., Li, H., and Huang, J. Z. (2008), "Variable Selection in Nonparametric Varying-coefficient Models for Analysis of Repeated Measurements," *Journal of the American Statistical Association*, 103, 1556–1569.

Wang, N. (2003), "Marginal Nonparametric Kernel Regression Accounting for Within Subject Correlation," *Biometrika*, 90, 43–52.

Wang, N., Carroll, R. J., and Lin, X. (2005), "Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data," *Journal of the American Statistical Association*, 100, 147–157.

Wang, Y. (1998), "Mixed effects smoothing spline analysis of variance," *Journal of The Royal Statistical Society Series B*, 60, 159–174.

Welsh, A. H., Lin, X., and Carroll, R. J. (2002), "Marginal Longitudinal Nonparametric Regression," *Journal of the American Statistical Association*, 97, 482–493.

Wild, C. J., and Yee, T. W. (1996), "Additive Extensions to Generalized Estimation Equation Methods," *Journal of the Royal Statistical Society, Ser. B*, 58, 711–725.

Wood, S. N. (2004), "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models," *Journal of the American Statistical Association*, 99, 673–686.

Wu, C. O., and Chiang, C. T. (2000), "Kernel Smoothing on Varying Coefficient Models with Longitudinal Dependent Variable," *Statistica Sinica*, 10, 433–456.

Wu, H., and Zhang, J.-T. (2006), *Nonparametric regression methods for longitudinal data analysis*, Hoboken, New Jersey: John Wiley and Sons.

Zeger, S. L., and Diggle, P. J. (1994), "Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699.

Zhang, D. W., Lin, X., Raz, J., and Sowers, M. (1998), "Semiparametric Stochastic Mixed Models for Longitudinal Data," *Journal of the American Statistical Association*, 93, 710–719.

Zhu, Z., Fung, W., and He, X. (2008), "On the Asymptotics of Marginal Regression Splines with Longitudinal Data," *Biometrika*, 95, 907–917.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., and Li, R. (2008), "One-step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36, 1509–1533.

APPENDIX A

TECHNICAL PROOFS OF CHAPTER III

Before proving Theorem 3.2.1, we first give a Lemma that will be used in the proof of that theorem.

Lemma 4.5.1. *Denote* $\tilde{L}_n(\boldsymbol{u}) = \sum_{t=p+1}^{n} h_t(|\epsilon_t - n^{-\frac{1}{2}} X_t^T \boldsymbol{u}| - |\epsilon_t|)$. *Then under Conditions* 1 *and* 2, *for any fixed* $\boldsymbol{u}$, *we have*

$$\tilde{L}_n(\boldsymbol{u}) = -\boldsymbol{u}^T T_n + f(0)\boldsymbol{u}^T \left(\frac{1}{n} \sum_{t=p+1}^{n} h_t X_t X_t^T \right) \boldsymbol{u} + o_p(1) \to -\boldsymbol{u}^T \Phi + f(0)\boldsymbol{u}^T \Sigma \boldsymbol{u} \quad \text{(A.1)}$$

in distribution, where $T_n = n^{-\frac{1}{2}} \sum_{t=p+1}^{n} h_t X_t \{I(\epsilon_t > 0) - I(\epsilon_t < 0)\}$, $\Phi \sim N(0, \Omega)$, *and* Σ *and* Ω *are presented in Lemma 3.2.1.*

Lemma 4.5.1 can be obtained from the proof of Theorem 1 in Ling (2005).

Proof of Theorem 3.2.1 We adopt an approach similar to a proof in Zou (2006). At first, we prove the asymptotic normality part. Denote

$$\begin{aligned} \tilde{V}_n(\boldsymbol{u}) &= L_{1n}(\boldsymbol{\phi}_0 + n^{-\frac{1}{2}}\boldsymbol{u}) - L_{1n}(\boldsymbol{\phi}_0) + \lambda_n \Sigma_{j=1}^{p} r_{1j}(|\phi_j^0 + n^{-\frac{1}{2}} u_j| - |\phi_j^0|) \\ &= \tilde{L}_n(\boldsymbol{u}) + n^{-\frac{1}{2}} \lambda_n \Sigma_{j=1}^{p} r_{1j} n^{\frac{1}{2}}(|\phi_j^0 + n^{-\frac{1}{2}} u_j| - |\phi_j^0|), \end{aligned} \quad \text{(A.2)}$$

where $\tilde{L}_n(\boldsymbol{u})$ has been defined in Lemma 4.5.1. Then we have $n^{\frac{1}{2}}(\hat{\boldsymbol{\phi}}_{1n} - \boldsymbol{\phi}_0) = \arg\min\{\tilde{V}_n(\boldsymbol{u})\}$. By Lemma 4.5.1 for each $\boldsymbol{u}$, we have the asymptotic property (A.1) for $\tilde{L}_n(\boldsymbol{u})$. Now consider the second part of (A.2). If $\phi_j^0 \neq 0$, then by the definition of r_{1j}, we have $r_{1j} \to |\phi_j^0|^{-\gamma}$ in probability. Furthermore, we have $n^{\frac{1}{2}}(|\phi_j^0 + n^{-\frac{1}{2}} u_j| - |\phi_j^0|) \to u_j \text{sgn}(\phi_j^0)$. Thus, by Slutsky's theorem and the condition that $\lambda_n n^{-\frac{1}{2}} \to 0$, we have

$$\lambda_n r_{1j}(|\phi_j^0 + n^{-\frac{1}{2}} u_j| - |\phi_j^0|) \to 0 \quad \text{(A.3)}$$

in probability. If $\phi_j^0 = 0$, then $n^{\frac{1}{2}}(|\phi_j^0 + n^{-\frac{1}{2}}u_j| - |\phi_j^0|) = |u_j|$ and $n^{\frac{1}{2}}\tilde{\phi}_{1j} = O_p(1)$, where $\tilde{\phi}_{1j}$ is the jth element of $\tilde{\phi}_{1n}$. Thus, we have

$$\begin{aligned} \lambda_n r_{1j}(|\phi_j^0 + n^{-\frac{1}{2}}u_j| - |\phi_j^0|) &= n^{-\frac{1}{2}}\lambda_n r_{1j}|u_j| \\ &= \lambda_n n^{\frac{\gamma-1}{2}}(|n^{\frac{1}{2}}\tilde{\phi}_{1j}|^{-\gamma})|u_j| \\ &\to \begin{cases} 0 & (u_j = 0), \\ \infty & (u_j \neq 0) \end{cases} \end{aligned} \tag{A.4}$$

in probability, where (A.4) follows because $\lambda_n n^{\frac{\gamma-1}{2}} \to \infty$.

Finally, by (A.1)–(A.4), using Slutsky's theorem, we have $\tilde{V}_n(\boldsymbol{u}) \to \tilde{V}(\boldsymbol{u})$ in distribution, where

$$\tilde{V}(\boldsymbol{u}) = \begin{cases} -\boldsymbol{u}_{\mathcal{S}}^T\Phi_{\mathcal{S}} + f(0)\boldsymbol{u}_{\mathcal{S}}^T\Sigma_{\mathcal{S}}\boldsymbol{u}_{\mathcal{S}} & (u_j = 0, j \in \mathcal{S}^c), \\ \infty & (\text{otherwise}), \end{cases}$$

where, as before, $\boldsymbol{u}_{\mathcal{S}}$ denotes the subvector of $\boldsymbol{u}$ corresponding to the non-zero coefficients. Since $\tilde{V}_n(\boldsymbol{u})$ is convex and has the unique minimum, following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\phi}}_{1\mathcal{S}} - \boldsymbol{\phi}_{\mathcal{S}}^0) \to \frac{1}{2f(0)}\Sigma_{\mathcal{S}}^{-1}\Phi_{\mathcal{S}} \tag{A.5}$$

in distribution and $n^{\frac{1}{2}}\hat{\phi}_{1\mathcal{S}^c} \to 0$ in distribution, where $\hat{\phi}_{1\mathcal{S}^c}$ is the subvector of $\hat{\phi}_{1n}$ corresponding to the zero coefficients. Since $\Phi_{\mathcal{S}} \sim N(0, \Omega_{\mathcal{S}})$, by (A.5), the asymptotic normality part is obtained.

Now we prove the consistent variable selection part. For all $j \in \mathcal{S}$, by the asymptotic normality (A.5), we have $\text{pr}\,(j \in \mathcal{S}_1^*) \to 1$ immediately. Then it suffices to show that for all $j \in \mathcal{S}^c$, $\text{pr}\,(j \in \mathcal{S}_1^*) \to 0$. For any $j \in \mathcal{S}^c$, if $j \in \mathcal{S}_1^*$, then we must have $\lambda_n r_{1j} \leq \sum_{t=p+1}^n h_t|X_{tj}|$, where X_{tj} is the jth element of X_t. Thus, it follows

immediately that

$$\mathrm{pr}\,(j \in \mathcal{S}_1^*) \leq \mathrm{pr}\,\big(\lambda_n |\tilde{\phi}_{1j}|^{-\gamma} \leq \sum_{t=p+1}^{n} h_t |X_{tj}|\big). \tag{A.6}$$

However,

$$\frac{1}{n-p}\sum_{t=p+1}^{n} h_t |X_{tj}| \leq \big(\frac{1}{n-p}\sum_{t=p+1}^{n} h_t^2 |X_{tj}|^2\big)^{\frac{1}{2}} \to \Omega_{jj}^{\frac{1}{2}} \tag{A.7}$$

almost surely as $n \to \infty$, where Ω_{jj} is the jth diagonal element of Ω. Moreover,

$$\frac{\lambda_n}{(n-p)|\tilde{\phi}_{1j}|^{\gamma}} = \frac{n}{n-p} \times \frac{\lambda_n}{n^{1-\frac{\gamma}{2}} |n^{\frac{1}{2}}\tilde{\phi}_{1j}|^{\gamma}} \to \infty, \tag{A.8}$$

where we have use the condition $\lambda_n n^{(\frac{\gamma}{2}-1)} \to \infty$ and the property $n^{\frac{1}{2}}\tilde{\phi}_{1j} = O_p(1)$.

Combining (A.6)–(A.8), we have $\mathrm{pr}\,(j \in \mathcal{S}_1^*) \to 0$. Thus, the variable selection consistency is obtained, which completes the proof of Theorem 3.2.1.

Proof of Theorem 3.2.5 The proof is similar to that of Theorem 3.2.1. Recall the definition of $W_n(\boldsymbol{u})$ in (3.9) and denote

$$\tilde{V}_{1n}(\boldsymbol{u}) = W_n(\boldsymbol{u}) + \lambda_n \Sigma_{j=1}^{p} r_{2j}(|\phi_j^0 + b_n^{-1}u_j| - |\phi_j^0|). \tag{A.9}$$

Then we have $b_n(\hat{\phi}_{2n} - \phi_0) = \arg\min\{\tilde{V}_{1n}(\boldsymbol{u})\}$. By Lemma 3.2.2, we have, for each u,

$$W_n(\boldsymbol{u}) \to W(\boldsymbol{u}) \tag{A.10}$$

in distribution, where $W(\boldsymbol{u})$ is defined in (3.10). For the second part of (A.9), by a discussion similar to that in the proof of Theorem 3.2.1, we have

$$\lambda_n r_{2j}(|\phi_j^0 + b_n^{-1}u_j| - |\phi_j^0|) \to \begin{cases} 0 & (\phi_j^0 \neq 0), \\ 0 & (\phi_j^0 = 0, u_j = 0), \\ \infty & (\phi_j^0 = 0, u_j \neq 0) \end{cases} \tag{A.11}$$

in probability. Thus, combining (A.10) and (A.11) and using Slutsky's theorem, we

have $\tilde{V}_{1n}(u) \to \tilde{V}_1(u)$ in distribution, where

$$\tilde{V}_1(u) = \begin{cases} W(u|_{u_{\mathcal{S}^c}=0}) & (u_j = 0, j \in \mathcal{S}^c), \\ \infty & (\text{otherwise}). \end{cases}$$

Following a discussion similar to that in Davis et al. (1992), it is readily seen that the conditions in Theorem 3.2.5 guarantee $W(\boldsymbol{u}|_{\boldsymbol{u}_{\mathcal{S}^c}=0})$ to have a unique minimum ξ_1 almost surely, thus the unique minimum of $\tilde{V}_1(\boldsymbol{u})$ is $(\xi_1^T, 0^T)^T$. Since $\tilde{V}_{1n}(\boldsymbol{u})$ is convex, following the epi-convergence results of Geyer (1994) and Knight and Fu (2000) again, we finally have

$$b_n(\hat{\boldsymbol{\phi}}_{2\mathcal{S}} - \boldsymbol{\phi}^0_{\mathcal{S}}) \to \xi_1 \tag{A.12}$$

in distribution and $b_n\hat{\phi}_{2\mathcal{S}^c} \to_D 0$, where $\hat{\phi}_{2\mathcal{S}^c}$ is the subvector of $\hat{\phi}_{2n}$ corresponding to the zero coefficients. Therefore, $b_n(\hat{\boldsymbol{\phi}}_{2\mathcal{S}} - \boldsymbol{\phi}^0_{\mathcal{S}}) = O_p(1)$.

Next we prove the variable selection consistency. For all $j \in \mathcal{S}$, by the asymptotic property (A.12), we have $\text{pr}\,(j \in \mathcal{S}_2^*) \to 1$ immediately. Then it suffices to show that for all $j \in \mathcal{S}^c$, $\text{pr}\,(j \in \mathcal{S}_2^*) \to 0$. For any $j \in \mathcal{S}^c$, if $j \in \mathcal{S}_2^*$, then we must have $\lambda_n r_{2j} \leq \sum_{t=p+1}^n |X_{tj}|$, where X_{tj} is the jth element of X_t. Thus, it follows immediately that

$$\text{pr}\,(j \in \mathcal{S}_2^*) \leq \text{pr}\,\big(\lambda_n|\tilde{\phi}_{2j}|^{-\gamma} \leq \sum_{t=p+1}^n |X_{tj}|\big). \tag{A.13}$$

Using the inequality $|x+y|^\delta \leq |x|^\delta + |y|^\delta$ for $0 < \delta < 1$, we have

$$\frac{1}{n-p}\Big(\sum_{t=p+1}^n |X_{tj}|\Big)^{\alpha/2} \leq \frac{1}{n-p}\Big(\sum_{t=p+1}^n |X_{tj}|^{\alpha/2}\Big) \to E(|y_t|^{\frac{\alpha}{2}}) < \infty \tag{A.14}$$

almost surely as $n \to \infty$, where the convergence make sense by the ergodic theorem. However,

$$\frac{1}{n-p}\left(\frac{\lambda_n}{|\tilde{\phi}_{2j}|^\gamma}\right)^{\alpha/2} = \frac{n}{n-p} \times n^{-1}\left(\frac{\lambda_n b_n^\gamma}{|b_n\tilde{\phi}_{2j}|^\gamma}\right)^{\alpha/2} \to \infty, \tag{A.15}$$

where we have use the condition $n^{-1}(\lambda_n b_n^{\gamma-1})^{\frac{\alpha}{2}} \to \infty$ and the property $b_n\tilde{\phi}_{2j} = O_p(1)$.

Combining (A.13)–(A.15), we have $\text{pr}\,(j \in \mathcal{S}_2^*) \to 0$. Thus, We have shown the variable selection consistency, completing the proof of Theorem 3.2.5.

APPENDIX B

TECHNICAL PROOFS OF CHAPTER IV

Proof of Theorem 4.3.1. For fixed $\boldsymbol{\lambda}$, let $\hat{\boldsymbol{\beta}}^{[-i]}$ be the minimizer of (4.3) using data without observations from the subject i. Consider the data set $\{(\boldsymbol{y}_l^*, X_l)\}, 1 \leq l \leq n$, where $\boldsymbol{y}_i^* = \mathbf{X}_i\hat{\beta}^{[-i]}$ and $\boldsymbol{y}_l^* = \boldsymbol{y}_l$ if $l \neq i$, $l = 1, \cdots, n$. Then, for any $\boldsymbol{\beta}$,

$$\begin{aligned}
pl(\boldsymbol{\beta}) &= \sum_{l=1}^{n}(\boldsymbol{y}_l^* - \mathbf{X}_l\boldsymbol{\beta})^T\mathbf{W}_l^{-1}(\boldsymbol{y}_l^* - \mathbf{X}_l\boldsymbol{\beta}) + \sum_{k=1}^{m}\lambda_k\boldsymbol{\beta}^T\mathbf{S}_k\boldsymbol{\beta} \\
&\geq \sum_{l\neq i}(\boldsymbol{y}_l^* - \mathbf{X}_l\boldsymbol{\beta})^T\mathbf{W}_l^{-1}(\boldsymbol{y}_l^* - \mathbf{X}_l\boldsymbol{\beta}) + \sum_{k=1}^{m}\lambda_k\boldsymbol{\beta}^T\mathbf{S}_k\boldsymbol{\beta} \\
&\geq \sum_{l\neq i}(\boldsymbol{y}_l^* - \mathbf{X}_l\hat{\boldsymbol{\beta}}^{[-i]})^T\mathbf{W}_l^{-1}(\boldsymbol{y}_l^* - \mathbf{X}_l\hat{\boldsymbol{\beta}}^{[-i]}) + \sum_{k=1}^{m}\lambda_k\hat{\boldsymbol{\beta}}^{[-i]T}\mathbf{S}_k\hat{\boldsymbol{\beta}}^{[-i]} \\
&= \sum_{l=1}^{n}(\boldsymbol{y}_l^* - \mathbf{X}_l\hat{\boldsymbol{\beta}}^{[-i]})^T\mathbf{W}_l^{-1}(\boldsymbol{y}_l^* - \mathbf{X}_l\hat{\boldsymbol{\beta}}^{[-i]}) + \sum_{k=1}^{m}\lambda_k\hat{\boldsymbol{\beta}}^{[-i]T}\mathbf{S}_k\hat{\boldsymbol{\beta}}^{[-i]}.
\end{aligned}$$

Hence, $\hat{\boldsymbol{\beta}}^{[-i]}$ is the minimizer of $pl(\boldsymbol{\beta})$ given data $\{(\boldsymbol{y}_l^*, \mathbf{X}_l)\}$, which implies

$$\mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]} = \mathbf{L}_i\mathbf{A}(\boldsymbol{\lambda})\boldsymbol{Y}^*,$$

where $\boldsymbol{Y}^* = (\boldsymbol{y}_1^{*T}, \cdots, \boldsymbol{y}_n^{*T})^T$, and $\mathbf{L}_i = [\mathbf{0}, \cdots, \mathbf{I}_{ii}, \cdots, \mathbf{0}]_{n_i\times N}$ with $\mathbf{I}_{ii}$ being the $n_i \times n_i$ identity matrix. By the definition of $\boldsymbol{Y}^*$ and using $\mathbf{A}_{ii} = \mathbf{L}_i\mathbf{A}\mathbf{L}_i^T$, we have that

$$\mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]} = \mathbf{L}_i\mathbf{A}\left\{\boldsymbol{Y} - \mathbf{L}_i^T\left(\boldsymbol{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]}\right)\right\} = \hat{\boldsymbol{y}}_i - \mathbf{A}_{ii}(\boldsymbol{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]}).$$

By some straightforward algebra, we have that

$$\left(\boldsymbol{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{[-i]}\right) = (\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i),$$

Plugging this identity into the definition of $\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda})$, we obtain

$$\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T(\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-T}(\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-1}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i),$$

which is the desired formula. □

Let $\lambda_{max}(\mathbf{M}) \geq \lambda_2(\mathbf{M}) \geq \cdots \geq \lambda_{min}(\mathbf{M})$ denote the eigenvalues of the $p \times p$ symmetric matrix $\mathbf{M}$. We present several useful lemmas.

Lemma 4.5.2. *For any positive semi-definite matrices* $\mathbf{M}_1$ *and* $\mathbf{M}_2$*,*

$$\lambda_i(\mathbf{M}_1)\lambda_p(\mathbf{M}_2) \leq \lambda_i(\mathbf{M}_1\mathbf{M}_2) \leq \lambda_i(\mathbf{M}_1)\lambda_1(\mathbf{M}_2), \quad i = 1, \cdots, p\,. \tag{B.1}$$

Proof. See Lemma 2.2.1 of Anderson and Gupta (1963) and Benasseni (2002). □

Lemma 4.5.3. *For any positive semi-definite matrices* $\mathbf{M}_1$ *and* $\mathbf{M}_2$*,*

$$tr(\mathbf{M}_1\mathbf{M}_2) \leq \lambda_{max}(\mathbf{M}_1)tr(\mathbf{M}_2), \tag{B.2}$$

Proof. The proof is trivial, using the eigen decomposition of $\mathbf{M}_1$. □

Lemma 4.5.4. *Eigenvalues of* $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ *and* $(\mathbf{I} - \mathbf{A})\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{A})^T$ *are bounded above by* $\xi(\boldsymbol{\Sigma}, \mathbf{W}) = \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})\lambda_{max}(\mathbf{W})$.

Proof. Recall that $\tilde{\mathbf{A}} = \mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{1/2}$. For $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$, by Lemma 4.5.2, we have that

$$\begin{aligned}\lambda_i(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T) &= \lambda_i(\mathbf{W}^{1/2}\tilde{\mathbf{A}}\mathbf{W}^{-1/2}\boldsymbol{\Sigma}\mathbf{W}^{-1/2}\tilde{\mathbf{A}}\mathbf{W}^{1/2}) \leq \lambda_i(\tilde{\mathbf{A}}\mathbf{W}\tilde{\mathbf{A}})\lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1}) \\ &\leq \lambda_i(\tilde{\mathbf{A}}^2)\lambda_{max}(\mathbf{W})\lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1}) \leq \xi(\boldsymbol{\Sigma}, \mathbf{W}).\end{aligned}$$

The last inequality follows from the fact that $\max_i\{\lambda_i(\tilde{\mathbf{A}}^2)\} \leq 1$. Similarly,

$$\begin{aligned}\lambda_i((\mathbf{I} - \mathbf{A})\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{A})^T) &= \lambda_i(\mathbf{W}^{1/2}(\mathbf{I} - \tilde{\mathbf{A}})\mathbf{W}^{-1/2}\boldsymbol{\Sigma}\mathbf{W}^{-1/2}(\mathbf{I} - \tilde{\mathbf{A}})^T\mathbf{W}^{1/2}) \\ &\leq \lambda_i((\mathbf{I} - \tilde{\mathbf{A}})^2)\lambda_{max}(\mathbf{W})\lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1}) \\ &\leq \xi(\boldsymbol{\Sigma}, \mathbf{W}),\end{aligned}$$

where we have used $\max_i\{\lambda_i((\mathbf{I}-\tilde{\mathbf{A}})^2)\} \leq 1$. □

Denote $\boldsymbol{e} = (\boldsymbol{e}_1^T, \cdots, \boldsymbol{e}_n^T)^T$, where $\boldsymbol{e}_i$'s are independent random vectors with length n_i, $E(\boldsymbol{e}_i) = 0$ and $Var(\boldsymbol{e}) = \mathbf{I}_i$ for $i = 1, \cdots, n$. For each i, define $z_{ij} = \boldsymbol{u}_{ij}^T \boldsymbol{e}_i \boldsymbol{e}_i^T \boldsymbol{v}_{ij}$ where $\boldsymbol{u}_{ij}$ and $\boldsymbol{v}_{ij}$ are vectors with the property $\boldsymbol{u}_{ij}^T \boldsymbol{u}_{ik} = \boldsymbol{v}_{ij}^T \boldsymbol{v}_{ik} = 1$ if $j = k$ and 0 otherwise, $j, k = 1, \cdots, n_i$.

Lemma 4.5.5. *If there exists a constant K such that $E(z_{ij}^2) \leq K$ holds for all $j = 1, \cdots, n_i$, $i = 1, \cdots, n$, then*

$$Var(\boldsymbol{e}^T \mathbf{B} \boldsymbol{e}) \leq 2tr(\mathbf{B}\mathbf{B}^T) + K\left\{\sum_{j=1}^{n_i} d_{ij}(\mathbf{B}_{ii})\right\}^2, \tag{B.3}$$

where $\mathbf{B}$ is any $N \times N$ matrix (not necessarily symmetric), $\mathbf{B}_{ii}$ is the ith $(n_i \times n_i)$ diagonal block of $\mathbf{B}$, and $d_{ij}(\mathbf{B}_{ii})$ is the jth singular value of $\mathbf{B}_{ii}$.

Proof. Since $E(\boldsymbol{e}^T \mathbf{B} \boldsymbol{e}) = tr(\mathbf{B})$, we have that

$$Var(\boldsymbol{e}^T \mathbf{B} \boldsymbol{e}) = E\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \boldsymbol{e}_i^T \mathbf{B}_{ij} \boldsymbol{e}_j^T \boldsymbol{e}_k^T \mathbf{B}_{lk}^T \boldsymbol{e}_l\right) - \{tr(\mathbf{B})\}^2.$$

Using the fact that $\boldsymbol{e}_i$'s are independent and $E(\boldsymbol{e}_i) = 0$, we obtain

$$\begin{aligned} Var(\boldsymbol{e}^T \mathbf{B} \boldsymbol{e}) &= \sum_{i=1}^{n} E(\boldsymbol{e}_i^T \mathbf{B}_{ii} \boldsymbol{e}_i)^2 + \sum_{i\neq j=1}^{n} E(\boldsymbol{e}_i^T \mathbf{B}_{ii} \boldsymbol{e}_i \boldsymbol{e}_j^T \mathbf{B}_{jj}^T \boldsymbol{e}_j) \\ &\quad + 2\sum_{i\neq j=1}^{n} E(\boldsymbol{e}_i^T \mathbf{B}_{ij} \boldsymbol{e}_j \boldsymbol{e}_j^T \mathbf{B}_{ij}^T \boldsymbol{e}_i) - \{tr(\mathbf{B})\}^2 \\ &= \sum_{i=1}^{n} E(\boldsymbol{e}_i^T \mathbf{B}_{ii} \boldsymbol{e}_i)^2 + \sum_{i\neq j=1}^{n} tr(\mathbf{B}_{ii}) tr(\mathbf{B}_{jj}^T) \\ &\quad + 2\sum_{i\neq j=1}^{n} tr(\mathbf{B}_{ij}\mathbf{B}_{ij}^T) - \{tr(\mathbf{B})\}^2. \end{aligned}$$

Notice that

$$tr(\mathbf{B}\mathbf{B}^T) = \sum_{i=1}^{n} tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T) + \sum_{i\neq j=1}^{n} tr(\mathbf{B}_{ij}\mathbf{B}_{ij}^T),$$

$$\{tr(\mathbf{B})\}^2 = \sum_{i=1}^{n}\{tr(\mathbf{B}_{ii})\}^2 + \sum_{i\neq j=1}^{n} tr(\mathbf{B}_{ii})tr(\mathbf{B}_{jj}^T),$$
$$\{E(\boldsymbol{e}_i^T\mathbf{B}_{ii}\boldsymbol{e}_i)\}^2 = \{tr(\mathbf{B}_{ii})\}^2.$$

Some straightforward algebra yield

$$Var(\boldsymbol{e}^T\mathbf{B}\boldsymbol{e}) = 2tr(\mathbf{B}\mathbf{B}^T) + \sum_{i=1}^{n} Var(\boldsymbol{e}_i^T\mathbf{B}_{ii}\boldsymbol{e}_i) - 2\sum_{i=1}^{n} tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T).$$

Consider the singular value decomposition $\mathbf{B}_{ii} = \mathbf{U}_i\mathbf{D}_i\mathbf{V}_i^T$. Let $d_{ij}(\mathbf{B}_{ii})$ be the jth singular value, and $\mathbf{u}_{ij}, \mathbf{v}_{ij}$ be the jth column of $\mathbf{U}_i$ and $\mathbf{V}_i$, respectively, $j = 1, \cdots, n_i$. Define $z_{ij} = \mathbf{u}_{ij}^T\boldsymbol{e}_i\boldsymbol{e}_i^T\mathbf{v}_{ij}$, then by the condition of this lemma, we have that $Cov(z_{ij}, z_{ik}) \leq \{Var(z_{ij})Var(z_{ik})\}^{1/2} \leq K$. By some algebra, we have

$$\begin{aligned} Var(\boldsymbol{e}_i^T\mathbf{B}_{ii}\boldsymbol{e}_i) &= Var\Big\{\sum_{j=1}^{n_i} d_{ij}(\mathbf{B}_{ii})z_{ij}\Big\} \\ &= \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} d_{ij}(\mathbf{B}_{ii})d_{ik}(\mathbf{B}_{ii})Cov(z_{ij}, z_{ik}) \\ &\leq K\big\{\sum_{j=1}^{n_i} d_{ij}(\mathbf{B}_{ii})\big\}^2. \end{aligned}$$

Therefore, we get

$$Var(\boldsymbol{e}^T\mathbf{B}\boldsymbol{e}) \leq 2tr(\mathbf{B}\mathbf{B}^T) + K\big\{\sum_{j=1}^{n_i} d_{ij}(\mathbf{B}_{ii})\big\}^2 - 2\sum_{i=1}^{n} tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T).$$

Since $tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T) \geq 0$, (B.3) holds. □

Proof of Theorem 4.2.1. In light of (4.10) and (4.12), it suffices to show that

$$L(\mathbf{W}, \boldsymbol{\lambda}) - R(\mathbf{W}, \boldsymbol{\lambda}) = o_p(R(\mathbf{W}, \boldsymbol{\lambda})), \tag{B.4}$$

$$\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I} - \mathbf{A})^T\boldsymbol{\epsilon} = o_p(R(\mathbf{W}, \boldsymbol{\lambda})), \tag{B.5}$$

and

$$\frac{2}{n}\big\{\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon} - tr(\mathbf{A}\boldsymbol{\Sigma})\big\} = o_p(R(\mathbf{W},\boldsymbol{\lambda})) \tag{B.6}$$

because, combining (B.4)–(B.6), we have

$$U(\mathbf{W},\boldsymbol{\lambda}) - L(\mathbf{W},\boldsymbol{\lambda}) - \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\mathbf{W},\boldsymbol{\lambda})).$$

We first prove (B.4). By (4.9), we have

$$Var(L(\mathbf{W},\boldsymbol{\lambda})) = \frac{1}{n^2}Var\big\{\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon} - 2\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{A}\boldsymbol{\epsilon}\big\}. \tag{B.7}$$

Define $\mathbf{B} = \boldsymbol{\Sigma}^{1/2}\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma}^{1/2}$. Then $\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon} = (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon})^T\mathbf{B}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon})$. Since $\mathbf{B}$ is positive semi-definite, $\sum_{j=1}^{n_i} d_{ij}(\mathbf{B}_{ii}) = tr(\mathbf{B}_{ii})$. Under Condition 1, applying Lemma 4.5.5 with $\boldsymbol{e} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}$ and $\mathbf{B} = \boldsymbol{\Sigma}^{1/2}\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma}^{1/2}$, we obtain

$$\frac{1}{n^2}Var(\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon}) \le \frac{2}{n^2}tr(\mathbf{B}^2) + \frac{K}{n^2}\sum_{i=1}^{n}\{tr(\mathbf{B}_{ii})\}^2, \tag{B.8}$$

for some $K > 0$ as defined in lemma 4.5.5. By Lemma 4.5.3 and Lemma 4.5.4, under Condition 3, we have

$$\begin{aligned}\frac{2}{n^2}tr(\mathbf{B}^2) &\le \frac{2\lambda_{max}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)}{n^2}tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)\\ &\le \frac{2\xi(\boldsymbol{\Sigma},\mathbf{W}))}{n}\frac{1}{n}tr(\mathbf{A}\Sigma\mathbf{A}^T)\\ &= o(R^2(\mathbf{W},\boldsymbol{\lambda})).\end{aligned} \tag{B.9}$$

Define $\mathbf{C}_{ii}$ as the ith diagonal block of $\tilde{\mathbf{A}}^2$. Then, under Condition 2(ii), $tr(\mathbf{C}_{ii}) \sim$

$o(1)$. Thus,

$$\begin{aligned} tr(\mathbf{B}_{ii}) &= tr(\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}\mathbf{W}^{-1/2}\tilde{\mathbf{A}}\mathbf{W}\tilde{\mathbf{A}}\mathbf{W}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T) \\ &\leq \lambda_{max}(\mathbf{W})tr(\tilde{\mathbf{A}}\mathbf{W}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}\mathbf{W}^{-1/2}\tilde{\mathbf{A}}) \\ &= \lambda_{max}(\mathbf{W})tr(\mathbf{C}_{ii}\mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1/2}) \\ &\leq \lambda_{max}(\mathbf{W})\lambda_{max}(\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1})tr(\mathbf{C}_{ii}) \\ &= o(1)\xi(\boldsymbol{\Sigma},\mathbf{W}). \end{aligned} \tag{B.10}$$

Since $\sum_{i=1}^n\{tr(\mathbf{B}_{ii})\} = tr(\mathbf{B}) = tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$, under Condition 3,

$$\begin{aligned} \frac{K}{n^2}\sum_{i=1}^n\{tr(\mathbf{B}_{ii})\}^2 &= o(1)\frac{K\xi(\boldsymbol{\Sigma},\mathbf{W})tr(\mathbf{B})}{n^2} \\ &= o(1)\frac{K\xi(\boldsymbol{\Sigma},\mathbf{W})}{n}\frac{1}{n}tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T) \\ &= o(R^2(\mathbf{W},\boldsymbol{\lambda})). \end{aligned} \tag{B.11}$$

Combining (B.8)–(B.11), we obtain

$$\frac{1}{n^2}Var(\boldsymbol{\epsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\epsilon}) \sim o(R^2(\mathbf{W},\boldsymbol{\lambda})).$$

Since $\lambda_{max}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T) \leq \xi(\boldsymbol{\Sigma},\mathbf{W})$, by Lemma 4.5.4, under Condition 3, we have

$$\begin{aligned} \frac{1}{n^2}Var\big\{\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{A}\boldsymbol{\epsilon}\big\} &= \frac{1}{n^2}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\ &\leq \frac{1\xi(\boldsymbol{\Sigma},\mathbf{W})}{n}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\ &= o(R^2(\mathbf{W},\boldsymbol{\lambda})). \end{aligned} \tag{B.12}$$

Combining (B.7)-(B.12) and using the Cauchy–Schwarz inequality, we obtain that $Var(L(\mathbf{W},\boldsymbol{\lambda})) = o(R^2(\mathbf{W},\boldsymbol{\lambda}))$, which proves (B.4).

To show (B.5), by Lemma (4.5.4) and Condition 3, we have

$$\begin{aligned}\frac{1}{n^2}Var\big\{\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\boldsymbol{\epsilon}\big\} &= \frac{1}{n^2}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu}\\ &\le \frac{\lambda_{max}(\boldsymbol{\Sigma})}{n}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu}\\ &\le \frac{\xi(\boldsymbol{\Sigma},\mathbf{W})}{n}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu}\\ &= o(R^2(\mathbf{W},\boldsymbol{\lambda})).\end{aligned}$$

The result follows from an application of the Chebyshev inequality.

To show (B.6), applying Lemma (4.5.4) with $\boldsymbol{e}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}$ and $\mathbf{B}=\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}$, we obtain

$$\begin{aligned}\frac{2}{n^2}Var(\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon}) &= \frac{2}{n^2}Var(\boldsymbol{e}^T\mathbf{B}\boldsymbol{e})\\ &\le \frac{2}{n^2}tr(\mathbf{B}\mathbf{B}^T)+K\sum_{i=1}^{n}\Big\{\sum_{j=1}^{n_i}d_{ij}(\mathbf{B}_{ii})\Big\}^2,\end{aligned} \tag{B.13}$$

where K is as in Lemma 4.5.5. By Lemma 4.5.3, under Condition 3, we have

$$\begin{aligned}\frac{2}{n^2}tr(\mathbf{B}\mathbf{B}^T) &= \frac{2}{n^2}tr(\mathbf{A}^T\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}) \le \frac{2\lambda_{max}(\boldsymbol{\Sigma})}{n}\frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma})\\ &\le \frac{2\xi(\boldsymbol{\Sigma},\mathbf{W})}{n}\frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma}) = o(R^2(\mathbf{W},\boldsymbol{\lambda})).\end{aligned}$$

By the definition of $d_{ij}(\mathbf{B}_{ii})$, using Lemma 4.5.2 repeatedly, we have

$$\begin{aligned}d_{ij}^2(\mathbf{B}_{ii}) &= \lambda_{ij}(\mathbf{B}_{ii}^T\mathbf{B}_{ii}) = \lambda_{ij}(\mathbf{A}_{ii}^T\boldsymbol{\Sigma}_i\mathbf{A}_{ii}\boldsymbol{\Sigma}_i)\\ &= \lambda_{ij}(\mathbf{W}_i^{-1/2}\tilde{\mathbf{A}}_{ii}\mathbf{W}_i^{1/2}\boldsymbol{\Sigma}_i\mathbf{W}_i^{1/2}\tilde{\mathbf{A}}_{ii}\mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i)\\ &\le \lambda_{max}(\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1})\lambda_{max}(\mathbf{W}_i)\lambda_{max}(\boldsymbol{\Sigma}_i)\lambda_{ij}(\tilde{\mathbf{A}}_{ii}^2)\\ &\le \xi^2(\boldsymbol{\Sigma},\mathbf{W})\lambda_{ij}^2(\tilde{\mathbf{A}}_{ii}).\end{aligned}$$

Under Conditions 2(i), 3 and 4, we have

$$\begin{aligned}\frac{K}{n^2}\sum_{i=1}^{n}\Big\{\sum_{j=1}^{n_i} d_{ij}(\mathbf{B}_{ii})\Big\}^2 &\le \xi^2(\mathbf{\Sigma},\mathbf{W})\frac{K}{n^2}\sum_{i=1}^{n}\big\{tr(\tilde{\mathbf{A}}_{ii})\big\}^2 \\ &= \frac{K_2\xi(\mathbf{\Sigma},\mathbf{W})}{n}\xi(\mathbf{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) \\ &= o(R^2(\mathbf{W},\boldsymbol{\lambda})).\end{aligned} \tag{B.14}$$

Therefore, combining (B.13)–(B.14) and noticing Conditions 1–4, we have

$$\frac{1}{n^2}Var(\boldsymbol{\epsilon}^T\mathbf{A}\boldsymbol{\epsilon}) \sim o(R^2(\mathbf{W},\boldsymbol{\lambda})).$$

Apply the Chebyshev inequality to obtain (B.6). □

To prove Theorem 4.2.2, it is easier to prove Theorem 4.3.2 first. To prove Theorem 4.3.2, we need the following lemma.

Lemma 4.5.6. *Let* $\mathbf{D} = diag\{\mathbf{D}_{11},\cdots,\mathbf{D}_{nn}\}$ *where* $\mathbf{D}_{ii}$*'s are* $n_i \times n_i$ *matrices and* $\max_{1\le i\le n}\{tr(\mathbf{D}_{ii}\mathbf{W}_i\mathbf{D}_{ii}^T\} \sim \lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)$*. Under Conditions 1–5, we have*

$$\frac{1}{n^2}Var\big\{\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\mathbf{Y}\big\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})).$$

Proof. Using the decomposition $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, we obtain

$$\begin{aligned}&\frac{1}{n^2}Var\big\{\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\mathbf{Y}\big\}) \\ &\quad= \frac{1}{n^2}Var\big\{\boldsymbol{\epsilon}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon} + 2\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon}\big\}.\end{aligned}$$

By a simple application of the Cauchy–Schwarz inequality, it suffices to show

$$\frac{1}{n^2}Var\big\{\boldsymbol{\epsilon}^T(\mathbf{I}-\mathbf{A})\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon})\big\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})), \tag{B.15}$$

and

$$\frac{1}{n^2}Var\big\{2\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon}\big\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{B.16}$$

We shall show (B.15) first. Using Lemma 4.5.5 with $\boldsymbol{e} = \mathbf{\Sigma}^{-1/2}\boldsymbol{\epsilon}$ and $\mathbf{B} =$

$\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}$ to yield

$$\frac{1}{n^2}Var\big\{\boldsymbol{\epsilon}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon}\big\} \le \frac{2}{n^2}tr(\mathbf{B}\mathbf{B}^T) + \frac{K}{n^2}\sum_{i=1}^{n}\big\{\sum_{j=1}^{n_i} d_{ij}(\mathbf{B}_{ii})\big\}^2. \quad \text{(B.17)}$$

Repeatedly using Lemma 4.5.3 and 4.5.4, and the fact that $\lambda_{max}((\mathbf{I}-\tilde{\mathbf{A}})^2) \le 1$, we have

$$\begin{aligned} tr(\mathbf{B}\mathbf{B}^T) &= tr\big\{\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{A})^T\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\big\} \\ &\le \lambda_{max}\big\{(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{W}^{-1/2}\boldsymbol{\Sigma}\mathbf{W}^{-1/2}(\mathbf{I}-\tilde{\mathbf{A}})\big\} \\ &\qquad \times \lambda_{max}\big\{(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{A})^T\big\}tr(\mathbf{D}\mathbf{W}\mathbf{D}^T) \\ &\le \lambda^2_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})\lambda_{max}(\mathbf{W})tr(\mathbf{D}\mathbf{W}\mathbf{D}^T). \end{aligned}$$

Noticing $\max_{1\le i\le n}\{tr(\mathbf{D}_{ii}\mathbf{W}_i\mathbf{D}_{ii}^T)\} = \lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)$ and using Conditions 3–4, we have

$$\frac{2}{n^2}tr(\mathbf{B}\mathbf{B}^T) = \frac{2\xi^2(\boldsymbol{\Sigma},\mathbf{W})}{n}O(n^{-2}tr(\mathbf{A})^2) = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \quad \text{(B.18)}$$

Note that $\mathbf{B}_{ii} = \mathbf{L}_i\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T$. Thus,

$$\begin{aligned} tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T) &= tr\big\{\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T \\ &\qquad \mathbf{L}_i\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T\big\} \\ &\le \lambda_{max}\big\{(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{W}^{-1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}\mathbf{W}^{-1/2}(\mathbf{I}-\tilde{\mathbf{A}})\big\} \\ &\quad \times tr\big\{\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}\mathbf{W}\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T\big\} \\ &\le \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})tr\big\{\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}\mathbf{W}\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T\big\}. \end{aligned} \quad \text{(B.19)}$$

Let $\mathbf{D}^* = \mathbf{W}^{1/2}\mathbf{D}\mathbf{W}\mathbf{D}^T\mathbf{W}^{1/2}$ be the block diagonal matrix with diagonal blocks $\mathbf{D}^*_{ii}$ and $\mathbf{C}_{ii}$ be the the ith diagonal block of $\tilde{\mathbf{A}}^2$. We have

$$\begin{aligned} \lambda_{max}(\mathbf{D}^*) &\le \max_{1\le i\le n}\{tr(\mathbf{D}^*_{ii})\} \le \lambda_{max}(\mathbf{W})\max_{1\le i\le n}\{tr(\mathbf{D}_{ii}\mathbf{W}_i\mathbf{D}_{ii})\} \\ &= \lambda^2_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2). \end{aligned}$$

By Condition 2, we have $tr(\mathbf{C}_{ii}) = o(1)$. Then

$$\begin{aligned} & tr\big\{\mathbf{L}_i\boldsymbol{\Sigma}^{1/2}(\mathbf{I}-\mathbf{A})^T\mathbf{D}\mathbf{W}\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}^{1/2}\mathbf{L}_i^T\big\} \\ &= tr\big\{\boldsymbol{\Sigma}_i^{1/2}\mathbf{W}_i^{-1/2}(\mathbf{D}_{ii}^* - \tilde{\mathbf{A}}_{ii}\mathbf{D}_{ii}^* - \mathbf{D}_{ii}^*\tilde{\mathbf{A}}_{ii} + \lambda_{max}(\mathbf{D}^*)\mathbf{C}_{ii})\mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i^{1/2}\big\} \\ &\le \lambda_{max}(\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1})\big\{tr(\mathbf{D}^*) + \lambda_{max}(\mathbf{D}^*)tr(\mathbf{C}_{ii})\big\} - tr(\mathbf{M}_{ii}) \\ &= \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})\lambda_{max}^2(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) - tr(\mathbf{M}_{ii}), \end{aligned} \tag{B.20}$$

where $\mathbf{M}_{ii} = \boldsymbol{\Sigma}_i^{1/2}\mathbf{W}_i^{-1/2}(\tilde{\mathbf{A}}_{ii}\mathbf{D}_{ii}^* + \mathbf{D}_{ii}^*\tilde{\mathbf{A}}_{ii})\mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i^{1/2}$. Observe that

$$tr\big\{\boldsymbol{\Sigma}_i^{1/2}\mathbf{W}_i^{-1/2}(\tilde{\mathbf{A}}_{ii} - \mathbf{D}_{ii}^*)^2\mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i^{1/2}\big\} \ge 0.$$

Under Condition 2,

$$\begin{aligned} tr(\mathbf{M}_{ii}) &\le tr\big\{\boldsymbol{\Sigma}_i^{1/2}\mathbf{W}_i^{-1/2}(\tilde{\mathbf{A}}_{ii}^2 + \mathbf{D}_{ii}^{*2})\mathbf{W}_i^{-1/2}\boldsymbol{\Sigma}_i^{1/2}\big\} \\ &\le \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})tr(\tilde{\mathbf{A}}_{ii}^2 + \mathbf{D}_{ii}^{*2}) \\ &= \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})O(n^{-2}tr(\mathbf{A})^2)\big\{1 + \lambda_{max}^4(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)\big\}. \end{aligned}$$

Since $\mathbf{W}$ is the working covariance matrix, $\lambda_{max}(\mathbf{W}) = O(1)$ if n_i's are bounded. It follows that, under Condition 5, $\lambda_{max}^2(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) = o(1)$, which leads to

$$tr(\mathbf{M}_{ii}) = \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})\lambda_{max}^2(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2). \tag{B.21}$$

(B.19)–(B.21) together imply that $tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T) = \xi^2(\boldsymbol{\Sigma}, \mathbf{W})O(n^{-2}tr(\mathbf{A})^2)$. Under Conditions 3–4, by the Jensen inequality, we have

$$\begin{aligned} \frac{K}{n^2}\sum_{i=1}^{n}\Big\{\sum_{j=1}^{n_i} d_{ij}(\mathbf{B}_{ii})\Big\}^2 &\le \frac{K}{n^2}\sum_{i=1}^{n}\Big\{n_i\sum_{j=1}^{n_i} d_{ij}^2(\mathbf{B}_{ii})\Big\} = \frac{K}{n^2}\sum_{i=1}^{n}\big\{n_i tr(\mathbf{B}_{ii}\mathbf{B}_{ii}^T)\big\} \\ &= \frac{\xi^2(\boldsymbol{\Sigma}, \mathbf{W})}{n}O(n^{-2}tr(\mathbf{A})^2) = o(R^2(\mathbf{W}, \boldsymbol{\lambda})). \end{aligned}$$

Using this result, (B.17), and (B.18), we obtain (B.15).

To show (B.16), note that

$$\lambda_{max}(\mathbf{D}\mathbf{W}\mathbf{D}^T) \leq \max_{1\leq i\leq n}\{tr(\mathbf{D}_{ii}\mathbf{W}_i\mathbf{D}_{ii}^T)\} = \lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2).$$

Use Lemma 4.5.4 and Conditions 3–4 to yield

$$\begin{aligned}
&\frac{1}{n^2}Var\big\{2\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\epsilon}\big\} \\
&\quad= \frac{4}{n^2}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{A})^T\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\
&\quad\leq \frac{4\lambda_{max}\big\{(\mathbf{I}-\tilde{\mathbf{A}})\mathbf{W}^{-1/2}\boldsymbol{\Sigma}\mathbf{W}^{-1/2}(\mathbf{I}-\tilde{\mathbf{A}})\big\}}{n^2}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}\mathbf{W}\mathbf{D}^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\
&\quad\leq \frac{4\lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})}{n}\lambda_{max}(\mathbf{D}\mathbf{W}\mathbf{D}^T)\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\
&\quad= \frac{\xi(\boldsymbol{\Sigma},\mathbf{W}))O(n^{-2}tr(\mathbf{A})^2)}{n}\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\
&\quad= o(R^2(\mathbf{W},\boldsymbol{\lambda})),
\end{aligned}$$

which is the desired result. □

Proof of Theorem 4.3.2. By Theorem 4.2.1, it suffices to show that

$$\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda}) = o_p(R(\mathbf{W},\boldsymbol{\lambda})),$$

which can be obtained by showing

$$E\big\{\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda})\big\}^2 = o_p(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{B.22}$$

Hence, it suffices to show that

$$E\big\{\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda})\big\} = o(R(\mathbf{W},\boldsymbol{\lambda})) \tag{B.23}$$

and

$$Var\big\{\text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda})\big\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{B.24}$$

Denote $\mathbf{A}_d = diag\{\mathbf{A}_{11},\cdots,\mathbf{A}_{nn}\}$ and $\tilde{\mathbf{A}}_d = diag\{\tilde{\mathbf{A}}_{11},\cdots,\tilde{\mathbf{A}}_{nn}\}$. It follows

that $\tilde{\mathbf{A}}_d = \mathbf{W}^{-1/2}\mathbf{A}_d\mathbf{W}^{1/2}$ and $n^{-1}tr(\tilde{\mathbf{A}}_d^2) = O(n^{-2}tr(\mathbf{A})^2)$ by Condition 2. Some algebra yields that

$$\text{LsoCV*}(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda}) = \frac{2}{n}\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T\mathbf{A}_d(\mathbf{I}-\mathbf{A})\mathbf{Y} - \frac{2}{n}tr(\mathbf{A}\boldsymbol{\Sigma}).$$

First consider (B.23). We have that

$$\begin{aligned} &E\big\{\text{LsoCV*}(\mathbf{W},\boldsymbol{\lambda}) - U(\mathbf{W},\boldsymbol{\lambda})\big\} \\ &\quad= \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{A}_d+\mathbf{A}_d^T)(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} + \frac{1}{n}tr\big\{\mathbf{A}^T(\mathbf{A}_d+\mathbf{A}_d^T)\mathbf{A}\boldsymbol{\Sigma}\big\} \\ &\qquad- \frac{2}{n}tr(\mathbf{A}_d^T\mathbf{A}_d\boldsymbol{\Sigma}) - \frac{2}{n}tr(\mathbf{A}_d^2\boldsymbol{\Sigma}). \end{aligned} \tag{B.25}$$

We shall show that each term in (B.25) is of the order $o(R(\mathbf{W},\boldsymbol{\lambda}))$.

Condition 2 says that $\max_{1\le i\le n} tr(\tilde{\mathbf{A}}_{ii}) = O(n^{-1}tr(\mathbf{A})) = o(1)$. Using Conditions 2 and 5, we have

$$\begin{aligned} tr(\mathbf{A}_{ii}+\mathbf{A}_{ii}^T)^2 &= 2tr(\mathbf{A}_{ii}^2 + \mathbf{A}_{ii}\mathbf{A}_{ii}^T) \\ &= 2tr(\tilde{\mathbf{A}}_{ii}^2 + \tilde{\mathbf{A}}_{ii}\mathbf{W}_i\tilde{\mathbf{A}}_{ii}\mathbf{W}_i^{-1}) \\ &\le 2tr(\tilde{\mathbf{A}}_{ii}^2)\big\{1+\lambda_{max}(\mathbf{W}_i^{-1})\lambda_{max}(\mathbf{W}_i)\big\} \\ &= \lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})O(n^{-2}tr(\mathbf{A})^2) = o(1), \end{aligned}$$

which implies that all eigenvalues of $(\mathbf{A}_d+\mathbf{A}_d^T)$ are of order $o(1)$, and hence

$$\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{A}_d+\mathbf{A}_d^T)(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(1)\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(R(\mathbf{W},\boldsymbol{\lambda})),$$

$$\frac{1}{n}tr\big\{\mathbf{A}^T(\mathbf{A}_d+\mathbf{A}_d^T)\mathbf{A}\boldsymbol{\Sigma}\big\} = o(1)\frac{1}{n}tr(\mathbf{A}^T\mathbf{A}\boldsymbol{\Sigma}) = o(R(\mathbf{W},\boldsymbol{\lambda})).$$

Under Condition 4, the third term in (B.25) can be bounded as

$$\begin{aligned}\frac{1}{n}tr(\mathbf{A}_d^T\mathbf{A}_d\boldsymbol{\Sigma}) &\le \lambda_{max}(\boldsymbol{\Sigma}\mathbf{W}^{-1})\frac{1}{n}tr(\tilde{\mathbf{A}}_d\mathbf{W}\tilde{\mathbf{A}}_d)\\ &\le \xi(\boldsymbol{\Sigma},\mathbf{W})\frac{1}{n}tr(\tilde{\mathbf{A}}_d^2)\\ &= \xi(\boldsymbol{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)\\ &= o(R(\mathbf{W},\boldsymbol{\lambda})).\end{aligned} \tag{B.26}$$

For the last term in equation (B.25), observe that

$$\frac{2}{n}tr(\mathbf{A}_d^2\boldsymbol{\Sigma}) = \frac{2}{n}tr(\tilde{\mathbf{A}}_d^2\mathbf{W}^{-1/2}\boldsymbol{\Sigma}\mathbf{W}^{1/2}) = \frac{1}{n}tr\big\{\tilde{\mathbf{A}}_d^2(\boldsymbol{\Sigma}^* + \boldsymbol{\Sigma}^{*T})\big\},$$

where $\boldsymbol{\Sigma}^* = \mathbf{W}^{-1/2}\boldsymbol{\Sigma}\mathbf{W}^{1/2}$. Let $\boldsymbol{\Sigma}_i^*$ be the ith diagonal block of $\boldsymbol{\Sigma}^*$. We have

$$\begin{aligned}tr\big\{(\boldsymbol{\Sigma}_i^* + \boldsymbol{\Sigma}_i^{*T})^2\big\} &= 2tr(\boldsymbol{\Sigma}_i^2 + \boldsymbol{\Sigma}_i\mathbf{W}_i\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1})\\ &\le 2n_i\lambda_{max}^2(\boldsymbol{\Sigma}_i) + 2\lambda_{max}(\mathbf{W}_i)tr(\mathbf{W}_i^{-1}\boldsymbol{\Sigma}^2\mathbf{W}_i^{-1}\mathbf{W}_i)\\ &\le 2n_i\lambda_{max}^2(\boldsymbol{\Sigma}_i) + 2n_i\lambda_{max}^2(\boldsymbol{\Sigma}_i\mathbf{W}_i^{-1})\lambda_{max}^2(\mathbf{W}_i)\\ &\le 4n_i\xi^2(\boldsymbol{\Sigma},\mathbf{W}),\end{aligned}$$

which implies that $\pm\max_{1\le i\le n}\{2\sqrt{n_i}\}\xi(\boldsymbol{\Sigma},\mathbf{W})$ are upper and lower bounds of eigenvalues of $\boldsymbol{\Sigma}^* + \boldsymbol{\Sigma}^{*T}$. Hence, under Condition 4, one has

$$\begin{aligned}\frac{2}{n}tr(\mathbf{A}_d^2\boldsymbol{\Sigma}) &\le \max_{1\le i\le n}\{2\sqrt{n_i}\}\xi(\boldsymbol{\Sigma},\mathbf{W})\frac{1}{n}tr(\tilde{\mathbf{A}}_d^2)\\ &= \xi(\boldsymbol{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) = o(R(\mathbf{W},\boldsymbol{\lambda})).\end{aligned}$$

Therefore, (B.23) has been proved.

To prove (B.24), since

$$tr(\mathbf{A}_{ii}\mathbf{W}_i\mathbf{A}_{ii}^T) = tr(\tilde{\mathbf{A}}_{ii}^2\mathbf{W}_i) \le \lambda_{max}(\mathbf{W}_i)\big\{tr(\mathbf{A}_{ii})\big\}^2,$$

we have $\max_{1\le i\le n} tr(\mathbf{A}_{ii}\mathbf{W}_i\mathbf{A}_{ii}^T) = \lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)$ by Condition 2. Under Conditions 3–4, (B.24) follows from Lemma 4.5.6 with $\mathbf{D} = \mathbf{A}_d$. □

Proof of Theorem 4.2.2. By Theorem 4.3.2, it suffices to show

$$\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda}) = o_p(L(\mathbf{W},\boldsymbol{\lambda})),$$

which can be proved by showing that

$$E\big\{\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})\big\}^2 = o_p(R^2(\mathbf{W},\boldsymbol{\lambda})).$$

It suffices to show

$$E\big\{\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})\big\} = o(R(\mathbf{W},\boldsymbol{\lambda})) \tag{B.27}$$

and

$$Var\big\{\text{LsoCV}(\mathbf{W},\boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W},\boldsymbol{\lambda})\big\} = o(R^2(\mathbf{W},\boldsymbol{\lambda})). \tag{B.28}$$

For each $i = 1,\ldots,n$, consider the eigen-decomposition $\tilde{\mathbf{A}}_{ii} = \mathbf{P}_i\boldsymbol{\Lambda}_i\mathbf{P}_i^T$, where $\mathbf{P}_i$ is a $n_i \times n_i$ orthogonal matrix and $\boldsymbol{\Lambda}_i = diag\{\lambda_{i1},\cdots,\lambda_{in_i}\}$, $\lambda_{ij} \geq 0$. Using this decomposition, we have

$$(\mathbf{I}_{ii} - \mathbf{A}_{ii})^{-1} = \mathbf{W}_i^{1/2}\mathbf{P}_i\boldsymbol{\Lambda}_i^*\mathbf{P}_i^T\mathbf{W}^{-1/2},$$

where $\boldsymbol{\Lambda}_i^*$ is a diagonal matrix with diagonal elements $(1-\lambda_{ij})^{-1}$, $j = 1,\cdots,n_i$. Since under Condition 2, $\max_{1\leq j\leq n_i}\{\lambda_{ij}\} \sim o(1)$, we have $(1-\lambda_{ij})^{-1} = \sum_{k=0}^{\infty}\lambda_{ij}^k$, which leads to

$$(\mathbf{I}_{ii} - \tilde{\mathbf{A}}_{ii})^{-1} = \sum_{k=0}^{\infty}\mathbf{P}_i\boldsymbol{\Lambda}_i^k\mathbf{P}_i^T = \sum_{k=0}^{\infty}\tilde{\mathbf{A}}_{ii}^k.$$

Define $\tilde{\mathbf{D}}^{(m)} = diag\{\tilde{\mathbf{D}}_{11}^{(m)},\cdots,\tilde{\mathbf{D}}_{nn}^{(m)}\}$, $m = 1,2$, where $\tilde{\mathbf{D}}_{ii}^{(1)} = \sum_{k=1}^{\infty}\tilde{\mathbf{A}}_{ii}^k$, and $\tilde{\mathbf{D}}_{ii}^{(2)} = \sum_{k=2}^{\infty}\tilde{\mathbf{A}}_{ii}^k$, $i = 1,\cdots,n$. It follows that, for each i and $m = 1,2$,

$$tr(\tilde{\mathbf{D}}_{ii}^{(m)}) = \sum_{k=m}^{\infty}tr(\tilde{\mathbf{A}}_{ii}^k) \leq \sum_{k=m}^{\infty}\big\{tr(\tilde{\mathbf{A}}_{ii})\big\}^k = \frac{\big\{tr(\tilde{\mathbf{A}}_{ii})\big\}^m}{1-tr(\tilde{\mathbf{A}}_{ii})}.$$

Since Condition 2(i) gives $\max_{1\le i\le n} tr(\mathbf{A}_{ii}) \sim O(n^{-1}tr(\mathbf{A}))$, we obtain that

$$\max_{1\le i\le n} tr(\tilde{\mathbf{D}}_{ii}^{(m)}) = O(n^{-m}tr(\mathbf{A})^m), \quad m = 1, 2. \tag{B.29}$$

Some algebra yields

$$\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda}) = \frac{1}{n}\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{D}^{(1)} + \mathbf{D}^{(2)})(\mathbf{I}-\mathbf{A})\mathbf{Y}$$

where $\mathbf{D}^{(1)} = \mathbf{W}^{-1/2}\tilde{\mathbf{D}}^{(1)}\mathbf{W}\tilde{\mathbf{D}}^{(1)}\mathbf{W}^{-1/2}$ and $\mathbf{D}^{(2)} = \mathbf{W}^{1/2}\tilde{\mathbf{D}}^{(2)}\mathbf{W}^{-1/2}$.

To show (B.27), note that

$$\begin{aligned} &E\big\{\text{LsoCV}(\mathbf{W}, \boldsymbol{\lambda}) - \text{LsoCV}^*(\mathbf{W}, \boldsymbol{\lambda})\big\} \\ &\quad= \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(1)}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} + \frac{1}{n}tr\big\{(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(1)}(\mathbf{I}-\mathbf{A})\Sigma\big\} \\ &\qquad+ \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} + \frac{1}{n}tr\big\{(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\Sigma\big\}. \end{aligned} \tag{B.30}$$

Using Lemma 4.5.2 and 4.5.3 repeatedly and noticing Condition 5, we have

$$\lambda_{max}(\mathbf{D}^{(1)}) \le \lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})O(n^{-2}tr(\mathbf{A})^2) = o(1).$$

Thus, the first terms (B.30) can be bounded as

$$\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(1)}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(1)\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(R(\mathbf{W}, \boldsymbol{\lambda})).$$

Using Condition 4 and (B.29), the second term of (B.30) can be bounded as

$$\begin{aligned} \frac{1}{n}tr\big\{(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(1)}(\mathbf{I}-\mathbf{A})\Sigma\big\} &\le \xi(\boldsymbol{\Sigma}, \mathbf{W})\frac{1}{n}tr(\tilde{\mathbf{D}}^{(1)2}) \\ &= \xi(\boldsymbol{\Sigma}, \mathbf{W})O(n^{-2}tr(\mathbf{A})^2) \\ &= o(R(\mathbf{W}, \boldsymbol{\lambda})). \end{aligned}$$

Now consider the third term in (B.30). Under Condition 5 and (B.29),

$$\begin{aligned} tr\big\{(\mathbf{D}_{ii}^{(2)}+\mathbf{D}_{ii}^{(2)T})^2\big\} &= 2tr(\tilde{\mathbf{D}}_{ii}^{(2)2}) + 2tr(\mathbf{D}_{ii}^{(2)}\mathbf{D}_{ii}^{(2)T}) \\ &= 2tr(\tilde{\mathbf{D}}_{ii}^{(2)2}) + 2tr(\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i^{-1}\tilde{\mathbf{D}}_{ii}^{(2)}\mathbf{W}_i) \\ &\le 2tr(\tilde{\mathbf{D}}_{ii}^{(2)2}) + 2\lambda_{max}(\mathbf{W}_i^{-1})\lambda_{max}(\mathbf{W}_i)tr(\tilde{\mathbf{D}}_{ii}^{(2)2}) \\ &= o(n^{-2}tr(\mathbf{A})^2), \end{aligned} \tag{B.31}$$

which implies that all eigenvalues of $\mathbf{D}_{ii}^{(2)}+\mathbf{D}_{ii}^{(2)T}$ are of the order $O(n^{-1}tr(\mathbf{A}))$, and thus $o(1)$. Then, under Conditions 1–5, we have

$$\begin{aligned} \frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} &= \frac{1}{2n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{D}^{(2)}+\mathbf{D}^{(2)T})(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} \\ &= o(1)\frac{1}{n}\boldsymbol{\mu}^T(\mathbf{I}-\mathbf{A})^T(\mathbf{I}-\mathbf{A})\boldsymbol{\mu} = o(R(\mathbf{W},\boldsymbol{\lambda})). \end{aligned}$$

To study the the fourth term in (B.30), define $\boldsymbol{\Sigma}^\dagger = \mathbf{W}^{-1/2}(\mathbf{I}-\mathbf{A}_d)\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{A}_d)^T\mathbf{W}^{1/2}$, where $\mathbf{A}_d$ is as defined in the proof of Theorem 4.3.2. Then

$$\begin{aligned} \frac{1}{n}tr\big\{(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\boldsymbol{\Sigma}\big\} &= \frac{1}{2n}tr\big\{\tilde{\mathbf{D}}^{(2)}(\boldsymbol{\Sigma}^\dagger+\boldsymbol{\Sigma}^{\dagger T})\big\} \\ &\quad + \frac{1}{n}tr\big\{\mathbf{D}^{(2)}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T - \mathbf{A}_d\boldsymbol{\Sigma}\mathbf{A}_d^T)\big\}. \end{aligned} \tag{B.32}$$

Let $\boldsymbol{\Sigma}_i^\dagger$ be the ith diagonal block of $\boldsymbol{\Sigma}^\dagger$. Using Lemma 4.5.4 to obtain

$$\begin{aligned} tr\big\{(\boldsymbol{\Sigma}_i^\dagger+\boldsymbol{\Sigma}_i^{\dagger T})^2\big\} &\le 2n_i\lambda_{max}^2\big\{(\mathbf{I}_{ii}-\mathbf{A}_{ii})\boldsymbol{\Sigma}_i(\mathbf{I}_{ii}-\mathbf{A}_{ii})^T\big\} + 2n_i\xi^2(\boldsymbol{\Sigma},\mathbf{W}) \\ &\le 4n_i\xi^2(\boldsymbol{\Sigma},\mathbf{W}), \end{aligned}$$

which means that $\pm\max_{1\le i\le n}\{2\sqrt{n_i}\}\xi(\boldsymbol{\Sigma},\mathbf{W})$ are the lower and upper bounds of eigenvalues of $\boldsymbol{\Sigma}^\dagger+\boldsymbol{\Sigma}^{\dagger T}$. Hence, application of Condition 4 and (B.29) gives

$$\begin{aligned} \frac{1}{2n}tr\big\{\tilde{\mathbf{D}}^{(2)}(\boldsymbol{\Sigma}^\dagger+\boldsymbol{\Sigma}^{\dagger T})\big\} &\le \max\{\sqrt{n_i}\}\xi(\boldsymbol{\Sigma},\mathbf{W})\frac{1}{n}tr(\tilde{\mathbf{D}}^{(2)}) \\ &= \xi(\boldsymbol{\Sigma},\mathbf{W})O(n^{-2}tr(\mathbf{A})^2) = o(R(\mathbf{W},\boldsymbol{\lambda})). \end{aligned} \tag{B.33}$$

It has been shown in (B.26) that $tr(\mathbf{A}_d\Sigma\mathbf{A}_d^T) = o(R(\mathbf{W},\boldsymbol{\lambda}))$. Using Lemma 4.5.3 and

(A.1), we have

$$\begin{aligned}\frac{1}{n}tr\big\{\mathbf{D}^{(2)}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T-\mathbf{A}_d\boldsymbol{\Sigma}\mathbf{A}_d^T)\big\} &= o(1)tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T+\mathbf{A}_d\boldsymbol{\Sigma}\mathbf{A}_d^T)\\ &= o(R(\mathbf{W},\boldsymbol{\lambda})).\end{aligned} \tag{B.34}$$

Using (B.32), (B.33) and (B.34), we have shown that the fourth term of (B.30) satisfies

$$\frac{1}{n}tr[(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(2)}(\mathbf{I}-\mathbf{A})\Sigma] = o(R(\mathbf{W},\boldsymbol{\lambda})).$$

Therefore, (B.27) has been proved.

Next, we proceed to prove (B.28). Since under Condition 5, we have

$$\begin{aligned}tr(\mathbf{D}_{ii}^{(1)}\mathbf{W}_i\mathbf{D}_{ii}^{(1)T}) &\le \lambda_{max}^2(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})\lambda_{max}^2(\tilde{\mathbf{D}}_{ii}^{(1)})tr(\tilde{\mathbf{D}}_{ii}^{(1)2})\\ &= \big\{\lambda_{max}(\mathbf{W})\lambda_{max}(\mathbf{W}^{-1})O(n^{-2}tr(\mathbf{A})^2)\big\}\lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)\\ &= \lambda_{max}(\mathbf{W})O(n^{-2}tr(\mathbf{A})^2)\end{aligned}$$

and

$$\begin{aligned}tr(\mathbf{D}_{ii}^{(2)}\mathbf{W}_i\mathbf{D}_{ii}^{(2)T}) &\le \lambda_{max}(\mathbf{W})tr(\tilde{\mathbf{D}}_{ii}^{(2)2}) = \lambda_{max}(\mathbf{W})O(n^{-4}tr(\mathbf{A})^4)\\ &= \lambda_{max}(\mathbf{W})o(n^{-2}tr(\mathbf{A})^2).\end{aligned}$$

By applying Lemma 4.5.6 with $\mathbf{D}=\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ respectively, we have

$$\frac{1}{n^2}Var\big\{\mathbf{Y}^T(\mathbf{I}-\mathbf{A})^T\mathbf{D}^{(m)}(\mathbf{I}-\mathbf{A})\mathbf{Y}\big\} = o_p(R^2(\mathbf{W},\boldsymbol{\lambda})),\ m=1,2,$$

and (B.28) follows by the Cauchy–Schwarz inequality. □

VITA

Ganggang Xu received his B.S. in Statistics in July 2002 from Zhejiang University, P.R.China. In August 2008, he received his M.S. in statistics from Texas A&M University, College Station. In December 2011, he received his Ph.D. in Statistics from Texas A&M University, College Station. The current research interests of his lie in a broad range of Statistics including nonparametric statistics, penalized methods, model selection, model averaging, machine learning, spatial statistics, measurement error, etc. His address is: Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143.

The typist for this thesis was Ganggang Xu.