The Pennsylvania State University
The Graduate School

STATISTICAL METHODS FOR DIFFERENT ULTRAHIGH DIMENSIONAL MODELS

A Dissertation in
Statistics
by
Jingyuan Liu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2013

The dissertation of Jingyuan Liu was reviewed and approved* by the following:

Rongling Wu
Professor of Public Health Sciences and Statistics
Dissertation Advisor, Chair of Committee

Runze Li
Distinguished Professor of Statistics and Professor of Public Health Sciences
Dissertation Advisor, Co-chair of Committee

Dennis K. J. Lin
Distinguished Professor of Statistics and Supply Chain

Vernon M. Chinchilli
Distinguished Professor of Public Health Sciences and Statistics

Liwang Cui
Professor of Entomology

David Hunter
Professor of Statistics, Department Head

*Signatures are on file in the Graduate School.

Abstract

This thesis studies feature screening and variable selection procedures for ultra-high dimensional varying coefficient models and partially linear models, and the extension of the methods to longitudinal data structure.

A new independence screening procedure is proposed for varying coefficient models based on the conditional correlation between each predictor and the response given the depending covariate (CCIS, for short). We establish and prove the ranking consistency and sure screening property of CCIS, and demonstrate them empirically through simulations. Furthermore, the iterative screening procedure (ICCIS) is developed to enhance the finite sample performance. In the Framingham Heart Study (FHS) example, we derive a new two-stage approach to select significant Single-nucleotide polymorphism (SNPs) for explaining body mass index (BMI), and the effect of SNPs may depend on the baseline age of patients. Firstly CCIS is applied to reduce the ultrahigh dimensionality to the scale under sample size, and secondly several penalized regression techniques are modified for varying coefficient models to further select important variables as well as estimate the coefficient functions.

Moreover, CCIS for varying coefficient models can be extended for the longitudinal data structure. Consider the time-varying coefficient model as an example, where multiple response values are observed for every subject. We apply CCIS in the first stage to the pooled sample, in which we treat all the observations as independent individuals, although those from the same subject are actually correlated. In this case, the within subject correlation is ignored in the screening stage. However, the simulation studies show that we do not lose ranking consistency and sure screening property by doing this. In the real data example, we use a modified two-stage approach to restudy the effect of SNPs on BMI using FHS data. The dynamic pattern of age instead of baseline age is considered to illustrate the longitudinal structure. If the efficiency of coefficient function estimators are of interest,

we can add one more step of a weighted least squared method after the variable selection stage, by incorporating the covariance matrix estimation procedure.

For partially linear models, another independence screening procedure is developed in this thesis based on the partial residual method (PRSIS, for short). The partially linear model can be converted to a linear model with transformed response and predictors, and then the traditional screening methods for linear models can be applied, such as sure independence screening (SIS, Fan and Lv, 2008). The desired theoretical properties are demonstrated through simulation studies. Soybean data analysis are provided to illustrate the two-stage approach based on PRSIS, using which the important markers are selected for explaining the dry biomass of soybean.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to gratefully and sincerely thank my thesis advisors, Dr. Rongling Wu and Runze Li, for their assistance and guidance in getting my research career started on the right track. Their encouragement helps me a lot to build up my self-confidence and to become an independent thinker. Their mentorship was paramount in providing a well rounded experience consistent with my long-term career goals. Most importantly, the friendship with them built during my graduate studies at Pennsylvania State University is a precious treasure for me.

I would like to thank the Department of Statistics, and the members of my doctoral committee, Dr. Dennis Lin, Dr. Vernon Chinchilli and Dr. Liwang Cui for their input, valuable discussions and accessibility. Without their knowledge and assistance this study would not have been successful. Additionally, I am very grateful to all of my classmates, with whom I worked together, puzzled over many problems, and spent an awesome five years.

Finally, I would like to thank my parents, Mr. Liangwei Liu and Ms. Lihua Yang for their encouragement and unwavering love. Their understanding and tolerance are the testament of their unyielding devotion, and it supports me to a significant extent all the time.

This thesis research is supported by National Institute on Drug Abuse (NIDA) grants P50-DA10075 and R01-CA168676, NSF grant IOS 0923975, and National Natural Science Foundation of China grant 11028103.

Chapter 1

Introduction

1.1 A Brief Overview of Variable Selection and High Dimensional Models

Variable selection is widely used to identify the underlying model structure when a large number of predictors are introduced at the initial stage of modeling but only a few of them are truly relevant to the response. Various techniques are well developed in the literature to select significant variables, including classical variable selections and penalized least squares methods. For classical variable selections, many criteria can be used to select the best subset of the full model, e.g. Adjusted R2, PRESS (Allen, 1974), Mallow's Cp (Mallows, 1973), AIC (Akaike, 1974), BIC (Schwarz, 1978), GIC (Nishii, 1984), among others. Although the sampling properties are well developed and studied in the literature, the classical variable selection criteria are not widely used in the modern scientific world due to its computational cost. Instead, the penalized regression methods, by which one can simultaneously select significant variables and estimate the coefficients, have gained a lot of popularity. Researchers have proposed a variety of penalty functions, such as the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Adaptive LASSO (Zou, 2006), MCP (Zhang, 2010), etc. Some of them are shown to possess nice theoretical properties such as sparsity, continuity, unbiasedness, consistency and oracle property.

However, the standard variable selection techniques may fail for high or ultra-

high dimensional data analysis, which tends to be increasingly frequent and important due to the rapid development of data collecting technique. For instance, genome-wide association studies (GWAS), which explore the genetic effect on certain phenotypes, have attracted great attention recently. Hundreds of thousands of single-nucleotide polymorphisms (SNPs) are genotyped, leading to ultrahigh dimensional data analysis, although it is often the case that only a small number of SNPs are truly associated with the phenotype of interest. This motivates the researchers to develop new statistical methods for ultrahigh dimensional data, which contain loosely two categories. One category comprises two stages, where we first reduce ultrahigh dimensionality to moderate dimensionality, and then apply a regularization method such as the penalized regression. For the first stage, Fan and Lv (2008) showed that Sure Independence Screening procedure (SIS) possesses sure screening property for linear models; Wang(2009) proposed the forward regression in the same model setting; Hall and Miller (2009) extended the linear model to nonlinear model using generalized empirical correlation learning; Fan and Song (2010) explored SIS in generalized linear model by ranking the maximum marginal likelihood or its estimate; Fan, Feng and Song (2011) explored the feature screening technique for the ultrahigh dimensional additive model, by ranking the magnitude of spline approximations for nonparametric components; Zhu, Li, Li and Zhu (2011) proposed a sure independence ranking and screening (SIRS) procedure to select important predictors in the multi-index model; Li, Zhong and Zhu (2012) studied a model-free sure independence screening procedure based on the distance correlation (DC-SIS), which can also be used directly to screen grouped predictor variables and multivariate response variables, among many other screening techniques. One issue with the aforementioned model-based screening methods is that the validity of screening procedures indeed depend on the model form, in other words, the screening results are no longer reliable if the underlying model structure is misspecified. The other category aims to simultaneously identify and estimate the underlying model structure through a single step of regularized regression. The examples are Dantzig Selector (Candes and Tao, 2007) and SCAD on high dimensions (Kim, Choi, and Oh, 2008).

1.2 Ultrahigh Dimensional Varying Coefficient Models and Partially Linear Models

Ultrahigh dimensional varying coefficient models have become great attraction to researchers as useful extension of linear models. They allow the number of predictors to be much larger than the sample size, and the corresponding regression coefficients to change over different subjects characterized by certain covariate. For example, in genetic research, one might be interested in the effect of singlenucleotide polymorphism (SNP) on body mass index (BMI), and the effect may depend on age of each individual. In this case, millions of SNPs are considered, leading to an ultrahigh dimensional problem. To guarantee the changing effect of SNPs, we need to allow the coefficients of SNPs to vary with age. More specifically, we can consider the coefficients as functions of age. However, the diverging dimensionality of the predictors and their changing effect make the ultrahigh dimensional varying coefficient models challenging to analyze. Therefore, in this paper, we propose a novel feature screening method specifically for these models to reduce dimensionality, and a two-stage approach based on the screening technique is introduced to select important predictors and depict their effect.

Some variable selection methods have been developed for varying coefficient models in literature. Li and Liang (2008) used a generalized likelihood ratio test to select significant nonparametric components based on SCAD penalty (Fan and Li, 2001); Wang et al. (2008) presented a regularized estimation procedure based on the basis function approximations and the SCAD penalty, which can simultaneously select significant variables and estimate the nonzero smooth coefficient functions; Wang and Xia (2009) proposed a shrinkage method incorporating local polynomial smoothing (Fan and Gijbels, 1996) and LASSO penalized regression (Tibshirani, 1996). Nevertheless, the existing techniques for varying coefficient models require fixed model dimension, thus they cannot be applied to ultrahigh dimensional cases.

To deal with ultrahigh dimensionality, as previously discussed, the two-stage approach with screening procedure largely depends on model specification, thus the existing methods are not applicable for the varying coefficient model setting. Therefore, we are motivated to develop a feature screening method specifically for

ultrahigh dimensional varying coefficient models.

We construct an conditioning-correlation independent screening (CCIS). The reason is that varying coefficient models are indeed linear models conditioning on u. More specifically, we rank the importance of predictors using the conditional correlation estimates between each predictor and the response given u. We define the conditional correlation parallel to the pearson correlation in the linear model setting, except that the expectation and variance are now substituted by the corresponding conditional expectation and conditional variance. Therefore, the problem of estimating conditional correlation is transformed to estimating several conditional means by nonparametric smoothing techniques.

Several desirable theoretical properties of CCIS are also systematically studied. We show that CCIS possesses the ranking consistency property (Zhu, Li, Li, and Zhu, 2011), which means with probability tending to 1, the important predictors rank before the unimportant ones. In addition, CCIS satisfies sure screening property (Fan and Lv, 2008) for varying coefficient models under mild technical conditions, which guarantees the probability that the model chosen by CCIS includes the true model tends to 1 as the sample size goes to infinity. Monte Carlo simulation studies are conducted to empirically verify these theoretical advantages, and the results indicate that CCIS significantly outperforms SIS under the varying coefficient model setting.

Furthermore, the varying coefficient model with longitudinal data structure is also studied. If the depending covariate u is a time vector instead of a scaler, a similar approach with CCIS can be applied by ignoring the within-subject correlation in the screening stage. The simulation results empirically show that by doing this we do not lose the desirable theoretical properties.

Another useful extension to linear model is called partially linear model, where the response is assumed to depend on certain variable in a nonparametric form, aside from the linear dependency of other variables. The estimation procedures and variable selection techniques have been well studied in literature. See Chen (1988), Engle et.al (1986), Heckman (1986), Speckman (1988), Robinson (1988), and Fan and Huang (2005) for details. However, due to the same issue as varying coefficient models, ultrahigh dimensional partially linear model remains an open area. We advocate a novel approach, called PRSIS, to reduce the dimensionality

based on the idea of partial residual method. Both Monte Carlo simulation results and the real data analysis are demonstrated.

At last, we consider the generalized varying coefficient model in the discussion part, where the response is binary or count instead of continuous data type. Conditional correlation is now substituted by conditional likelihood as a screening score. Simulation results illustrate the validity of the methods.

1.3 Organization of the Dissertation

The rest of the dissertation is organized as follows. In Chapter 2, we give a detailed review of the existing methods in literature about four foregoing topics: the traditional variable selection criteria, high or ultrahigh dimensional data analysis, the statistical methods for varying coefficient models, and the estimation methods for partially linear models. In Chapter 3, we develop the feature screening procedure CCIS for varying coefficient models, with Monte Carlo simulations to illustrate the performance of it and Framingham Heart Study data analysis to show its application. In addition, we show that the finite sample performance is improved by adopting iterative feature screening procedure. In Chapter 4, two theoretical properties, ranking consistency and sure screening property, are proved. The longitudinal version of ultrahigh dimensional varying coefficient model is studied in Chapter 5. And PRSIS, the screening method for ultrahigh dimensional partially linear models are presented in Chapter 6. In Chapter 7, we summarize the research in this thesis and discuss the possible extension of our methods to generalized varying coefficient models.

Chapter 2

Literature Review

As mentioned before, we mainly focus on three topics. Thus we divide the review of literature into four parts: First we briefly introduce several standard variable selection methods for linear models, ranging from classical variable selection to the penalized regression approach; Second, we discuss how to extend the variable selection technique to the high or ultrahigh dimensions; Third, we study the statistical methods used in the varying coefficient model analysis; At last, we review the estimation procedures for partially linear models.

2.1 Variable Selection for Linear Models

Variable selection is a fundamental technique of finding important explanatory factors for predicting the response. To reduce the model bias, we tend to introduce large amount of potential predictors at first, most of which, nevertheless, are redundant and cause problems in model interpretation, computation, prediction, etc. Therefore, how to find a parsimonious model which contains only a few predictors but still gives a good fit becomes one of the most important tasks in the statistical field.

In this section, we mainly focus on the variable selection for the linear regression model. Suppose that $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, is a random sample from the linear model:

$$y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon, \tag{2.1}$$

where $y \in \mathbb{R}^1$ is the response, $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ is the d-dimensional predictor, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}} \in \mathbb{R}^d$ is the d-dimensional coefficient vector, and $\varepsilon \in \mathbb{R}^1$ is the independently and identically distributed (i.i.d.) random noise with mean zero.

For the matrix form, denote $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, then the linear model (2.1) is reexpressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.2}$$

2.1.1 Classical Variable Selection Criteria

The classical variable selection aims to find the best subset of the full model using some criteria computed from each submodel. There is large amount of literature on the topic of subset selection. See Miller (2002) for details.

1. *Mallows's C_p and AIC.*

   The C_p statistic (Mallows, 1973) is defined as

   $$C_p = \frac{RSS_p}{\sigma^2} - (n - 2p).$$

   When σ^2 is unknown, it is replaced by the residual mean squares under the full model $\mathcal{F}$ in practice:

   $$\hat{\sigma}^2 = \frac{RSS_F}{n - d}, \tag{2.3}$$

   An equivalent criteria with Mallow's C_p is called An Information Criterion (AIC, Akaike, 1974), which is developed based on the Kullback-Leibler distance. For linear regression models with least square estimators, AIC is defined by

   $$AIC_p = RSS_p + 2p\sigma^2, \tag{2.4}$$

   where σ^2 is unbiasly estimated by (2.3). Thus, AIC_p is equivalent to Mallow's C_p in the linear model setting.

2. *BIC.*

   For linear models, Bayesian Information Criterion (BIC, Schwarz, 1978) is to minimize

   $$BIC_p = RSS_p + \log(n)p\sigma^2.$$

   Compared with AIC in (2.4), BIC assigns more weight to the degree of freedom of the submodel as long as $\log(n) > 2$, thus it tends to choose smaller models than AIC does. And when the sample size goes to infinity, BIC is able to determine the true model, while AIC often overfits the model.

3. *Other variable selection criteria.*

   There are many other variable selection criteria similar with AIC and BIC, summarized by the Generalized Information Criterion (GIC, Nishii, 1984)

   $$GIC_p = RSS_p + \kappa_n p\sigma^2.$$

   For example, $\kappa_n = 2$ yields AIC, while $\kappa_n = \log(n)$ indicates BIC. Hannan and Quinn (1979) advocated the ϕ-criterion by taking $\kappa_n = c\log(\log(n))$, and the Risk Inflation Criterion (RIC, Foster and George, 1994) requires $\kappa_n = 2\log(d)$.

Based on these criteria, several algorithms for subset selection can be applied to choose the optimal submodel. The first, called the *best subset selection* where all the 2^d possible subsets of the full model are considered, is quite computationally expensive. The second one is the *forward regression*, where we start from the intercept-only model and add one variable with the largest F-value at a time, until none of the variables is significant in terms of F-statistics. Therefore, this method requires at most d steps. The third one is the *backward elimination*, where on the contrary, we start from the full model and delete one variable with the smallest F-value at a time until all the remaining are significant. For these two methods, once a variable is added or deleted, it cannot be reconsidered. To address this issue, the fourth algorithm, called the *stepwise regression*, is advocated, which requires iterative realizations of the forward regression and the backward elimination.

2.1.2 Variable Selection via Penalized Least Squares

The subset selection procedures based on the classical criteria admit nice sampling properties (Barron, Birge and Massart, 1999). However, the computation of them is infeasible when d is large, and they ignore stochastic errors during the selection process. To solve these problems, penalized regression approaches are proposed, where the variable selection is accomplished through the coefficient estimation. In this section, we mainly focus on the penalized least squares problem (PLS), which coincides with the penalized likelihood in linear model setting.

Consider the penalized least square function

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.5}$$

where $p_\lambda(\cdot)$ is the penalty function and $\lambda \geq 0$ is a tuning parameter controlling the model complexity. By minimizing (2.5), we can simultaneously select variables and estimate their associated regression coefficients, i.e. the coefficient of insignificant variables are automatically estimated to be 0.

2.1.2.1 Penalty Functions

One natural question arises with PLS problem: what kind of penalty functions we should use? Fan and Li (2001) advocated three desirable properties of a PLS estimator:

1) *Unbiasedness:* The estimator is nearly unbiased for the truly large coefficients, to reduce model bias.

2) *Sparsity:* The estimator automatically sets small estimated coefficients to zero, to reduce model complexity.

3) *Continuity:* The estimator is continuous in the data, in order to guarantee the model prediction to be stable.

Moreover, Antoniadis and Fan (2001) advocated three conditions to guarantee the above properties for a penalty function $p_\lambda(t)$:

1) *Approximate Unbiasedness* if $p'_\lambda(t) = 0$ for large t;

2) *Sparsity* if $\min_{t\geq 0}\{t + p'_\lambda(t)\} > 0$;

3) *Continuity* if and only if $\operatorname{argmin}_{t\geq 0}\{t + p'_\lambda(t)\} = 0$.

Besides, the penalty function $p_\lambda(t)$ is required to be nondecreasing, continuously differentiable on $[0, \infty)$, and singular at the origin (i.e., $p'_\lambda(0+) > 0$). The function $-t - p'_\lambda(t)$ is strictly unimodal on $(0, \infty)$. Some widely used penalty functions are listed below:

1. *L_0 or Entropy penalty:*

$$p_\lambda(|t|) = \frac{\lambda^2}{2} I(|t| \neq 0). \tag{2.6}$$

   Since $\sum_{j=1}^{d} I(|\beta_j| \neq 0) = p$ in linear model setting, this penalty leads to the classical variable selection criteria by assigning different values to λ, e.g., when $\lambda = \sigma\sqrt{2/n}$, (2.6) yields AIC which is equivalent to Mallow's C_p for linear regression models; If $\lambda = \sigma\sqrt{\log(n)/n}$, this penalty gives BIC; RIC is obtained by taking $\lambda = \sigma\sqrt{\log(d)/n}$. However, L_0 penalty is not even continuous, thus does not possess the aforementioned properties.

2. *Hard thresholding penalty:*

$$p_\lambda(|t|) = \lambda^2 - (\lambda - |t|)_+^2, \tag{2.7}$$

   Hard thresholding penalty (Antonialdis, 1996) is smoother than the Entropy penalty, but results in the same estimates when the design matrix is orthonormal, that is, the best subset selection coincides with the hard thresholding rule for orthonormal designs, which is illustrated in Figure 2.2.

3. *L_2 penalty:*

$$p_\lambda(|t|) = \frac{\lambda}{2}|t|^2. \tag{2.8}$$

   L_2 penalty yields the ridge regression $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + n\lambda I_d)^{-1}\mathbf{X}^T\mathbf{y}$ (Hoerl and Kennard, 1970), which can be applied to deal with collinearity in the predictors. The advantage of this estimator lies in its easy implementation and the

explicit form of its solution. Yet ridge regression does not produce sparse estimators, thus it cannot be used for variable selection. Besides, the resulting estimator $\hat{\boldsymbol{\beta}}$ is biased.

4. *L_1 penalty:*

$$p_\lambda(|t|) = \lambda|t|. \tag{2.9}$$

The L_1 penalty yields the Leat Absolute Shrinkage and Selection Operator (LASSO, Tibshirani, 1996). With this penalty, the penalized least square function $Q(\boldsymbol{\beta})$ in (2.5) becomes

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\sum_{j=1}^{d}|\beta_j|. \tag{2.10}$$

Minimizing $Q(\boldsymbol{\beta})$ above is equivalent to minimizing the RSS subject to the constraint $\sum_{j=1}^{d}|\beta_j| < s$, by which the model size is controlled and the sparsity is guaranteed. When the design matrix is orthonormal, the solution coincides with the soft thresholding rule (Donoho and Johnstone, 1994; Donoho, Jognstone, Kerkyacharian and Picard, 1995). Yet the drawback of LASSO estimator is that it equally penalizes all the coefficients, resulting in the biasness of large coefficients.

5. *L_q penalty:*

$$p_\lambda(|t|) = \frac{\lambda}{q}|t|^q. \tag{2.11}$$

The L_q penalty, $0 \leq q \leq 2$ contains the L_2, L_1 and L_0 penalties as special cases, and produces bridge regression estimates (Frank and Friedman, 1993). If $q < 1$, the solution is sparse but not continuous, while the opposite occurs for $q > 1$. When $q = 1$ which results in LASSO estimates, sparsity and continuity can be simultaneously satisfied, yet not the unbiasness.

6. *SCAD penalty:*

The SCAD penalty is proposed by Fan and Li (2001) with the form

$$\begin{aligned} p_\lambda(|t|) &= \lambda|t|I(|t|<\lambda) + \frac{a\lambda|t| - (|t|^2+\lambda^2)/2}{a-1} I(\lambda \le |t| < a\lambda) \\ &\quad + \frac{(a+1)\lambda^2}{2} I(|t| > a\lambda) \end{aligned} \tag{2.12}$$

or

$$p'_\lambda(|t|) = \lambda I(|t| \le \lambda) + \frac{(a\lambda - |t|)_+ I(|t| > \lambda)}{a-1}, \tag{2.13}$$

where $a > 2$, and often, $a = 3.7$ based on a Baysian argument. The SCAD penalty possesses all the three properties, i.e. unbiasedness, sparsity and continuity. Furthermore, Fan and Li (2001) proved the SCAD penalized estimator satisfies the oracle property. That is, the procedure works as well as if the true model is known in asymptotic sense.

7. *Adaptive LASSO penalty:*

   The form of Adaptive LASSO penalty (Zou, 2006) is the same as L_1 penalty, but it adopts different tuning parameters for different β_j's. More explicitly, the penalized loss function in (2.5) becomes

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{d} \omega_j |\beta_j|, \tag{2.14}$$

   where ω_j is used to assign different weights for the coefficients, aiming to penalize small coefficients more than large coefficients. Specifically, Adaptive LASSO chooses $\omega_j = 1/|\hat{\beta}_j^0|^\gamma$, where $\gamma > 0$ and $\hat{\beta}_j^0$ is a $\sqrt{n}$-consistent estimate of β_j such as the least square estimate. The adaptive LASSO is a favorable alternative for LASSO since it is shown to possess the oracle property.

8. *Group LASSO penalty:*

   Group LASSO (Yuan and Lin, 2006) is a useful extension to LASSO when the potential predictors are grouped in advance, such as the dummy variables in analysis-of-variance (ANOVA) problems, and basis functions of nonparametric models, among others. The penalty possesses the same form with

LASSO, with $|\beta_j|$ replaced by the Euclidean norm $\|\boldsymbol{\beta}_j\|$, where $\boldsymbol{\beta}_j$ is the jth group of predictors. That is, the penalized loss function (2.5) is modified as

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{d} \|\boldsymbol{\beta}_j\|, \tag{2.15}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)^{\mathrm{T}}$. Therefore, based on this penalty, we select groups of variables instead of single variables.

9. *MCP:*

   The minimax concave penalty (MCP) is advocated by Zhang (2010) and defined by

$$p_\lambda(|t|) = \lambda(|t| - |t|^2/2a\lambda)I(|t| < a\lambda) + \frac{a\lambda^2}{2} I(|t| \geq a\lambda) \tag{2.16}$$

   where $a > 0$. The MCP shares similar spirit with the SCAD penalty, including the aforementioned three properties and the oracle property.

There are many other penalized regression techniques in literature, e.g., the Elastic Net (Zou and Hastie, 2005) which linearly combines L_1 and L_2 penalties to reduce bias while keep sparsity, and the Nonnegative Garotte (Breiman, 1995) which is closely related to adaptive LASSO, etc. Some penalties with their first order derivatives are depicted in Figure 2.1.

Figure 2.2 illustrates the relationship between the penalized least square estimates (PLSE) and the ordinary least square estimates (OLSE) for some commonly used penalties when the design matrix $\mathbf{X}$ is orthonormal. Since the OLSE is unbiased, we expect the large coefficient estimates by PLS equal or close to OLSE to achieve unbiasness; The sparsity requires small PLSE to be estimated as 0; And we also want the PLSE to be continuous for the sake of stability. From the plot, only SCAD penalty guarantees these three properties simultaneously. Generally, among all the penalties, nonconvex penalties are often more favorable since they are more likely to satisfy such properties.

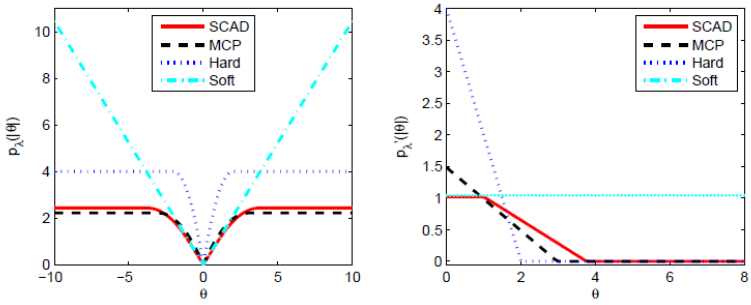


Figure 2.1. The left panel consists of the penalty functions of hard thresholding penalty, L_1 penalty, SCAD penalty and MCP, and the right panel includes their derivatives. (Fan and Lv, 2009)

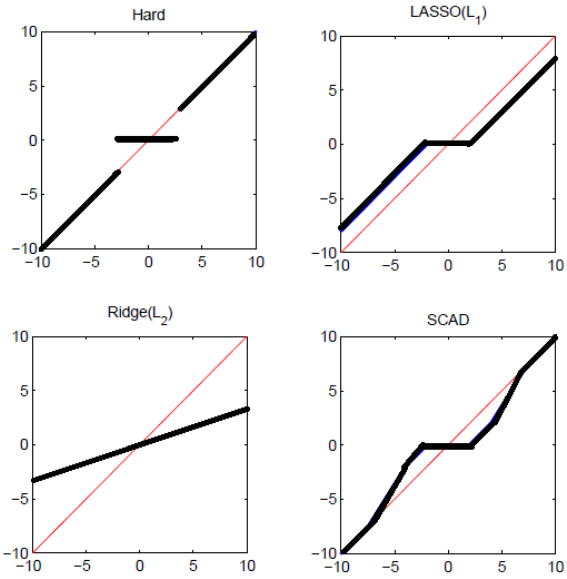


Figure 2.2. The relationship between PLSE and OLSE when the design matrix is orthonormal, where the x-axis is OLSE, and the y-axis is PLSE.

2.1.2.2 Computation and Implementation of PLS

Most PLS problems with penalties satisfying the aforementioned three properties are nonconvex and difficult to optimize directly. Nevertheless, we show in this section that they can be approximated by some convex functions, thus the nonconvex problem can be solved via convex optimization algorithms.

1. *Local Quadratic Approximation (LQA):*

Fan and Li (2001) proposed a unified local quadratic approximation (LQA) algorithm for optimizing nonconvex penalized least squares, the idea of which is to locally and iteratively approximate $Q(\boldsymbol{\beta})$ in (2.5) by a quadratic function. Since the first term of $Q(\boldsymbol{\beta})$ is already convex, we only need to consider the nonconvex penalty function $p_\lambda(|\beta_j|)$. Suppose the current value of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_d^0)^T$, then the penalty function can be approximated by

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^0|) + \frac{1}{2}\frac{p'_\lambda(\beta_j^0)}{|\beta_j^0|}(\beta_j^2 - (\beta_j^0)^2) \quad \text{for } \beta_j \approx \beta_j^0. \quad (2.17)$$

Then the minimization problem of (2.5) is reduced to a quadratic optimization program which can be solved by iteratively computing ridge regression

$$\hat{\boldsymbol{\beta}}_1 = \{\mathbf{X}^{\mathrm{T}}\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}^0)\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \quad (2.18)$$

where $\Sigma_\lambda(\beta^{\mathbf{0}}) = \mathrm{diag}\{p'_{j,\lambda}(|\beta_1^0|)/|\beta_1^0|, \ldots, p'_{j,\lambda}(|\beta_d^0|)/|\beta_d^0|\}$. Considering that the ridge regression cannot select significant variables automatically, we set the estimated coefficient $\hat{\beta}_j = 0$ if it is very close to 0.

Furthermore, LQA can be extended to any smooth loss function, denoted by $l(\boldsymbol{\beta})$, other than least square loss, where the objective function becomes

$$l(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_{j,\lambda}(|\beta_j|), \quad (2.19)$$

In this case, we need first locally approximate $l(\boldsymbol{\beta})$ by the quadratic function

$$\begin{aligned} &l(\boldsymbol{\beta}^0) + \nabla l(\boldsymbol{\beta}^0)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^{\mathrm{T}}\nabla^2 l(\boldsymbol{\beta}^0)(\boldsymbol{\beta} - \boldsymbol{\beta}^0) \\ &+\frac{1}{2}n\boldsymbol{\beta}^{\mathrm{T}}\Sigma_\lambda(\boldsymbol{\beta}^0)\boldsymbol{\beta}, \end{aligned} \quad (2.20)$$

then apply Newton-Raphson algorithm to get the minimizer of (2.19). Specifically, the LQA algorithm for (2.19) is described as follows:

1) Set the initial value $\boldsymbol{\beta}^0$ of $\hat{\boldsymbol{\beta}}$, e.g. unpenalized least square estimate.
2) For β_j close to β_j^0, and β_j^0 is not close to 0, approximate the objective function (2.19) based on (2.20) and (2.17).

3) Apply the Newton-Raphson algorithm to the quadratic function obtained in 2) to update β_j^0, and delete a variable x_j when $|\beta_j^0| < \epsilon$. Often $\epsilon = 0.5 \times \textit{Standard Error}$.

4) Iterate between 2) and 3) until convergence.

The LQA algorithm converges in a quadratic rate, which is the same as that of the modified EM (Lange, 1995).

2. *Local Linear Approximation (LLA)*:

A better approximation to the nonconvex penalty functions is the local linear approximation (LLA, Zou and Li, 2008):

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^0|) + p'_\lambda(|\beta_j^0|)(|\beta_j| - |\beta_j^0|) \quad \text{for } \beta_j \approx \beta_j^0, \qquad (2.21)$$

since it is the tightest convex majorant of the concave function on $[0, \infty)$. Figure 2.3 demonstrates the approximations to SCAD penalties with LQA and LLA.

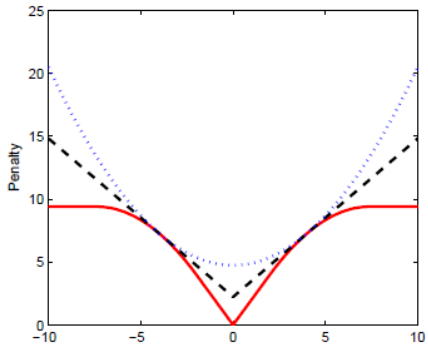


Figure 2.3. LQA (dotted) and LLA (dashed) to the SCAD penalty. (Fan and Lv, 2009)

With LLA, the nonconvex PLS can be transformed to an iteratively weighted penalized L_1 regression, which can be solved by the quadratic programming (Osborne, Presnell and Turlach, 2000), or a fast and efficient Least Angle Regression (LARS) algorithm (Efron et al., 2004). The idea of LARS is that it starts from a large λ that selects only one predictor having the largest correlation r_1 with the response, and decreases λ until the second predictor

is selected, where the selected variable has the same correlation r_1 with the current working residual, and so on so forth. It is shown to generate the same piecewise linear solution path in λ as LASSO, and can be modified for Adaptive LASSO.

There are many other powerful algorithms for the PLS problems, e.g., Fu (1998), Daubechies, Defrise and De Mol (2004), and Wu and Lang (2008) proposed a coordinate descent algorithm (CD) that can be applied to the Group LASSO, penalized likelihood, etc.; Kim, Choi, and Oh (2008) developed a DC-CCCP algorithm for SCAD penalty on high dimensions, which will be reviewed in next section; Zhang (2009) advocated the PLUS algorithm for computing a solution path for the PLS with quadratic spline penalties such as SCAD and MCP; The Iterative Coordinate Ascent algorithm (ICA, Fan and Lv, 2009) and the Iterative Conditional Maximization algorithm (ICM, Zhang and Li, 2009) were introduced to deal with general nonconvex penalties, among others.

2.1.2.3 Choice of Tuning Parameters for PLS

The choice of tuning parameters plays a substantial role of controlling the model complexity in PLS. For instance, if we set $\lambda = 0$, the PLS is reduced to OLS and all the variables are selected into the model; If $\lambda = \infty$, the penalty part of (2.5) becomes ∞ and none of the variables is selected. To choose a best λ between these two extreme cases, the classical criteria introduced in section 2.1.1 are often used. Specifically, for a given λ, compute the PLSE $\hat{\boldsymbol{\beta}}_\lambda$ of the coefficients based on a penalty function in section 2.1.2.1 and an algorithm in section 2.1.2.2, then calculate the value of certain selector using $\hat{\boldsymbol{\beta}}_\lambda$ based on the classical variable selection criteria. See details in Craven and Wahba (1979), Fan and Li (2001), Li et al. (2006), and Wang et al. (2007b). Some of the selectors are listed below:

AIC selector: $AIC(\lambda) = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2 + 2df_\lambda\hat{\sigma}^2$.

BIC selector: $BIC(\lambda) = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2 + \log(n)df_\lambda\hat{\sigma}^2$.

GCV selector: $GCV(\lambda) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2/(1 - df_\lambda/n)^2$.

In the selectors above, the degree of freedom of the estimated model df_λ is generally defined by

$$df_\lambda = tr\left(\mathbf{X}_\lambda(\mathbf{X}_\lambda^{\mathrm{T}}\mathbf{X}_\lambda + n\Sigma_\lambda)^{-1}\mathbf{X}_\lambda^{\mathrm{T}}\right),$$

where $\mathbf{X}_\lambda$ is the design matrix of the model corresponding to a given λ, and $\Sigma_\lambda = \mathrm{diag}_{\hat{\beta}_{j,\lambda}\neq 0}\{p'_\lambda(|\hat{\beta}_{j,\lambda}|)/|\hat{\beta}_{j,\lambda}|\}$, with $\hat{\beta}_{j,\lambda}$ being the jth component of the PLSE $\hat{\boldsymbol{\beta}}_\lambda$. Another way to model the degree of freedom is simply using the number of nonzero predictors

$$df_\lambda = \sum_{i=1}^{d} I(\hat{\beta}_{j,\lambda} \neq 0).$$

Therefore, for a series of grid points $(\lambda_1, \ldots, \lambda_M)$ and a selector, we can obtain M values of the selector, and the optimal λ is chosen to minimize them.

2.2 Statistical Methods for High or Ultrahigh Dimensions

High or ultrahigh dimensional data analysis has gained much popularity in the modern scientific field with the development of data collecting technology, ranging from genome-wide association studies (GWAS) to economics and finance. In this thesis, by high dimension we refer to the dimensionality of covariates $p = O(n^\alpha)$ for some $\alpha > 0$, while the ultrahigh dimension means $p = O(\exp(an))$ for some $a > 0$. And from this section on, the letter p no longer stands for the submodel size from the full d-dimensional model as in the previous section, rather it becomes the high or ultrahigh dimension which is far larger than the sample size n. And we use d as the moderate dimensionality, often $d = o(n)$.

When the dimension of predictors are much larger than the sample size, the traditional statistical procedures are challenged in terms of statistical accuracy, model interpretability and computational complexity, and the penalized regression methods introduced in last section may fail for high or ultrahigh dimensional data analysis. To address these issues, recall that the existing literature can be loosely divided into two categories: One is the two-stage approach, and the other aims to simultaneously identify and estimate the underlying model structure through a single step of regularized regression.

2.2.1 Independent Screening Procedures

In this section, we focus on the first step of the two-stage approach, where independent screening procedures are used to screen the ultrahigh dimensional data, by ranking features of predictors according to their marginal utilities.

2.2.1.1 Sure Independence Screening (SIS)

Fan and Lv (2008) proposed a sure screening method based on a correlation learning, called the Sure Independence Screening (SIS), and developed its sure screening property in linear model setting.

Consider the linear model (2.1), but now the dimension of $\mathbf{x}$ becomes $p >> n$, leading to ultrahigh dimensionality. The goal of SIS is to reduce dimensionality p from a huge scale (e.g. $\exp(O(n^{\xi}))$ for some $\xi > 0$) to a moderate scale d (e.g. $o(n)$) by a fast and efficient method described below.

Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^{\mathrm{T}}$ be the p-vector obtained by the componentwise regression

$$\boldsymbol{\omega} = \mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{2.22}$$

For the sake of simplicity, assume all predictors are standardized to have sample mean 0 and standard deviation 1, hence $\boldsymbol{\omega}$ is really a vector of marginal correlations of predictors with the response variable. SIS defines a submodel

$$\mathcal{M}_{\gamma} = \{1 \leq j \leq p : |\omega_j| \text{ is among the first } [\gamma n] \text{ largest of all}\} \tag{2.23}$$

for any given $\gamma \in (0, 1)$. Then the full model $\{1, \ldots, p\}$ is shrunken down to the submodel $\mathcal{M}_{\gamma}$ with size $d = [\gamma n] < n$ by ranking the marginal correlations $\omega_1, \ldots, \omega_p$.

SIS is shown to possess the *sure screening property* that the submodel chosen by the aforementioned method contains the true model with probability tending to one. More precisely, let $\mathcal{M}_* = \{1 \leq j \leq p : \beta_j \neq 0\}$ be the true sparse model with nonsparsity rate $s = |\mathcal{M}_*|$, and $\log p = O(n^{\xi})$ for some $\xi > 0$. Suppose the following regularity conditions are satisfied:

C1. Denote $\mathbf{z}_i = \Sigma^{-1/2}\mathbf{x}_i$ and $\mathbf{Z} = \mathbf{X}\Sigma^{-1/2}$, where $\Sigma = \mathrm{cov}(\mathbf{x}_i)$. Then $\mathbf{Z}$ has a

spherically symmetric distribution and satisfies the concentration property, i.e., there exist some c, $c_1 > 1$ and $C_1 > 0$ such that

$$P(\lambda_{\max}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}) > c_1 \text{ and } \lambda_{\min}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}) < 1/c_1) \leq e^{-C_1 n}$$

holds for any $n \times \tilde{p}$ submatrix $\tilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $cn < \tilde{p} \leq p$, where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ are the largest and smallest eigenvalues of a matrix, respectively.

C2. $\mathrm{var}(y) = O(1)$ and for some $\kappa \geq 0$ and c_2, $c_3 > 0$,

$$\min_{j \in \mathcal{M}_*} |\beta_j| \geq \frac{c_2}{n^\kappa} \text{ and } \min_{j \in \mathcal{M}_*} |\mathrm{cov}(\beta_j^{-1} y, x_j)| \geq c_3.$$

C3. There are some $\tau \geq 0$ and $c_4 > 0$ such that $\lambda_{\max}(\Sigma) \leq c_4 n^\tau$, i.e., the strong collinearity is ruled out.

C4. The random noise $\varepsilon \sim N(0, \sigma^2)$ for some $\sigma > 0$.

Theorem 1. *(sure screening property) Under the regularity conditions, and if $2\kappa + \tau < 1$, then there exists some $\theta < 1 - 2\kappa - \tau$ such that when $\gamma \sim cn^{-\theta}$ with $c > 0$, we have*

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O(\exp(-Cn^{1-2\kappa}/\log n))$$

for some $C > 0$.

The sure screening property implies $P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1$ as $n \to \infty$.

Therefore, it is reasonable to work on $\mathcal{M}_\gamma$ to further select significant variables and identify the underlying model structure in the second step of the two-stage approach. Specifically, based on the submodel

$$\mathbf{y} = \mathbf{X}_{\mathcal{M}_\gamma} \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$, becomes d-dimensional, all the well-developed moderate dimensional techniques can be applied, such as SCAD (Fan and Li, 2001), adaptive LASSO (Zou, 2006), among others. The simulation results show that SIS-SCAD performs the best and generates much smaller and more accurate models.

Remark: SIS does have some drawbacks, for instance, some unimportant predictors are selected due to its strong correlation with other important predictors, or an important predictor might be missed because it is only jointly correlated with the response but not marginally correlated, or the issue of collinearity, etc. To overcome these problems, an Iterative SIS (ISIS) is proposed to select the d predictors in l steps, where in each step, we use the regression residual from the model selected in the previous iteration as the new response, and conduct SIS. All the SIS based variable selection methods can be then modified by ISIS based variable selection.

2.2.1.2 Aggressive Betting Using SIS

The SIS requires strong regularity conditions, for instance, the design matrix should satisfy the concentration property, the smallest nonzero component of the signal should be greater than a threshold, the random noise is assumed normality, etc. To relax the conditions, Xue and Zou (2011) introduced a new method called Aggressive Betting (AB) using the SIS for both sparse noiseless signal recovery and sparse recovery with noise. They showed that AB possesses the exact recovery property for the sparse noiseless signal recovery, and enjoys the sure screening property for the contaminated linear system when applied together with robust compressed sensing (Candes et al., 2006).

First consider the underdetermined linear equations system without noise

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1},$$

where $p >> n$ and $\mathbf{X}$ is referred to as the sensing matrix. The sparsest but computationally infeasible solution is

$$\min \|\boldsymbol{\beta}\|_{L_0} \quad \text{subject to} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta}. \tag{2.24}$$

Xue and Zou (2011) advocated the Aggressive Betting (AB) based on SIS, which yields the equivalent solution to (2.24) and is computationally efficient. The algorithm is divided into two steps:

Step1. Define $\boldsymbol{\omega}$ as in (2.22) and find the index set $\mathcal{M}_1$ by letting $\gamma = 1$ in (2.23).

Then we obtain a new $n \times n$ sensing matrix $\mathbf{X}_{\mathcal{M}_1}$ and the new corresponding coefficient vector $\boldsymbol{\beta}_{\mathcal{M}_1}$, where $\mathbf{X}_{\mathcal{M}_1}$ is invertible almost surely if $\mathbf{X}$ is a random sensing matrix (Candes and Tao, 2005). Thus, we can rewrite the linear system as

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{\mathcal{M}_1}\boldsymbol{\beta}_{\mathcal{M}_1}.$$

Step2. Compute $\mathbf{Z} = \mathbf{X}_{\mathcal{M}_1}^{-1}\mathbf{y}$, then under mild conditions, $\mathbf{Z}$ is exactly equal to $\boldsymbol{\beta}_{\mathcal{M}_1}$, and $\boldsymbol{\beta}_{\mathcal{M}_1^c} = \mathbf{0}$. This is called the *Exact Recovery Property* of AB.

In statistics, instead of the foregoing noiseless system, a contaminated linear system is often considered:

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}. \tag{2.25}$$

Candes et al. (2006) advocated the sparsest but computationally expensive solution, called Robust Compressed Sensing, given by

$$\min \|\boldsymbol{\beta}\|_{L_1}, \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{L_2} \leq \nu, \tag{2.26}$$

where ν denotes the size of the error term $\boldsymbol{\varepsilon}$. Xue and Zou (2011) proposed a new computationally efficient method for stable signal recovery from model (2.25) in the following two steps:

Step1. (*Aggressive Betting*)
Implement the first step in AB to get the new model

$$\mathbf{y} = \mathbf{X}_{\mathcal{M}_1}\boldsymbol{\beta}_{\mathcal{M}_1} + \boldsymbol{\varepsilon}. \tag{2.27}$$

Step2. (*Robust Compressed Sensing*)
Apply the robust compressed sensing (2.26) to the reduced model (2.27).

The authors showed that with high probability $\mathcal{M}_1$ is a secure bet such that $\mathcal{M}_* \subset \mathcal{M}_1$, where $\mathcal{M}_* = \{1 \leq i \leq p : \beta_i \neq 0\}$ is the truly relevant index set. In other words, the AB possesses the sure screening property. To establish this

property under more relaxed conditions compared to SIS, first they introduced two definitions.

Def1. A random variable z is sub-Gaussian if $E(\exp(tz)) \leq \exp(ct^2)$ for some $c > 0$. (Buldygin and Kozachenko, 1980).

Def2. An $n \times p$ matrix $\mathbf{X}$ is a sub-Gaussian sensing matrix with a scale factor σ if its entries are independent random variables with mean 0 and variance ν and $E(\exp(tx_{ij})) \leq \exp(t^2\sigma^2/2)$.

Sub-Gaussian sensing matrix comprises a broad class, such as Bernoulli and Gaussian sensing matrices. The mean and variance can be achieved by transformation. The sure screening property is described as follows:

Theorem 2. *(Sure Screening Property) Suppose* $\mathbf{X}$ *is a sub-Gaussian sensing matrix with factor* σ *and* $E(x_{ij}^2) = n^{-1}$*, where* $\sigma^2 = dn^{-1}$ *and* $d > 2^{-7/2}$*. Let* $\boldsymbol{\varepsilon}$ *be a vector of sub-Gaussian errors with a common scale factor* σ_ε *and* $\sigma_\varepsilon^2 = d_\varepsilon n^{-1}$ *for some* $d_\varepsilon > 2^{-7/2}$*. Define* $\kappa = d_\varepsilon d^{-1}(\sum_{j\in\mathcal{M}_*}\beta_j^2)^{-1}$*. Then for any constant* $\delta \in (0, 0.5)$*,*

$$P(\mathcal{M}_* \subset \mathcal{M}_1) \geq 1 - se^{-c_1 n} - 2pe^{-c_2 c_3 n/s},$$

where $s = |\mathcal{M}_*|$*,* $c_1 = 2^{-9/2}d^{-2}(\delta - 0.5)^2$*,* $c_2 = 0.25[\{1 + 4\delta^2 d^{-2}(1+\kappa)^{-1}\}^{1/2} - 1]$*, and* $c_3 = (\min_{j\in\mathcal{M}_*}\beta_j^2)(\sum_{j\in\mathcal{M}_*}\beta_j^2/s)^{-1}$*.*

Remark: Compared with SIS, AB relaxes conditions significantly and gains more theoretical insight. First, the design matrix for the SIS is required to satisfy a concentration property which is not easy to verify in practice, and the rows are i.i.d. random vectors; But the AB only requires the entries of $\mathbf{X}$ to be independent but not identically distributed, and they may be non-Gaussian. Second, if $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are both Gaussian, κ becomes the noise-to-signal ratio. And by simple calculations, $n >> (8\kappa + 10)c_3^{-1} s \log p$ is sufficient to ensure the sure screening property holds with high probability, which provides some insight into the simulation findings in the SIS.

2.2.1.3 Screening Based on Forward Regression (FR)

Wang (2009) proposed another popular and classical variable screening procedure for ultrahigh dimensional linear models based on the Forward Regression (FR), by

which all truly relevant predictors are consistently included in the submodel. The FR screening procedure starts form an empty model, and updates the models by selecting one predictor at a time with the smallest residual sum of squares defined below, until the number of predictors reaches the sample size n. Among the n sequentially selected models, the BIC criterion (Chen and Chen, 2008) is applied to choose the best model for further variable selection.

More explicitly, consider the linear model (2.1) with ultrahigh dimensionality p and standardized $\mathbf{x}$. For an arbitrary candidate model $\mathcal{M} = \{j_1, \ldots, j_{\mathcal{M}}\}$, denote $\mathbf{X}_{(\mathcal{M})}$ to be the subdesign matrix corresponding to $\mathcal{M}$, and $\mathcal{F}$ to be the full model. The FR algorithm is implemented in the following three steps:

1) Set the initial model $\mathcal{M}^{(0)} = \emptyset$;

2) For the mth step, construct a series of candidate models $\mathcal{M}_j^* = \mathcal{M}^{(m-1)} \bigcup \{j\}$ for all $j \in \mathcal{F} \backslash \mathcal{M}^{(m-1)}$, and compute the residual sum of squares $RSS_j^* = \mathbf{y}^{\mathrm{T}}(I_n - H_j^*)\mathbf{y}$, where the projection matrix $H_j^* = \mathbf{X}_{\mathcal{M}_j^*}(\mathbf{X}_{\mathcal{M}_j^*}^{\mathrm{T}}\mathbf{X}_{\mathcal{M}_j^*})^{-1}\mathbf{X}_{\mathcal{M}_j^*}$ and $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix. Choose the model with the smallest RSS_j^*, denoted by $\mathcal{M}^{(m)}$.

3) Iterate 2) for n times, and obtain the solution path $\mathbb{S}$ consisting of n nested candidate models, i.e. $\mathbb{S} = \{\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(n)}\}$.

4) For each of the n models in $\mathbb{S}$, compute the corresponding BIC score (Chen and Chen, 2008):

$$BIC(\mathcal{M}) = \log(RSS_{\mathcal{M}}) + n^{-1}|\mathcal{M}|(\log n + 2\log p)$$

where $RSS_{\mathcal{M}}$ is the RSS for model $\mathcal{M}$, and $|\mathcal{M}|$ is the model size of $\mathcal{M}$. The final model is selected among $\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(n)}$ to minimize $BIC(\mathcal{M})$, denoted by $\mathcal{M}_\gamma$.

As to the theoretical properties, Wang (2009) proved the screening consistency of the solution path $\mathbb{S}$ and the sure screening property of the BIC based selection. Specifically, denote $\mathcal{M}_*$ to be the true model as in the SIS, then under some regularity conditions, we have $P(\mathcal{M}_* \subset \mathcal{M}^{(m)} \in \mathbb{S} \text{ for some } 1 \leq m \leq n) \to 1$, and $P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1$.

2.2.1.4 Sure Independence Screening for Generalized Linear Models

The SIS (Fan and Lv, 2008) and the FR technique (Wang, 2009) provide a novel track for dimensionality reduction through fast and efficient screening procedures in linear model setting. However, the method cannot be directly used in the case of discrete covariates such as GWAS or ANOVA problems, where a Generalized Linear Regression Model (GLIM) is appropriate. Therefore, a natural question is how to extend the SIS to GLIM. Fan and Song (2009) proposed a general version of screening procedure by ranking the maximum marginal likelihood or its estimates, which is suitable for GLIM.

Consider the GLIM with the canonical link. The conditional density function is given by

$$f(y|\mathbf{x}) = \exp\{y\theta(\mathbf{x}) - b(\theta(\mathbf{x})) + c(y)\},$$

where $b(\cdot)$ and $c(\cdot)$ are known functions, and $\theta(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$. The $p+1$ dimensional predictor $\mathbf{x}$ contains as intercept in the first column, and all the other columns are standardized without loss of generality. We take the dispersion parameter $\phi = 1$ for simplicity, since we are only interested in the conditional mean function of y, i.e.,

$$E(y|\mathbf{x}) = b'(\theta(\mathbf{x})) = b'(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}).$$

Fan and Song (2009) defined the maximum marginal likelihood estimator (MMLE) $\hat{\boldsymbol{\beta}}_j^M$ as the minimizer of the componentwise regression

$$\hat{\boldsymbol{\beta}}_j^M = (\hat{\boldsymbol{\beta}}_{j,0}^M, \hat{\boldsymbol{\beta}}_j^M) = \mathrm{argmin}_{\beta_0,\beta_j} \sum_{i=1}^{n} l(\beta_0 + \beta_j x_{ij}, y_i),\ j = 1, \ldots, p$$

where $l(\theta(\mathbf{x}), y) = b(\theta(\mathbf{x})) - y\theta(\mathbf{x})$. Then we can select a set of variables with large $\hat{\boldsymbol{\beta}}_j^M$ values based on a predefined threshold γ:

$$\mathcal{M}_\gamma = \{1 \le j \le p :\ |\hat{\boldsymbol{\beta}}_j^M| \ge \gamma\},$$

where γ is chosen to guarantee the sure screening property below. To dig into the rationale of the method, Fan and Song (2009) showed that under mild conditions, if $|\mathrm{cov}(x_j, y)| \ge c_1 n^{-\kappa}$, $j \in \mathcal{M}_*$ for some given constants $c_1 > 0$ and $0 < \kappa < 1/2$,

where $\mathcal{M}_*$ is the true model, then there exists a constant c_2 such that

$$\min_{j \in \mathcal{M}_*} |\boldsymbol{\beta}_j^M| \geq c_2 n^{-\kappa}.$$

In other words, the signal $\hat{\boldsymbol{\beta}}_j^M$ can be detected as long as x_j and y are somehow marginally correlated. Therefore, although the interpretations and implications of the marginal models are biased from the joint model, this screening criterion still maintains the nonsparse information about the joint model, and hence is suitable for selecting significant variables.

Furthermore, the authors proved the *sure screening property* of this screening method, that is, under regularity conditions, if $\gamma = c_3 n^{-\kappa}$, we have

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1 \text{ as } n \to \infty.$$

2.2.1.5 Screening via Generalized Correlation

The aforementioned three independent screening procedures are all model-based, that is, the results may be misleading if the assumed underlying model structure is wrong. This issue motives researchers to explore model free screening techniques. Hall and Miller (2009) advocated an approach based on raking the generalized empirical correlation between the predictors and response, which can identify significant variables without assuming an underlying model.

For the random sample $\{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$ from the population $(\mathbf{x}, y)$, where $\mathbf{x} = (x_1, \ldots, x_p)^{\mathbf{T}}$, the generalized empirical correlation between x_j and y is defined by

$$\hat{\psi}_j = \sup_{h \in \mathcal{H}} \frac{\sum_i (h(x_{ij}) - \bar{h}_j)(y_i - \bar{y})}{\sqrt{\sum_i (h(x_{ij})^2 - \bar{h}_j^2) \cdot \sum_i (y_i - \bar{y})^2}} \tag{2.28}$$

as an estimation of the population-version generalized correlation

$$\psi_j = \sup_{h \in \mathcal{H}} \frac{\text{cov}(h(x_j), y)}{\sqrt{\text{var}(h(x_j))\text{var}(y)}},$$

where $\mathcal{H}$ is a vector space of functions including all linear functions, and

$$\bar{h}_j = \frac{1}{n}\sum_{i=1}^{n} h(x_{ij}), \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}.$$

Note that if $\mathcal{H}$ is comprised of constant and linear functions, the generalized correlation reduces to the Pearson correlation in SIS. Therefore, this criterion can be viewed as the generalization of SIS to the nonlinear case.

Considering that (2.28) is not easy to compute in practice, Hall and Miller simplifies the problem in a wide range of cases where $\mathcal{H}$ is a finite-dimensional function space including the constant function. In this case, ranking $\hat{\psi}_j,\ j = 1, \ldots, p$ is equivalent to ranking

$$\hat{\varphi}_j = -\min_{h\in\mathcal{H}} \sum_{i=1}^{n} (y_i - h(x_{ij}))^2,\ j = 1, \ldots, p.$$

In this fashion, we obtain the submodel $\mathcal{M}_\gamma$ consisting of the top-ranking $\hat{\varphi}_j$'s. To determine the cutoff, i.e., the submodel size, Hall and Miller (2009) introduced a bootstrap procedure. First draw B resamples $\{(\mathbf{x}_i^{(k)}, y_i^{(k)}),\ i = 1, \ldots, n\},\ k = 1, \ldots, B$, from the original sample $\{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$, say, $B = 1000$. For each resample $\{(\mathbf{x}_i^{(k)}, y_i^{(k)}),\ i = 1, \ldots, n\}$, compute $\hat{\varphi}_j$ corresponding to $x_j,\ j = 1, \ldots, p$, and record their rankings $r_j^{(k)},\ j = 1, \ldots, p$. Therefore, for each predictor x_j, a set of rankings $r_j^{(1)}, \ldots, r_j^{(B)}$ is obtained based on the B resamples, and so is its $1-\alpha$ approximated confidence interval $(r_{j,-}, r_{j,+})$, where the approximated argument comes from the discreteness of rankings. The predictor x_j is selected into $\mathcal{M}_\gamma$ if $r_{j,+} < \gamma p$, where $0 < \gamma < 1$ is a small fraction, e.g. $\gamma = 1/8$.

Regarding to the theoretical properties, Hall and Miller (2009) stated and proved the ranking consistency of the Generalized-Correlation-based screening procedure. Specifically, Denote $\mathcal{I}_1 = \{j : 1 \le j \le p, |\mathrm{cov}(x_j, y)| \le c_1(n^{-1}\log n)^{1/2}\}$ and $\mathcal{I}_2 = \{j : 1 \le j \le p, |\mathrm{cov}(x_j, y)| \ge c_2(n^{-1}\log n)^{1/2}\}$. Then under regularity conditions, if c_1 is sufficiently small and c_2 is sufficiently large, all the indices in $\mathcal{I}_2$ are listed before any of the indices in $\mathcal{I}_1$ with probability tending to one.

In addition to the four methods introduced above, there exist many other screening methods in literature. For instance, Meinshausen and Yu (2009) demonstrated that

LASSO can be used as the screening procedure under mild conditions; Fan, Samworth and Wu (2009) extended Iterative SIS (ISIS) to a general pseudo-likelihood framework without requiring explicit definition of residuals; Zhu, Li, Li and Zhu (2011) proposed a model-free feature screening, called the Sure Independent Ranking and Screening (SIRS), among others. All these procedures provide powerful tools to reduce the ultrahigh dimensionality p down to a moderate scale d at the first step of the two-stage approach. Subsequent variable selection methods can now be applied to refine the submodel chosen by these procedures as well as to estimate the coefficients. The two-stage approach is illustrated in Figure 2.4.

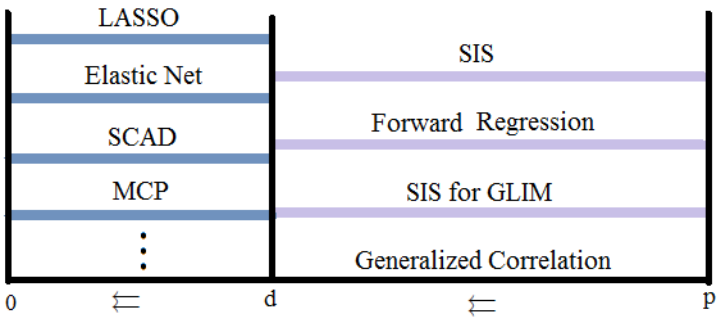


Figure 2.4. Illustrate the two-stage approach for ultrahigh dimensional variable selection.

2.2.2 Single-Step Regularized Regression for Ultrahigh Dimensional Problems

In addition to the two-stage approach above, another efficient procedure to deal with high or ultrahigh dimensionality is the single-step regularized regression, through which we can simultaneously identify and estimate the underlying model directly from the full model with dimensionality p.

2.2.2.1 Dantzig Selector (DS)

Candes and Tao (2007) proposed the Dantzig Selector (DS) for the linear regression model (2.2) with ultrahigh dimensionality p and Gaussian random noise $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$, which is defined as a solution of the following minimization prob-

lem:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \ \text{ subject to } \ \|\mathbf{X}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|_\infty \le \lambda, \tag{2.29}$$

where $\lambda > 0$ is the tuning parameter, and $\|\cdot\|_\infty$ refers to the L_∞ norm, by which the above expression measures the maximum absolute covariance between a predictor and the residual vector $\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$ to control model fitting, and is indeed a relaxation of the normal equation $\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$.

Due to the convexity of DS program, its computation is tractable by recasting it as a linear program:

$$\begin{aligned} \min \sum_{j=1}^{p} u_j \quad \text{subject to} \qquad & -u_j \le \beta_j \le u_j, \ j = 1, \ldots, p \ \text{ and} \\ & -\lambda\sigma\mathbf{1} \le \mathbf{X}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \le \lambda\sigma\mathbf{1}, \end{aligned}$$

where the optimization variables become u and $\mathbf{1}$ is a p-dimensional vector of ones.

To dig into the theory behind DS, the *Uniform Uncertainty Principle* (UUP, Candes and Tao, 2006) is required, which informally states that all $n \times m$ submatrices of design matrix $\mathbf{X}$ are uniformly close to orthonormal matrices, where $m \le S$, and S is comparable to the number of truly nonzero coefficients. Under UUP and other mild conditions, the authors proved the oracle inequality of DS, indicating that DS estimate $\hat{\boldsymbol{\beta}}$ mimics the risk of the oracle estimator up to a logarithmic factor of p: If we choose $\lambda = \sqrt{2\log p/n}$, then with large probability, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \le c^2\lambda^2(\sigma^2 + \sum_{i=1}^{p} \min(\beta_i^2, \sigma^2)),$$

where $\boldsymbol{\beta}_0$ is the vector of true coefficients, and c is a constant.

2.2.2.2 SCAD on High Dimensions

The SCAD penalized regression has been reviewed in last section, which possesses many desirable properties, including unbiasedness, sparsity, continuity, and oracle property. Kim, Choi, and Oh (2008) further studied SCAD on high dimensions where $p = O(n^\alpha)$ for some $\alpha > 0$. They developed an efficient optimization

algorithm for finding a local minimum of the PLS with SCAD penalty, and showed that SCAD still has nice properties on high dimensions.

To address the issue of computational complexity for high dimensional variable selection, the authors carefully studied the difference convex penalties such as SCAD, where the penalty functions can be expressed as the difference of two convex functions. Specifically, the SCAD penalty (2.12) can be decomposed as

$$p_\lambda(|\beta|) = \lambda|\beta| - J_\lambda(|\beta|),$$

where the second term, which turns out to be convex, is obtained by subtracting $p_\lambda(|\beta|)$ from the L_1 penalty $\lambda|\beta|$. The original SCAD penalty and its convex decomposition are demonstrated in Figure 2.5.

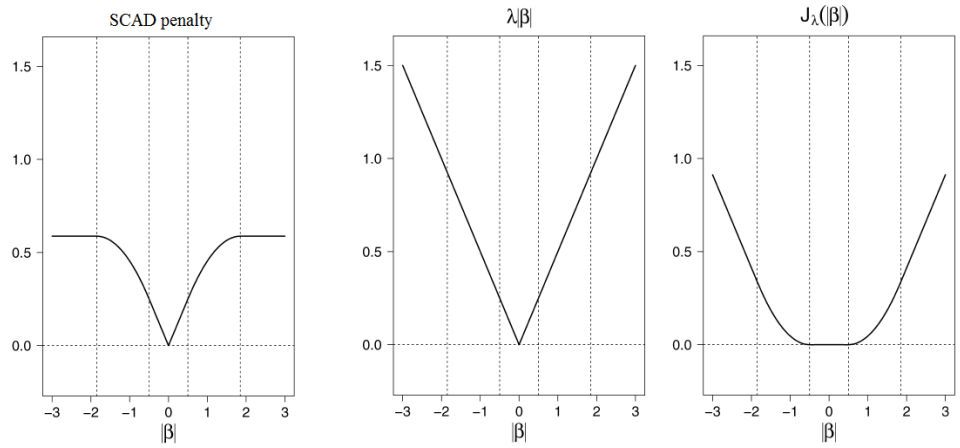


Figure 2.5. The SCAD penalty and its convex decomposition when $a = 3.7$ and $\lambda = 0.5$. (Kim, Choi, and Oh, 2008)

By the decomposition above, the penalized loss function in (2.5) can be written as

$$Q(\boldsymbol{\beta}) = \left[\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1\right] - \left[\sum_{j=1}^{p} J_\lambda(|\beta_j|)\right], \tag{2.30}$$

which is also a difference convex function. In (2.30), We use p instead of d in (2.5) to indicate the high dimensionality. The authors developed a Difference Convex Algorithm based on the CCCP algorithm proposed by An and Tao (1997)

(DC-CCCP), which can be applied to iteratively update the solution by locally approximating (2.30) with

$$Q(\boldsymbol{\beta}) \approx \left[\frac{1}{n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\|\boldsymbol{\beta}\|_1\right]-\left[\sum_{j=1}^{p}\nabla J_\lambda(|\beta_j^c|)\beta_j\right], \tag{2.31}$$

where β_j^c is a current solution of β_j in each iteration. Specifically, the DC-CCCP algorithm for SCAD penalty is described as follows:

1) Take the initial value $\boldsymbol{\beta}^c = 0$;

2) Find the minimizer $\boldsymbol{\beta}$ of $Q(\boldsymbol{\beta})$ in (2.31), and set $\boldsymbol{\beta}^c = \boldsymbol{\beta}$;

3) Iterate 2) until convergence.

One appealing property of DC-CCCP algorithm is that the computing time and convergence do not rely on the initial value, while many other algorithms such as LQA and LLA do require a good initial estimator. And it turns out that DC-CCCP might also be used in any other PLS problem as long as the penalty is difference convex function, such as MCP.

Although by DC-CCCP, we only obtain the local minimizer of $Q(\boldsymbol{\beta})$ in (2.31), Kim, Choi, and Oh (2008) proved that SCAD on high dimensions based on this algorithm still possesses the oracle property, in the sense that the oracle estimator defined below has a large probability to lie in the set of local minimizers. More explicitly, assume the true coefficient vector $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}, \mathbf{0})$, where $\boldsymbol{\beta}_{10} = (\beta_{10}, \ldots, \beta_{q0})$ with dimension q is the truly relevant coefficient. And correspondingly divide the fixed design matrix $\mathbf{X}$ into two parts $(\mathbf{X}_1, \mathbf{X}_2)$, the oracle estimator is then defined by $\hat{\boldsymbol{\beta}}_o = (\hat{\boldsymbol{\beta}}_o^{(1)}, \mathbf{0})$, where $\hat{\boldsymbol{\beta}}_o^{(1)} = (\mathbf{X}_1^{\mathrm{T}}\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{y}$. Moreover, denote $C(i,j) = \mathbf{X}_i^{\mathrm{T}}\mathbf{X}_j/n$ for $i,j = 1,2$, and $\mathbf{x}_{(j)} \in \mathbb{R}^n,\ j = 1,\ldots,p$, to be the jth column of the design matrix $\mathbf{X}$. The following regularity conditions are needed to prove the oracle property:

C1. There exists $M_1 > 0$ such that

$$\frac{1}{n}\mathbf{x}_{(j)}^{\mathrm{T}}\mathbf{x}_{(j)} \le M_1, \quad \text{for all} \quad j = 1,\ldots,p \quad \text{and all} \quad n.$$

C2. There exists $M_2 > 0$ such that $\alpha^{\mathrm{T}}C(1,1)\alpha \ge M_2$ for all α such that $\|\alpha\|_2^2 = 1$.

C3. $q = O(n^{c_1})$ for some $0 < c_1 < 1$.

C4. There exist constants c_2 and M_3, $c_1 < c_2 \leq 1$ such that

$$n^{(1-c_1)/2} \min_{j=1,\ldots,q} |\beta_{j0}| \geq M_3.$$

Under the foregoing regularity conditions, the oracle property of SCAD on high dimensions where $p = O(n^{\alpha})$ is established as follows.

Oracle Property. Let $\mathcal{A}$ be the set of local minima of (2.30) with the SCAD penalty. Assume that $E(\varepsilon_i)^{2k} < \infty$ for an integer $k > 0$, $\lambda = o(n^{-(1-(c_1-c_2))/2})$ and $p(\sqrt{n}\lambda)^{-2k} \to 0$, then

$$P(\hat{\boldsymbol{\beta}}_o \in \mathcal{A}) \to 1.$$

The authors further pointed out two stronger conclusions by adding some additional conditions. First, provided that ε_i's are i.i.d. Gaussian random variables, the SCAD also has the oracle property on ultrahigh dimensions where $p = O(\exp(c_3 n))$ for some $c_3 > 0$. Another conclusion is drawn under moderate dimensionality with $p \leq n$, where the global minima of (2.30) with SCAD penalty equates the oracle estimator with probability tending to one under mild conditions, i.e. $P(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_o) \to 1$, where $\hat{\boldsymbol{\beta}}$ is the global minima of the PLS based on SCAD.

Many other regularized regression techniques for high or ultrahigh dimensional problems were developed recently in literature, besides the Danzig selector and SCAD on high dimensions introduced here. Greenshtein and Ritov (2004) showed that the LASSO-type procedures are persistent for high dimensionality; Meinshausen (2007) presented similar results for ultrahigh dimensionality with finite nonsparsity size; The consistency results of LASSO on high dimensions were established in Donoho, Elad and Temlyakov (2006) Meinshausen and Buhlmann (2006), Wainwright (2006), etc.; Huang, Ma and Zhang (2008) considered Adaptive LASSO when a consistent initial estimator is available; Candes, Wakin and Boyd (2007) proposed weighted L_1 minimization to enhance the sparsity of the Dantzig selector; Efron, Hastie and Tibshirani (2007), and Bickel, Ritov and Tsybakov (2008) showed the asymptotic equivalence of LASSO and Dantzig selector under a sparsity scenario, among others.

2.3 Statistical Methods for Varying Coefficient Models

In the last two sections, we reviewed various techniques for selecting significant variables and estimating the associated coefficients mainly under linear model setting. However, linear models are often unrealistic in applications, and mis-specification of the data mechanism by a linear model could lead to large bias. To address these issues, nonparametric modeling and semiparametric modeling are developed in literature, among which varying coefficient modeling arises in many contexts.

The varying coefficient model is a natural extension of linear model to enhance the model flexibility and interpretability, where the coefficients of the linear models are replaced by smooth nonparametric functions and hence the regression coefficients are allowed to vary as functions of certain covariate. The model is defined as follows:

$$y = \mathbf{x}^T \boldsymbol{\beta}(u) + \varepsilon \tag{2.32}$$

where y is the response, $\mathbf{x}$ is the d-dimensional predictor, u is the univariate index variable, and $\varepsilon \in \mathbb{R}^1$ is the random noise satisfying $E(\varepsilon|\mathbf{x}, u) = 0$ almost surely. $\boldsymbol{\beta}(u) = \{\beta_1(u), \ldots, \beta_d(u)\}^{\mathrm{T}} \in \mathbb{R}^d$ is the coefficient vector, which is a smooth function of u to be estimated.

In this section, we review some standard statistical methods for the varying coefficient model, including the estimation procedure, hypothesis testing, and variable selection. These classical methods serve as the basis of extending the high or ultrahigh dimension techniques for linear models to varying coefficient models.

2.3.1 Preliminary: Nonparametric Smoothing

The estimation procedures for varying coefficient models are based on the preliminary knowledge about the nonparametric smoothing. Therefore, we briefly introduce the smoothing techniques for nonparametric regression, which further relaxes the model structure by assuming no parametric form of the regression function. Explicitly, suppose $(x_i, y_i), i = 1, \ldots, n$, is a bivariate random sample

from the nonparametric model

$$y = m(x) + \varepsilon \tag{2.33}$$

where ε is random noise with $E(\varepsilon|x) = 0$ and $\text{var}(\varepsilon|x) = \sigma^2(x)$. $m(x)$ is the mean function we are interested in. There are two commonly used local smoothing techniques to estimate $m(x)$ without assuming its parametric form: *kernel smoothing* and *local polynomial smoothing*.

2.3.1.1 Kernel Smoothing

Kernel Smoothing is also referred to as local constant approximation, since the idea is to treat $m(x)$ as a constant θ, where θ certainly relies on x. Define the estimator $\hat{\theta}$ to be

$$\hat{\theta} = \text{argmin}_\theta \sum_{i=1}^{n} (y_i - \theta)^2 \omega_i(x), \tag{2.34}$$

where $\omega_i(x)$, depending on x, is used to assign different weights to different data points, and usually a point closer to x has more information about $m(x)$ hence is worth taking more weight. One can easily get the solution of (2.34):

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \omega_i(x) y_i}{\sum_{i=1}^{n} \omega_i(x)}. \tag{2.35}$$

We need choose the weight function in (2.35) to get the explicit form of the solution. There exist in literature two popular kernel regression estimators with different choices of weight:

1. *NW estimator:*
   Nadaraya (1964) and Watson (1964) advocated the Nadaraya-Watson (NW) estimator based on the weight function

$$\omega_i(x) = K_h(x_i - x), \tag{2.36}$$

hence the NW kernel regression estimator for $m(x)$ is

$$\hat{\theta} = \frac{\sum_{i=1}^{n} K_h(x_i - x) y_i}{\sum_{i=1}^{n} K_h(x_i - x)}, \tag{2.37}$$

where $K_h(u) = K(u/h)/h$ with bandwidth h, and $K(u)$ is a Kernel function often satisfying $\int K(u)du = 1$. The frequently used Kernel functions are listed below:

- Gaussian Kernel: $K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$
- Uniform Kernel: $K(t) = I(|t| < 1/2)$
- Epanechikov Kernel: $K(t) = 0.75(1 - t^2)_+,\ t \in [-1, 1]$, where $a_+ = aI(a > 0)$
- Biweight Kernel: $K(t) = 0.9375(1 - t^2)_+^2,\ t \in [-1, 1]$
- Triweight Kernel: $K(t) = 1.09375(1 - t^2)_+^3,\ t \in [-1, 1]$.

For example, by taking the weight function to be uniform kernel, the NW estimator becomes the running local average, resulting in the unweighted K-nearest neighbor (KNN) estimator. However, as is shown in Marron and Nolan (1988), the kernel functions is not essential in the nonparametric estimation, yet the choice of bandwidth h, which will be discussed shortly, is crucial to the performance of the estimator.

2. *GM estimator:*
   Gasser and Müller (1984) introduced another choice of weight function:

$$\omega_i(x) = \int_{s_i - 1}^{s_i} K_h(u - x)du, \tag{2.38}$$

where $s_i = (x_{(i)} + x_{(i+1)})/2$, with $x_{(i)}$ being the ith order statistic of x. For convention, $x_{(0)} = -\infty$ and $x_{(n+1)} = +\infty$. Note that the denominator of (2.35) is 1 with this definition of weight, hence the GM estimator of $m(x)$ is

$$\hat{\theta} = \sum_{i=1}^{n} \left(\int_{s_i - 1}^{s_i} K_h(u - x)du \right) y_i. \tag{2.39}$$

The details for GM estimator can be found in Müller (1988).

2.3.1.2 Local Polynomial Regression

In kernel smoothing or local constant smoothing, we approximate $m(x)$ with a constant determined by x. A natural extension is to use polynomial approximation instead of constant, resulting in the local polynomial regression, which is systematically studied by Fan and Gijbels (1996).

Suppose the mean function $m(x)$ of the nonparametric model (2.33) is smooth and its $(q+1)$th derivative exists, and we want to estimate $m(x)$ for a given point x_0. We can locally approximate $m(x)$ using Taylor expansion at x_0:

$$\begin{aligned} m(x) &\approx m(x_0) + m'(x_0)(x - x_0) + \ldots + \frac{m^{(q)}(x)}{q!}(x - x_0)^q \\ &\triangleq \beta_0 + \beta_1(x - x_0) + \ldots + \beta_q(x - x_0)^q \\ &\triangleq \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}. \end{aligned} \tag{2.40}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $\mathbf{z} = \{1, x - x_0, \ldots, (x - x_0)^q\}^{\mathrm{T}}$, and

$$m^{(j)}(x_0) = j!\beta_j. \tag{2.41}$$

Thus $m(x_0)$ and its derivatives are estimated through $\hat{\boldsymbol{\beta}}$, which is defined similar to the NW estimator:

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 K_h(x_i - x_0), \tag{2.42}$$

where $\mathbf{z}_i = \{1, x_i - x_0, \ldots, (x_i - x_0)^q\}^{\mathrm{T}}$. This is really a weighted least squares problems, with the equivalent matrix form

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}), \tag{2.43}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^{\mathrm{T}}$, and $\mathbf{W}$ is a $n \times n$ diagonal matrix with the ith diagonal component $K_h(x_i - x_0)$. The solution of (2.43) is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{W}\mathbf{y}.$$

Therefore, not only $m(x_0)$, but also its derivatives can be estimated through $\hat{\boldsymbol{\beta}}$ us-

ing (2.41), which is one appealing feature of local polynomial smoothing over kernel smoothing. When $q = 1$ in (2.40), the local polynomial regression reduces to local linear approximation, one of the most frequently used nonparametric smoothing techniques in literature.

2.3.1.3 Choice of Bandwidth

As was mentioned before, the choice of bandwidth h in nonparametric regression is essential. Specifically, if h is chosen too small, we pay too much attention to the data in the local neighborhood and undersmoothly estimate the mean function; While if h is too large, we tend to miss some fine features of the data and yield an oversmoothed estimate. Then how to choose an optimal bandwidth? A popular method is the Mean Squared Error (MSE) criterion, which measures how far apart of an estimate from its true value, and can be decomposed into the summation of variance and the squared bias. Mathematically, suppose θ is the parameter of interest with the estimate $\hat{\theta}$, then the MSE is defined to be

$$MSE(\hat{\theta}) \triangleq E(\hat{\theta} - \theta)^2 = \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta}),$$

where $bias(\hat{\theta}) = E(\hat{\theta}) - \theta$. In the case of nonparametric regression where $m(x)$ is the function to be estimated,

$$MSE(\hat{m}(x)) \triangleq E(\hat{m}(x) - m(x))^2 = \text{var}(\hat{m}(x)) + \text{bias}^2(\hat{m}(x)).$$

By this definition, however, the MSE only measures the performance of $\hat{m}(x)$ at a fixed point x. To take advantage of all the data points, the Mean Integrated Square Error (MISE) is proposed to integrate the MSE over x's:

$$\text{MISE}(\hat{m}(\cdot)) = \int \text{MSE}(\hat{m}(x)) f(x) dx,$$

where $f(x)$ is the density function of x, which can be estimated by a Kernel Density Estimate (KDE):

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x_i - x).$$

Clearly MISE is determined by the bandwidth h. Therefore, the optimal bandwidth h is chosen to minimize MISE. To be more specific, for each bandwidth h, compute the corresponding MISE. Then for a series of grid points $(h_1, \ldots, h_M)$, the optimal h is chosen to be

$$h_{opt} = \text{argmin}_{h_1,\ldots,h_M} \text{MISE}(\hat{m}(\cdot)). \tag{2.44}$$

In practice, we calculate MSE or MISE based on the pointwise asymptotic variance and bias, which are summarized in Table 2.1 for some nonparametric estimators introduced in the previous two sections.

Table 2.1. Pointwise asymptotic bias and variance.

Method	Bias	Variance
NW estimator	$\{m''(x) + \frac{2m'(x)f'(x)}{f(x)}\}b_n$	V_n
GM estimator	$m''(x)b_n$	$1.5V_n$
Local linear	$m''(x)b_n$	V_n

In Table 2.1, b_n and V_n are defined by

$$b_n = \frac{1}{2}h^2 \int u^2 K(u)du, \quad V_n = \frac{\sigma^2(x)}{f(x)nh} \int K^2(u)du.$$

In addition, $m'(x)$, $m''(x)$, and $\sigma^2(x)$ can be estimated by pre-modeling. For example, we first fit a "wrong" but convenient model by approximating $m(x)$ with a polynomial function, i.e.

$$y = m(x) = \alpha_0 + \alpha_1 x + \ldots + \alpha_q x^q + \epsilon. \tag{2.45}$$

When $q = 4$, which is sufficient in practice,

$$\hat{m}'(x) = \hat{\alpha}_1 + 2\hat{\alpha}_2 x + 3\hat{\alpha}_3 x^2 + 4\hat{\alpha}_4 x^3, \quad \hat{m}''(x) = 2\hat{\alpha}_2 + 6\hat{\alpha}_3 x + 12\hat{\alpha}_4 x^2,$$

and $\widehat{\sigma^2}(x)$ is taken to be the mean squared error of model (2.45). In this fashion, MISE can be estimated and the optimal bandwidth is obtained by (2.44).

2.3.2 Estimation Methods for Varying Coefficient Models

From this section on, we get back to the varying coefficient model (2.32). The primary task is to estimate its coefficient function $\boldsymbol{\beta}$. There are basically three estimating methods: The first one is the Kernel-local Polynomial Smoothing, which is studied by Wu et.al (1998), Hoover et al. (1998), Fan and Zhang (1999), Kauermann and Tutz (1999), and related to the *local polynomial regression* reviewed in section 2.3.1.2. The second technique is the *polynomial spline* by Huang et al. (2002, 2004) and Huang and Shen (2004). The last is *smoothing spline*, proposed by Hastie and Tibshirani (1993) and studied in Hoover et al. (1998) and Chiang et al. (2001).

In this thesis, we mainly introduce the first approach, *kernel-local polynomial smoothing*, which contains two categories. One is to estimate the coefficient functions through a single step of local polynomial regression, which is simple and useful, but implicitly assumes the same degree of smoothness of all β_j's. Therefore, the other category, called two-step estimation, arises to address this issue.

2.3.2.1 One-step Estimation

Suppose $\{(\mathbf{x}_i, y_i, u_i),\ i = 1, \ldots, n\}$ is the random sample from the varying coefficient model (2.32). For each given u, approximate the coefficient functions $\beta_j(u_i),\ j = 1, \ldots, d$ locally as

$$\begin{aligned} \beta_j(u) &\approx \beta_j(u) + \beta_j^{'}(u_i - u) \\ &\triangleq a_j + b_j(u_i - u) \end{aligned}$$

for u_i in a neighborhood of u. Similar as the Local Polynomial Regression (2.42), we consider the following weighted least squares problem:

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{d} \{a_j + b_j(u_i - u)\} x_{ij} \right\}^2 K_h(u_i - u).$$

Then $\hat{\boldsymbol{\beta}}(u)$ is obtained corresponding to $\hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_d)^{\mathrm{T}}$. Fan and Zhang (1999) provided the matrix form of solution:

$$\hat{\boldsymbol{\beta}}(u) = (I_d, \mathbf{0}_d)(\Gamma^{\mathrm{T}}\mathbf{W}\Gamma)^{-1}\Gamma^{\mathrm{T}}\mathbf{W}\mathbf{y},$$

where I_d is the $d \times d$ identity matrix, $\mathbf{0}_d$ is the $d \times d$ matrix with each entry being 0, and other quantities are defined by

$$\begin{aligned} &\mathbf{W} = \mathrm{diag}(K_h(u_1 - u), \ldots, K_h(u_n - u)), \quad \Gamma = (\mathbf{X}, \mathbf{U}\mathbf{X}) \\ &\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}, \quad \mathbf{U} = \mathrm{diag}(u_1 - u, \ldots, u_n - u). \end{aligned} \tag{2.46}$$

The bandwidth h of the kernel function is chosen to minimize the MISE comprised of bias and variance, which are derived by Carroll et al. (1998) and Fan and Zhang (1999):

$$bias(\hat{\boldsymbol{\beta}}(u)) = \mu_2\hat{\boldsymbol{\beta}}''(u)h^2/2, \quad \mathrm{var}(\hat{\boldsymbol{\beta}}(u)) = \{nhf(u)E(\mathbf{x}\mathbf{x}^{\mathrm{T}}|u)\}^{-1}\nu_0\sigma^2(u).$$

where $\mu_i = \int u^i K(u)du$ and $\nu_i = \int u^i K^2(u)du$. Their estimates were computed by Fan and Zhang (2000) and systematically studied by Fan and Zhang (2008):

$$\begin{aligned} &\widehat{bias}(\hat{\boldsymbol{\beta}}(u)) = (I_d, \mathbf{0}_d)(\Gamma^{\mathrm{T}}\mathbf{W}\Gamma)^{-1}\Gamma^{\mathrm{T}}\mathbf{W}\hat{\boldsymbol{\tau}} \\ &\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}(u)) = (I_d, \mathbf{0}_d)(\Gamma^{\mathrm{T}}\mathbf{W}\Gamma)^{-1}(\Gamma^{\mathrm{T}}\mathbf{W}\Gamma)(\Gamma^{\mathrm{T}}\mathbf{W}\Gamma)^{-1}(I_d, \mathbf{0}_d)^{\mathrm{T}}\hat{\sigma}^2(u), \end{aligned} \tag{2.47}$$

where the ith element of $\hat{\boldsymbol{\tau}}$ is $\frac{1}{2}\mathbf{x}_i^{\mathrm{T}}\{\boldsymbol{\beta}''(u)(u_i - u)^2 + \frac{1}{3}\boldsymbol{\beta}^{(3)}(u)(u_i - u)^3\}$, with $\boldsymbol{\beta}''(u)$ and $\boldsymbol{\beta}^{(3)}$ estimated by local cubic fitting with a pilot bandwidth h^*; and $\hat{\sigma}^2(u)$ is obtained by

$$\hat{\sigma}^2(u) = \frac{\mathbf{y}^{\mathrm{T}}\{\mathbf{W}^* - \mathbf{W}^*\Gamma^*(\Gamma^{*\mathrm{T}}\mathbf{W}^*\Gamma^*)^{-1}\Gamma^{*\mathrm{T}}\mathbf{W}^*\}\mathbf{y}}{tr\{\mathbf{W}^* - (\Gamma^{*\mathrm{T}}\mathbf{W}^*\Gamma^*)^{-1}(\Gamma^{*\mathrm{T}}\mathbf{W}^{*2}\Gamma^*)\}},$$

where $\mathbf{W}^*$ is $\mathbf{W}$ in (2.46) with h replaced by h^*, and $\Gamma^* = (\mathbf{X}, \mathbf{U}\mathbf{X}, \mathbf{U}^2\mathbf{X}, \mathbf{U}^3\mathbf{X})$.

Furthermore, Zhang and Lee (2000) also proved the asymptotic normality of $\hat{\boldsymbol{\beta}}(u)$, that is, under regularity conditions,

$$\mathrm{var}^{-1/2}(\hat{\boldsymbol{\beta}}(u))\left\{\hat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}(u) - bias(\hat{\boldsymbol{\beta}}(u))\right\} \xrightarrow{D} N(\mathbf{0}, I_d)$$

2.3.2.2 Two-step Estimation

Although the single step estimation procedure is simple and useful, it by default assume the coefficient functions $\beta_j(u)$'s admit the same degree of smoothness, which is unrealistic under certain circumstances. Intuitively we need larger bandwidth for the smoother component, aiming to obtain smoother estimate, while for the less smoother component, smaller bandwidth is preferable. To achieve this goal, Fan and Zhang (1999) developed a two-step estimation procedure. Without loss of generality, we assume $\beta_d(u)$, which has fourth derivative, is smoother than any $\beta_j(u),\ j = 1, \ldots, d-1$, with existence of second derivatives. The two-step estimation is sketched as follows.

Step1

1. Apply the one-step estimation procedure to the original model with a small bandwidth h, then for any u, we get an initial estimator of $\boldsymbol{\beta}(u)$:

$$\tilde{\boldsymbol{\beta}}(u) = (I_d, \mathbf{0}_d)(\Gamma^{\mathrm{T}}\mathbf{W}\Gamma)^{-1}\Gamma^{\mathrm{T}}\mathbf{W}\mathbf{y},$$

   where all the quantities are defined as (2.46).

2. Write the varying coefficient model (2.32) as

$$y = \sum_{j=1}^{d-1} \beta_j(u)x_j + \beta_d(u)x_d + \varepsilon. \tag{2.48}$$

   For $j = 1, \ldots, d-1$, replace $\beta_j(u)$ in model (2.48) by the jth component of $\tilde{\boldsymbol{\beta}}(u)$ above, and obtain the synthetic model

$$y - \sum_{j=1}^{d-1} \tilde{\beta}_j(u)x_j = \beta_d(u)x_d + \varepsilon \tag{2.49}$$

   with the new response $\tilde{y} = y - \sum_{j=1}^{d-1} \tilde{\beta}_j(u)x_j$. Model (2.49) is ready for the second step: reestimating $\beta_d(u)$ by a smoother function.

Step2

1. Since $\beta_d(u)$ has the fourth derivative, we approximate $\beta_d(u)$ by a cubic

function

$$\beta_d(u_i) = \sum_{k=0}^{3} \frac{\beta_d^{(k)}(u)}{k!}(u_i - u)^k \triangleq \sum_{k=0}^{3} a_{d,k}(u_i - u)^k$$

for u_i in the neighborhood of u, and conduct the local cubic regression with the objective function

$$\sum_{i=1}^{n} \left\{ \tilde{y}_i - x_{id} \sum_{k=0}^{3} a_{d,k}(u_i - u)^k \right\}^2 K_{h_2}(u_i - u),$$

where optimization variable is $(a_{p,0}, a_{p,1}, a_{p,2}, a_{p,3})$, and the bandwidth h_2 is chosen to be larger than h in the first step.

2. The final estimator of $\beta_d(u)$, which corresponds to $\hat{a}_{p,0}$, is computed as

$$\hat{\beta}_d(u) = e_{1,4}^{\mathrm{T}}(\mathbf{G}^{\mathrm{T}}\mathbf{W}_2\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{W}_2\tilde{\mathbf{y}},$$

where $e_{k,m}$ refers to the unit vector of length m with 1 at the kth position, and other quantities are given as follows:

$$\begin{aligned}
&\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)^{\mathrm{T}}, \quad \mathbf{G} = \mathrm{diag}(x_{1d}, \ldots, x_{nd})\mathbf{Q}, \\
&\mathbf{W}_2 = \mathrm{diag}\left(K_{h_2}(u_1 - u), \ldots, K_{h_2}(u_n - u)\right), \\
&\mathbf{Q} = \begin{pmatrix} 1 & u_1 - u & (u_1 - u)^2 & (u_1 - u)^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & u_n - u & (u_n - u)^2 & (u_n - u)^3 \end{pmatrix}.
\end{aligned}$$

Remark: In the second step, there is another way to get the estimate of $\beta_d(u)$ which is easier to implement, where we simply smooth $\tilde{\beta}_d(u_i)$ against u_i by local cubic modeling with bandwidth h_2. Specifically, treat $\tilde{\beta}_d(u)$ as the new response, denoted by y^*, and estimate the mean function of the nonparametric model

$$y^* = \beta_d(u) + \varepsilon.$$

By the local polynomial regression with $q = 3$, the resulting final estimator of $\beta_d(u)$ is

$$\beta_d^*(u) = e_{1,4}^{\mathrm{T}}(\mathbf{Q}^{\mathrm{T}}\mathbf{W}_2\mathbf{Q})^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{W}_2\mathbf{y}^*,$$

where $\mathbf{y}^* = (\tilde{\beta}_d(u_1), \ldots, \tilde{\beta}_d(u_n))^{\mathrm{T}}$, and other quantities are defined as before.

In addition, the bandwidth h for the initial estimation is chosen to be small, aiming to reduce bias in the first step. Then apply higher order smoothing with a larger bandwidth h_2 to get the final estimator of the smoother component $\beta_d(u)$, in which fashion the variance can be reduced. This is the core reason that the two-step approach always outperforms the one-step approach when estimating smoother components.

2.3.3 Confidence Band and Hypothesis Test for Varying Coefficient Models

The confidence band is of interest for nonparametric inference. To construct the $1-\alpha$ confidence band $[g_{1j}(u),\ g_{2j}(u)]$ for varying coefficient models, i.e.

$$P(g_{1j}(u) \le \beta_j(u) \le g_{2j}(u), \quad \text{for any } u \in D) = 1-\alpha,\ j = 1, \ldots, d,$$

where D is a compact set, the most important and challenging part is to derive the distribution of the maximum discrepancy between the estimate and the true coefficient function. Here we only consider the one-step estimation. Fan and Zhang (2000) proved that the asymptotic distribution is related to $\exp\{-2\exp(-x)\}$, based on which the $1-\alpha$ confidence band for $\beta_j(u),\ j = 1, \ldots, d$ is constructed as

$$\hat{\beta}_j(u) - \widehat{bias}(\hat{\beta}_j(u)) \pm \triangle_{j,\alpha}(u), \tag{2.50}$$

where
$$\triangle_{j,\alpha}(u) = \left\{d_{\nu,n} + (-2\log h)^{-1/2}[\log 2 - \log(-\log(1-\alpha))]\right\}\left\{\widehat{\mathrm{var}}(\hat{\beta}_j(u))\right\}^{1/2},$$
$$d_{\nu,n} = (-2\log h)^{1/2} + (-2\log h)^{-1/2}\log\left\{\frac{\int (K'(u))^2 du}{4\pi\int K^2(u)du}\right\},$$

and $\widehat{bias}(\hat{\beta}_j(u))$, $\widehat{\mathrm{var}}(\hat{\beta}_j(u))$ are obtained by (2.47).

In addition to the confidence band, one might be interested in testing if certain

coefficient functions are significant in varying coefficient models, or whether they do vary with the index variable u, or if they have some known function form. These problems amount the null hypotheses to be parametric while the alternatives to be nonparametric. For such tests, traditional testing can not be applied, but there are several approaches specifically for them, among which two are introduced in this section.

1) *Generalized Likelihood Ratio Test:*
Fan, Zhang, and Zhang (2001) advocated the Generalized Likelihood Ratio Test (GLRT) for the varying coefficient model (2.32), where the model under null hypothesis H_0 is nested within that under alternative H_1. Suppose $\hat{\boldsymbol{\beta}}^0(u)$ and $\hat{\boldsymbol{\beta}}^1(u)$ are the estimated coefficient functions under H_0 and H_1, respectively. The residual sum of squares RSS_0 for the reduced model and RSS_1 for the full model are defined by

$$RSS_0 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^0(u))^2 \quad \text{and} \quad RSS_1 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^1(u))^2.$$

Then the generalized likelihood ratio

$$\Omega = l_n(H_1) - l_n(H_0) = \frac{n}{2}\log\frac{RSS_0}{RSS_1} \approx \frac{n}{2}\frac{RSS_0 - RSS_1}{RSS_1}.$$

The challenge of the GLRT is the approximate distribution of Ω due to the infinite dimension of nonparametric functions. Fan, Zhang, and Zhang (1999) unveiled the new Wilks phenomenon with the new definition of degree of freedom for nonparametric model. Explicitly, under mild conditions, we have

$$\gamma_K\Omega \sim \chi^2_\delta \tag{2.51}$$

where $\delta = \gamma_K|\mathbb{U}|(p_1 - p_0)\{K(0) - 0.5\int K^2(u)du\}/h$, $|\mathbb{U}|$ is the range of the index variable support $\mathbb{U}$, p_1 and p_0 are the number of nonzero estimated coefficients under H_1 and H_0, respectively; and γ_K is a constant which is determined by the kernel function $K(\cdot)$. The values of γ_K for the aforementioned kernels are listed in Table 2.2. Therefore, we may reject the null

hypothesis or not based on the p-value from the distribution (3.14).

Table 2.2. The values of γ_K.

Kernel	**Gaussian**	**Uniform**	**Epanechnikov**	**Biweight**	**Triweight**
γ_K	2.5375	1.2000	2.1153	2.3061	2.3797

2) *Maximum Discrepancy*

Fan and Zhang (2000) took another testing approach based on the asymptotic distribution of the maximum discrepancy between $\hat{\boldsymbol{\beta}}(u)$ and the true function $\boldsymbol{\beta}(u)$. First consider the test

$$H_0: \ \beta_j(u) = \beta_j^0(u), \quad \text{v.s.} \quad H_1: \ \beta_j(u) \neq \beta_j^0(u) \tag{2.52}$$

where $\beta_j^0(u)$ is a known function. A natural test is to check whether $\beta_j^0(u)$ falls in the confidence band (2.50) or not. The authors constructed an equivalent test statistic:

$$M = \sqrt{-2\log h}\left\{\left\|\frac{\hat{\beta}_j(u) - \beta_j^0(u) - \widehat{bias}(\hat{\beta}_j(u))}{\sqrt{\widehat{\text{var}}\hat{\beta}_j(u)}}\right\|_\infty - d_{\nu,n}\right\} \tag{2.53}$$

where $\|g(u)\|_\infty = \sup_{u\in\mathbb{U}}|g(u)|$, and $\mathbb{U}$ is the support of u. They showed H_0 should be rejected if $M > -\log(-\frac{1}{2}\log(1-\alpha))$.

In addition, one might also be interested in testing whether certain coefficient function does vary with u, i.e.

$$H_0: \ \beta_j(u) = c \quad \text{v.s.} \quad \beta_j(u) \neq c. \tag{2.54}$$

The difference between hypothesis (2.52) and (2.54) lies in that c is unknown in (2.54), since it only implies the constancy of $\beta_j(u)$. The test statistic is constructed in the same way as (2.53) with the same rejection criterion, but $\beta_j^0(u)$ is replaced by

$$\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j(u_i).$$

2.3.4 Variable Selection for Varying Coefficient Models

To conduct variable selection for the varying coefficient model (2.32), Wang and Xia (2009) proposed a shrinkage estimation procedure, called Kernel LASSO method (KLASSO), based on the local polynomial smoothing (Fan and Gijbels, 1996) and the LASSO penalized regression (Tibshirani, 1996), which can simultaneously select significant variables and estimate the coefficient function $\boldsymbol{\beta}(u)$. However, since we cannot directly estimate the whole coefficient function by an explicit form, we are specifically interested in estimating $\boldsymbol{\beta}(u)$ when $u = u_1, \ldots, u_n$, i.e. the matrix

$$\mathbf{B} = \{\boldsymbol{\beta}(u_1), \ldots, \boldsymbol{\beta}(u_n)\}^{\mathrm{T}} = (\mathbf{b}_1, \ldots, \mathbf{b}_d)\,,$$

where $\mathbf{b}_j \in \mathbb{R}^{n\times 1}$ is the jth column of $\mathbf{B}$. Wang and Xia (2009) suggested the penalized loss function

$$Q_\lambda(\mathbf{B}) = \sum_{t=1}^{n}\sum_{i=1}^{n}\{y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(u_t)\}^2 K_h(u_t - u_i) + n\sum_{j=1}^{d}\lambda_j\|\mathbf{b}_j\| \qquad (2.55)$$

where $\|\cdot\|$ is the Euclidean norm. They also proposed an iterative algorithm with the local quadratic approximation to the penalty function (Fan and Li, 2001) to obtain the minimizer of $Q_\lambda(\mathbf{B})$, $\widehat{\mathbf{B}}_\lambda$. That is, the $(m+1)$th-step value of $\widehat{\mathbf{B}}_\lambda$ is

$$\widehat{\mathbf{B}}_\lambda^{(m+1)} = \{\hat{\boldsymbol{\beta}}_\lambda^{(m+1)}(u_1), \ldots, \hat{\boldsymbol{\beta}}_\lambda^{(m+1)}(u_n)\}^{\mathrm{T}}$$

with the tth row $\left(\hat{\boldsymbol{\beta}}_\lambda^{(m+1)}(u_t)\right)^{\mathrm{T}}$, where

$$\hat{\boldsymbol{\beta}}_\lambda^{(m+1)}(u_t) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}K_h(u_t - u_i) + D^{(m)}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i K_h(u_t - u_i)\right),$$

$D^{(m)}$ is a $d\times d$ diagonal matrix with the jth diagonal component $\lambda_j/\|\hat{\mathbf{b}}_{\lambda,j}^{(m)}\|$, and $\hat{\mathbf{b}}_{\lambda,j}^{(m)}$ is the jth column of the mth-step value $\widehat{\mathbf{B}}_\lambda^{(m)}$.

Remark: The initial value $\widehat{\mathbf{B}}_\lambda^{(0)}$ of the above algorithm is set to be the unpenalized estimator $\tilde{\mathbf{B}} = \mathrm{argmin}_{\mathbf{B}\in\mathbb{R}^{n\times d}}\left\{\sum_{t=1}^{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}(u_t))^2 K_h(u_t - u_i)\right\}$. And the tuning parameter vector $\lambda = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$ is chosen following the idea of Zou (2006), Zhang and Lu (2007), Wang et al. (2007a), Zou and Li (2008). That is,

$\lambda_j = n^{1/2}\lambda_0/\|\tilde{\mathbf{b}}_j\|$ with λ_0 selected by the BIC-type criterion, where $\tilde{\mathbf{b}}_j$ is the jth column of $\tilde{\mathbf{B}}$.

To develop the theoretical properties of the KLASSO estimates, the following regularity conditions are needed:

(C1) For $s > 2$, $E|y_i|^{2s} < \infty$ and $E\|\mathbf{x}_i\|^{2s} < \infty$.

(C2) The density function of u_i, denoted by $f(u)$, is continuous and positively bounded away from 0 on $\mathbb{U}$, where $\mathbb{U}$ is the support of u_i.

(C3) $\Omega(u) = E(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}|u_i = u)$ is nonsingular and has bounded second order derivatives on $\mathbb{U}$. $E(\|\mathbf{x}_i\|^4|u_i = u)$ is also bounded.

(C4) The second order derivative of $f(u)$ and $\sigma^2(u) = E(\varepsilon_i^2|u_i = u)$ are bounded.

(C5) $K(u)$ is a symmetric density function with a compact support.

(C6) The second order derivatives of coefficients $\beta_{0j}(u), j = 1, \ldots, d$ are continuous, where $\boldsymbol{\beta}_0(u) = \{\beta_{01}(u), \ldots, \beta_{0d}(u)\}^{\mathrm{T}} \in \mathbb{R}^d$ is the true coefficient vector.

Under regularity conditions above, the authors stated the sparsity and oracle property of KLASSO estimates. To establish the sparsity, we assume the true model contains only a small number of predictors. Without loss of generality, assume the first d_0 predictors are truly important but others are not. Define the KLASSO estimate of the relevant and irrelevant coefficient vector by

$$\hat{\boldsymbol{\beta}}_{a,\lambda}(u) = \{\hat{\boldsymbol{\beta}}_{\lambda,1}(u), \ldots, \hat{\boldsymbol{\beta}}_{\lambda,d_0}(u)\}^{\mathrm{T}} \in \mathbb{R}^{d_0}$$
$$\hat{\boldsymbol{\beta}}_{b,\lambda}(u) = \{\hat{\boldsymbol{\beta}}_{\lambda,d_0+1}(u), \ldots, \hat{\boldsymbol{\beta}}_{\lambda,d}(u)\}^{\mathrm{T}} \in \mathbb{R}^{d-d_0},$$

respectively. Let $a_n = \max\{\lambda_j : 1 \le j \le d_0\}$ and $b_n = \min\{\lambda_j : d_0 < j \le d\}$.

Theorem 3. *(Sparsity) Under regularity conditions (C1)-(C6), suppose $h \propto n^{-1/5}$, $n^{11/10}a_n \to 0$, and $n^{11/10}b_n \to \infty$, then we have*

$$P(\sup_{u\in\mathbb{U}} \|\hat{\boldsymbol{\beta}}_{\lambda,b}(u)\| = 0) \to 1$$

for any $d_0 < j \le d$.

Hence the sparse solutions can be consistently produced for every irrelevant predictors over $\mathbb{U}$ uniformly. To establish the oracle property, first define the oracle estimator (i.e. the unpenalized estimator obtained under the true model) by

$$\hat{\boldsymbol{\beta}}_{ora}(u) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{ia}\mathbf{x}_{ia}^{\mathrm{T}}K_h(u-u_i)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{ia}y_iK_h(u-u_i)\right),$$

where $\mathbf{x}_{ia} = (x_{i1},\ldots,x_{id_0})^{\mathrm{T}} \in \mathbb{R}^{d_0}$ and $\mathbf{x}_{ib} = (x_{i(d_0+1)},\ldots,x_{id})^{\mathrm{T}} \in \mathbb{R}^{d-d_0}$.

Theorem 4. *(Oracle Property) Under regularity conditions (C1)-(C6), suppose* $h \propto n^{-1/5}$, $n^{11/10}a_n \to 0$, *and* $n^{11/10}b_n \to \infty$, *we then have*

$$\sup_{u\in\mathbb{U}} \|\hat{\boldsymbol{\beta}}_{a,\lambda}(u) - \hat{\boldsymbol{\beta}}_{ora}(u)\| = o_p(n^{-2/5}).$$

Therefore, the difference between the KLASSO estimate and the oracle estimate is negligible uniformly over $\mathbb{U}$, and consequently $\hat{\boldsymbol{\beta}}_{a,\lambda}(u)$ shares the same asymptotic distribution and efficiency as the oracle estimate $\hat{\boldsymbol{\beta}}_{ora}(u)$.

2.4 Estimation Procedures for Partially Linear Models

Partially linear models are useful extension of linear models, where we allow the response to depend on certain variable in a nonparametric form, aside from the linear dependency of other variables. And they are essentially special cases of varying coefficient models, with all but one coefficients not varying with the index variable. Specifically, consider the partially linear model with the form

$$y = \alpha(u) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \varepsilon, \tag{2.56}$$

where y is the univariate response, $\alpha(u)$ is the nonparametric component, indicating that y is partially explained by u, but we do not put any assumptions to the form of $\alpha(u)$. $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^{\mathrm{T}} \in \mathbb{R}^p$ is the p-dimensional coefficient vector corresponding to the linear predictor $\mathbf{x} = (x_1,\ldots,x_p)^{\mathrm{T}} \in \mathbb{R}^p$.

Although the statistical methods for varying coefficient models can still be

applied for partially linear models, we often adopt different strategies to analyze these special models to enhance the accuracy and precision of the estimations. In this subsection, we study the main estimation procedures for partially linear models. See Chen (1988), Engle et.al (1986), Heckman (1986), Speckman (1988), Robinson (1988), and Fan and Huang (2005) for details. We briefly review the following estimation methods for partially linear model (2.56).

2.4.1 Difference Based Method

The only difference between model (2.56) and the well-studied linear model is the nonparametric component $\alpha(u)$, which is assumed to be a smooth and continuous function of u. Thus the intuition of difference based method (Fan and Li, 2004) is to get rid of $\alpha(u)$ and transform the partially linear model to the corresponding linear model. Assume $\{(u_i, \mathbf{x}_i, y_i), i = 1, \ldots, n\}$ is a random sample from model (2.56), and u_i's are dense, which means $u_{i+1} - u_i$ is negligibly small. The random error ε_i's are independently and identically distributed with mean 0, and they are independent with u_i. The procedure is as follows:

1. Consider the sample version of model (2.56)

$$y_i = \alpha(u_i) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i. \tag{2.57}$$

   Sort the sample by u such that $u_{(1)} \leq u_{(2)} \leq \ldots \leq u_{(n)}$. So the sorted sample version of model (2.56) becomes

$$y_{(i)} = \alpha(u_{(i)}) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{(i)} + \varepsilon_{(i)}. \tag{2.58}$$

2. Consider the $(i+1)$th subject:

$$y_{(i+1)} = \alpha(u_{(i+1)}) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{(i+1)} + \varepsilon_{(i+1)}. \tag{2.59}$$

   Subtracting model (2.58) from (2.59):

$$y_{(i+1)} - y_{(i)} = \{\alpha(u_{(i+1)}) - \alpha(u_{(i)})\} + \boldsymbol{\beta}^{\mathrm{T}}\{\mathbf{x}_{(i+1)} - \mathbf{x}_{(i)}\} + \{\varepsilon_{(i+1)} - \varepsilon_{(i)}\}.$$

We assume that $\alpha(u)$ is a continuous and smooth function of u, and $u_{(i+1)} - u_{(i)} \approx 0$, hence $\alpha(u_{(i+1)}) - \alpha(u_{(i)}) \approx 0$ by Taylor's theorem. Define $y_i^* = y_{(i+1)} - y_{(i)}$, $\mathbf{x}_i^* = \mathbf{x}_{(i+1)} - \mathbf{x}_{(i)}$, and $\varepsilon_i^* = \varepsilon_{(i+1)} - \varepsilon_{(i)}$, then the model above becomes

$$y_i^* = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i^* + \varepsilon_i^*, \tag{2.60}$$

with $\boldsymbol{\beta}$ remaining the same while $\alpha(u)$ is canceled.

3. Apply the least squared method to estimate $\boldsymbol{\beta}$ in model (2.60):

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{*\mathrm{T}} \mathbf{X}^*)^{-1} \mathbf{X}^{*\mathrm{T}} \mathbf{y}^*, \tag{2.61}$$

where $\mathbf{X}^*$ is the design matrix for model (2.60) and $\mathbf{y}$ is the corresponding response vector.

4. To estimate $\alpha(\cdot)$, plug the estimate $\widehat{\boldsymbol{\beta}}$ back into (2.57) and treat $y_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{x}_i \equiv y_i^{**}$ as the new response, that is, the model of interest in this step is

$$y_i^{**} = \alpha(u_i) + \varepsilon_i. \tag{2.62}$$

Then we can apply any one dimensional nonparametric estimating technique to get $\widehat{\alpha}(\cdot)$, such as NW estimator, local linear estimator, local polynomial estimator, among others, as introduced in last section.

The difference based method is simple to use; it does not need to specify the smoothing matrix and select bandwidth when estimating the parametric part $\boldsymbol{\beta}$. However, it puts relatively strong assumptions on u and $\alpha(\cdot)$. Furthermore, the estimation of $\boldsymbol{\beta}$ is not asymptotically efficient by this method.

2.4.2 Back Fitting Algorithm

The idea of this algorithm (Fan and Li, 2004) is to iteratively estimate the nonparametric part $\alpha(\cdot)$ and the parametric part $\boldsymbol{\beta}$. For the ease of presentation, the same notations for y in the following procedure as the difference based method may have different meanings.

1. Specify an initial value of $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}^{(0)}$. It can be obtained by the difference based method introduced above. Thus model (2.57) is transformed to

$$y_i^* = \alpha(u_i) + \varepsilon_i, \tag{2.63}$$

where $y_i^* = y_i - \widehat{\boldsymbol{\beta}}^{(0)\mathrm{T}}\mathbf{x}_i$. Then $\widehat{\alpha}^{(0)}(\cdot)$, the initial estimator of $\alpha(\cdot)$, can be obtained in the same fashion as step 4 in the difference based method.

2. Based on the initial estimator of $\alpha(\cdot)$, we modify the original model (2.57) as

$$y_i^{**} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i, \tag{2.64}$$

where $y_i^{**} = y_i - \widehat{\alpha}^{(0)}(u_i)$. Then $\boldsymbol{\beta}$ can again be estimated by the least square method based on the reduced model (2.64):

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}^{**}. \tag{2.65}$$

3. Iterate the above two steps until convergence.

The back fitting algorithm is a modified estimation procedure of difference based method; it does not put strong assumptions on $\alpha(\cdot)$ and u as difference based method, and still easy to implement. The algorithm, however, is still not efficient in terms of the variance of the estimation for $\boldsymbol{\beta}$.

2.4.3 Profile Least Square and Profile Likelihood Approach

In the partially linear model setting, profile least square method and profile likelihood method (Fan and Li, 2004) are identical. The procedure is described as follows.

1. First estimate $\alpha(\cdot)$ by treating $\boldsymbol{\beta}$ as a nuisance parameter. Specifically, define $y_i^* = y_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$, where $\boldsymbol{\beta}$ is only a dummy variable. This is one difference between profile least square method and back fitting algorithm – the latter

specifies a concrete vector value as the initial estimator of $\boldsymbol{\beta}$. But the transformed model has the same form (2.63) as back fitting algorithm. Apply any linear smoother S_h to estimate $\alpha(\cdot)$ by

$$\widehat{\boldsymbol{\alpha}}_\beta \equiv (\widehat{\alpha}_\beta(u_1), \ldots, \widehat{\alpha}_\beta(u_n))^{\mathrm{T}} = S_h \mathbf{y}^*.$$

2. With $\widehat{\boldsymbol{\alpha}}_\beta = S_h \mathbf{y}^* = S_h(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, the estimation problem now is equivalent to minimizing

$$Q(\boldsymbol{\beta}) \equiv \|\mathbf{y} - \boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|(i - S_h)\mathbf{y} - (I - S_h)\mathbf{X}\boldsymbol{\beta}\|^2,$$

where I is the identity matrix. It is easy to see

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{**\mathrm{T}}\mathbf{X}^{**})^{-1}\mathbf{X}^{**\mathrm{T}}\mathbf{y}^{**},$$

with $\mathbf{X}^{**} = (I - S_h)\mathbf{X}$ and $\mathbf{y}^{**} = (I - S_h)\mathbf{y}$.

3. We can refit the nonparametric function $\alpha(\cdot)$ by plugging in the updated estimate $\widehat{\boldsymbol{\beta}}$ above by

$$\widehat{\boldsymbol{\alpha}} = S_h(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}).$$

In practice, however, it is difficult to derive the explicit form of S_h. A popular solution is to consider $S_h(\mathbf{y})$ as $E(\mathbf{y}|u)$, the conditional mean of $\mathbf{y}$ given u. Then it can be estimated by any nonparametric estimating procedure like the local linear estimation. The profile method provide the efficient estimates of $\boldsymbol{\beta}$.

Chapter 3

Statistical Methods for Ultrahigh Dimensional Varying Coefficient Models

This chapter considers the problem of feature screening and variable selection for ultrahigh dimensional varying coefficient models. A new conditioning-correlation independence screening procedure (CCIS) is proposed specifically for these models. The ranking consistency and sure screening property of CCIS are established, and they are verified empirically through the simulation studies. Furthermore, the iterative conditioning-correlation screening procedure is developed to enhance the finite sample performance. In the real data example, a two-stage approach for varying coefficient models is derived – firstly CCIS is applied to reduce the ultrahigh dimensionality to the scale under sample size, and secondly several penalized regression techniques are modified for varying coefficient models to further select important variables as well as to estimate the coefficient functions.

3.1 Introduction

The variable selection for ultrahigh dimensional varying coefficient models is attractive to researchers yet challenging to deal with, because these models assume the number of predictors to be much larger than the sample size, and the re-

gression coefficients to change over subjects characterized by a certain covariate. For example, in genetic research, one might be interested in selecting the significant single-nucleotide polymorphisms (SNPs) for explaining the body mass index (BMI), whose effects may depend on the age of each individual. In this case, millions of SNPs are studied, leading to an ultrahigh dimensional problem; to guarantee the changing effect of SNPs, their coefficients are rendered to be nonparametric functions of age. Therefore, the variable selection for such models aims to recover the sparse nonzero coefficient functions corresponding to the significant predictors.

Some variable selection methods have been developed for varying coefficient models in literature. Li and Liang (2008) used a generalized likelihood ratio test to select significant nonparametric components based on SCAD penalty (Fan and Li, 2001). Wang et al. (2008) presented a regularized estimation procedure based on the basis function approximations and the SCAD penalty, which can simultaneously select significant variables and estimate the nonzero smooth coefficient functions. Wang and Xia (2009) proposed a shrinkage method incorporating local polynomial smoothing (Fan and Gijbels, 1996) and LASSO penalized regression (Tibshirani, 1996), among others. Nevertheless, most existing techniques for varying coefficient models require fixed model dimension, thus they cannot be applied to the ultrahigh dimensionality.

To deal with the ultrahigh dimensionality, one appealing method is the two-stage approach. First, a computationally efficient screening procedure is applied to reduce the ultrahigh dimensionality to a moderate scale under sample size; second, the final sparse model is recovered by a regularization method, such as the penalized regression approach. In the first stage, several screening techniques were advocated for various models. Fan and Lv (2008) showed that the sure independence screening (SIS) possesses sure screening property in the linear model setting. Hall and Miller (2009) extended the methodology from linear models to nonlinear models using generalized empirical correlation learning, but it is not trivial to choose a reasonable transformation function. Fan and Song (2010) extended SIS to the generalized linear model by ranking the maximum marginal likelihood estimates. Fan, Feng and Song (2011) explored the feature screening technique for ultrahigh dimensional additive models, by ranking the magnitude of spline approx-

imations of the nonparametric components. Zhu, Li, Li and Zhu (2011) proposed a sure independence ranking and screening procedure to select important predictors in the multi-index model setting. Li, Zhong and Zhu (2012) studied a model-free screening procedure using distance correlation learning, which can deal with multiple response problem and grouped predictors. However, much less have been done for varying coefficient models. Therefore, in this paper, we develop a novel feature screening method specifically for these models to reduce dimensionality, and a two-stage approach based on this screening technique to select the final model and to depict the effect of significant predictors.

The main focus of the paper is the feature screening technique. Notice that the varying coefficient models are indeed linear models conditioning on the depending covariate (denoted by u afterwards), where SIS can be applied by ranking the magnitude of pearson correlations. This motivates us to define the conditional correlation for varying coefficient models parallel to the pearson correlation for linear models, except that the expectations and variances are now substituted by the conditional expectations and conditional variances. Subsequently by averaging out the effect of u, an unconditioned-squared correlation between each predictor and the response is obtained, whose ranks can represent the importance of the predictor. The whole screening procedure is referred to as the conditioning-correlation independence screening (CCIS).

Several desirable theoretical properties of CCIS are systematically studied. We show that CCIS possesses the ranking consistency property (Zhu, Li, Li, and Zhu, 2011), which means with probability tending to 1, the important predictors rank before the unimportant ones. In addition, CCIS satisfies the sure screening property (Fan and Lv, 2008) for varying coefficient models, which guarantees the probability that the model chosen by CCIS includes the true model tends to 1 as the sample size goes to infinity. Monte Carlo simulation studies are conducted to empirically verify these theoretical advantages.

3.2 Methodology

Suppose $\{(u_i, \mathbf{x}_i, y_i), i = 1, \ldots, n\}$ is a random sample from the varying coefficient model:

$$y = \beta_0(u) + \mathbf{x}^T \boldsymbol{\beta}(u) + \varepsilon. \tag{3.1}$$

In the model (3.1), y is the response and $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ is the p-dimensional predictor. $\boldsymbol{\beta}(u) = \{\beta_1(u), \ldots, \beta_p(u)\}^{\mathrm{T}} \in \mathbb{R}^p$ is the coefficient vector, where $\beta_j(u)$'s are unknown smooth functions of the depending variable $u \in \mathbb{R}^1$. The random noise $\varepsilon \in \mathbb{R}^1$ satisfies $E(\varepsilon|\mathbf{x}, u) = 0$ almost surely.

3.2.1 Conditional Correlations and Their Estimations

To define the conditional correlation between each predictor and the response, first consider the conditional covariance between two generic variables z and w given u:

$$\mathrm{cov}(z, w|u) = E(zw|u) - E(z|u)E(w|u).$$

Then the conditional correlation between x_j, $j = 1, \ldots, p$, and y given u is defined as

$$\rho(x_j, y|u) = \frac{\mathrm{cov}(x_j, y|u)}{\sqrt{\mathrm{cov}(x_j, x_j|u)\mathrm{cov}(y, y|u)}}. \tag{3.2}$$

Elementary calculation shows that $\rho(x_j, y|u)$ is essentially a function of five conditional means $E(y|u)$, $E(y^2|u)$, $E(x_j|u)$, $E(x_j^2|u)$ and $E(x_j y|u)$, which can be estimated through nonparametric smoothing techniques. In this paper, the local constant estimation (Fan and Gijbels, 1996) is applied due to its parsimony and desirable properties. The conditional mean of a random scalar z given u is estimated based on the sample $\{(u_i, z_i),\ i = 1, \ldots, n\}$ by the weighted average

$$\widehat{E}(z|u) = \sum_{i=1}^{n} \omega_i(u) z_i. \tag{3.3}$$

The u-dependent weight $\omega_i(u)$ is the normalized kernel function

$$\omega_i(u) = \frac{K_h(u_i - u)}{\sum_{i=1}^{n} K_h(u_i - u)}, \tag{3.4}$$

where $K_h(t) = h^{-1}K(t/h)$, $K(t)$ is a kernel function, and h is the tuning bandwidth selected to minimize the mean integrated squared error of the estimator.

In our setting, the five conditional mean estimations are accomplished by assigning z to be y, y^2, x_j, x_j^2 and $x_i y$. Then the conditional covariance and conditional correlation are naturally estimated by

$$\begin{aligned} \widehat{\mathrm{cov}}(x_j, y|u) &= \widehat{E}(x_j y|u) - \widehat{E}(x_j|u)\widehat{E}(y|u) \\ \widehat{\rho}(x_j, y|u) &= \frac{\widehat{\mathrm{cov}}(x_j, y|u)}{\sqrt{\widehat{\mathrm{cov}}(x_j, x_j|u)\widehat{\mathrm{cov}}(y, y|u)}}. \end{aligned} \tag{3.5}$$

In practice, various nonparametric smoothing methods can be used to estimate the aforementioned five conditional means. However, since the conditional variance and covariance are estimated through conditional means, the mean estimators are required to guarantee the following elementary inequalities:

$$\widehat{\mathrm{cov}}(z, w|u) \leq \widehat{\mathrm{cov}}(z, z|u) \cdot \widehat{\mathrm{cov}}(w, w|u) \quad \text{and} \quad \widehat{\mathrm{cov}}(z, z|u) \geq 0. \tag{3.6}$$

Furthermore, the bandwidths of the smoothing techniques for estimating all the five conditional means need to be the same to ensure (3.6).

3.2.2 Conditioning-Correlation Independence Screening

In this section, we study the methodology and implementation of CCIS for the ultrahigh dimensional varying coefficient models. In the ultrahigh dimensional context, the dimension p of the predictor $\mathbf{x}$ is allowed to increase at an exponential rate of the sample size n, while the number of x_j's that are truly important to y is assumed to be small relative to n.

The substantive idea brought by SIS (Fan and Lv, 2008) is using the pearson correlation between each predictor x_j and y as the marginal utility to filter out the predictors with weak signals under the linear model setting. Since the varying coefficient model (3.1) becomes a linear model after conditioned on u, we are inspired

to develop a feature screening criterion analogous to the pearson correlation, based on the conditioning-correlation learning.

Notice that $\rho(x_j, y|u)$ in (3.2) depends on the value of u. For each predictor x_j, we can obtain all the n sample conditional correlations $\rho(x_j, y|u_i)$, $i = 1, \ldots, n$, by taking u to be u_i, $i = 1, \ldots, n$. However, a unified score of each x_j is needed to represent the importance of x_j. To average out the effect of u, we define the unconditioned-squared correlation ρ_j^* between x_j and y and its estimate $\widehat{\rho}_j^*$ as

$$\rho_j^* = \frac{1}{n}\sum_{i=1}^{n} \rho^2(x_j, y|u_i), \quad \text{and} \quad \widehat{\rho}_j^* = \frac{1}{n}\sum_{i=1}^{n} \widehat{\rho}^2(x_j, y|u_i). \tag{3.7}$$

We take the square of each sample conditional correlation to avoid the counteraction between positive and negative correlation effects. The CCIS procedure requires to sort $\widehat{\rho}_j^*$, $j = 1, \ldots, p$ in a decreasing order, and yields the screened submodel

$$\widehat{\mathcal{M}} = \{j : 1 \le j \le p,\ \widehat{\rho}_j^* \text{ ranks among the first } d\},$$

where the submodel size d is taken to be smaller than the sample size n. Thus the ultrahigh dimensionality p is reduced to the moderate scale d. To determine d, the hard threshold $d = [n/\log(n)]$ is often used in literature, where $[a]$ refers to the integer part of a. For CCIS, however, we modify the threshold as $d = [n^{4/5}/\log(n^{4/5})]$ because the effective sample size becomes nh instead of n due to the nonparametric estimation procedure, and the optimal bandwidth h has the rate $O(n^{-1/5})$. In practice, one can always get a more conservative submodel with size $d = k[n^{4/5}/\log(n^{4/5})]$, $k = 2, 3, \ldots$, to enlarge the probability of including the truly important predictors.

3.3 Theoretical Properties

In this section, we study the theoretical properties of CCIS. To begin with, some notations and regularity conditions are introduced.

3.3.1 Notations and Regularity Conditions

For the univariate depending variable u, suppose $\mathbb{U} = [a, b]$ is its bounded support, where a and b are finite constants. Define the true model index set and its complement to be

$$\begin{aligned} \mathcal{M}_* &= \{1 \le j \le p : \ \beta_j(u) \neq 0 \text{ for some } u \in \mathbb{U}\}, \\ \mathcal{M}_*^c &= \{1 \le j \le p : \ \beta_j(u) \equiv 0 \text{ for any } u \in \mathbb{U}\}. \end{aligned}$$

Denote $\rho_{j0}^* = E_u \rho^2(x_j, y|u)$ to be the population version of the unconditioned-squared correlation, then by definition, $\rho_{j0}^* = E_u \rho_j^*$. The following regularity conditions are imposed.

(C1) The density function $f(u)$ of u satisfies $\sup_{u\in\mathbb{U}} f(u) \le M_1$, $\sup_{u\in\mathbb{U}} |f'(u)| \le M_2$ and $sup_{u\in\mathbb{U}}|f''(u)| \le M_3$ for some finite constants M_1, M_2 and M_3, where $f'(u)$ and $f''(u)$ are the first and second order derivatives of $f(u)$.

(C2) The kernel function $K(\cdot)$ is bounded uniformly: $\sup_{u\in\mathbb{U}} |K(u)| \le M_4 < \infty$. And $\mu_1(K) = \int tK(t)dt < \infty$, $\mu_2(K) = \int t^2 K(t)dt < \infty$.

(C3) The random variables x_j and y satisfy the sub-exponential tail probability uniformly in p, i.e. there exists $s_0 > 0$, such that for $0 < s < s_0$,

$$\begin{aligned} &\sup_{u\in\mathbb{U}} \max_{1\le j\le p} E\{\exp(2sx_j^2|u)\} < \infty, \quad \sup_{u\in\mathbb{U}} E\{\exp(2sy^2|u)\} < \infty, \\ &\sup_{u\in\mathbb{U}} \max_{1\le j\le p} E\{\exp(2sx_j y|u)\} < \infty. \end{aligned}$$

(C4) The five conditional means $E(y|u)$, $E(y^2|u)$, $E(x_j|u)$, $E(x_j^2|u)$ and $E(x_j y|u)$ have finite first and second order derivatives uniformly in $u \in \mathbb{U}$, and the conditional variances of x_j and y are uniformly positive in $u \in \mathbb{U}$, i.e.

$$\inf_{u\in\mathbb{U}} \min_{1\le j\le p} \text{var}(x_j|u) > 0, \quad \inf_{u\in\mathbb{U}} \text{var}(y|u) > 0.$$

Condition (C1) and (C2) put mild constraints on the density function $f(u)$ of u and the kernel function $K(\cdot)$, which can be guaranteed by most commonly used

distributions and kernels. Moreover, (C2) implies that $K(\cdot)$ has every finite moment, i.e. $E(|K(u)|^r) < \infty, \ \ \forall r > 0$. Condition (C3) is relatively strong and only used to facilitate the technical proofs. Furthermore, if condition (C3) holds, the moment generating functions of x_i's and y are also finite, resulting in the existence of finite moments in any order, especially,

$$\begin{aligned} &\sup_{u\in\mathbb{U}} \max_{1\le j\le p} E(x_j|u) < \infty, \ \ \sup_{u\in\mathbb{U}} E(y|u) < \infty, \ \ \sup_{u\in\mathbb{U}} \max_{1\le j\le p} E(x_j y|u) < \infty, \\ &\sup_{u\in\mathbb{U}} \max_{1\le j\le p} E(x_j^2|u) < \infty, \ \ \sup_{u\in\mathbb{U}} E(y^2|u) < \infty. \end{aligned} \tag{3.8}$$

Condition (C4) puts more constraints on the five conditional means in addition to (3.8), and the nonnegative variances guarantees that the conditional correlation is well defined.

3.3.2 Ranking Consistency Property

In this section we study the ranking consistency property (Zhu, Li, Li and Zhu, 2011) of CCIS, which ensures that with an overwhelming probability, all the truly important predictors rank above the unimportant ones. Another condition is required to ensure this property.

(C5)

$$\liminf_{n\to\infty} \{ \min_{j\in\mathcal{M}_*} \rho_{j0}^* - \max_{j\in\mathcal{M}_*^c} \rho_{j0}^* \} > 0.$$

Condition (C5) provides a clear separation between the important and unimportant predictors in terms of the population level unconditioned-squared correlation ρ_{j0}^*. This condition rules out the situation when certain unimportant predictors have large ρ_{j0}^*'s and rank high just because they are highly correlated with the true ones, while some important predictors with weaker signals are left unselected.

Theorem 5. *(Ranking Consistency Property) Under conditions (C1)-(C5), for* $p = o\{\exp(an)\}$ *with some* $a > 0$*, we have*

$$\liminf_{n\to\infty} \{ \min_{j\in\mathcal{M}_*} \widehat{\rho}_j^* - \max_{j\in\mathcal{M}_*^c} \widehat{\rho}_j^* \} > 0 \ \ \textit{in probability.}$$

Theorem 1 states that the truly important predictors should have larger $\widehat{\rho}^*$'s than the unimportant ones, leading to the model selection consistency of CCIS for ultrahigh dimensional varying coefficient models.

3.3.3 Sure Screening Property

Next we develop the sure screening property (Fan and Lv, 2008) of CCIS, which guarantees that the true model has an overwhelming probability to be included in the chosen model. This property relies further on the following condition in addition to (C1) – (C4):

(C6) There exist some $c_3 > 0$ and $0 \leq \kappa < 1/5$, such that

$$\min_{j \in \mathcal{M}_*} \rho^*_{j0} \geq 2c_3 n^{-\kappa}.$$

Condition (C6) requires the true unconditioned-squared correlations between the important x_j's and y to be bounded away from 0. However, the lower bound depends on n, thus ρ^*_{j0}'s are allowed to go to 0 in the asymptotic sense. This condition rules out the situation where the predictors are marginally uncorrelated with y but jointly correlated.

Theorem 6. *(Sure Screening Property) Under condition (C1)-(C4), we have*

$$P\left(\max_{1 \leq j \leq p} |\widehat{\rho}^*_j - \rho^*_{j0}| > c_3 \cdot n^{-\kappa}\right) \leq O\{np \exp(-n^{\frac{1}{5}-\kappa}/\xi)\},$$

and under one more condition (C6),

$$P(\mathcal{M}_* \subset \widehat{\mathcal{M}}) \geq 1 - O\{ns_n \exp(-n^{\frac{1}{5}-\kappa}/\xi)\},$$

where ξ is some positive constant determined by c_3, and s_n is the cardinality of $\mathcal{M}_*$, which is sparse and may vary with n.

The proofs of Theorem 5 and Theorem 6 are studied in next chapter.

3.4 Monte Carlo Simulations

To evaluate the performance of CCIS, we conduct several Monte Carlo simulations. For each example below, the univariate index variable u_i and the covariate $\mathbf{x}_i$ are generated i.i.d in the following fashion:

$$(u_i^*, \mathbf{x}_i) \sim MVN(\mathbf{0}, \Sigma), \quad \text{where} \quad \Sigma_{jk} = \rho^{|j-k|}, \quad j, k = 1, \ldots, p+1,$$
$$u_i = \Phi(u_i^*), \quad i = 1, \ldots, n,$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard Normal distribution. Then $u_i \sim \mathrm{U}(0,1)$ and is correlated with the covariate $\mathbf{x}_i$. And the random noise $\varepsilon_i \sim N(0,1)$ independently. We set $p = 1000$, $n = 200$, and repeat the experiment 100 times. In each of the 100 simulations, we choose d to be d_0, $2d_0$ and $3d_0$, where $d_0 = [n^{4/5}/\log(n^{4/5})]$ according to the hard threshold introduced in section 2.2, hence we reduce the dimensionality from $p = 1000$ to d. For each example, two models are considered with $\rho = 0.8$ and $\rho = 0.4$, respectively.

The following criteria are used to assess the performance of the screening procedures:

- p_j: The proportion of the jth predictor being selected into the model with size d.
- p_a: The proportion that all active predictors are selected into the model.
- $rank_j$: The ranking of $\widehat{\rho}_j^*$ in a decreasing order.
- M: the minimum size of the model which contains all the true predictors. We report the 5%, 25%, 50%, 75% and 95% quantiles of M from 100 simulations.

The above criteria can be used to empirically verify two theoretical properties of a screening procedure. Fan and Lv (2008) proposed the sure screening property, which states that the probability for the screened submodel $\mathcal{M}_\gamma$ to contain the true model $\mathcal{M}_*$ tends to one when the sample size goes to infinity, that is, p_j and p_a are close to one when the submodel size d is sufficiently large. Moreover, the ranking consistency refers to the property that the screening scores of true predictors rank among the top, hence a reasonable screening procedure is expected to guarantee

that $rank_j$ for the true predictors and the minimum model size M are small. The following two examples are designed to assess the performance of CCIS and to compare it with SIS.

Example 1. This example is a linear regression model where SIS is expected to work well. The nonzero components of the coefficient $\boldsymbol{\beta}(u)$ is generated as follows:

$$\beta_2(u) = 1, \quad \beta_{100}(u) = 0.8, \quad \beta_{400}(u) = 1.2, \quad \beta_{600}(u) = -0.8, \quad \beta_{1000}(u) = -1.2.$$

Table 3.1 depicts the individual selecting rates $p_j,\ j = 2, 100, 400, 600, 1000$, and the overall selecting rate p_a given different d's for both SIS and CCIS, by which the sure screening property is verified. And ranking consistency is illustrated in Table 3.2 and Table 3.3, where the rankings of $\widehat{\rho}_j^*$'s for the truly active predictors and the summary of minimum model sizes M are reported. From the three tables one can see that the performances of SIS and CCIS are similar for this example, indicating that our method works well in linear model setting. Yet if the underlying model is known to be linear, SIS is preferred due to its computational efficiency.

Table 3.1. The proportions p_j and p_a for Example 1.

	SIS						CCIS					
d	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
	$\rho = 0.4$											
d_0	0.99	0.94	1.00	1.00	1.00	0.94	0.99	0.93	1.00	0.97	1.00	0.93
$2d_0$	1.00	0.99	1.00	1.00	1.00	0.99	0.99	0.96	1.00	0.99	1.00	0.96
$3d_0$	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.97	1.00	1.00	1.00	0.97
	$\rho = 0.8$											
d_0	0.95	0.82	1.00	0.85	1.00	0.82	0.87	0.87	1.00	0.88	1.00	0.87
$2d_0$	0.99	0.95	1.00	0.97	1.00	0.95	0.99	0.96	1.00	0.97	1.00	0.96
$3d_0$	0.99	0.98	1.00	0.99	1.00	0.98	0.99	0.98	1.00	0.99	1.00	0.98

Table 3.2. $rank_j$ of each true predictor x_j for Example 1.

	SIS					CCIS				
ρ	x_2	x_{100}	x_{400}	x_{600}	x_{1000}	x_2	x_{100}	x_{400}	x_{600}	x_{1000}
0.4	3.27	5.75	1.76	4.73	1.77	4.02	8.32	1.75	5.21	1.80
0.8	5.44	15.48	2.13	11.55	2.00	9.66	13.87	1.82	10.80	1.74

Table 3.3. The minimum model size M for Example 1.

	SIS					CCIS				
ρ	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.4	5.00	5.00	5.00	6.25	17.05	5.0	5.0	5.0	8.0	38.1
0.8	7.00	11.00	14.00	18.25	41.25	7.00	11.00	15.00	19.25	50.00

In addition, comparing the two models with different ρ's for both screening methods, the ones with $\rho = 0.4$ perform slightly better than those with $\rho = 0.8$, which is because when the predictors are highly correlated, say $\rho = 0.8$, the screening score $\widehat{\rho}^*$'s of some insignificant variables are inflated by their adjacent significant variables, hence these unimportant predictors may be selected due to their strong correlation with the active predictors. This issue can be fixed by the iterative algorithm discussed in section 5.

Example 2. In this example we consider a varying coefficient model where the coefficients are smooth functions of the index variable u, which are defined by

$$\beta_2(u) = 2I(u > 0.4), \quad \beta_{100}(u) = 1 + u, \quad \beta_{400}(u) = (2 - 3u)^2$$
$$\beta_{600}(u) = 2\sin(2\pi u), \quad \beta_{1000}(u) = \exp\{u/(u + 1)\}, \quad \text{other } \beta(u)\text{'s are 0.}$$

The results are shown in Table 3.4, Table 3.5, and Table 3.6. From the outputs, the SIS does not perform well for this model. The individual selecting rates p_j and the rankings of $\widehat{\rho}_j^*$ show that it fails to detect $\beta_{600}(u)$. The reason is that $\beta_{600}(u) = 2\sin(2\pi u)$ has mean 0 when $u \sim \mathrm{U}(0, 1)$, making the sample Pearson correlation between x_{600} and response y close to 0, although x_{600} is functionally important when modeling y. CCIS works well for this model, where the individual selection rate p_j for each significant variable and the overall selecting rate p_a are close to one. Moreover, the $\widehat{\rho}^*$'s corresponding to the true predictors indeed rank on the top, consequently the minimum model size M is small. Therefore, the sure screening property and ranking consistency of the proposed method are verified through the three tables. Again, both methods work better for the models with $\rho = 0.4$ than with $\rho = 0.8$ for the same reason as Example 1.

Table 3.4. The proportions p_j and p_a for Example 2.

	SIS						CCIS					
d	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
	$\rho = 0.4$											
d_0	0.99	1.00	0.92	0.02	1.00	0.02	1.00	1.00	1.00	1.00	0.99	0.99
$2d_0$	0.99	1.00	0.98	0.05	1.00	0.05	1.00	1.00	1.00	1.00	1.00	1.00
$3d_0$	1.00	1.00	0.98	0.07	1.00	0.07	1.00	1.00	1.00	1.00	1.00	1.00
	$\rho = 0.8$											
d_0	0.89	1.00	0.86	0.00	1.00	0.00	0.95	1.00	0.97	0.98	0.98	0.95
$2d_0$	0.99	1.00	0.94	0.01	1.00	0.01	1.00	1.00	1.00	1.00	1.00	1.00
$3d_0$	0.99	1.00	0.98	0.01	1.00	0.01	1.00	1.00	1.00	1.00	1.00	1.00

Table 3.5. $rank_j$ of each true predictor x_j for Example 2.

	SIS					CCIS				
ρ	x_2	x_{100}	x_{400}	x_{600}	x_{1000}	x_2	x_{100}	x_{400}	x_{600}	x_{1000}
0.4	3.75	1.46	7.28	436.60	2.37	2.68	2.03	3.48	3.81	3.54
0.8	8.55	1.63	10.89	499.53	3.26	6.21	1.94	6.17	6.20	4.16

Table 3.6. The minimum model size M for Example 2.

	SIS					CCIS				
ρ	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.4	38.80	197.00	416.50	685.75	874.25	5.0	5.0	5.0	5.0	7.0
0.8	89.30	204.75	495.00	769.50	952.60	5.0	7.0	10.0	13.0	21.1

Example 3. In the previous example, the significant variables can be viewed as independent with spatial-power correlation structure. In this example, however, we set $x_1, x_2, \ldots, x_5$ to be important predictors, which are highly correlated, and the associated coefficient functions are constructed in the same fashion as Example 2:

$$\beta_1(u) = 2I(u > 0.4), \quad \beta_2(u) = 1 + u, \quad \beta_3(u) = (2 - 3u)^2$$
$$\beta_4(u) = 2\sin(2\pi u), \quad \beta_5(u) = \exp\{u/(u+1)\}, \quad \text{other } \beta(u)\text{'s are 0.}$$

From the outputs given by Table 3.7, Table 3.8, and Table 3.9, both SIS and CCIS work well for this model. The reason is that since the significant variables are

highly correlated, the values of their marginal utilities are inflated by the adjacent predictors. In other words, the important predictors become "more significant" because they are also correlated with other important variables, thus although their own signals might be weak, they can still be detected due to the strong signal of their neighbors. We also notice that when $\rho = 0.8$, both methods work better than when $\rho = 0.4$, resulting from the fact that stronger correlation between predictors will reinforce the marginal utility more.

Table 3.7. The proportions p_j and p_a for Example 3.

	SIS						CCIS					
d	p_1	p_2	p_3	p_4	p_5	p_a	p_1	p_2	p_3	p_4	p_5	p_a
	$\rho = 0.4$											
d_0	1.00	1.00	1.00	0.95	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00
$2d_0$	1.00	1.00	1.00	0.98	1.00	0.98	1.00	1.00	1.00	1.00	1.00	1.00
$3d_0$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\rho = 0.8$											
d_0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$2d_0$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$3d_0$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 3.8. $rank_j$ of each true predictor x_j for Example 3.

	SIS					CCIS				
ρ	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
0.4	2.71	1.08	2.41	6.39	3.93	2.45	1.11	2.63	4.25	4.56
0.8	3.02	1.13	2.03	3.94	4.88	3.48	1.11	2.01	3.80	4.60

Table 3.9. The minimum model size M for Example 3.

	SIS					CCIS				
ρ	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.4	5.0	5.0	5.0	5.0	16.1	5.0	5.0	5.0	5.0	5.0
0.8	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0

3.5 Iterative Feature Screening for Varying Coefficient Models

As is known, ordinary feature screening procedures may fail to detect some active covariates due to the association among the potential predictors, illustrated by the following example. To fix this issue we propose an iterative screening method called the iterative conditioning-correlation independence screening (ICCIS). The procedure for choosing d predictors comprises the following steps:

1. Apply CCIS to each column of X, where X is the $n \times p$ matrix containing all the candidate covariates. Select d_1 predictors with the highest d_1 $\widehat{\rho}^*$ values, denoted by $\mathcal{X}_1 = \{x_{1_1}, \ldots, x_{1_{d_1}}\}$, where $d_1 \leq d$.

2. Denote $X_s = (x_{1_1}, \ldots, x_{1_{d_1}})$ to be the $n \times d_1$ matrix of selected predictors, and X_r to be the complement of X_s with dimension $n \times (p - d_1)$. Then compute the projection of X_r onto the orthogonal complement space of X_s, $X_{proj} = (I_n - X_s(X_s^T X_s)^{-1} X_s^T) X_r$.

3. Apply CCIS to each column of X_{proj}, and select another d_2 predictors $\mathcal{X}_2 = \{x_{2_1}, \ldots, x_{2_{d_2}}\}$ in the same fashion as step 1, where $d_1 + d_2 \leq d$.

4. Repeat 2. and 3. until the kth step where $d_1 + d_2 + \ldots + d_k \geq d$. And the selected predictors are $\mathcal{X}_1 \bigcup \mathcal{X}_2 \bigcup \ldots \bigcup \mathcal{X}_k$.

In the algorithm, $d_1, \ldots, d_k$ are chosen by users according to the model complexity. Two steps are often sufficient in practice to achieve satisfactory result. If $d_1 = d$, the procedure becomes CCIS. To illustrate how ICCS outperforms CCIS in some cases, we consider the example below.

Example 4. This example demonstrates one possible issue of CCIS that when some covariates are jointly active in the presence of other covariates but marginally unassociated with the response, the CCIS may fail to detect them. We generate $(\mathbf{x}_i, u_i)$ as in section 3, and set $\rho = 0.4$. The coefficient functions have the following forms:

$$\beta_1(u) = 2 + \cos\left\{\frac{\pi(6u-5)}{3}\right\}, \quad \beta_2(u) = 3 - 3u, \quad \beta_3(u) = -2 + \frac{(2-3u)^3}{4}$$

$$\beta_4(u) = \sin(\frac{9u^2}{2}) + 1, \quad \beta_5(u) = \exp\{3u^2/(3u^2+1)\}, \quad \beta_6(u) = \ldots = \beta_{1000}(u) \equiv 0$$

In this model setting, the conditional correlation between x_3 and y is small for $u \sim \mathrm{U}(0,1)$, but x_3 is still jointly correlated with y. Table 3.10 and Table 3.11 compare the performances of CCIS and ICCIS for this model setting in terms of sure screening property and ranking consistency. The rankings of $\widehat{\rho}^*$'s are not reported because in each iteration, the $\widehat{\rho}^*$'s of the remaining predictors will change after the previously chosen predictors are removed. From the tables one can see that the ICCIS procedure is able to select x_3 which only jointly contributes to modeling y and is easily overlooked by CCIS.

Table 3.10. The proportions p_j and p_a for Example 4.

	CCIS						ICCIS					
d	p_1	p_2	p_3	p_4	p_5	p_a	p_1	p_2	p_3	p_4	p_5	p_a
d_0	1.00	1.00	0.36	1.00	1.00	0.36	1.00	1.00	1.00	0.99	1.00	0.99
$2d_0$	1.00	1.00	0.47	1.00	1.00	0.47	1.00	1.00	1.00	1.00	1.00	1.00
$3d_0$	1.00	1.00	0.59	1.00	1.00	0.59	1.00	1.00	1.00	1.00	1.00	1.00

Table 3.11. The minimum model size M for Example 4.

	5%	25%	50%	75%	95%
CCIS	6.0	12.0	37.5	193.0	609.0
ICCIS	5.95	11.00	11.00	11.00	11.00

3.6 Two-Stage Approach and the Application to Framingham Heart Study

In this section, we analyze a GWAS data set from Framingham Heart Study (FHS). FHS is a cardiovascular study beginning in 1948 under the direction of the National Heart, Lung and Blood Institute (NHLBI), by recruiting originally 5,209 men and women between the ages of 30 and 62 from the town of Framingham, Massachusetts. Recently, $550,000$ SNPs from 24 chromosomes have been genotyped from the cohort study (Jaquish 2007) with 418 males and 559 females. The

SNPs with rare allele frequency $< 10\%$ were removed from the analysis, leaving $349,985$ non-rare SNPs of interest. In addition, the body mass indexes (BMI) and the ages of the subjects are measured. The goal of our analysis is to detect significant non-rare SNPs that are associated with BMI.

3.6.1 Statistical Model

In this chapter, we only focus on the baseline measurements of Age and BMI. One may argue that the effect of SNPs on BMI might change with age, hence the following varying coefficient model is appropriate, with baseline Age being the univariate index variable u and baseline BMI being the response y:

$$y = \beta_0(u) + \mathbf{x}_a^T \boldsymbol{\beta}_a(u) + \mathbf{x}_d^T \boldsymbol{\beta}_d(u) + \varepsilon, \tag{3.9}$$

where ε is the random noise that is assumed to follow $N(0, \sigma^2)$; The smooth functions $\boldsymbol{\beta}_a(u) = (\beta_{a1}(u), \ldots, \beta_{ap}(u))^{\mathrm{T}}$ and $\boldsymbol{\beta}_d(u) = (\beta_{d1}(u), \ldots, \beta_{dp}(u))^{\mathrm{T}}$ are the additive and dominant effects of the non-rare SNPs with $p = 349,985$; $\mathbf{x}_a$ and $\mathbf{x}_d$ are the indicator vectors of the additive and dominant effects of SNPs. More explicitly, consider a SNP $\mathbf{A}$ with two alleles A and a, generating three genotypes AA, Aa and aa, then the jth element of $\mathbf{x}_a$ and $\mathbf{x}_d$ are defined as

$$\mathbf{x}_{aj} = \begin{cases} 2, & \text{if the genotype of SNP } j \text{ is } AA \\ 1, & \text{if the genotype of SNP } j \text{ is } Aa \\ 0, & \text{if the genotype of SNP } j \text{ is } aa, \end{cases}$$

$$\mathbf{x}_{dj} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } Aa \\ 0, & \text{if the genotype of SNP } j \text{ is } AA \text{ or } aa. \end{cases}$$

The model (3.9) can then be unified in the following form:

$$y = \mathbf{x}^T \boldsymbol{\beta}(u) + \varepsilon, \tag{3.10}$$

where $\mathbf{x} = (\mathbf{x}_a^{\mathrm{T}}, \mathbf{x}_d^{\mathrm{T}})^{\mathrm{T}}$, and $\boldsymbol{\beta}(u) = (\boldsymbol{\beta}_a(u)^{\mathrm{T}}, \boldsymbol{\beta}_d(u)^{\mathrm{T}})^{\mathrm{T}}$. Since we consider both additive and dominant effects the total dimension becomes $2p = 699,970 >> n = 977$,

resulting in ultrahigh dimensionality. To deal with these types of problems, we use the aforementioned two-stage approach: In the first stage, a feature screening method is applied to reduce dimensionality, and in the second stage, some regularization method is implemented to select important variables and estimate the coefficients. Various techniques are used to reduce dimensionality of SNPs in the first stage, such as the single SNP analysis, the preconditioning technique, etc. Yet these dimension reduction techniques fail to consider the varying coefficient model structure. In addition,not much has been done in literature about the variable selection for varying coefficient models in the second stage. To address these issues, we first use CCIS to reduce dimensionality, by which the varying coefficient structure is taken into account; And in the second stage, we apply several penalized regression procedures to choose significant variables as well as to estimate the coefficient functions.

3.6.2 Two-Stage Approach

Stage 1: feature screening procedure

Recall the varying coefficient model (3.10). The predictor $\mathbf{x}$ with dimension $2p = 699,970$ consists of both additive and dominant effects of the non-rare SNPs. In this stage, we compute the screening statistic $\widehat{\rho}_j^*, \quad j = 1, \ldots, 2p$ for each SNP, and sort $\widehat{\rho}_j^*$'s in a decreasing order. Then we obtain a submodel with size d:

$$\mathcal{M}_\gamma = \{j : 1 \leq j \leq 2p,\ \widehat{\rho}_j^* \text{ is among the first } d \text{ largest of all } \widehat{\rho}^*\text{'s}\}.$$

where $d = [n^{4/5}/\log(n^{4/5})] = 45$ according to the hard threshold. By CCIS, the ultrahigh dimension $2p$ is reduced to the moderate dimension d.

Stage 2: Post-screening variable selection

In this stage, we conduct the penalized regression procedures to select important variables. Wang and Xia (2009) advocated Kernel LASSO technique (KLASSO) as the regularization method for varying coefficient models, incorporating local polynomial smoothing (Fan and Gijbels, 1996) and LASSO regression (Tibshirani, 1996). We extend this idea to SCAD penalty, and apply LASSO, adaptive LASSO

and SCAD penalized regression to the screened data with dimension $d = 45$.

Consider the submodel in the second stage,

$$y = \mathbf{x}^T \boldsymbol{\beta}(u) + \varepsilon,$$

where $\boldsymbol{\beta}(u)$ is now the d-dimensional coefficient function to be estimated. However, since we cannot directly estimate the whole coefficient function with an explicit form, we focus on estimating $\boldsymbol{\beta}(u)$ when $u = u_1, \ldots, u_n$, i.e. the matrix

$$\mathbf{B} = \{\boldsymbol{\beta}(u_1), \ldots, \boldsymbol{\beta}(u_n)\}^{\mathrm{T}} = (\mathbf{b}_1, \ldots, \mathbf{b}_d),$$

where $\mathbf{b}_j \in \mathbb{R}^{n \times 1}$ is the jth column of $\mathbf{B}$.

To obtain the estimate $\widehat{\mathbf{B}}_\lambda$ of $\mathbf{B}$, we need to minimize the penalized loss function

$$Q_\lambda(\mathbf{B}) = \sum_{t=1}^{n} \sum_{i=1}^{n} \{y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}(u_t)\}^2 K_h(u_t - u_i) + n \sum_{j=1}^{d} p_\lambda(\|\mathbf{b}_j\|) \quad (3.11)$$

where $p_\lambda(\cdot)$ is the penalty function, $\|\cdot\|$ is the Euclidean norm, and $K_h(u) = h^{-1}K(u/h)$ with $K(\cdot)$ being any kernel function. For the sake of simplicity, we take Epanechikov kernel $K(t) = 0.75(1 - t^2)I(|t| \leq 1)$.

An iterative algorithm based on the local quadratic approximation (LQA) to the penalty function (Fan and Li, 2001) is applied to get the minimizer $\widehat{\mathbf{B}}_\lambda$ of $Q_\lambda(\mathbf{B})$. Set the initial value $\widehat{\mathbf{B}}_\lambda^{(0)}$ to be the unpenalized estimator (Fan and Zhang, 2000b)

$$\widehat{\mathbf{B}}_\lambda^{(0)} = \tilde{\mathbf{B}} = \mathrm{argmin}_{\mathbf{B} \in \mathbb{R}^{n \times d}} \left\{ \sum_{t=1}^{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}(u_t))^2 K_h(u_t - u_i) \right\},$$

then the $(m+1)$th-step value of $\widehat{\mathbf{B}}_\lambda$ is

$$\widehat{\mathbf{B}}_\lambda^{(m+1)} = \{\hat{\boldsymbol{\beta}}_\lambda^{(m+1)}(u_1), \ldots, \hat{\boldsymbol{\beta}}_\lambda^{(m+1)}(u_n)\}^{\mathrm{T}}$$

with the tth row $\left(\hat{\boldsymbol{\beta}}_{\lambda}^{(m+1)}(u_t)\right)^{\mathrm{T}}$, where

$$\hat{\boldsymbol{\beta}}_{\lambda}^{(m+1)}(u_t) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}K_h(u_t-u_i) + D^{(m)}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i K_h(u_t-u_i)\right) \quad (3.12)$$

$D^{(m)}$ is a $d\times d$ diagonal matrix with the jth diagonal component given by

$$D_{jj}^{(m)} = \frac{p_{\lambda}'(\|\hat{\mathbf{b}}_{\lambda,j}^{(m)}\|)}{2\|\hat{\mathbf{b}}_{\lambda,j}^{(m)}\|},$$

and $\hat{\mathbf{b}}_{\lambda,j}^{(m)}$ is the jth column of the mth-step value $\widehat{\mathbf{B}}_{\lambda}^{(m)}$. Therefore, the difference among LASSO, Adaptive LASSO and SCAD estimates lies only in $D^{(m)}$.

1. *LASSO for varying coefficient model:*

   The LASSO penalty $p_{\lambda}(\|\mathbf{b}_j\|) = \lambda\|\mathbf{b}_j\|$ can be locally approximated by a quadratic form based on LQA in the mth step (Fan and Li, 2001; Hunter and Li, 2005), i.e.

   $$p_{\lambda}(\|\mathbf{b}_j\|) \approx \frac{\lambda\|\mathbf{b}_j\|^2}{\|\hat{\mathbf{b}}_{\lambda,j}^{(m)}\|},$$

   where $\lambda = \sqrt{n}\lambda_0$ and λ_0 is the tuning parameter. Hence the $(m+1)$th-step value of $\hat{\boldsymbol{\beta}}_{\lambda}(u_t)$ is computed by (3.12) with $D_{jj}^{(m)} = \lambda/\|\hat{\mathbf{b}}_{\lambda,j}^{(m)}\|$.

2. *Adaptive LASSO for varying coefficient model:*

   Adaptive LASSO can be applied to reduce the bias of LASSO estimates, by using different tuning parameters for different predictors. More explicitly, the penalty

   $$p_{\lambda}(\|\mathbf{b}_j\|) \approx \lambda_j\|\mathbf{b}_j\|^2/\|\hat{\mathbf{b}}_{\lambda,j}^{(m)}\|$$

   where $\lambda_j = \sqrt{n}\lambda_0/\|\tilde{\mathbf{b}}_j\|$ with $\tilde{\mathbf{b}}_j$ being the jth column of the unpenalized estimate $\tilde{\mathbf{B}}$, and λ_0 is the tuning parameter. Similar with LASSO, the jth diagonal component of $D^{(m)}$ for Adaptive LASSO is $\lambda_j/\|\hat{\mathbf{b}}_{\lambda,j}^{(m)}\|$.

3. *SCAD for varying coefficient model:*

The SCAD penalty is defined by

$$p_\lambda(\|\mathbf{b}_j\|) = \lambda\|\mathbf{b}_j\|I(0 \le \|\mathbf{b}_j\| < \lambda) + \frac{a\lambda\|\mathbf{b}_j\| - (\|\mathbf{b}_j\|^2 + \lambda^2)/2}{a-1} I(\lambda \le \|\mathbf{b}_j\| < a\lambda) + \frac{(a+1)\lambda^2}{2} I(\|\mathbf{b}_j\| > a\lambda),$$

with the first order derivative

$$p'_\lambda(\|\mathbf{b}_j\|) = \lambda I(\|\mathbf{b}_j\| \le \lambda) + \frac{(a\lambda - \|\mathbf{b}_j\|)_+ I(\|\mathbf{b}_j\| > \lambda)}{a-1},$$

where $a = 3.7$ as suggested by Fan and Li (2001), $\lambda = \sqrt{n}\lambda_0$, and λ_0 is the tuning parameter. Thus the jth diagonal component of $D^{(m)}$ is

$$D_{jj}^{(m)} = \frac{1}{2\|\hat{\mathbf{b}}_j^{(m)}\|} \left\{ \lambda I(\|\hat{\mathbf{b}}_j^{(m)}\| \le \lambda) + \frac{(a\lambda - \|\hat{\mathbf{b}}_j^{(m)}\|)_+ \cdot I(\|\hat{\mathbf{b}}_j^{(m)}\| > \lambda)}{a-1} \right\},$$

and the estimates are obtained by iteratively computing (3.12).

To implement each of the three penalized regression techniques, we need to choose tuning parameter λ. Here we use three criteria, AIC, BIC, and GCV defined as follows (Craven and Wahba, 1979, Fan and Li, 2001, Li et al., 2006, and Wang et al., 2007b):

$$AIC = \log(RSS_\lambda) + \frac{2df_\lambda}{n},$$
$$BIC = \log(RSS_\lambda) + df_\lambda \cdot \frac{\log(n)}{n},$$
$$GCV = \frac{RSS_\lambda}{(1 - \frac{df_\lambda}{n})^2},$$

where df_λ is the degree of freedom for nonparametric models defined by (Fan, Zhang, and Zhang, 1999)

$$df_\lambda = \gamma_K |\mathbb{U}| d_\lambda \{K(0) - 0.5 \int K^2(u)du\}/h, \tag{3.13}$$

$|\mathbb{U}|$ is the range of the index variable support $\mathbb{U}$, d_λ is the number of nonzero coefficient functions in the model, and γ_K is a constant which is determined by the

kernel function $K(\cdot)$. For Epanechnikov kernel, $\gamma_K = 2.1153$, and RSS_λ is defined by

$$RSS_\lambda = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_\lambda(u_i)\}^2,$$

where $\hat{\boldsymbol{\beta}}_\lambda(u)$ is the coefficient estimate for a given λ, using any of the three penalized regression technique.

Figure 3.1 illustrates how the values of the three criteria change with the tuning parameter λ, and in each plot, the optimal λ is chosen to minimize the corresponding criterion value. AIC and GCV perform quite similar for each penalized regression model and choose the same tuning parameter, which tends to be smaller than that selected by BIC, hence AIC and GCV tend to generate more conservative models than BIC. More specifically, the sizes of the nine models chosen based on different penalties and different tuning parameter selection criteria are reported in Table 3.12, where the same size indicates identical model. Therefore, one can see that SCAD penalized regressions give the sparsest model of size 34, while LASSO-AIC and LASSO-GCV produce the most conservative model with 43 predictors.

Table 3.12. The sizes of the nine models

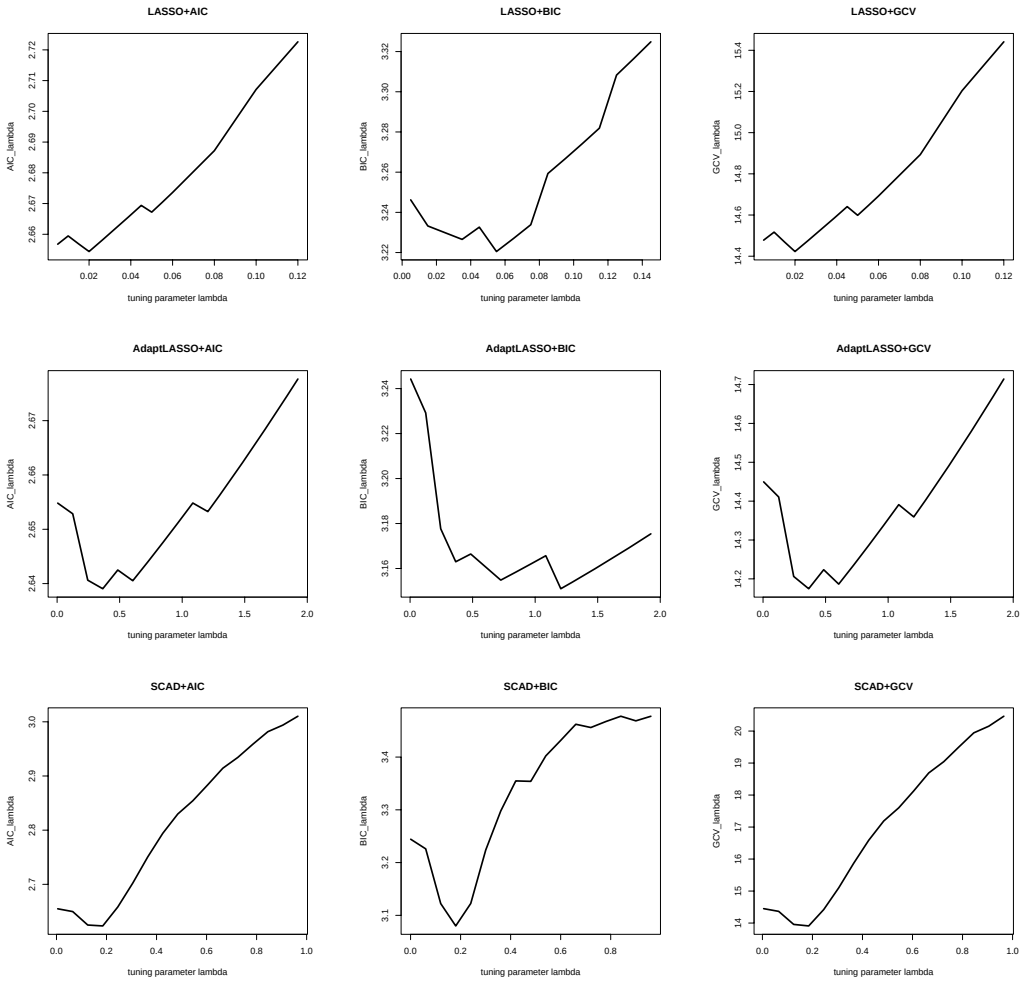
	CCIS+LASSO	CCIS+AdaptiveLASSO	CCIS+SCAD
AIC	43	40	34
BIC	42	38	34
GCV	43	40	34

Table 3.13 compares the performances of the above models in terms of median squared prediction error (MSPE). The median of MSPE's based on 100 simulations are reported. The SCAD penalized regressions give the smallest MSPE.

Table 3.13. Median of MPSE

	CCIS+LASSO	CCIS+AdaptiveLASSO	CCIS+SCAD
AIC	0.405	0.401	0.380
BIC	0.395	0.400	0.380
GCV	0.405	0.401	0.380

Figure 3.1. Tuning parameter selection for with three penalties and three criteria



3.6.3 Generalized Likelihood Ratio Tests

In Table 3.12, the same model size indicates the identical model, thus the nine methods produce 5 different models: lasso-AIC, lasso-BIC, Adeptive LASSO-AIC (Alasso-AIC, for short), Adaptive LASSO-BIC (Alasso-BIC), and SCAD. In addition, we can also obtain the unpenalized regression model (UP) with size 45. The 6 models are nested, which motivates us to conduct pairwise generalized likelihood ratio tests (Fan and Zhang, 2001) to compare their performances. Let RSS_1 be the residual sum of squares of full model with estimated coefficient vector $\hat{\boldsymbol{\beta}}_F(u)$,

and RSS_0 be that of reduced model with $\hat{\boldsymbol{\beta}}_R(u)$, i.e.

$$RSS_1 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_F(u_i))^2, \qquad RSS_0 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_R(u_i))^2.$$

Then the test statistic is constructed as

$$\Omega \approx \frac{n}{2} \cdot \frac{RSS_0 - RSS_1}{RSS_1}.$$

Fan and Zhang (2001) showed that under the null hypothesis, approximately we have

$$\gamma_K \Omega \sim \chi^2_\delta \tag{3.14}$$

where the degree of freedom δ has the same form as df_λ in (3.13), but with d_λ replaced by $p_1 - p_0$, where p_1 is the full model size and p_0 is the reduced model size.

Based on (3.14), the p-values of all the pairwise tests are computed and reported in Table 3.14. Note that we did not contain the unpenalized model in the reduced model part, and did not contain SCAD model in the full model part, since they are the largest and smallest model. Each p-value is large enough that we cannot reject the corresponding reduced model. Consequently, the sparsest model chosen by SCAD is sufficient for modeling BMI.

Table 3.14. The p-values of the pairwise generalized likelihood ratio tests

		H_1				
		Unpenalized	lasso-AIC	lasso-BIC	Alasso-AIC	Alasso-BIC
H_0	lasso-AIC	0.9952	·	·	·	·
	lasso-BIC	0.9999	0.9462	·	·	·
	Alasso-AIC	0.9999	0.9998	0.9995	·	·
	Alasso-BIC	0.9999	0.9967	0.9854	0.7481	·
	SCAD	0.9999	0.9991	0.9965	0.9516	0.9268

3.6.4 The Results

In Table 3.15, the names, positions, and effects (additive or dominant) of the significant SNPs in the SCAD model are tabulated, where "Additive" indicates that the additive effect of the corresponding SNP is significant, and "Dominant" indicates that the dominant effect is significant.

Figure 3.2 is the plot of the estimated coefficient functions versus the univariate index variable *Age*, which depicts the age-dependent effects of the 34 chosen SNPs in Table 3.15. From the plot, one can see that $\hat{\beta}_k(Age)$'s indeed vary with *Age*, indicating the necessity of varying-coefficient structure.

Figure 3.2. The estimated coefficient functions of significant SNPs

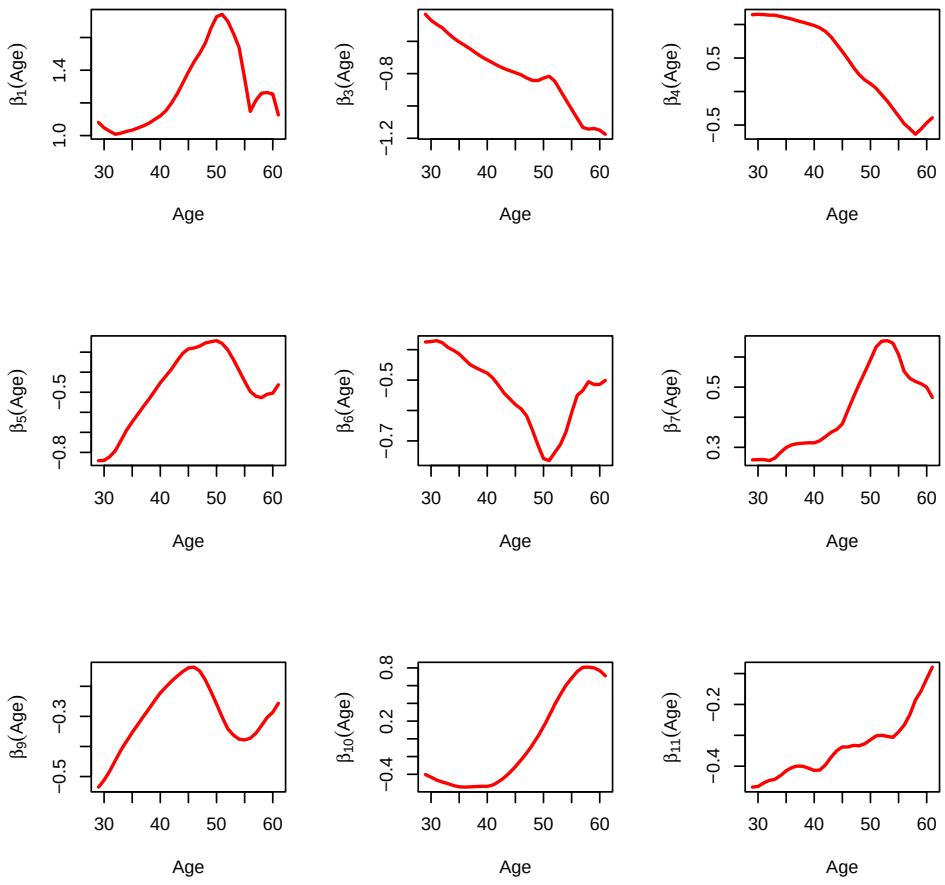


Table 3.15. Information of the significant SNPs

Chromosome	SNP Name	Position	Effect
1	ss66379476	181239647	Additive
1	ss66516012	198313489	Additive
2	ss66282476	47658001	Additive
2	ss66085516	10151206	Dominant
3	ss66266272	29713029	Dominant
4	ss66346937	92071818	Additive
4	ss66137328	94451805	Additive
4	ss66159949	105978188	Additive
4	ss66353634	15504889	Dominant
4	ss66354801	115353605	Dominant
5	ss66078741	34192815	Dominant
5	ss66164865	99237174	Dominant
7	ss66524659	44465215	Additive
7	ss66155306	134464951	Additive
7	ss66261449	46646583	Dominant
8	ss66236850	32389924	Additive
8	ss66143305	122601829	Additive
8	ss66445258	14959676	Dominant
8	ss66517429	15818735	Dominant
9	ss66319388	16387155	Additive
11	ss66153510	13262887	Additive
11	ss66110771	13267430	Additive
11	ss66112931	103378701	Dominant
12	ss66470239	51255904	Additive
12	ss66323107	117659019	Additive
13	ss66041456	107914455	Dominant
14	ss66404926	24422783	Additive
15	ss66058021	44940166	Dominant
16	ss66064472	5022290	Dominant
19	ss66435333	5178008	Dominant
20	ss66272727	2700340	Additive
20	ss66176990	2723332	Dominant
21	ss66511535	15841940	Additive
22	ss66305798	16578327	Additive

Figure 3.2. (Cont'd.) The estimated coefficient functions of significant SNPs

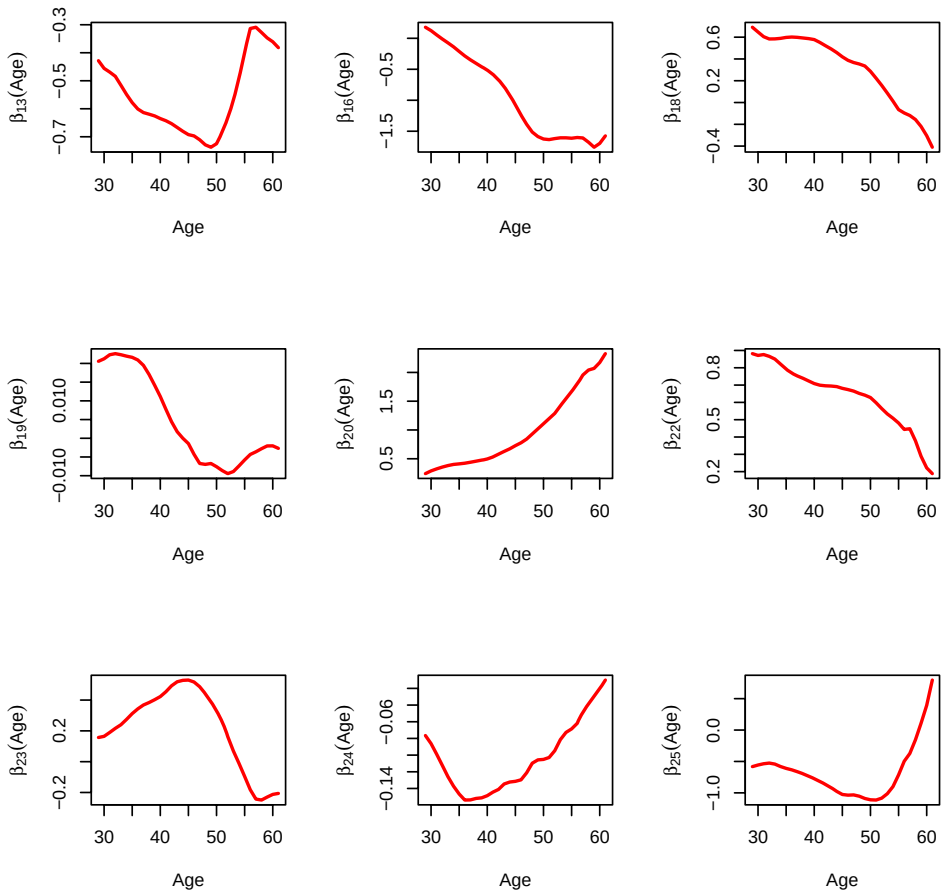


Figure 3.2. (Cont'd.) The estimated coefficient functions of significant SNPs

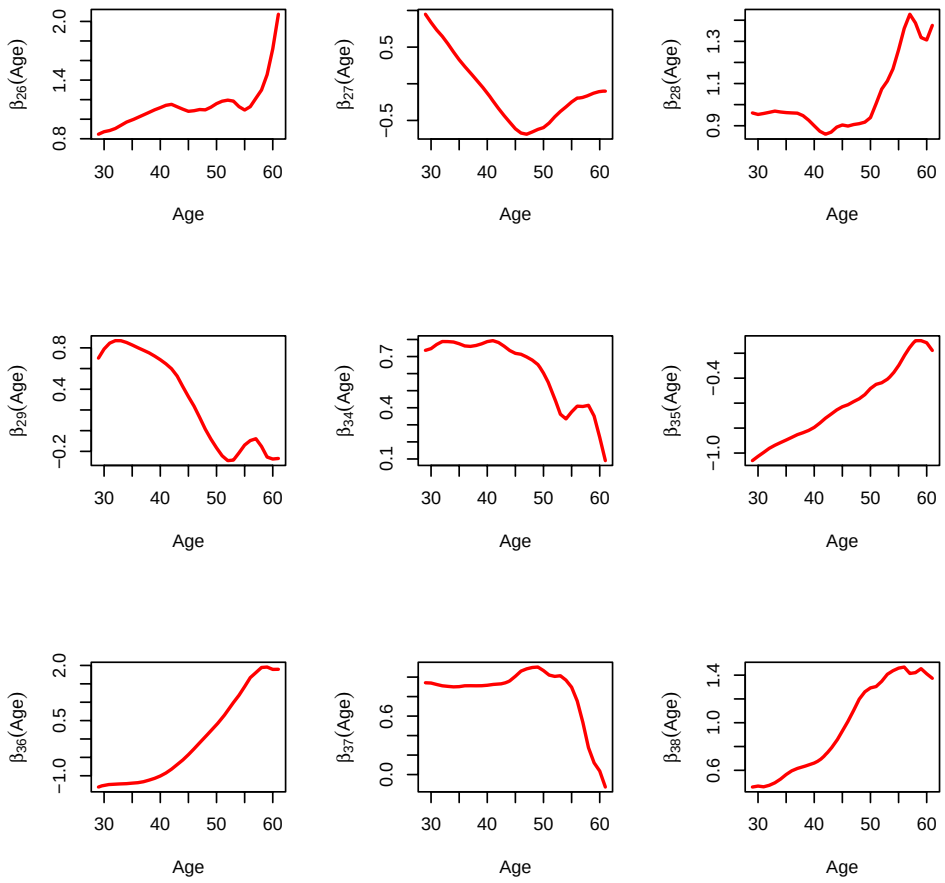
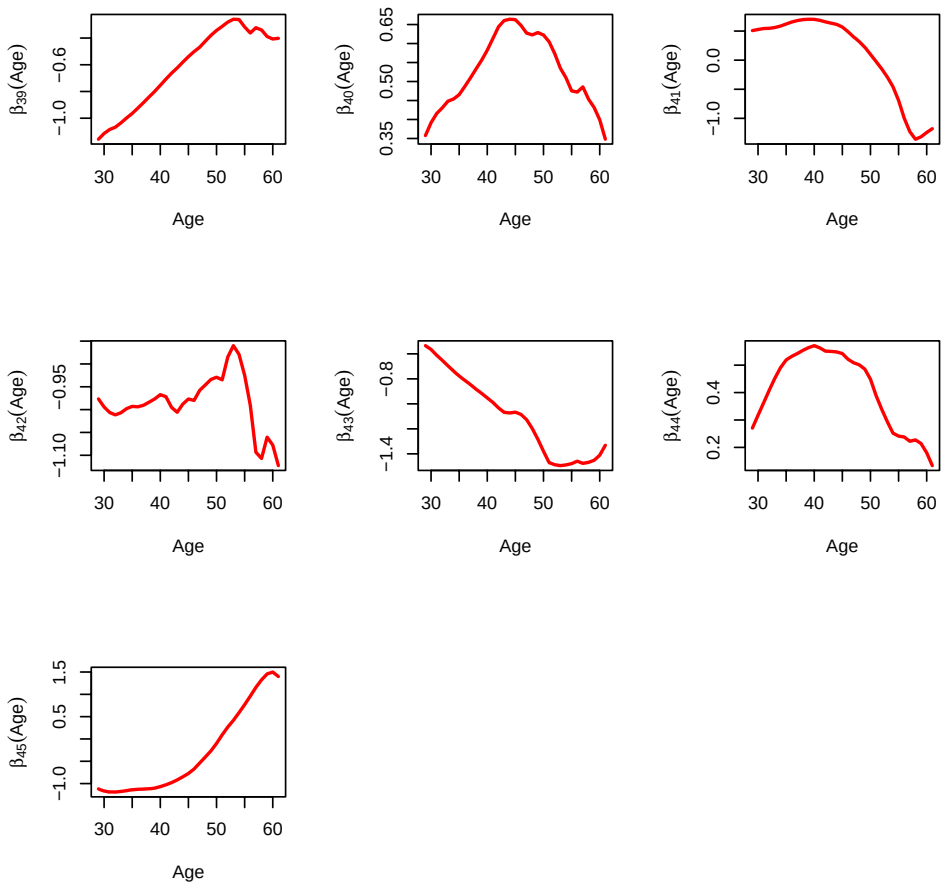


Figure 3.2. (Cont'd.) The estimated coefficient functions of significant SNPs



Chapter 4

Proofs of Theoretical Properties of CCIS

In this chapter, we study the technical proofs of the aforementioned two desirable theoretical properties ranking consistency and sure screening property of the screening procedure CCIS for varying coefficient models.

4.1 Proof of Theorem 5

In this section, we prove the ranking consistency of CCIS stated in Theorem 5, by studying the consistency of the estimated screening criterion. To begin with, two lemmas are introduced.

Lemma 1. *(Hoeffding's inequality) Assume the independent random sample* $\{X_i,\ i = 1, \ldots, n\}$ *satisfies* $P(X_i \in [a_i, b_i]) = 1$ *for some* a_i *and* b_i, $\forall i = 1, \ldots, n$. *Then, for any* $\varepsilon > 0$, *we have*

$$P(|\overline{X} - \mathrm{E}(\overline{X})| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right), \tag{4.1}$$

where $\overline{X} = (X_1 + \cdots + X_n)/n$.

Lemma 2. *Assume* $a(u)$ *and* $b(u)$ *are two uniformly-bounded functions of* u, *that*

is, there exist $M_5 > 0$, $M_6 > 0$ *such that*

$$\sup_{u\in\mathbb{U}} |a(u)| \le M_5, \quad \sup_{u\in\mathbb{U}} |b(u)| \le M_6.$$

For a given $u \in \mathbb{U}$, $\hat{A}(u)$ *and* $\hat{B}(u)$ *are estimates of* $a(u)$ *and* $b(u)$ *based on a sample with size* n. *Suppose for an arbitrary* $\varepsilon > 0$, *there exist positive constants* c_1, c_2 *and* s, *such that*

$$\begin{aligned} P(|\hat{A}(u) - a(u)| \ge \varepsilon) &\le c_1(1 - \frac{\varepsilon s}{c_1})^n, \\ P(|\hat{B}(u) - b(u)| \ge \varepsilon) &\le c_2(1 - \frac{\varepsilon s}{c_2})^n. \end{aligned} \tag{4.2}$$

Furthermore, assume $b(u)$ *is uniformly bounded away from* 0, *that is, there is* $M_7 > 0$ *such that* $\inf_{u\in\mathbb{U}} |b(u)| > M_7$. *Then* $\hat{A}(u)\hat{B}(u)$, $\hat{A}(u) - \hat{B}(u)$, $\hat{A}(u)/\hat{B}(u)$ *and* $\sqrt{\hat{B}(u)}$, *if well defined, all have the same form of inequality as* (4.2).

Proof of Lemma 2:

For all the proof, we denote C as a generic constant depending on the context, which can vary from line to line. First notice that $\hat{A}(u)$ and $\hat{B}(u)$ are bounded in probability. More specifically, for any $\varepsilon > 0$, since $\sup_{u\in\mathbb{U}} |a(u)| \le M_5$,

$$\begin{aligned} P\left(|\hat{A}(u)| \ge M_5 + \varepsilon\right) &\le P\left(|\hat{A}(u) - a(u)| + |a(u)| \ge M_5 + \varepsilon\right) \\ &\le P\left(|\hat{A}(u) - a(u)| \ge \varepsilon\right) \\ &\le c_1(1 - \frac{\varepsilon s}{c_1})^n \end{aligned}$$

by (4.2). In the same fashion, we can prove $P\left(|\hat{B}(u)| \ge M_6 + \varepsilon\right) \le c_2(1 - \varepsilon s/c_2)^n$.

1. Consider $\hat{A}(u)\hat{B}(u)$. For a given u and any $\varepsilon > 0$,

$$\begin{aligned} & P\left(|\hat{A}(u)\hat{B}(u) - a(u)b(u)| \ge \varepsilon\right) \\ = \; & P\left(|\hat{A}(u)\hat{B}(u) - \hat{A}(u)b(u) + \hat{A}(u)b(u) - a(u)b(u)| \ge \varepsilon\right) \\ \le \; & P\left(|\hat{A}(u)| \cdot |\hat{B}(u) - b(u)| + |b(u)| \cdot |\hat{A}(u) - a(u)| \ge \varepsilon\right) \\ \le \; & P\left(|\hat{A}(u)| \cdot |\hat{B}(u) - b(u)| \ge \frac{\varepsilon}{2}\right) + P\left(|b(u)| \cdot |\hat{A}(u) - a(u)| \ge \frac{\varepsilon}{2}\right), \end{aligned} \tag{4.3}$$

where the first term

$$\begin{aligned}
& P\left(|\hat{A}(u)|\cdot|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2}\right)\\
= \;& P\left(|\hat{A}(u)|\cdot|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2},\ |\hat{A}(u)|\geq M_5+\varepsilon\right)\\
& +P\left(|\hat{A}(u)|\cdot|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2},\ |\hat{A}(u)|<M_5+\varepsilon\right)\\
\leq \;& P\left(|\hat{A}(u)|\geq M_5+\varepsilon\right)+P\left((M_5+\varepsilon)\cdot|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2}\right)\\
\leq \;& c_1(1-\frac{\varepsilon s}{c_1})^n+P\left(|\hat{B}(u)-b(u)|\geq\frac{\varepsilon}{2(M_5+\varepsilon)}\right)\\
\leq \;& c_1(1-\frac{\varepsilon s}{c_1})^n+c_2(1-\frac{\varepsilon s}{2c_2(M_5+\varepsilon)})^n\\
\leq \;& c_3(1-\frac{\varepsilon s}{c_3})^n,\quad \text{where } c_3=\max\{c_1+c_2,\ 2c_2(M_5+\varepsilon)\}
\end{aligned}$$

and the second term

$$\begin{aligned}
& P\left(|b(u)|\cdot|\hat{A}(u)-a(u)|\geq\frac{\varepsilon}{2}\right)\\
\leq \;& P\left(|\hat{A}(u)-a(u)|\geq\frac{\varepsilon}{2M_6}\right)\\
\leq \;& c_2(1-\frac{\varepsilon s}{2c_2M_6})^n.
\end{aligned}$$

Therefore, (4.3) becomes

$$\begin{aligned}
& P\left(|\hat{A}(u)\hat{B}(u)-a(u)b(u)|\geq\varepsilon\right)\\
\leq \;& c_3(1-\frac{\varepsilon s}{c_3})^n+c_2(1-\frac{\varepsilon s}{2c_2M_6})^n\\
\leq \;& C(1-\frac{\varepsilon s}{C})^n,\quad \text{where } C=\max\{c_3+c_2,2c_2M_6\}.
\end{aligned}\tag{4.4}$$

In addition, by setting $\hat{B}(u)\equiv\hat{A}(u)$ and $b(u)\equiv a(u)$, (4.4) indicates

$$P\left(|\hat{A}(u)^2-a(u)^2|\geq\varepsilon\right)\leq C(1-\frac{\varepsilon s}{C})^n \ \text{ for some } \ C>0.$$

2. Consider $\hat{A}(u) - \hat{B}(u)$. For a given u and any $\varepsilon > 0$,

$$\begin{aligned} & P\left(|\{\hat{A}(u) - \hat{B}(u)\} - \{a(u) - b(u)\}| \geq \varepsilon\right) \\ \leq\ & P\left(|\hat{A}(u) - a(u)| + |\hat{B}(u) - b(u)| \geq \varepsilon\right) \\ \leq\ & P\left(|\hat{A}(u) - a(u)| \geq \frac{\varepsilon}{2}\right) + P\left(|\hat{B}(u) - b(u)| \geq \frac{\varepsilon}{2}\right) \\ \leq\ & c_1(1 - \frac{\varepsilon s}{2c_1})^n + c_2(1 - \frac{\varepsilon s}{2c_2})^n \\ \leq\ & C(1 - \frac{\varepsilon s}{C})^n, \quad \text{where } C = \max\{2c_1,\ 2c_2,\ c_1 + c_2\}. \end{aligned}$$

3. Consider $\hat{A}(u)/\hat{B}(u)$. To guarantee it to be well defined, the denominator $\hat{B}(u)$ need to be bounded away from 0, i.e. there exist M_8 such that $\inf_{u\in\mathbb{U}} |\hat{B}(u)| > M_8$. Then for a given u and any $\varepsilon > 0$,

$$\begin{aligned} & P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)} - \frac{a(u)}{b(u)}\right| \geq \varepsilon\right) \\ \leq\ & P\left(\left|\frac{\hat{A}(u)}{\hat{B}(u)} - \frac{a(u)}{\hat{B}(u)}\right| + \left|\frac{a(u)}{\hat{B}(u)} - \frac{a(u)}{b(u)}\right| \geq \varepsilon\right) \\ \leq\ & P\left(\frac{1}{|\hat{B}(u)|} \cdot |\hat{A}(u) - a(u)| \geq \frac{\varepsilon}{2}\right) + P\left(\frac{|a(u)|}{|\hat{B}(u)| \cdot |b(u)|} \cdot |\hat{B}(u) - b(u)| \geq \frac{\varepsilon}{2}\right) \\ \leq\ & P\left(|\hat{A}(u) - a(u)| \geq \frac{\varepsilon M_8}{2}\right) + P\left(|\hat{B}(u) - b(u)| \geq \frac{\varepsilon M_7 M_8}{2M_5}\right) \\ \leq\ & c_1(1 - \frac{\varepsilon s M_8}{2c_1})^n + c_2(1 - \frac{\varepsilon s M_7 M_8}{2c_2 M_5})^n \\ \leq\ & C(1 - \frac{\varepsilon s}{C})^n, \quad \text{where } C = \max\{c_1 + c_2,\ \frac{2c_1}{M_8},\ \frac{2c_2 M_5}{M_7 M_8}\}. \end{aligned}$$

4. Consider $\sqrt{\hat{B}(u)}$, if well defined. For any $\varepsilon > 0$,

$$\begin{aligned} & P\left(|\sqrt{\hat{B}(u)} - \sqrt{b(u)}| \geq \varepsilon\right) \\ =\ & P\left(\frac{|\hat{B}(u) - b(u)|}{\sqrt{\hat{B}(u)} + \sqrt{b(u)}} \geq \varepsilon\right) \\ \leq\ & P\left(|\hat{B}(u) - b(u)| \geq \varepsilon(\sqrt{M_7} + \sqrt{M_8})\right) \end{aligned}$$

$$\begin{aligned} &\leq c_2(1-\frac{\varepsilon s(\sqrt{M_7}+\sqrt{M_8})}{c_2})^n \\ &\leq C(1-\frac{\varepsilon s}{C})^n, \quad \text{where } C=\max\{c_2,\ \frac{c_2}{\sqrt{M_7}+\sqrt{M_8}}\}. \end{aligned}$$

To summarize, the four estimators have the same form of convergence as $\hat{A}(u)$ and $\hat{B}(u)$. □

Proof of Theorem 5:

Now we go back to the proof of ranking consistency property. We divide the proof into the following three steps.

Step I. Prove for any $\varepsilon > 0$, $u \in \mathbb{U}$ and $1 \leq j \leq p$, we have

$$P(|\hat{\rho}(x_j, y|u) - \rho(x_j, y|u)| \geq \varepsilon) \leq C(1-\frac{s\varepsilon}{C})^n.$$

With elementary calculation, $\hat{\rho}(x_j, y|u)$ can be represented by

$$\hat{\rho}(x_j, y|u) = \frac{Z_1 Z_6 - Z_2 Z_3}{\sqrt{(Z_1 Z_4 - Z_3^2)(Z_1 Z_5 - Z_2^2)}},$$

where

$$\begin{aligned} Z_1 &= \frac{1}{n}\sum_{i=1}^n K(\frac{u_i-u}{h}), \quad Z_2 = \frac{1}{n}\sum_{i=1}^n y_i K(\frac{u_i-u}{h}), \\ Z_3 &= \frac{1}{n}\sum_{i=1}^n x_{ij} K(\frac{u_i-u}{h}), \quad Z_4 = \frac{1}{n}\sum_{i=1}^n x_{ij}^2 K(\frac{u_i-u}{h}), \\ Z_5 &= \frac{1}{n}\sum_{i=1}^n y_i^2 K(\frac{u_i-u}{h}), \quad Z_6 = \frac{1}{n}\sum_{i=1}^n x_{ij} y_i K(\frac{u_i-u}{h}). \end{aligned} \tag{4.5}$$

1. First prove $P(|Z_1 - h \cdot f(u)| \geq \varepsilon) \leq 4(1 - s\varepsilon/4)^n$, where $f(u)$ is the density function of u. For any $t > 0$, by Markov's Inequality,

$$\begin{aligned} P\left(Z_1 - h \cdot f(u) \geq \epsilon\right) &= P\left(\exp\{t(Z_1 - h \cdot f(u))\} \geq \exp(t\varepsilon)\right) \\ &\leq E[\exp\{tZ_1 - thf(u)\}]/\exp(t\varepsilon) \\ &= \exp(-t\varepsilon) \cdot \exp\{-thf(u)\} \cdot E\{\exp(tZ_1)\}, \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} E\{\exp(tZ_1)\} &= E\left[\exp\left\{t\cdot\frac{1}{n}\sum_{i=1}^{n}K(\frac{u_i-u}{h})\right\}\right] \\ &= E\left[\prod_{i=1}^{n}\exp\left\{\frac{t}{n}K(\frac{u_i-u}{h})\right\}\right] \\ &= \left[E\left\{\exp\left(\frac{t}{n}K(\frac{u_i-u}{h})\right)\right\}\right]^n. \end{aligned}$$

Set the arbitrary positive t to be $t = ns$ where the constant $s > 0$ is specified later, and define $\varphi_1(s) = E\{\exp(s \cdot K(\frac{u_i-u}{h}))\}$. Then (4.6) becomes

$$P\left(Z_1 - h\cdot f(u) \geq \epsilon\right) \leq \left[\exp(-s\varepsilon)\cdot\exp\{-shf(u)\}\cdot\varphi_1(s)\right]^n. \tag{4.7}$$

Now we deal with the last two terms of (4.7):

$$\begin{aligned} & \exp\{-shf(u)\}\cdot\varphi_1(s) \\ = \;& \exp\{-shf(u)\}\cdot E\left\{\exp\left(s\cdot K(\frac{u_i-u}{h})\right)\right\} \\ = \;& E\left[\exp\left\{s\left(K(\frac{u_i-u}{h}) - hf(u)\right)\right\}\right] \end{aligned} \tag{4.8}$$

By Taylor's expansion, for x close to 0, we have

$$\exp(x) = 1 + x + o(|x|) \leq 1 + x + |x| \leq 1 + 2|x|. \tag{4.9}$$

The constant s in (4.8) is chosen small enough that (4.9) can be applied. Based on the conditions (C1) and (C2), (4.8) satisfies the inequality

$$E\left[\exp\left\{s\left(K(\frac{u_i-u}{h}) - hf(u)\right)\right\}\right] \leq 1 + 2s\left|E\left\{(K(\frac{u_i-u}{h})\right\} - hf(u)\right| \tag{4.10}$$

To simplify (4.10), notice that Z_1/h is indeed the kernel density estimate of $f(u)$. Therefore, under (C1) and (C2), the bias of Z_1/h is

$$E(\frac{Z_1}{h}) - f(u) = E\left\{\frac{1}{h}K(\frac{u_i-u}{h})\right\} - f(u) = \frac{f''(u)}{2}\mu_2(K)h^2 + o(h^2) = O(h^2),$$

where $\mu_2(K) = \int t^2 K(t)dt < \infty$. See Wand and Jones (1995) for details. Then for any $\varepsilon > 0$, (4.10) can be simplified by

$$E\left[\exp\left\{s\left(K(\frac{u_i-u}{h}) - hf(u)\right)\right\}\right] \leq 1 + shO(h^2) < 1 + \frac{\varepsilon s}{4}$$

for large n. The last inequality is because the bandwidth h chosen by minimizing the MISE has the rate $h = O(n^{-1/5})$. Therefore, (4.7) now becomes

$$\begin{aligned} P\left(Z_1 - h \cdot f(u) \geq \epsilon\right) &\leq \{\exp(-\varepsilon s)(1+\varepsilon s/4)\}^n \\ &\leq \{(1-\varepsilon s + o(\varepsilon s))(1+\varepsilon s/4)\}^n \\ &\leq \{(1-\varepsilon s + \varepsilon s/2)(1+\varepsilon s/4)\}^n \\ &= \{1-\varepsilon s/4 - (\varepsilon s)^2/8\}^n \\ &\leq (1-\varepsilon s/4)^n \end{aligned}$$

Similarly, $P\left(Z_1 - h \cdot f(u) \leq -\epsilon\right) \leq (1 - s\varepsilon/4)^n$, then we have

$$P\left(|Z_1 - h \cdot f(u)| \geq \epsilon\right) \leq 2(1-s\varepsilon/4)^n \leq 4(1-s\varepsilon/4)^n.$$

2. Next prove $P(|Z_2 - hf(u)m(u)| \geq \varepsilon) \leq 2(1-s\varepsilon/2)^n$, where $m(u) = E(y|u)$. Using the same techniques as (4.6) – (4.10), we have

$$\exp(-shf(u)m(u)) \cdot \varphi_2(s) \leq 1 + 2s\left|E\left\{y_i K(\frac{u_i-u}{h}) - hf(u)m(u)\right\}\right| \quad (4.11)$$

under conditions (C1) – (C3), where $\varphi_2(s) = E\{\exp(s \cdot y_i \cdot K(\frac{u_i-u}{h}))\}$. Now consider the second term in (4.11). By the iterative expectation principal,

$$\begin{aligned} & E\left\{y_i K(\frac{u_i-u}{h}) - hf(u)m(u)\right\} \quad (4.12) \\ = \; & E_{u_i}\left\{E(y_i|u_i)K(\frac{u_i-u}{h}) - hf(u)m(u)\right\} \\ = \; & E_{u_i}\left\{m(u_i)K(\frac{u_i-u}{h}) - hf(u)m(u)\right\}, \end{aligned}$$

where $m(u_i) = m(u) + m'(u)(u_i-u) + m''(u^*)(u_i-u)^2/2$, u^* is between u

and u_i, for u_i close to u. Thus (4.12) is expanded as

$$\begin{aligned}
& E_{u_i}\left\{m(u_i)K(\frac{u_i-u}{h})-hf(u)m(u)\right\} \\
= \; & E_{u_i}\left\{m(u)K(\frac{u_i-u}{h})-hf(u)m(u)\right\}+m'(u)E_{u_i}\left\{(u_i-u)K(\frac{u_i-u}{h})\right\} \\
& +\frac{1}{2}E_{u_i}\left\{m''(u^*)(u_i-u)^2K(\frac{u_i-u}{h})\right\},
\end{aligned}$$

where the first term

$$\begin{aligned}
& E_{u_i}\left\{m(u)K(\frac{u_i-u}{h})-hf(u)m(u)\right\} \\
= \; & m(u)\left\{\int K(\frac{x-u}{h})f(x)dx\right\}-hf(u)m(u) \\
= \; & h\cdot m(u)\left\{\int K(t)f(u+ht)dt\right\}-hf(u)m(u) \quad \text{where} \quad t=(x-u)/h \\
= \; & h\cdot m(u)\left[\int K(t)\{f(u)+f'(u)ht+t\cdot o(h)\}dt\right]-hf(u)m(u) \\
= \; & hf(u)m(u)+h^2f'(u)m(u)\left\{\int tK(t)dt\right\}+o(h^2)m(u)\left\{\int tK(t)dt\right\}-hf(u)m(u) \\
= \; & h^2f'(u)m(u)\mu_1(K)+o(h^2),
\end{aligned}$$

the second term

$$\begin{aligned}
& m'(u)E_{u_i}\left\{(u_i-u)K(\frac{u_i-u}{h})\right\} \\
= \; & m'(u)\int(x-u)K(\frac{x-u}{h})f(x)dx \\
= \; & m'(u)\int thK(t)f(u+th)hdt \quad \text{where} \quad t=(x-u)/h \\
= \; & m'(u)h^2\int tK(t)\{f(u)+f'(u)th+t\cdot o(h)\}dt \\
= \; & h^2f(u)m'(u)\mu_1(K)+o(h^2),
\end{aligned}$$

and the last term

$$\frac{1}{2}E_{u_i}\left\{m''(u^*)(u_i-u)^2K(\frac{u_i-u}{h})\right\}$$

$$
\begin{aligned}
&\leq \frac{1}{2}C \cdot E\left[(u_i-u)^2 K(\frac{u_i-u}{h})\right], \quad \text{where } C \text{ s.t. } |\sup_{u\in\mathbb{U}} m''(u)| \leq C \\
&= \frac{1}{2}C \cdot \int (x-u)^2 K(\frac{x-u}{h}) f(x)dx \\
&= \frac{1}{2}Ch^3 \cdot \int t^2 K(t) f(u+ht)dt \quad \text{where } \ t=(x-u)/h \\
&= o(h^2).
\end{aligned}
$$

Hence (4.12) is rewritten as

$$E\{y_i K(\frac{u_i-u}{h}) - hf(u)m(u)\} \leq h^2\mu_1(K)\{f'(u)m(u)+f(u)m'(u)\} + o(h^2) = O(h^2),$$

leading (4.11) to be

$$\exp(-shf(u)m(u)) \cdot \varphi_2(s) \leq 1 + sO(h^2)$$

Therefore, by the same method as 1, we can obtain

$$P(|Z_2 - hf(u)m(u)| \geq \varepsilon) \leq 4(1-s\varepsilon/4)^n.$$

3. In the same fashion, we get all the other inequalities:

$$
\begin{aligned}
&P(|Z_3 - hf(u)E(x_j|u)| \geq \varepsilon) \leq 4(1-s\varepsilon/4)^n; \\
&P(|Z_4 - hf(u)E(x_j^2|u)| \geq \varepsilon) \leq 4(1-s\varepsilon/4)^n; \\
&P(|Z_5 - hf(u)E(y^2|u)| \geq \varepsilon) \leq 4(1-s\varepsilon/4)^n; \\
&P(|Z_6 - hf(u)E(x_j y|u)| \geq \varepsilon) \leq 4(1-s\varepsilon/4)^n.
\end{aligned}
$$

Therefore, by Lemma 2, there exists some $C>0$ such that

$$P\left(|\hat{\rho}(x_j,y|u) - \rho(x_j,y|u)| \geq \varepsilon\right) \leq C \cdot (1-\frac{s\varepsilon}{C})^n.$$

Step II. For any $\varepsilon > 0$, derive the upper bound of $P(|\widehat{\rho}_j^* - \rho_{j0}^*| \geq \varepsilon)$. The notations are introduced in section 2. Notice that

$$P(|\widehat{\rho}_j^* - \rho_{j0}^*| \geq \varepsilon) \quad \leq \quad P(|\widehat{\rho}_j^* - \rho_j^*| + |\rho_j^* - \rho_{j0}^*| \geq \varepsilon)$$

$$\leq \quad P(|\widehat{\rho}_j^* - \rho_j^*| \geq \varepsilon/2) + P(|\rho_j^* - \rho_{j0}^*| \geq \varepsilon/2). \qquad (4.13)$$

The first term of (4.13)

$$\begin{aligned} P\left(|\widehat{\rho}_j^* - \rho_j^*| \geq \varepsilon/2\right) &= P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\rho}^2(x_j, y|u_i) - \frac{1}{n}\sum_{i=1}^{n}\rho^2(x_j, y|u_i)\right| \geq \varepsilon/2\right) \\ &\leq P\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{\rho}^2(x_j, y|u_i) - \rho^2(x_j, y|u_i)\right| \geq \varepsilon/2\right) \\ &= P\left(\sum_{i=1}^{n}\left|\hat{\rho}^2(x_j, y|u_i) - \rho^2(x_j, y|u_i)\right| \geq n\varepsilon/2\right) \\ &\leq \sum_{i=1}^{n} P\left(\left|\hat{\rho}^2(x_j, y|u_i) - \rho^2(x_j, y|u_i)\right| \geq \varepsilon/2\right) \\ &\leq nC(1 - \frac{s\varepsilon}{C})^n. \qquad (4.14) \end{aligned}$$

The last inequality in (4.14) is indicated by step I and Lemma 2. And Lemma 1 renders the second term of (4.13)

$$P(|\rho_j^* - \rho_{j0}^*| \geq \varepsilon/2) \leq 2\exp(-\frac{n\varepsilon^2}{8}). \qquad (4.15)$$

Thus (4.13) becomes

$$P(\widehat{\rho}_j^* - \rho_{j0}^*| \geq \varepsilon) \leq nC(1 - \frac{s\varepsilon}{C})^n + 2\exp(-\frac{n\varepsilon^2}{8}).$$

Step III. Prove $P\left(\liminf_{n\to\infty}\{\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^* - \max_{j\in\mathcal{M}_*^c}\widehat{\rho}_j^*\} > 0\right) = 1$. Under condition (C5), there exists some $\delta > 0$ such that $\min_{j\in\mathcal{M}_*}\rho_{j0}^* - \max_{j\in\mathcal{M}_*^c}\rho_{j0}^* = \delta$. Then we have

$$\begin{aligned} & P\left(\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^* \leq \max_{j\in\mathcal{M}_*^c}\widehat{\rho}_j^*\right) \\ = \quad & P\left(\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^* - \min_{j\in\mathcal{M}_*}\rho_{j0}^* + \delta \leq \max_{j\in\mathcal{M}_*^c}\widehat{\rho}_j^* - \max_{j\in\mathcal{M}_*^c}\rho_{j0}^*\right) \\ = \quad & P\left(\{\max_{j\in\mathcal{M}_*^c}\widehat{\rho}_j^* - \max_{j\in\mathcal{M}_*^c}\rho_{j0}^*\} - \{\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^* - \min_{j\in\mathcal{M}_*}\rho_{j0}^*\} \geq \delta\right) \end{aligned}$$

$$
\begin{aligned}
&\leq P\left(\max_{j\in\mathcal{M}_*^c}|\widehat{\rho}_j^*-\rho_{j0}^*|+\max_{j\in\mathcal{M}_*}|\widehat{\rho}_j^*-\rho_{j0}^*|\geq\delta\right)\\
&\leq P\left(2\max_{1\leq j\leq p}|\widehat{\rho}_j^*-\rho_{j0}^*|\geq\delta\right)\\
&= P\left(\max_{1\leq j\leq p}|\widehat{\rho}_j^*-\rho_{j0}^*|\geq\delta/2\right)\\
&\leq \sum_{j=1}^{p}P\left(|\widehat{\rho}_j^*-\rho_{j0}^*|\geq\delta/2\right)\\
&\leq pnC(1-\frac{\delta s}{2C})^n+2p\exp(-\frac{n\delta^2}{32}).
\end{aligned}
$$

The last inequality is the direct result from step II, and it goes to 0 as $n\to\infty$, for $p=o(\exp(an))$ where $a<\min\{\log(2C/(2C-\delta s)),\delta^2/32\}$. Then by Fatou's Lemma,

$$
P\left(\liminf_{n\to\infty}\{\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^*-\max_{j\in\mathcal{M}_*^c}\widehat{\rho}_j^*\}\leq 0\right)\leq\lim_{n\to\infty}P\left(\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^*-\max_{j\in\mathcal{M}_*^c}\widehat{\rho}_j^*\leq 0\right)=0.
$$

In other words,

$$
P\left(\liminf_{n\to\infty}\{\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^*-\max_{j\in\mathcal{M}_*^c}\widehat{\rho}_j^*\}>0\right)=1.
$$

Therefore, the ranking consistency is proved. □

4.2 Proof of Theorem 6

Again, to prove Theorem 6, we need to introduce the following three lemmas.

Lemma 3. *Under the same conditions as Lamma 2, suppose for $\forall\varepsilon>0$ and a given $u\in\mathbb{U}$,*

$$
\begin{aligned}
P(|\hat{A}(u)-a(u)|\geq\varepsilon)&\leq c_1\exp(-\varepsilon/h),\\
P(|\hat{B}(u)-b(u)|\geq\varepsilon)&\leq c_2\exp(-\varepsilon/h).
\end{aligned}
$$

Then $\hat{A}(u)\hat{B}(u)$, $\hat{A}(u)-\hat{B}(u)$, $\hat{A}(u)/\hat{B}(u)$ and $\sqrt{\hat{B}(u)}$, if well defined, all have the same forms of inequality.

The proof of Lemma 3 is a straightforward extension of Lemma 2.

Lemma 4: Suppose X is a random variable with $E(e^{a|X|})<\infty$ for some $a>0$.

Then for any $M > 0$, there exist positive constant b and c such that

$$P(|X| \geq M) \leq be^{-cM}.$$

Proof of Lemma 4:
For any nondecreasing and nonnegative function $g(x)$ and any real number x,

$$\begin{aligned} P(X \geq x) &\leq P\{g(X) \geq g(x)\} = E\{I(g(X) \geq g(x))\} \\ &\leq E\left\{\frac{g(X)}{g(x)} \cdot I(g(X) \geq g(x))\right\} \\ &= \frac{1}{g(x)} E\{g(X) \cdot I(g(X) \geq g(x))\} \\ &\leq \frac{Eg(X)}{g(x)}. \end{aligned}$$

Take $g(x) = e^{ax}$, then we have

$$\begin{aligned} P(|X| \geq M) &= P(|X| \geq M, X \geq 0) + P(|X| \geq M, X < 0) \\ &\leq P(X \geq M) + P(-X \geq M) \\ &\leq \frac{Ee^{aX}}{e^{aM}} + \frac{Ee^{-aX}}{e^{aM}} \\ &= be^{-cM}, \end{aligned}$$

where $b > 0$ such that $Ee^{a|X|} \leq b/2$, and $c = a$. □

Lemma 4 is used to control the tail distribution of x_j and y. In addition, we also need to impose Lemma 5 below (Zhu, Li, Li and Zhu, 2011) based on Hoeffding's inequality to prove Theorem 6.

Lemma 5: *Suppose X is a random variable with $P(a \leq X \leq b) = 1$, then*

$$E[\exp\{s(X - E(X))\}] \leq \exp\{s^2(b-a)^2/8\} \quad \textit{for } \forall s > 0.$$

Based on the above three lemmas, we can now prove Theorem 6. To begin with, we need to redefine the chosen set $\widehat{\mathcal{M}}$ based on an explicit cutoff $c_3 \cdot n^{-\kappa}$ for the sure screening property (Fan and Lv, 2008), i.e.

$$\widehat{\mathcal{M}} = \{j : \ \widehat{\rho}_j^* \geq c_3 \cdot n^{-\kappa}, \ 1 \leq j \leq p\}.$$

Proof of Theorem 6:

We also accomplish the proof with three steps.

Step I. For any $\varepsilon > 0$, $u \in \mathbb{U}$ and $1 \le j \le p$, prove

$$P(|\hat{\rho}(x_j, y|u) - \rho(x_j, y|u)| \ge \varepsilon) \le C \cdot \exp(-\varepsilon/h).$$

1. First consider Z_1 defined in (4.5). By the same argument as the proof of Theorem 5, for any $t > 0$,

$$P(Z_1 - hf(u) \ge \varepsilon) \le \exp(-t\varepsilon) \left(E\left[\exp\left\{ \frac{t}{n} \left(K(\frac{u_i - u}{h}) - hf(u) \right) \right\} \right] \right)^n,$$

where

$$\begin{aligned} & E\left[\exp\left\{ \frac{t}{n} \left(K(\frac{u_i - u}{h}) - hf(u) \right) \right\} \right] \\ = \; & E\left[\exp\left\{ \frac{t}{n} \left(K(\frac{u_i - u}{h}) - EK(\frac{u_i - u}{h}) \right) + \frac{t}{n} \left(EK(\frac{u_i - u}{h}) - hf(u) \right) \right\} \right] \\ = \; & E\left[\exp\left\{ \frac{t}{n} \left(K(\frac{u_i - u}{h}) - EK(\frac{u_i - u}{h}) \right) \right\} \right] \cdot \exp\left[\frac{t}{n} \left\{ EK(\frac{u_i - u}{h}) - hf(u) \right\} \right]. \end{aligned}$$

According to Lemma 5, since $K(\cdot)$ is bounded by condition (C2), the first term is bounded by $\exp\{M_4^2 t^2/(2n^2)\}$, and from the proof of Theorem 5, the second term

$$\exp\left[\frac{t}{n} \left\{ EK(\frac{u_i - u}{h}) - hf(u) \right\} \right] = \exp\left[\frac{t}{n} \left\{ \frac{f''(u)}{2} \mu_2(K) h^3 + o(h^3) \right\} \right].$$

Therefore,

$$\begin{aligned} P(Z_1 - hf(u) \ge \varepsilon) &\le \exp\left\{ -t\varepsilon + \frac{M_4^2 t^2}{2n} + \frac{th^3 f''(u)}{2} \mu_2(K) + o(th^3) \right\} \\ &= \exp\{-\varepsilon/h + o(1)\} \;\; \text{by setting} \;\; t = 1/h \\ &\le C \exp(-\varepsilon/h). \end{aligned}$$

Similarly, $P(Z_1 - hf(u) \leq -\varepsilon) \leq C\exp(-\varepsilon/h)$, thus,

$$P(|A - hf(u)| \geq \varepsilon) \leq C\exp(-\varepsilon/h).$$

2. Next consider Z_2 in (4.5). Notice that

$$P(|Z_2 - hf(u)m(u)| \geq \varepsilon) \leq P\left(|Z_2 - EZ_2| + |EZ_2 - hf(u)m(u)| \geq \varepsilon\right) \quad (4.16)$$

In the proof of Theorem 5, we derived $|EZ_2 - hf(u)m(u)| = O(h^2) \leq \varepsilon/2$ provided $h = O(n^{-1/5})$, thus (4.16) is simplified as

$$P(|Z_2 - hf(u)m(u)| \geq \varepsilon) \leq P(|Z_2 - EZ_2| \geq \varepsilon/2). \quad (4.17)$$

Since the i.i.d random elements in Z_2 are not necessarily bounded almost surely, Lemma 1 is not directly applicable to deal with (4.17). However, according to condition (C3) and Lemma 4, there exist some positive constant m_1, m_2, m_3 and m_4 such that for any $M > 0$,

$$P(\max_{1\leq j\leq p} |x_j| \geq M) \leq m_1\exp(-m_2M), \quad \text{and} \quad P(|y| \geq M) \leq m_3\exp(-m_4M) \quad (4.18)$$

Therefore, for any $M > 0$,

$$\begin{aligned}
& P(|Z_2 - EZ_2| \geq \varepsilon/2) \\
= \;& P(|Z_2 - EZ_2| \geq \varepsilon/2, |y_i| \leq M, \forall i) + P(|Z_2 - EZ_2| \geq \varepsilon/2, |y_i| \geq M \text{ for some } i) \\
\leq \;& P\left(|Z_2 - EZ_2| \geq \varepsilon/2 \mid \{|y_i| \leq M,\ \forall i\}\right) \cdot P(|y_i| \leq M,\ \forall i) + P(|y_i| \geq M \text{ for some } i) \\
\leq \;& P\left(|Z_2 - EZ_2| \geq \varepsilon/2 \mid \{|y_i| \leq M,\ \forall i\}\right) \cdot \{P(|y| \leq M)\}^n + nP(|y| \geq M). \quad (4.19)
\end{aligned}$$

Lemma 1 can now be applied to the first term in (4.19) as y_i's are bounded by M, and $K(\cdot)$ is bounded by M_4 under condition (C2). Together with (4.18), the above inequality (4.19) is further computed as

$$\begin{aligned}
& P(|Z_2 - EZ_2| \geq \varepsilon/2) \\
\leq \;& 2\exp(-\frac{n\varepsilon^2}{8M^2M_4^2}) \cdot (1 - m_3\exp(-m_4M))^n + nm_3\exp(-m_4M)
\end{aligned}$$

$$\begin{aligned} &\leq 2\exp(-\frac{n\varepsilon^2}{8M^2M_4^2}) + nm_3\exp(-m_4M) \\ &\leq 2\exp(-\frac{\varepsilon}{h}\frac{n\varepsilon h}{8M^2M_4^2}) + m_3\exp(-\frac{\varepsilon}{h}\frac{h(Mm_4-\log(n))}{\varepsilon}) \end{aligned} \tag{4.20}$$

Now consider the first and second term of (4.20) separately. Now take $M = O(n^\tau)$, where $1/5 < \tau < 2/5$,

$$\frac{n\varepsilon h}{8M^2M_4^2} = \frac{n\varepsilon C n^{-1/5} n^{-2\tau}}{8M_4^2} = Cn^{\frac{4}{5}-2\tau} = O(n^{\frac{4}{5}-2\tau}) \to \infty.$$

Hence

$$2\exp(-\frac{\varepsilon}{h}\frac{n\varepsilon h}{8M^2M_4^2}) \leq 2\exp(-\frac{\varepsilon}{h}).$$

Similarly,

$$\frac{h}{\varepsilon}(Mm_4 - \log(n)) = Cn^{-1/5}(n^\tau - \log n) = C(n^{\tau-\frac{1}{5}} - n^{-\frac{1}{5}}\log n) = O(n^{\tau-\frac{1}{5}}) \to \infty.$$

Hence

$$m_3\exp(-\frac{\varepsilon}{h}\frac{h(Mm_4-\log(n))}{\varepsilon}) \leq m_3\exp(-\frac{\varepsilon}{h}).$$

Thus (4.20) is simplified as

$$P(|Z_2 - EZ_2| \geq \varepsilon/2) \leq C\cdot\exp(-\frac{\varepsilon}{h}).$$

Furthermore, (4.17) indicates

$$P(|Z_2 - hf(u)m(u)| \geq \varepsilon) \leq C\cdot\exp(-\frac{\varepsilon}{h}).$$

3. By the same argument as proof of Theorem 5,

$$P(|\hat{\rho}(x_j, y|u) - \rho(x_j, y|u)| \geq \varepsilon) \leq C\cdot\exp(-\varepsilon/h).$$

Step II: Prove $P\left(\max_{1\le j\le p}|\widehat{\rho}_j^* - \rho_{j0}^*| > c_3\cdot n^{-\kappa}\right) \le O\{np\exp(-n^{\frac{1}{5}-\kappa}/\xi)\}$. Using the aforementioned techniques in Appendix A,

$$\begin{aligned} P(|\widehat{\rho}_j^* - \rho_{j0}^*| \ge c_3 n^{-\kappa}) &\le P(|\widehat{\rho}_j^* - \rho_j^*| \ge c_3 n^{-\kappa}/2) + P(|\rho_j^* - \rho_{j0}^*| \ge c_3 n^{-\kappa}/2) \\ &\le nC\exp(-\frac{c_3 n^{-\kappa}}{Ch}) + 2\exp(-c_3^2\frac{n^{1-2\kappa}}{8}). \end{aligned}$$

Hence, the maximum satisfies

$$\begin{aligned} P\left(\max_{1\le j\le p}|\widehat{\rho}_j^* - \rho_{j0}^*| \ge c_3 n^{-\kappa}\right) &\le \sum_{j=1}^{p} P(|\widehat{\rho}_j^* - \rho_{j0}^*| \ge c_3 n^{-\kappa}) \\ &\le p\cdot nC\exp\left(-\frac{c_3 n^{-\kappa}}{Ch}\right) + 2p\exp(-c_3^2\frac{n^{1-2\kappa}}{8}) \\ &= O\{np\exp(-n^{\frac{1}{5}-\kappa}/\xi)\}, \end{aligned}$$

where $0 \le \kappa < 1/5$, and ξ is determined by c_3 and C. The last equation is because the first term dominates the second when $h = O(n^{-1/5})$. Therefore, the first part of Theorem 6 is proved.

Step III: Furthermore under condition (C6), prove

$$P(\mathcal{M}_* \subset \widehat{\mathcal{M}}) \ge 1 - O\{ns_n\exp(-n^{\frac{1}{5}-\kappa}/\xi_{b,c})\}.$$

According to the definition of $\widehat{\mathcal{M}}$ in (4.16) and condition (C6),

$$\begin{aligned} P(\mathcal{M}_* \subset \widehat{\mathcal{M}}) &= P\left(\min_{j\in\mathcal{M}_*}\widehat{\rho}_j^* \ge c_3 n^{-\kappa}\right) \\ &= P\left(\min_{j\in\mathcal{M}_*}\rho_{j0}^* - \min_{j\in\mathcal{M}_*}\widehat{\rho}_j^* \le \min_{j\in\mathcal{M}_*}\rho_{j0}^* - c_3 n^{-\kappa}\right) \\ &\ge P\left(\min_{j\in\mathcal{M}_*}\rho_{j0}^* - \min_{j\in\mathcal{M}_*}\widehat{\rho}_j^* \le 2c_3 n^{-\kappa} - c_3 n^{-\kappa}\right) \\ &\ge P\left(\max_{j\in\mathcal{M}_*}|\rho_{j0}^* - \widehat{\rho}_j^*| \le c_3 n^{-\kappa}\right) \\ &= 1 - P\left(\max_{j\in\mathcal{M}_*}|\rho_{j0}^* - \widehat{\rho}_j^*| \ge c_3 n^{-\kappa}\right) \\ &\ge 1 - s_n P\left(|\rho_{j0}^* - \widehat{\rho}_j^*| \ge c_3 n^{-\kappa}\right) \end{aligned}$$

$$\geq \quad 1 - O\{ns_n \exp(-n^{\frac{1}{5}-\kappa}/\xi)\}.$$

□

Chapter 5

Statistical Methods for Ultrahigh Dimensional Varying Coefficient Models with Longitudinal Structure

5.1 Methodology

In the last chapter, we focus on the univariate varying coefficient model, that is, the index variable u is one dimensional, making the observations $(u_i, \mathbf{x}_i, y_i),\ i = 1, \ldots, n$ independent with each other. However, one might be interested in the dynamic pattern of some predictor effects in practice. For example, the effect of SNP on BMI may depend on not only the baseline age values, but also the dynamic pattern of age, that is, the coefficients associated with SNP might change as the person gets older. One of the most widely used technique for this type of problems is time-varying coefficient models, or longitudinal data structure.

Consider the repeated measurements $\{(t_{ij}, \mathbf{x}_{ij}, y_{ij}),\ i = 1, \ldots, n \text{ and } j = 1, \ldots, m_i\}$, where y_{ij} is the jth outcome from the ith subject corresponding to the time point t_{ij}, $\mathbf{x}_{ij} = (x_{ij1}, ..., x_{ijp})^T$ is the ultrahigh dimensional covariate vector corresponding to t_{ij}. The model with longitudinal data structure is represented as

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}(t_{ij}) + \epsilon_i(t_{ij}),\ i = 1, \ldots, n;\quad j = 1, \ldots, m_i, \tag{5.1}$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$ is the vector of smooth parameter function. The challenge of the above model compared to the univariate varying coefficient model lies in that although the measurements are assumed to be independent for different subjects, they can be correlated among different time points within each subject. Hence the sample $\{(t_{ij}, \mathbf{x}_{ij}, y_{ij}),\ i = 1, \ldots, n;\ j = 1, \ldots, m_i\}$ is no longer independently and identically distributed. However, we will show that our screening procedure still work well under this circumstance.

Specifically, we pool all the observations from different subjects together and treat them as independent. Based on the definition of the conditional correlation defined in chapter 3, we first compute $\hat{\rho}(x_k, y|t_{ij}),\ i = 1, \ldots, n;\ j = 1, \ldots, m_i$ for each predictor x_k, then define the final screening criterion ρ^* corresponding to x_k as

$$\rho_k^* = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \hat{\rho}^2(x_k, y|t_{ij}),$$

where $N = \sum_{i=1}^{n} m_i$.

After ρ^* value for each predictor x_k is obtained, we can rank them in a decreasing order and pick the top d as our screened model. However, since all the N observations are treated as independent sample points, the total sample size becomes N instead of n. Hence the moderate size d is now defined as $d = \nu \cdot [N^{4/5}/\log(N^{4/5})]$ for $\nu = 1, 2, 3$, etc. The simulation results below illustrate the validity of our screening method.

5.2 Monte Carlo Simulations

We conduct the following Monte Carlo simulations to assess the performance of the proposed screening technique. In each example, we set $p = 1000$ and $n = 50$. For the ith subject $i = 1, \ldots, n$, we draw m_i correlated observations, where m_i's are integers between 2 and 10, and the time points t_{ij}'s are standardized such that $t_{ij} \in [0, 1],\ i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. More specifically, we draw m_i and t_{ij} from the uniform distributions:

$$m_i \sim U\{2, 3, \ldots, 10\}, \quad \text{and} \quad t_{ij} \sim U[0, 1].$$

To make the m_i observations correlated with each other, the m_i-dimensional random error $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_{i1}), \ldots, \varepsilon_i(t_{im_i}))^T$ is drawn from

$$\boldsymbol{\varepsilon}_i \sim MVN(\mathbf{0}, \Sigma_\varepsilon), \text{ where } (\Sigma_\varepsilon)_{jk} = 0.5 \cdot \rho_\varepsilon^{|j-k|}.$$

Here we set $\rho_\varepsilon = 0.5$. In addition, we generate p correlated predictors based on AR(1) covariance structure with $\rho = 0.4$ and $\rho = 0.8$, and we apply different generating techniques in different examples. In total, there exist two types of correlation: One is that among observations within each subject, and the other is that among different predictors. We repeat the experiment 100 times, and in each experiment, three submodels are chosen according to the cutoff $d = \nu \cdot d_0$, $\nu = 1, 2, 3$, where $d_0 = [N^{4/5}/\log(N^{4/5})]$, $N = \sum_{i=1}^n m_i$. Therefore, the value of total sample size N varies with simulation.

To evaluate the performance, we still use the previously introduced criteria in chapter 3, and report the comparison between our screening method and SIS (Fan and Lv, 2008).

Example 1. In this example, we generate the predictors in the following fashion. Notice that we need to make the x's correlated within each of the n subjects. First we draw n i.i.d random vectors from the multivariate normal distribution:

$$\mathbf{x}_i \sim MVN(\mathbf{0}, \Sigma), \ i = 1, \ldots, n, \quad \text{where } \Sigma_{jk} = \rho^{|j-k|}, \quad j, k = 1, \ldots, p.$$

Thus the p predictors are correlated. Then treat each row $\mathbf{x}_i$ as the first observation for the ith subject, and generate the remaining observations for the same subject iteratively: Denote the current observation as $\mathbf{x}_1$, the corresponding time point as t_1, and the next observation as $\mathbf{x}_2$ with the time point t_2, then $\mathbf{x}_2$ is generated by

$$\mathbf{x}_2 = \mathbf{x}_1 + (t_2 - t_1)\boldsymbol{\delta}, \quad \text{where } \boldsymbol{\delta} \sim MVN(\mathbf{0}, I).$$

Therefore, the observations within each subject are highly correlated. The nonzero coefficient functions are

$$\beta_2(t) = 5I(t > 0.4), \quad \beta_{100}(t) = 3 + t, \quad \beta_{400}(t) = 3(2 - 3t)^2$$

$$\beta_{600}(t) = 6\sin(2\pi t), \quad \beta_{1000}(t) = 3\exp\{t/(t+1)\}, \quad \text{other } \beta(t)\text{'s are 0.}$$

Based on $\mathbf{x}$, $\boldsymbol{\beta}(t)$ and the random noise $\boldsymbol{\varepsilon}$, the response y is then generated according to the true model

$$y_{ij} = \beta_2(t_{ij})x_{ij2}+\beta_{100}(t_{ij})x_{ij100}+\beta_{400}(t_{ij})x_{ij400}+\beta_{600}(t_{ij})x_{ij600}+\beta_{1000}(t_{ij})x_{ij1000}+\varepsilon_{ij}.$$

Table 5.1, 5.2 and 5.3 illustrate the comparison between the two screening procedures in terms of the aforementioned criteria.

Table 5.1. The proportions p_j and p_a for Example 1.

	SIS						Feature screening for VCM					
d	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
	$\rho = 0.4$											
d_0	0.75	0.90	0.84	0.07	0.99	0.04	0.97	0.92	1	0.99	0.99	0.87
$2d_0$	0.88	0.95	0.92	0.13	0.99	0.09	0.97	0.94	1	1.00	0.99	0.90
$3d_0$	0.90	0.96	0.95	0.18	0.99	0.12	1.00	0.96	1	1.00	0.99	0.95
	$\rho = 0.8$											
d_0	0.80	0.92	0.88	0.05	0.99	0.02	0.94	0.92	0.97	0.97	1	0.82
$2d_0$	0.86	0.97	0.93	0.12	1.00	0.09	0.97	0.98	1.00	0.99	1	0.94
$3d_0$	0.88	0.99	0.93	0.19	1.00	0.14	0.98	0.99	1.00	0.99	1	0.96

Table 5.2. $rank_j$ of each true predictor x_j for Example 1.

	SIS					Feature screening for VCM				
ρ	x_2	x_{100}	x_{400}	x_{600}	x_{1000}	x_2	x_{100}	x_{400}	x_{600}	x_{1000}
0.4	51.46	21.73	29.18	468.64	5.36	9.95	20.37	5.68	6.85	9.23
0.8	61.18	16.38	38.86	483.73	3.71	17.06	15.17	7.27	11.62	3.99

Table 5.3. The minimum model size M for Example 1.

	SIS					Feature screening for VCM				
ρ	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.4	88.25	282.75	448.50	696.00	921.10	5.00	7.00	14.00	28.50	156.25
0.8	80.85	290.00	510.50	739.00	926.10	5.00	9.00	22.50	40.00	113.80

From Table 5.1, the sure screening property (Fan and Lv, 2008) of our proposed method is visualized. That is, although the observations are now not independent any more, we are still able to cover most of the important variables with an overwhelming probability. However, since the empirical mean value for β_{600} is about

0, SIS fails to detect the corresponding predictor for most of the simulation runs. Furthermore, we can see that the output for $\rho = 0.4$ and $\rho = 0.8$ do not differ much from the table, indicating that our screening procedure is robust against the correlation between predictors.

The ranking consistency (Zhu, Li, Li and Zhu, 2011) can be obtained from Table 5.2, where all the important predictors rank in the top in terms of the ρ^* values with our method, while x_{600} in SIS does not. The minimum model size M reported in Table 5.3 captures both sure screening property and ranking consistency.

Example 2. This example mimics the SNP data. Specifically, the predictors are categorical, taking values of 0, 1 or 2, and they have identical value for different time points of the same subject. To simulate the SNP data, we first draw n i.i.d. random vectors $\mathbf{x}^*$'s using the strategy for generating the i.i.d $\mathbf{x}$'s in Example 1, i.e.

$$\mathbf{x}_i^* \sim MVN(\mathbf{0}, \Sigma),\ i = 1, \ldots, n, \quad \text{where } \Sigma_{jk} = \rho^{|j-k|}, \quad j, k = 1, \ldots, p.$$

Then we recode the x_{ij} as 0, 1, 2, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, according to the 25% and 75% empirical quantiles of x_{ij}^*, q_1 and q_3:

$$x_{ij} = \begin{cases} 2, & \text{if } x_{ij}^* \leq q_1; \\ 1, & \text{if } q_1 < x_{ij}^* \leq q_3; \\ 0, & \text{if } x_{ij}^* > q_3. \end{cases}$$

By doing this we guarantee $P(x = 0) = P(x = 2) = 1/4$, and $P(x = 1) = 1/2$. Since the SNP values stay the same for different time points within each subject, we have $x_{ij} = x_{ij'},\ j, j' = 1, \ldots, m_i$. All the other quantities are generated in the same fashion as Example 1. Table 5.4, 5.5 and 5.6 shows the performance of both out method and SIS.

The performance of the two methods in this example is almost identical with the previous one, thus our procedure is valid not only for continuous predictors, but also for categorical x values. SIS still does not work well for x_{600}.

Example 3. This example assesses how the screening procedure works when the

Table 5.4. The proportions p_j and p_a for Example 2.

	SIS						Feature screening for VCM					
d	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
	$\rho = 0.4$											
d_0	0.78	0.88	0.86	0.03	0.97	0.01	0.97	0.93	0.98	1	0.97	0.85
$2d_0$	0.88	0.96	0.91	0.10	0.99	0.05	0.99	0.95	1.00	1	1.00	0.94
$3d_0$	0.91	0.98	0.95	0.18	0.99	0.15	0.99	0.98	1.00	1	1.00	0.97
	$\rho = 0.8$											
d_0	0.60	0.87	0.81	0.03	0.99	0.03	0.91	0.88	1	0.98	1	0.78
$2d_0$	0.76	0.91	0.89	0.09	0.99	0.07	0.94	0.93	1	0.99	1	0.86
$3d_0$	0.85	0.93	0.90	0.13	0.99	0.10	0.98	0.93	1	1.00	1	0.91

Table 5.5. $rank_j$ of each true predictor x_j for Example 2.

	SIS					Feature screening for VCM				
ρ	x_2	x_{100}	x_{400}	x_{600}	x_{1000}	x_2	x_{100}	x_{400}	x_{600}	x_{1000}
0.4	39.57	22.59	43.77	483.87	7.42	10.23	20.31	6.31	6.49	5.86
0.8	91.71	30.79	70.05	505.07	6.56	20.89	33.12	6.56	8.78	4.82

Table 5.6. The minimum model size M for Example 2.

	SIS					Feature screening for VCM				
ρ	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.4	106.75	255.00	522.50	710.50	892.55	6.00	9.00	16.00	34.25	112.05
0.8	81.80	300.75	509.00	757.50	975.35	6.95	13.00	22.50	46.50	255.90

important predictors are adjacent, namely, they are highly correlated with each other. Under the same data generating technique as Example 1, we specify the nonzero coefficient functions as

$$\beta_1(t) = 5I(t > 0.4), \quad \beta_2(t) = 3 + t, \quad \beta_3(t) = 3(2 - 3t)^2$$
$$\beta_4(t) = 6\sin(2\pi t), \quad \beta_5(t) = 3\exp\{t/(t+1)\}, \quad \text{other } \beta(t)\text{'s are 0.}$$

In this fashion, x_1 to x_5 are active and highly correlated with adjacent correlation $\rho = 0.4$ and 0.8. The following tables 5.7, 5.8 and 5.9 demonstrate the two screening procedure in this example.

From the above three tables, we can see that if the important predictors are correlated with each other, their signals can be reinforced to a large degree due to the signals from other adjacent active variables. Therefore, it is much easier to

Table 5.7. The proportions p_j and p_a for Example 3.

	SIS						Feature screening for VCM					
d	p_1	p_2	p_3	p_4	p_5	p_a	p_1	p_2	p_3	p_4	p_5	p_a
	$\rho = 0.4$											
d_0	0.98	0.98	0.98	0.30	0.99	0.28	0.99	0.99	1	1	1	0.98
$2d_0$	1.00	0.99	1.00	0.41	1.00	0.41	1.00	0.99	1	1	1	0.99
$3d_0$	1.00	0.99	1.00	0.45	1.00	0.45	1.00	0.99	1	1	1	0.99
	$\rho = 0.8$											
d_0	1	1	1	0.85	1	0.85	1	1	1	1	1	1
$2d_0$	1	1	1	0.90	1	0.90	1	1	1	1	1	1
$3d_0$	1	1	1	0.92	1	0.92	1	1	1	1	1	1

Table 5.8. $rank_j$ of each true predictor x_j for Example 3.

	SIS					Feature screening for VCM				
ρ	x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
0.4	9.49	5.51	7.44	336.12	3.82	3.90	5.22	3.21	6.10	4.12
0.8	4.46	2.42	2.25	41.38	2.82	3.29	2.48	2.38	5.31	2.75

Table 5.9. The minimum model size M for Example 3.

	SIS					Feature screening for VCM				
ρ	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.4	10.90	48.00	228.00	630.25	938.50	5.00	5.00	6.00	11.00	31.35
0.8	5.00	5.00	9.00	22.25	223.55	5.00	5.00	5.00	6.00	12.05

include all the important predictors than when they fall apart. Especially when the correlation is high, i.e. $\rho = 0.8$, even SIS has a relatively large probability to include the true model. But our method still outperforms SIS in terms of the three criteria.

5.3 FHS: Longitudinal Data Structure

In the last chapter, we assume the effect of SNP on BMI depend on the value of age, and the coefficient plots support our statement. However, we only focus on the baseline age, namely, we only use a single value of age for the varying coefficient model and assume each subject only possesses one age value. Observing the data, however, each subject was followed by 5 to 26 times, hence we are able to study the dynamic pattern of the effect of SNP on the response BMI. Therefore

in this section, we consider the longitudinal structure model (5.1), where the time points t_{ij}'s are now characterized by different age values u_{ij} of each subject. More specifically, the model studied here can be represented as:

$$y_{ij} = \beta_0(u_{ij}) + (\mathbf{x}_a)_i^T \boldsymbol{\beta}_a(u_{ij}) + (\mathbf{x}_d)_i^T \boldsymbol{\beta}_d(u_{ij}) + \varepsilon_i(u_{ij}), \tag{5.2}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, and m_i is between 5 and 26. The definition of $\boldsymbol{\beta}_a(u)$, $\boldsymbol{\beta}_d(u)$, $\mathbf{x}_{aj}$ and $\mathbf{x}_{dj}$ are the same as before, and the x-values stays identical for different ages of each subject.

To reduce ultrahigh model size and select important SNPs for dynamically explaining BMI, we apply the aforementioned two-stage approach.

Stage 1: feature screening procedure

In the first stage, the feature screening method for longitudinal data introduced in last subsection is applied. Note that since the pooled sample size is now $N = 6,590$, so the moderate sample size d is defined as $d = [N^{4/5}/\log(N^{4/5})] = 161$, that is, the submodel is chosen as

$$\mathcal{M}_\gamma = \{k : 1 \leq k \leq 2p,\ \rho_k^* \text{ is among the first 161 largest of all } \rho^*\text{'s}\}.$$

Then we obtain the submodel of size 161 as follows:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}(u_{ij}) + \varepsilon_{ij},\ i = 1, \ldots, n;\ j = 1, \ldots, m_i,$$

where $\mathbf{x}_{ij}$ is now a 161-dimensional vector, and $\boldsymbol{\beta}(u) = (\beta_1(u), \ldots, \beta_{161}(u))^T$.

Stage 2: Post-screening variable selection

In the variable selection stage, we first treat the N observations from n subjects as independent, and conduct three penalized regression techniques, LASSO, Adaptive LASSO and SCAD for further selecting important SNPs based on the submodel above. As is known, misspecification of covariance structure does not affect the consistency of the estimations. And after we select the final significant SNPs, the profile weighted least squares approach (Fan, Huang and Li, 2007) cooperating the covariance matrix estimation is applied to improve the efficiency of the estimated

coefficients and predicting trajectories of individuals.

First, pooling all the N observations from different subjects together, the sub-model chosen in the first stage becomes

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}(u_i) + \varepsilon_i, \ i = 1, \ldots, N, \tag{5.3}$$

where $\boldsymbol{\beta}(u) = (\beta_1(u), \ldots, \beta_{161}(u))^T$, the X-matrix consists of the corresponding columns from the original SNP data, and the pooled sample size $N = \sum_{i=1}^{n} m_i = 6,590$.

Before applying any variable selection technique to the model, we first look at the 161 plots for the unpenalized coefficient estimations for the 161 SNPs together with their confidence bands, as shown in Figure 5.1. We are able to visually detect some significant SNPs whose confidence bands do not contain 0, and their corresponding coefficients do differ with age. Hence a variable selection technique incorporating varying coefficient structure are desirable.

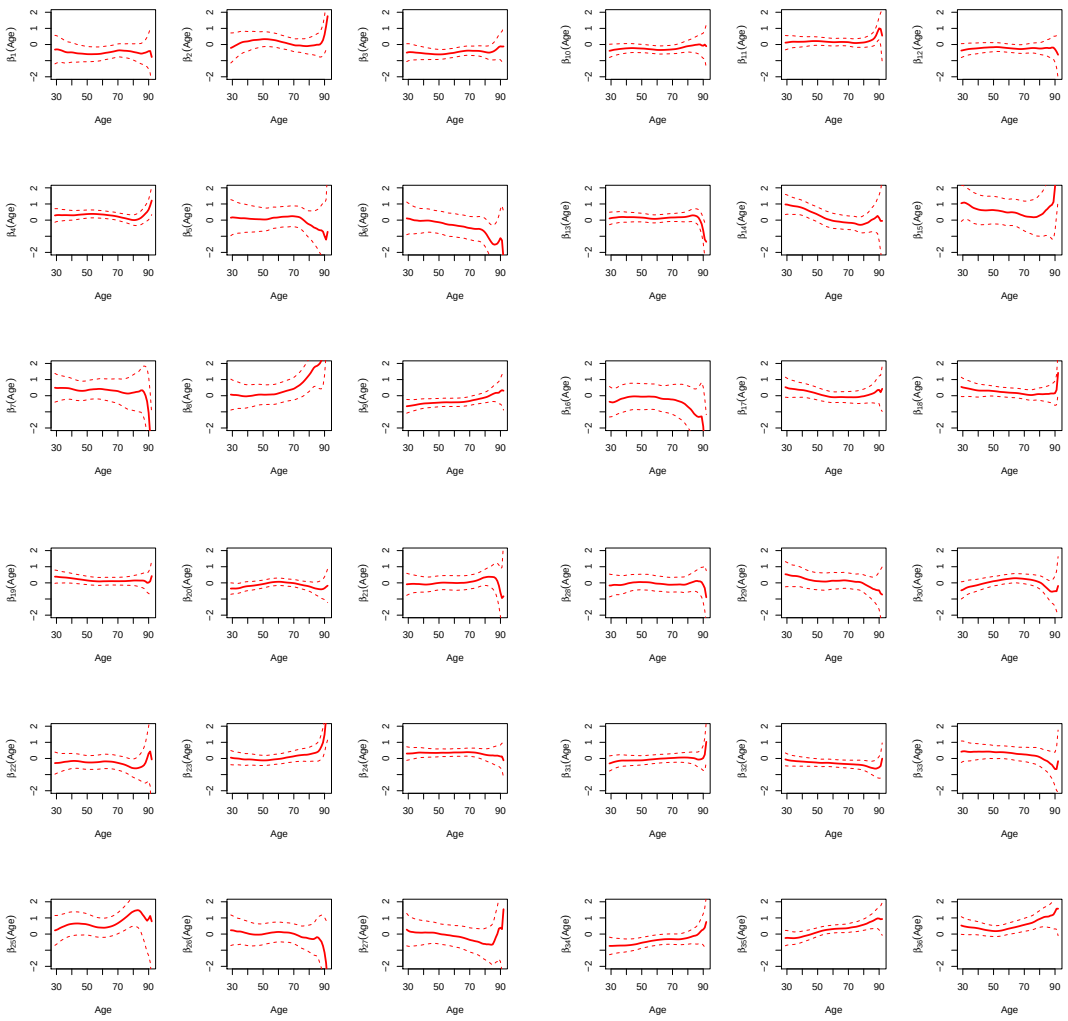
Following the idea of last chapter, we still apply three penalized regression methods with LASSO, Adaptive LASSO and SCAD penalties. However, when we choose the tuning parameter λ, since we originally have relatively large model size 161 here, the traditional criteria AIC, BIC and GCV are too conservative to get a sparse model, because even for the least conservative BIC, it tends to assign the weight proportional to the model size; AIC and GCV tend to be even more conservative and they yield identical models. The chosen model size and the tuning parameter selection plots are demonstrated in Table 5.10 and Figure 5.2.

Table 5.10. The chosen model sizes based on AIC, BIC, GCV.

	screening+LASSO	screening+AdaptiveLASSO	screening+SCAD
AIC	161	161	155
BIC	135	115	102
GCV	161	161	155

To obtain more sparse models, we adopt two new tuning parameter selection techniques here, the extended Bayesian information criteria (EBIC, Chen and Chen, 2008) and the modified Bayesian information criteria (MBIC, Bogdan et.al,

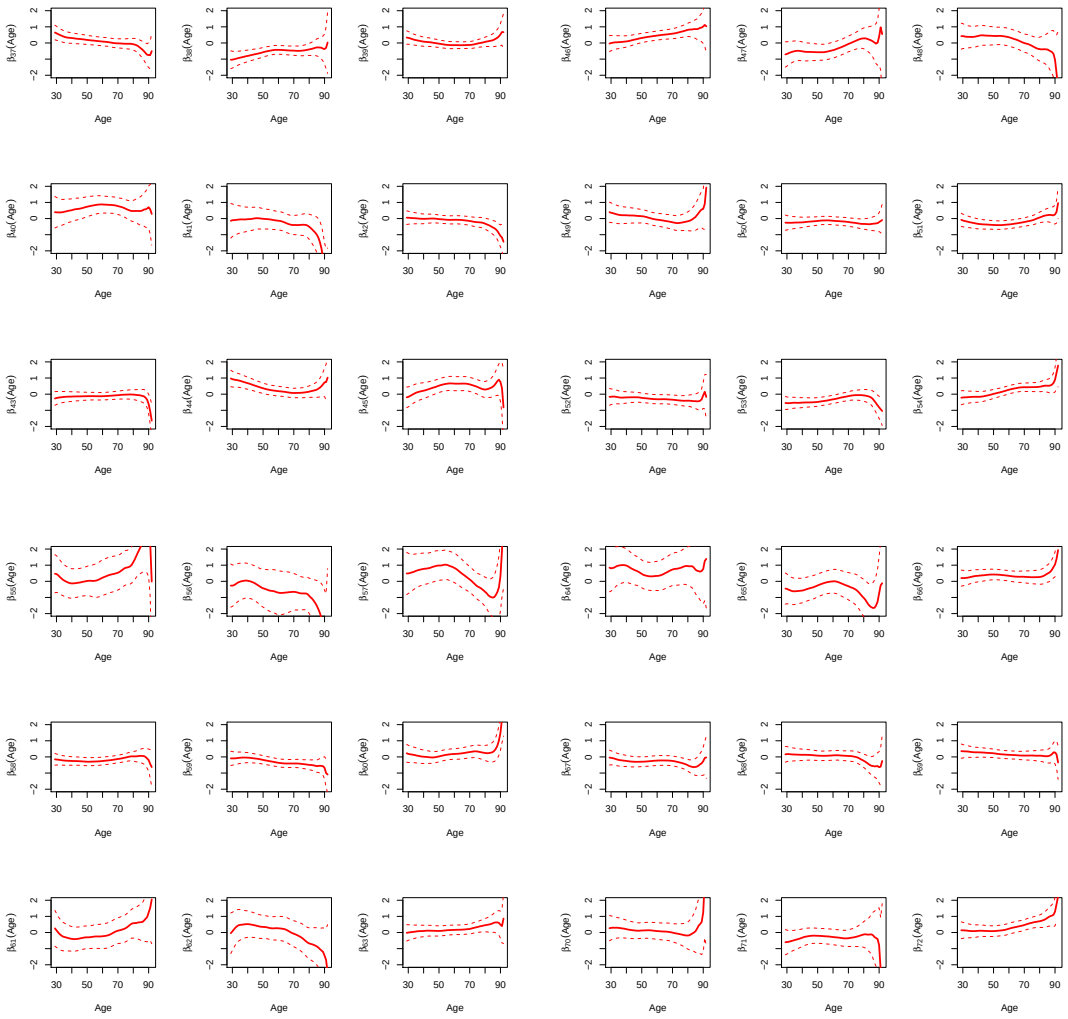
Figure 5.1. The estimated coefficient functions of the screened model



2004), which are defined as follows:

$$EBIC = N\log(RSS) + df_\lambda \log(N) + 2\gamma \log \binom{df_0}{df_\lambda}$$
$$MBIC = N\log(RSS) + df_\lambda \log(N) + 2df_\lambda \log(df_0/2.2 - 1),$$

Figure 5.1. (Cont'd.) The estimated coefficient functions of the screened model



where df_λ is defined as (3.13) in last chapter (Fan, Zhang, and Zhang, 1999):

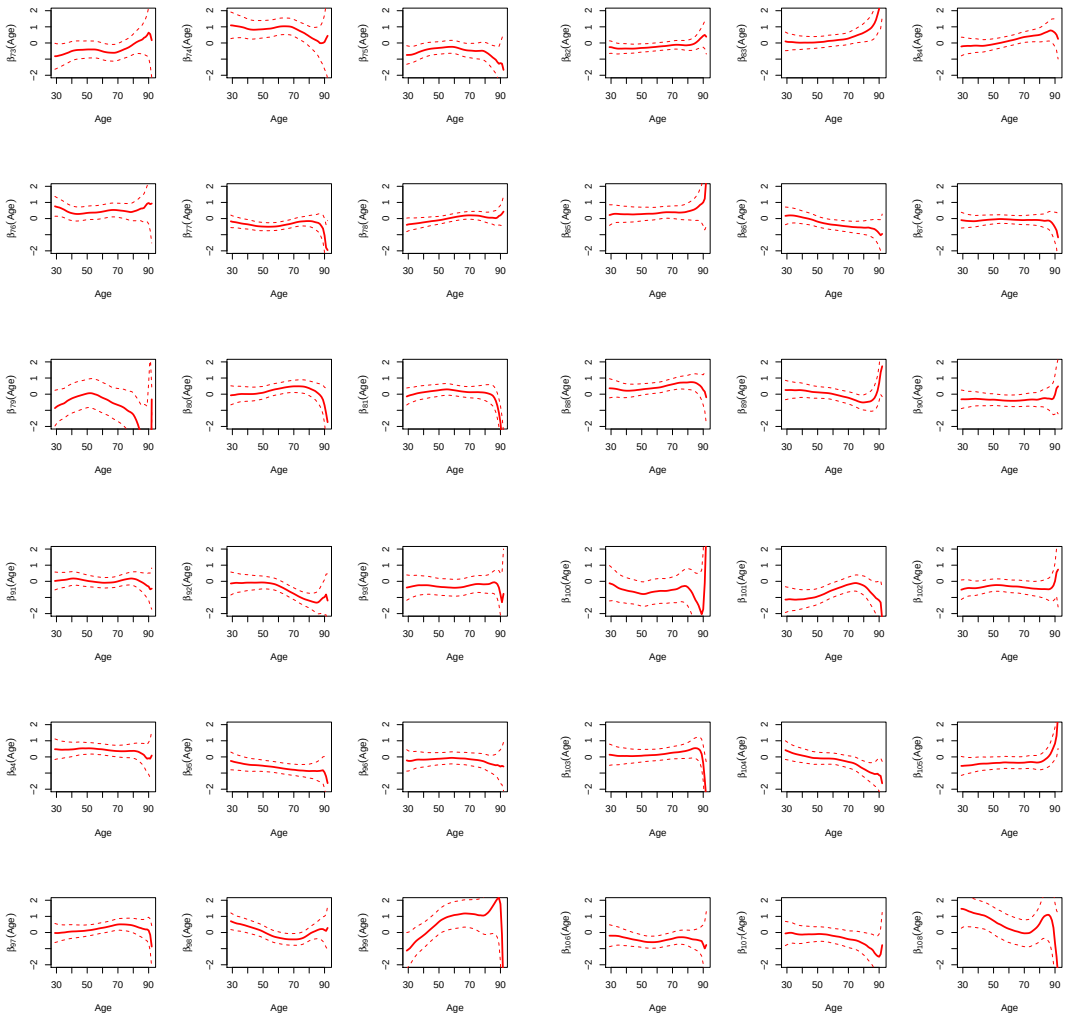
$$df_\lambda = \gamma_K |\mathbb{U}| d_\lambda \{K(0) - 0.5 \int K^2(u)du\}/h,$$

and df_0 is the corresponding degree of freedom for the original full model:

$$df_0 = \gamma_K |\mathbb{U}| d \{K(0) - 0.5 \int K^2(u)du\}/h,$$

with $d = 161$ here. To obtain the residual sum of squares RSS above, we need

Figure 5.1. (Cont'd.) The estimated coefficient functions of the screened model



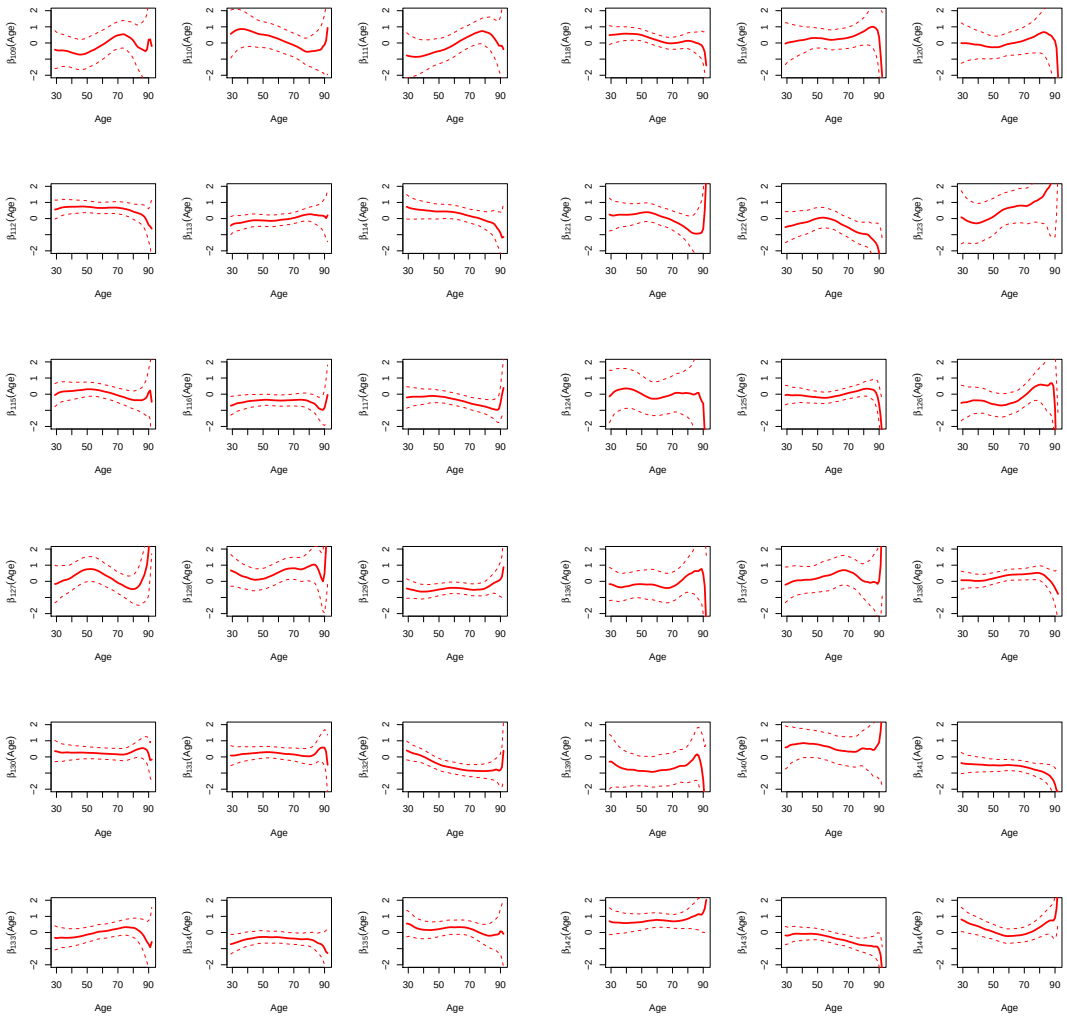
first select significant variables given a λ, then fit an unpenalized varying coefficient model based on the chosen predictors and compute the RSS, that is,

$$RSS_\lambda = \frac{1}{N}\sum_{i=1}^{N}\{y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\lambda 0}(u_i)\}^2,$$

where $\hat{\boldsymbol{\beta}}_{\lambda 0}(u)$ is the unpenalized coefficient estimation based on the chosen predictors corresponding to λ.

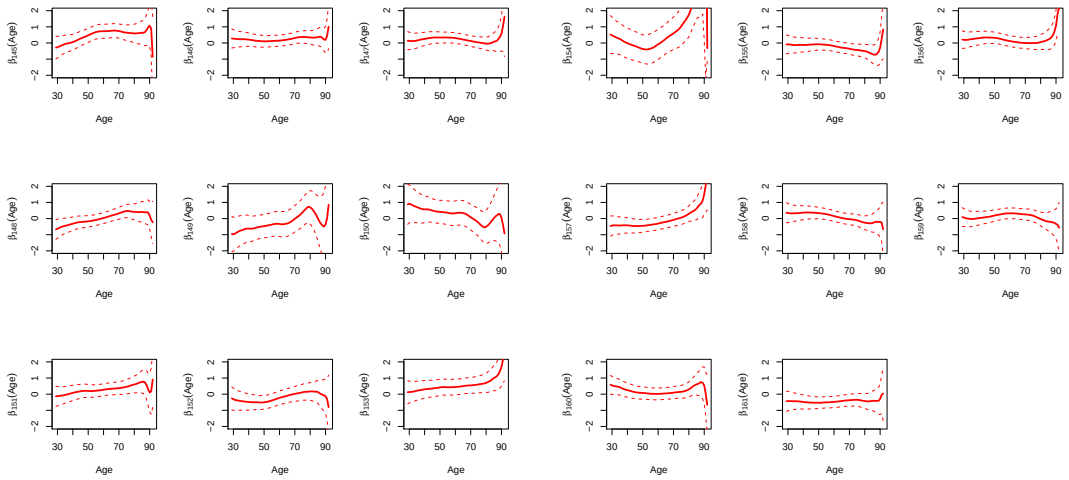
Figure 5.3 and Table 5.11 report some model selection results from these two

Figure 5.1. (Cont'd.) The estimated coefficient functions of the screened model



methods. Figure 5.3 depicts how the two criteria change over different λ's, and we adopt the λ that minimizes the corresponding criterion. From Table 5.11, both EBIC and MBIC yield much sparser models than AIC, BIC, and GCV. And between these two, MBIC is less conservative. The SCAD penalty together with MBIC produces the sparsest model with only 48 significant SNPs. Furthermore, among the 6 chosen models, smaller models are nested within larger ones, and based on the generalized likelihood ratio test (Fan and Zhang, 2001) mentioned in last chapter, the sparsest model is already sufficient to explain the response BMI.

Figure 5.1. (Cont'd.) The estimated coefficient functions of the screened model



Therefore, the remaining results are based on this model, with the nonzero coefficients $\beta_k(u),\ k \in \mathcal{S}$, and the significant index set $\mathcal{S}$ is the subset of $\{1, \ldots, 161\}$ from model (5.3), where

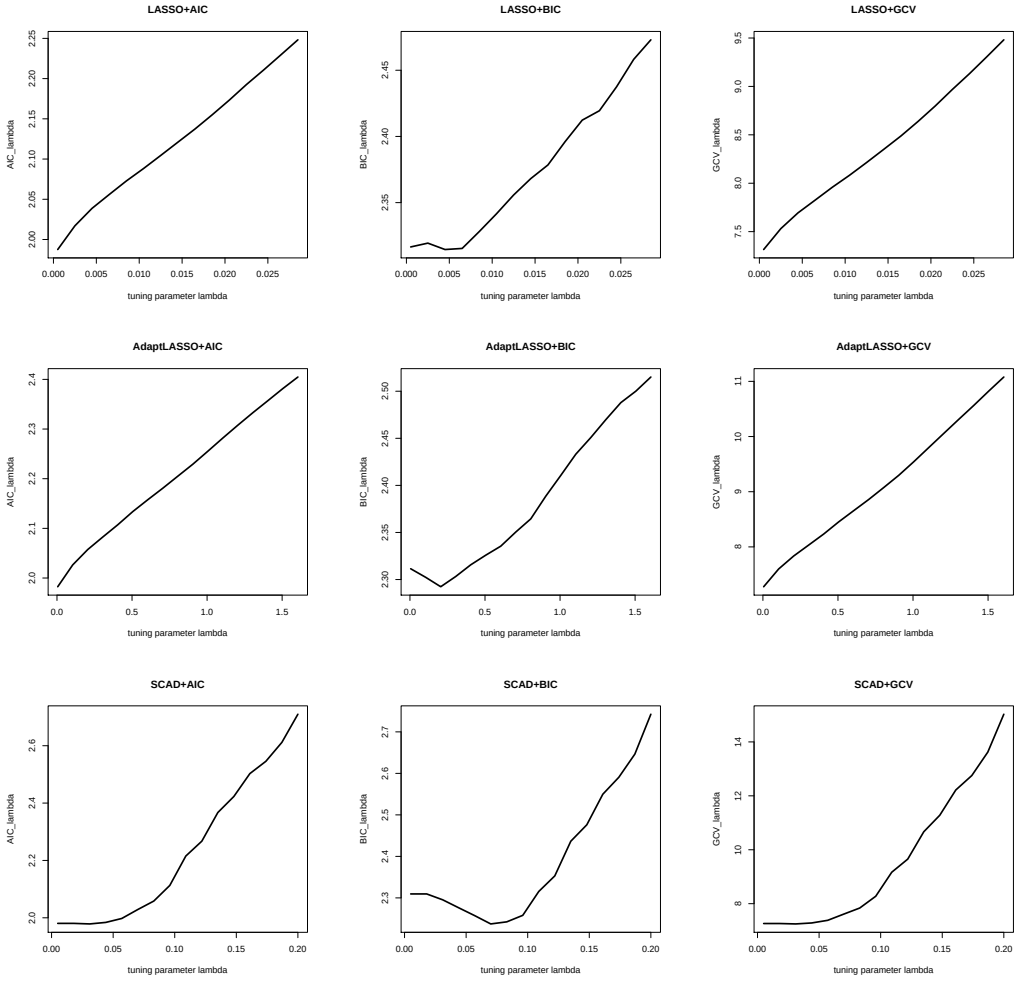
$$\begin{aligned}\mathcal{S} \;=\; &\{1, 3, 4, 9, 15, 24, 25, 32, 34, 35, 36, 38, 40, 45, 57, 59, 64, 65, 74, 77, 79, 88, 90, 92,\\ &94, 95, 99, 101, 105, 106, 108, 112, 114, 116, 117, 123, 128, 129, 132, 139, 140, 141,\\ &142, 143, 145, 153, 154, 161\}.\end{aligned}$$

Table 5.11. The chosen model sizes based on EBIC and MBIC.

	screening+LASSO	screening+AdaptiveLASSO	screening+SCAD
EBIC	118	78	106
MBIC	71	55	48

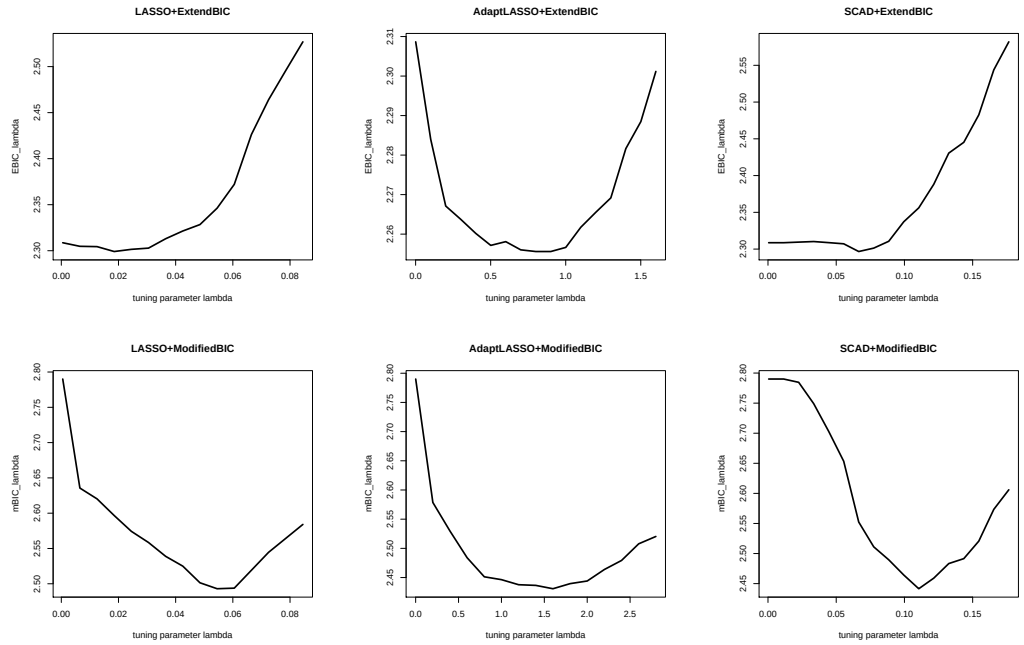
After we obtain the final sparse model with 48 SNPs, we can check the coefficient plots of the significant SNPs, as described in Figure 5.4. Recall that the penalized regression techniques enable us to estimate the coefficient functions at the same time as variable selection, and the red curves (the first and fourth columns) are the corresponding penalized estimations for the coefficient functions.

Figure 5.2. Tuning parameter selection based on AIC, BIC, GCV.



Meanwhile, we can also select significant variables first and then refit a model with the chosen SNPs without penalizing any coefficient. The results are more stable and reliable since we use the same information to estimate less parameters. The blue curves (the second and fifth columns) depict these unpenalized coefficient estimations. And the green curves (the third and sixth columns) are merely copy of the unpenalized estimation of the same SNPs from the original full model with size 161. For each significant SNP, the three estimated coefficient functions along with their confidence bands do not differ much. Here we take the unpenalized estimations (the second and fifth columns)for our further study.

Figure 5.3. Tuning parameter selection based on EBIC and MBIC.



From Figure 5.4, we notice that the confidence bands of some coefficient functions contain constant lines, that is, we can treat the corresponding coefficient functions as constants without varying with age, although they are significantly different from 0. More specifically, the chosen model is now modified as a semi-parametric varying-coefficient partially linear model:

$$y_{ij} = \mathbf{x}_{1,ij}^T \boldsymbol{\beta}_1(u_{ij}) + \mathbf{x}_{2,ij}^T \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_i(u_{ij}),\ i = 1, \ldots, n;\ j = 1, \ldots, m_i,$$

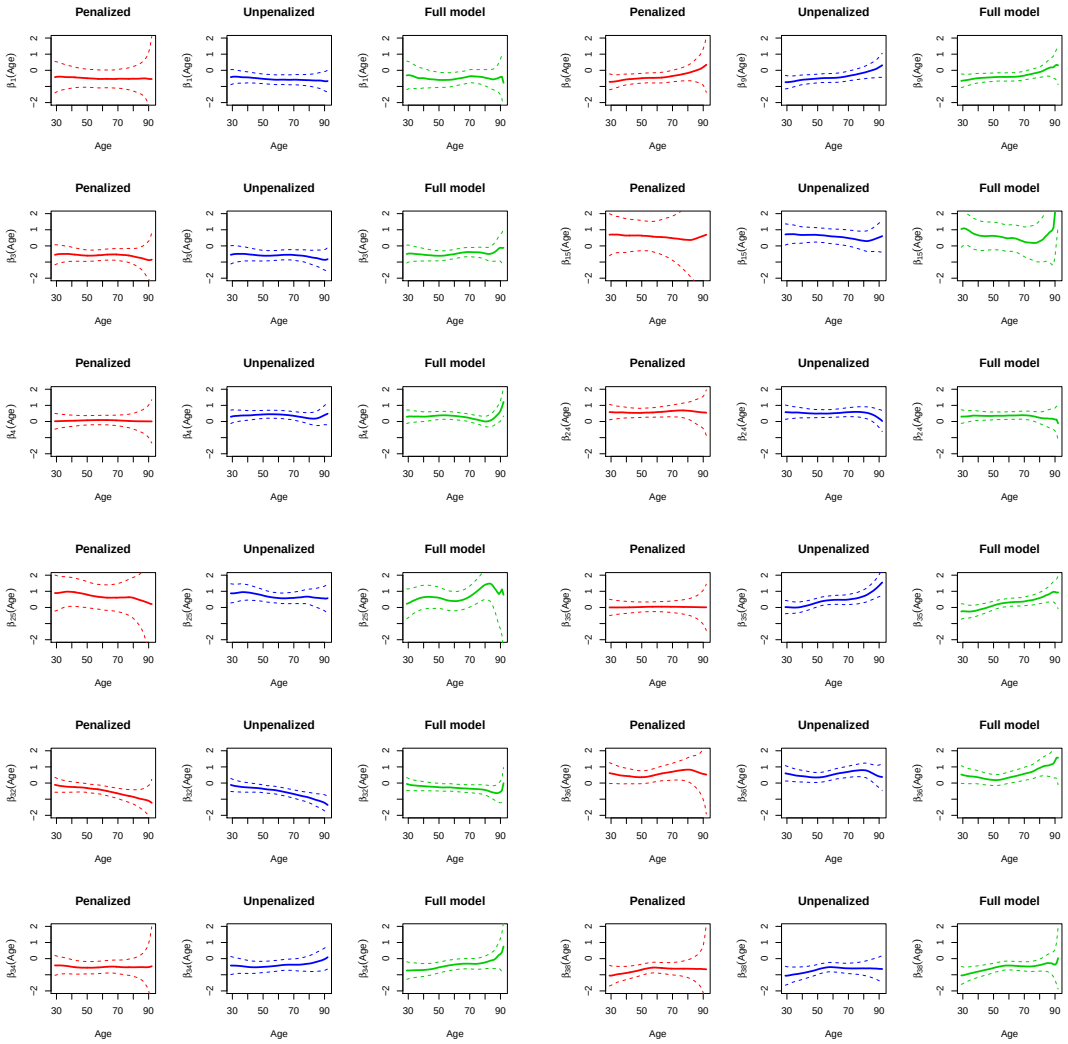
where the non-constant coefficient vector $\boldsymbol{\beta}_1(u)$ consists of $\beta_k(u),\ k \in \mathcal{F}$ with

$$\begin{aligned} \mathcal{F} \ = \ & \{9, 24, 32, 35, 40, 45, 57, 59, 65, 77, 88, 90, 99, 112, 114, 116, 117, 123, 128, \\ & 132, 140, 141, 143, 145, 161\}, \end{aligned}$$

the constant coefficient vector $\boldsymbol{\beta}_2$ contains $\beta_k,\ k \in \mathcal{C}$, with

$$\mathcal{C} = \mathcal{S}/\mathcal{F} = \{1, 2, 3, 15, 25, 34, 36, 38, 64, 79, 92, 94, 95, 101, 105, 106, 108, 129, 139, 142\},$$

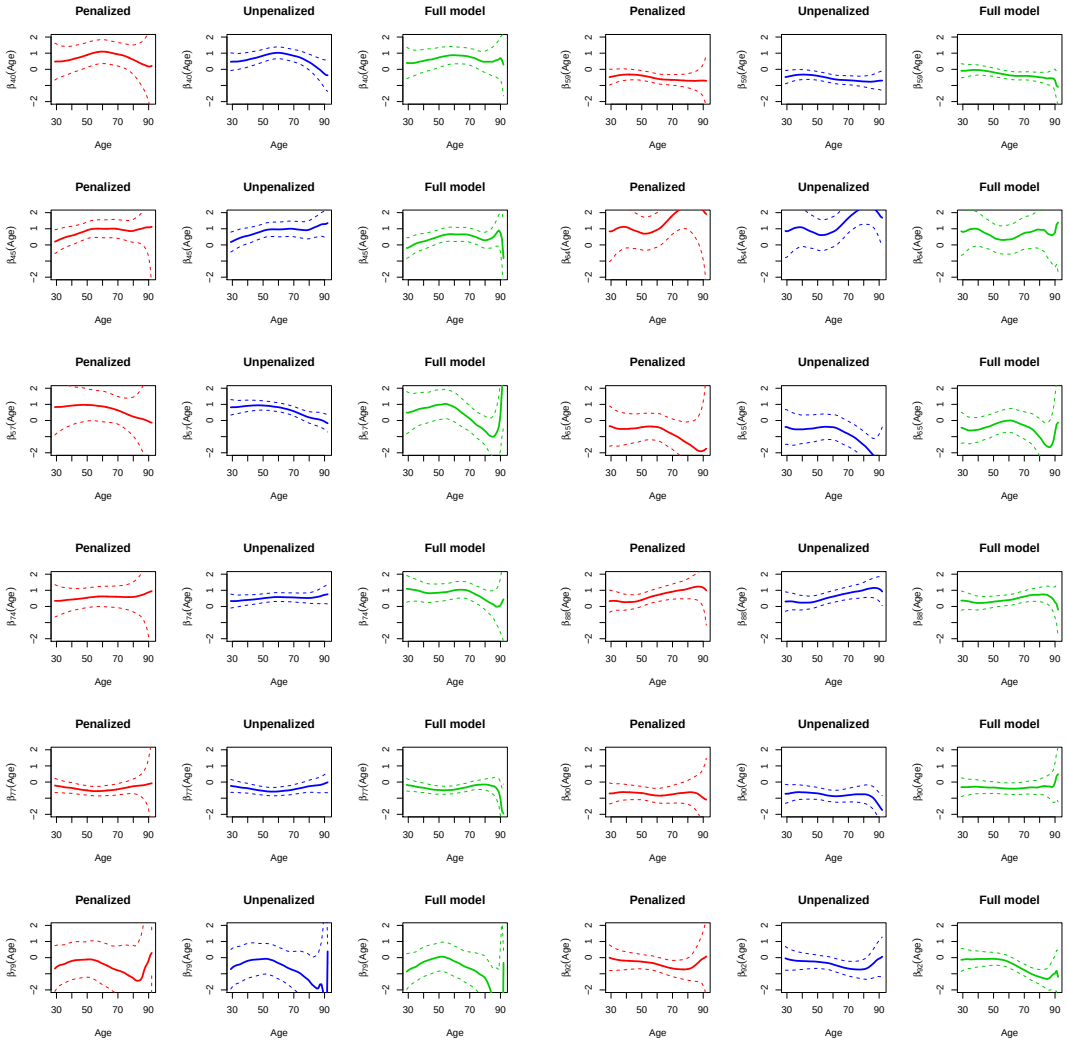
Figure 5.4. Estimated coefficients of penalized, unpenalized, and full model.



and the SNP vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ follow the same index sets.

We now apply the profile weighted least squares approach (PWLS, Fan, Huang and Li, 2007) to enhance the efficiency of our coefficient estimation, by taking into account the covariance structure which are estimated using minimum generalized variance (MGV) method. Assume $\boldsymbol{\varepsilon}_i$ has mean 0 and covariance matrix Σ_i. Then Σ_i is fully determined by the variance and correlation within subjects. Here we denote $\text{var}\{\varepsilon_i(u)|\mathbf{x}_1, \mathbf{x}_2\} = \sigma^2(u)$ and consider the ARMA(1,1) structure as the within subject correlation, i.e. $\text{corr}(\varepsilon_i(u), \varepsilon_i(v)) = \gamma\rho^{|u-v|}$. Hence the correlation

Figure 5.4 (Cont'd). Estimated coefficients of penalized, unpenalized, and full model.

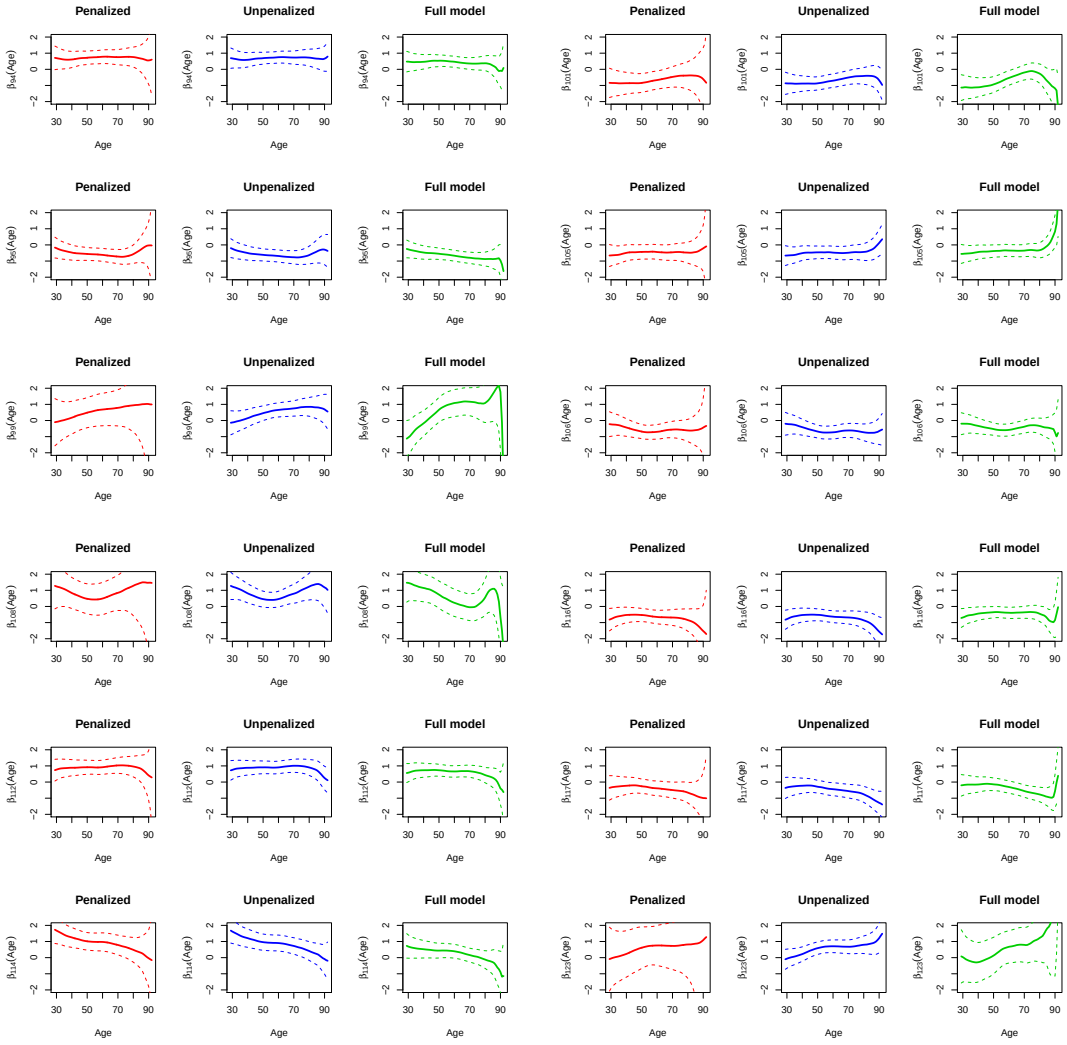


parameter vector to be estimated is $\boldsymbol{\theta} = (\gamma, \rho)$.

Since the estimations of $\sigma^2(u)$ and $\boldsymbol{\theta}$ depend on $\widehat{\boldsymbol{\beta}}_1(u)$ and $\widehat{\boldsymbol{\beta}}_2$, and on the other hand, improving the efficiency of $\widehat{\boldsymbol{\beta}}_1(u)$ and $\widehat{\boldsymbol{\beta}}_2$ relies on the $\widehat{\sigma}^2(u)$ and $\widehat{\boldsymbol{\theta}}$. Therefore, the estimation procedure must be done in steps:

Step 1. Conduct PWLS approach to estimate $\boldsymbol{\beta}_1(u)$ and $\boldsymbol{\beta}_2$ iteratively by ignoring the within subject correlation for the moment:

Figure 5.4 (Cont'd). Estimated coefficients of penalized, unpenalized, and full model.

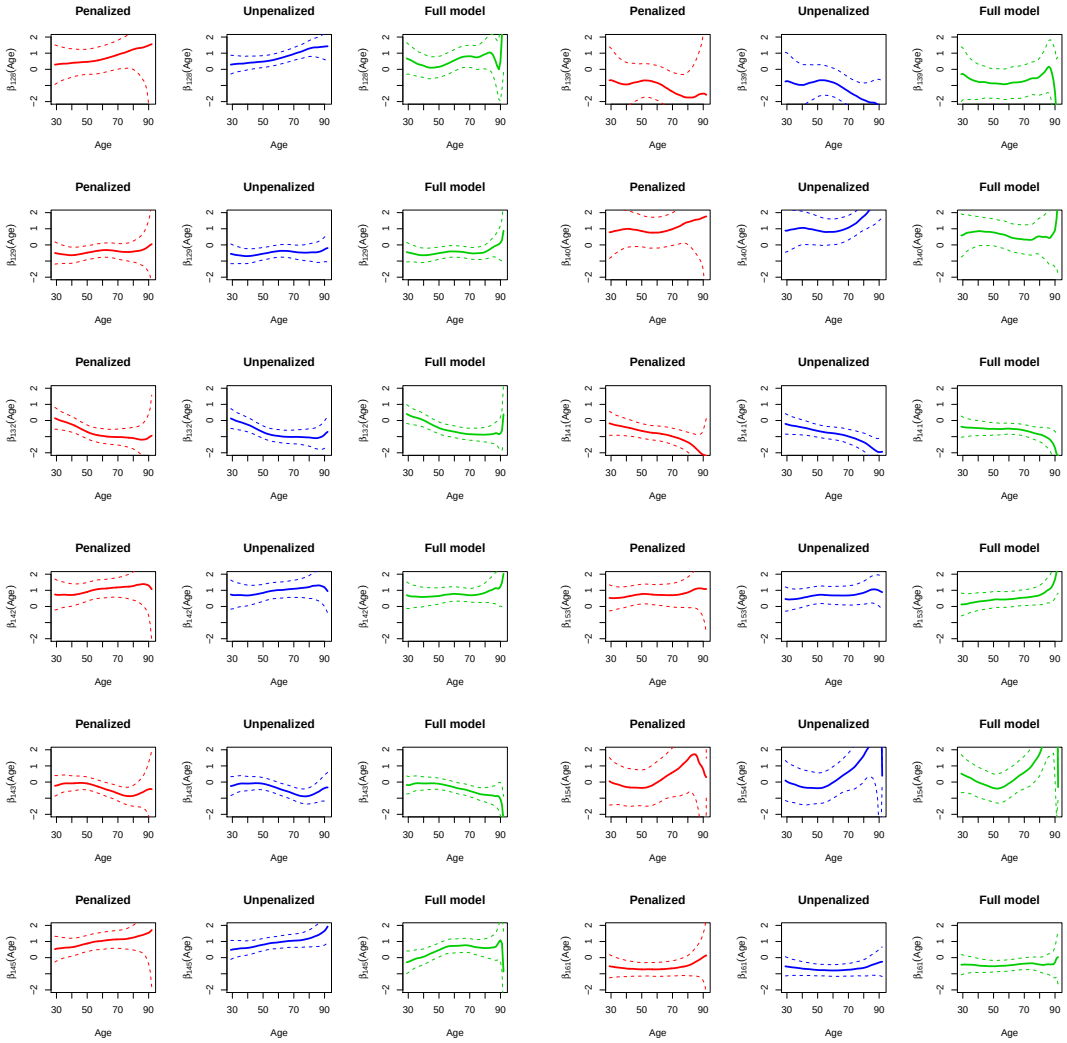


(1) Given a current estimate for the constant coefficient vector $\widehat{\boldsymbol{\beta}}_2$, define $y_{ij}^* = y_{ij} - \mathbf{x}_{2,ij}^T \widehat{\boldsymbol{\beta}}_2$, and the corresponding vector $\mathbf{y}^* = (\mathbf{y}_1^{*T}, \ldots, \mathbf{y}_n^{*T})^T$ where $\mathbf{y}_i^* = (y_{i1}^*, \ldots, y_{im_i}^*)^T$. Then the varying coefficient vector $\boldsymbol{\beta}_1(u)$ for a given value u is computed as

$$\widehat{\boldsymbol{\beta}}_1(u) = (\mathbf{I}_d, \mathbf{0}_d)(\Gamma^T \mathbf{K} \Gamma)^{-1} \Gamma^T \mathbf{K} \mathbf{y}^*,$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix, $\mathbf{0}_d$ is the $d \times d$ zero matrix, $\Gamma =$

Figure 5.4 (Cont'd). Estimated coefficients of penalized, unpenalized, and full model.



$(\mathbf{X}_1, \mathbf{U}\mathbf{X}_1)$, with $\mathbf{U} = \text{diag}(u_{11} - u, \ldots, u_{nm_n} - u)$, $\mathbf{X}_1 = (\mathbf{X}_{11}^T, \ldots, \mathbf{X}_{1n}^T)^T$, $\mathbf{X}_{1i} = (\mathbf{x}_{1,i1}, \ldots, \mathbf{x}_{1,im_i})^T$, and $\mathbf{K} = \text{diag}(K_h(u_1 - u), \ldots, K_h(u_n - u))$ where $K_h(\cdot)$ is the transformed kernel function as introduced before. The bandwidth can be selected by the multifold cross-validation.

(2) Given the current $\widehat{\boldsymbol{\beta}}_1(u)$, we can estimate the constant coefficient vector $\boldsymbol{\beta}_2$

with the idea of weighted least squares:

$$\widehat{\boldsymbol{\beta}}_2 = \{\mathbf{X}_2^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{X}_2\}^{-1}\mathbf{X}_2^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{y},$$

where $\mathbf{X}_2 = (\mathbf{X}_{21}^T, \ldots, \mathbf{X}_{2n}^T)^T$ with $\mathbf{X}_{2i} = (\mathbf{x}_{2,i1}, \ldots, \mathbf{x}_{2,im_i})^T$. $\mathbf{W}$ is the working covariance matrix, taken to be identity matrix for the moment.

Step 2. After estimated the coefficient functions in the initial step, we are now able to apply MGV method to estimate $\sigma^2(u)$ and $\boldsymbol{\theta}$:

(1) We estimate $\sigma^2(u)$ by the kernel estimator:

$$\widehat{\sigma}^2(u) = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} \hat{\gamma}_{ij}^2 K_{h_1}(u-u_{ij})}{\sum_{i=1}^n \sum_{j=1}^{m_i} K_{h_1}(u-u_{ij})},$$

where $\hat{\gamma}_{ij}$'s are residuals from the model estimated in Step 1, and the bandwidth h_1 in $K_{h_1}(\cdot)$ is chosen to minimize MISE.

(2) The estimated correlation parameter vector $\widehat{\boldsymbol{\theta}}$ is obtained to minimize the estimated variance of the constant coefficient $\boldsymbol{\beta}_2$, that is,

$$\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} |\widehat{G}(\widehat{\sigma}^2, \boldsymbol{\theta})|,$$

where $|\cdot|$ is the determinant of a matrix, and

$$\widehat{G}(\widehat{\sigma}^2, \boldsymbol{\theta}) = \widehat{\text{cov}}(\widehat{\boldsymbol{\beta}}_2 | u_{ij}, \mathbf{x}_{1,ij}, \mathbf{x}_{2,ij}) = \mathbf{D}^{-1}\widehat{\mathbf{V}}\mathbf{D}^{-1},$$

with $\mathbf{D} = \mathbf{X}_2^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{X}_2$, $\widehat{\mathbf{V}} = \mathbf{X}_2^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}\mathbf{R}\mathbf{W}^T(\mathbf{I}-\mathbf{S})\mathbf{X}_2$, $\mathbf{R} = \text{diag}(\mathbf{r}_1\mathbf{r}_1^T, \ldots, \mathbf{r}_n\mathbf{r}_n^T)$, and $\mathbf{r}_i = (r_{i1}, \ldots, r_{im_i})$.

Step 3. After obtained the estimations for the variance $\sigma^2(u)$ and correlation parameter $\boldsymbol{\theta}$ and hence Σ_i, we update the working covariance $\mathbf{W}$ with $\widehat{\Sigma}_i$'s, and repeat the first step. Therefore, we incorporate the covariance structure into the estimation of coefficient vectors and enhance their efficiency by reducing the variation.

However, if our goal is more about detecting significant SNPs for explaining the dynamic BMI change, we can stop at the variable selection stage, because using the misspecified covariance matrix does not affect the consistency of the penalized estimators.

Chapter 6

Partial Residual Two-Stage Approach for Partially Linear Models with Longitudinal Data Structure

6.1 Methodology

In this section, we present a new two-stage approach, called the partial residual two-stage approach for ultrahigh dimensional partial linear models (PLM). A fast screening procedure is proposed in the first stage specifically for PLM to reduce dimensionality, referred to as partial residual sure independence screening (PRSIS). In the second stage, we combine the partial residual method and standard variable selection for linear models to further select important variables.

Model and notation

Suppose the random sample $\{(t_{ij}, \mathbf{x}(t_{ij}), y(t_{ij})),\ i = 1, \ldots, n;\ j = 1, \ldots, T_i\}$ is from the partial linear model:

$$y(t) = \alpha(t) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}(t) + \varepsilon(t) \tag{6.1}$$

where n is the number of subjects (sample size), T_i is the number of observations for subject i, t_{ij} is the time point for the jth observation of the ith subject, $y(t)$ is the response variable, $\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))^{\mathrm{T}}$ is the p-dim covariate vector

at time t, $\alpha(t)$ is an unspecified baseline function of t , $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$ is the p-dim coefficient vector independent of t, and $\varepsilon(t)$ is the stochastic random noise with mean 0. In this paper, we assume $p >> n$, leading to the ultrahigh dimensionality.

Define $x_j(t)$ as a relevant or important (irrelevant or unimportant) predictor if $\beta_j \neq 0$ ($\beta_j = 0$). The true model is denoted as $\mathcal{M}$, i.e.

$$\mathcal{M} = \{j : 1 \leq j \leq p, \beta_j \neq 0\}. \tag{6.2}$$

Denote the number of truly relevant predictors to be $d_0 = |\mathcal{M}|$, where $|\mathcal{M}|$ is the cardinality of set $\mathcal{M}$. We impose the sparsity assumption that is standard for high and ultrahigh dimensional models: the number of relevant predictors d_0 to be much less than p, hopefully less than the sample size n.

Stage 1: partial residual sure independence screening for PLM

We first present a new screening approach, called partial residual sure independence screening (PRSIS) for partial linear models (PLM) in stage 1, which reduces the ultrahigh dimension p of the predictors to a moderate scale $d < n$. Consider the sample version of the partial linear model (6.1):

$$y(t_{ij}) = \alpha(t_{ij}) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}(t_{ij}) + \varepsilon(t_{ij}),\ i = 1, \ldots, n;\ j = 1, \ldots, T_i. \tag{6.3}$$

Although the n subjects are independent, the observations within each subject are correlated, referred to as within-subject correlation. However, in the screening stage, we need a fast and efficient algorithm to reduce dimensionality, hence we ignore the within-subject correlation in this stage. Consequently, we may pool the

data into the following formation:

$$\mathbf{t} = (t_i)_{N\times 1} = \begin{pmatrix} t_{11} \\ \vdots \\ t_{1T_1} \\ t_{21} \\ \vdots \\ t_{2T_2} \\ \vdots \\ t_{n1} \\ \vdots \\ t_{nT_n} \end{pmatrix} \quad \mathbf{y} = (y_i)_{N\times 1} = \begin{pmatrix} y(t_{11}) \\ \vdots \\ y(t_{1T_1}) \\ y(t_{21}) \\ \vdots \\ y(t_{2T_2}) \\ \vdots \\ y(t_{n1}) \\ \vdots \\ y(t_{nT_n}) \end{pmatrix} \quad \mathbf{X} = (\mathbf{x}_i^{\mathrm{T}})_{N\times p} = \begin{pmatrix} \mathbf{x}(t_{11})^{\mathrm{T}} \\ \vdots \\ \mathbf{x}(t_{1T_1})^{\mathrm{T}} \\ \mathbf{x}(t_{21})^{\mathrm{T}} \\ \vdots \\ \mathbf{x}(t_{2T_2})^{\mathrm{T}} \\ \vdots \\ \mathbf{x}(t_{n1})^{\mathrm{T}} \\ \vdots \\ \mathbf{x}(t_{nT_n})^{\mathrm{T}} \end{pmatrix} \tag{6.4}$$

where $N = \sum_{i=1}^{n} T_i$ is the pooled sample size, t_i and y_i are the ith elements of $\mathbf{t}$ and $\mathbf{y}$, and $\mathbf{x}_i^{\mathrm{T}}$ is the ith row of $\mathbf{X}$. Therefore, the model becomes

$$y_i = \alpha(t_i) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i, \quad i = 1, \ldots, N. \tag{6.5}$$

Based on model (6.5), Fan and Huang (2005) showed that if we consider the partial residual of y and $\mathbf{x}$ after regressing on t, the partial linear model can be easily transformed to a linear regression model. Explicitly, take conditional expectation of both sides of model (6.5) given t_i,

$$E(y|t_i) = \alpha(t_i) + \boldsymbol{\beta}^{\mathrm{T}}E(\mathbf{x}|t_i) + E(\varepsilon|t_i), \tag{6.6}$$

then subtract (6.6) from (6.5), the nonparametric component $\alpha(t_i)$ is canceled and the model becomes

$$y_i^* = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^* + \varepsilon_i^*, \quad i = 1, \ldots, N, \tag{6.7}$$

where $y_i^* = y_i - E(y|t_i)$, $\mathbf{x}_i^* = \mathbf{x}(t_i) - E(\mathbf{x}|t_i)$ and $\varepsilon_i^* = \varepsilon_i - E(\varepsilon|t_i)$. Therefore, the model (6.5) is transformed to the linear model (6.7), where various independence screening techniques for linear models can be applied. See Fan and Lv (2008),

Wang (2009), etc.

However, to implement this procedure in practice, we need to estimate the unknown quantities $E(Y|t_i)$ and $E(\mathbf{X}|t_i)$. The kernel smoothing technique is used to compute the NW estimates (Nadaraya, 1964; Watson, 1964) of these conditional expectations. Specifically, for any given t, the NW estimator of the mean of a random variable z is defined as

$$\widehat{E}(z|t) = \sum_{k=1}^{N} \omega_k(t) z_k, \quad \text{where} \quad \omega_k(t) = \frac{K_h(t_k - t)}{\sum_{k=1}^{N} K_h(t_k - t)} \tag{6.8}$$

with $K_h(\cdot) = h^{-1}K(\cdot/h)$ and $K(\cdot)$ is a nonnegative kernel function. The bandwidth h is chosen to minimize the mean squared error of the estimator. Since the resulting estimator of local smoothing technique is robust against the choice of kernel function, we take $K(\cdot)$ to be the Epanechnikov kernel function for simplicity, i.e. $K(t) = 0.75(1 - t^2)I(|t| < 1)$, where $I(E)$ is the indicator function taking value 1 if the statement E is true and 0 otherwise. Therefore, $E(y|t_i)$ and $E(\mathbf{x}|t_i)$ are estimated, and hence y_i^* and $\mathbf{x}_i^*$. For notation simplicity, we still denote the estimates as y_i^* and $\mathbf{x}_i^*$.

Then based on model (6.7), the standard independence screening procedures for linear models can be applied. In this paper, we employ sure independence screening (SIS, Fan and Lv, 2008) to rank the predictors according to the marginal sample pearson correlation criterion $\widehat{\rho}$. To compute $\widehat{\rho}_j$ for each $1 \leq j \leq p$, write the components in (6.7) in matrix formation:

$$\mathbf{y}^* = (y_i^*)_{N\times 1} = \begin{pmatrix} y_1 - E(y|t_{11}) \\ \vdots \\ y_{T_1} - E(y|t_{1T_1}) \\ \vdots \\ y_N - E(y|t_{nT_n}) \end{pmatrix} \tag{6.9}$$

$$\mathbf{X}^* = (\mathbf{x}_i^{*\mathrm{T}})_{N\times p} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} - E(\mathbf{x}^{\mathrm{T}}|t_{11}) \\ \vdots \\ \mathbf{x}_{T_1}^{\mathrm{T}} - E(\mathbf{x}^{\mathrm{T}}|t_{1T_1}) \\ \vdots \\ \mathbf{x}_N^{\mathrm{T}} - E(\mathbf{x}^{\mathrm{T}}|t_{nT_n}) \end{pmatrix} \triangleq (\mathfrak{X}_1^*, \ldots, \mathfrak{X}_p^*), \tag{6.10}$$

where y_i^* is the ith element of $\mathbf{y}^*$, $\mathbf{x}_i^{*\mathrm{T}}$ is the ith row of $\mathbf{X}^*$, and $\mathfrak{X}_j^*$ is the jth column of $\mathbf{X}^*$. Therefore, $\widehat{\rho}_j$ is defined as

$$\widehat{\rho}_j = \frac{(\mathfrak{X}_j^* - \overline{\mathfrak{X}}_j^*)^{\mathrm{T}}(\mathbf{y}^* - \overline{\mathbf{y}}^*)}{\sqrt{(\mathfrak{X}_j^* - \overline{\mathfrak{X}}_j^*)^{\mathrm{T}}(\mathfrak{X}_j^* - \overline{\mathfrak{X}}_j^*)(\mathbf{y}^* - \overline{\mathbf{y}}^*)^{\mathrm{T}}(\mathbf{y}^* - \overline{\mathbf{y}}^*)}}, \tag{6.11}$$

where the sample mean $\overline{\mathbf{z}}$ of a N-dim vector $\mathbf{z}$ is computed by $\overline{\mathbf{z}} = \sum_{i=1}^N z_i$ with z_i being the ith element of $\mathbf{z}$. In the above calculation, $\mathbf{z}$ is taken to be $\mathbf{y}^*$ and $\mathfrak{X}_j^*$.

Therefore, we rank the $\widehat{\rho}$ scores and select the top d predictors. The submodel index set $\widehat{\mathcal{M}}$ is taken to be

$$\widehat{\mathcal{M}} = \{j : 1 \le j \le p,\ \widehat{\rho}_j \text{ ranks among the top } \ d\}.$$

Following Fan and Lv (2008), we take $d = \lfloor N^{4/5}/\log(N^{4/5})\rfloor$, where $\lfloor a\rfloor$ is the integer part of a for $a > 0$. We will show in the simulation study that the chosen model $\widehat{\mathcal{M}}$ has an overwhelming probability to include the true model $\mathcal{M}$ defined in (6.2). Furthermore, we can always take more conservative d to be $d = \nu\lfloor N^{4/5}/\log(N^{4/5})\rfloor$ in practice, where ν is an integer larger than 1, to enlarge the probability of selecting all the relevant predictors. We name this screening technique as partial residual sure independence screening (PRSIS).

Stage 2: post-screening variable selection for PLM

In the first stage, we reduce the ultrahigh dimension p down to the moderate scale d by PRSIS, which is typically less than the sample size n. However, although all the truly relevant predictors are included in the model with large probability, there are still many irrelevant predictors in the chosen submodel $\widehat{\mathcal{M}}$. To address this issue, we adopt the second stage, where the standard shrinkage estimation procedures are applied to further select the important variables.

For easy presentation, we use the same model formation as (6.5) for the chosen submodel:

$$y_i = \alpha(t_i) + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i, \quad i = 1, \ldots, N, \tag{6.12}$$

but the dimension of $\boldsymbol{\beta}$ and $\mathbf{x}_i$ becomes $d < n$ instead of ultrahigh $p >> n$. To further select important variables in partial linear model, the profile method is applied based on the idea of Fan and Li (2004), where we iteratively estimate $\alpha(t)$ and select important predictors x. The algorithm is briefly described as follows.

(1) (Initial estimate of $\boldsymbol{\beta}$.) Based on model (6.12), we have the corresponding linear model with the same form of (6.7) using partial residual method, i.e. $y_i^* = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^* + \varepsilon_i^*$ where $\boldsymbol{\beta}$ and $\mathbf{x}_i$ are now d-dim. Compute the unpenalized least square estimate $\widehat{\boldsymbol{\beta}}^{(0)}$ from this model as initial value of $\boldsymbol{\beta}$. That is,

$$\widehat{\boldsymbol{\beta}}^{(0)} = (\mathbf{X}^{*\mathrm{T}}\mathbf{X}^*)^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{y}^*. \tag{6.13}$$

(2) (Initial estimate of $\alpha(t)$.) Plug $\widehat{\boldsymbol{\beta}}^{(0)}$ into model (6.12), and obtain a new model,

$$y_i^{(0)} = \alpha(t_i) + \varepsilon_i, \quad \text{where } \; y_i^{(0)} = y_i - \widehat{\boldsymbol{\beta}}^{(0)\mathrm{T}}\mathbf{x}_i. \tag{6.14}$$

Model (6.14) is indeed a nonparametric model which can be represented as $E(y^{(0)}|t) = \alpha(t)$. Therefore, for any given t, $\alpha(t)$ can be estimated using the same local smoothing technique as (6.8), except that the bandwidth needs to be chosen based on $y^{(0)}$. Write the resulting estimate as $\widehat{\alpha}^{(0)}(\cdot)$ and the chosen bandwidth as h_2.

(3) (Compute partial residual.) Refit $E(y|t)$ and $E(\mathbf{x}|t)$ using h_2 chosen in step (2), and hence get new $\mathbf{y}^*$ and $\mathbf{X}^*$, referred to as $\mathbf{y}^{(1)}$ and $\mathbf{X}^{(1)}$.

(4) (Shrinkage estimation of $\boldsymbol{\beta}$.) Estimate $\boldsymbol{\beta}$ by minimizing the penalized loss function $Q(\boldsymbol{\beta})$ based on $\mathbf{y}^{(1)}$ and $\mathbf{X}^{(1)}$:

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \mathrm{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\|\mathbf{y}^{(1)} - \mathbf{X}^{(1)}\boldsymbol{\beta}\|^2 + N \cdot \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|) \right\} \tag{6.15}$$

where β_j is the jth element in $\boldsymbol{\beta}$. The jth predictor x_j is selected into the final model if $\widehat{\beta}_j \neq 0$. Various penalty functions are deeply explored in literature to solve for the minimization problem (6.15). See Tibshirani (1996), Zou (2006), Fan and Li (2001), among others.

(5) (Refit $\alpha(t)$.) We can update the estimate of $\alpha(t)$ by the same technique as (2) with bandwidth h_2 and selected predictors.

With the two-stage approach, we obtain the final sparse model along with the estimated baseline function $\alpha(\cdot)$ and regression coefficient $\boldsymbol{\beta}$. We name the whole procedure as partial residual two-stage approach. In the following section, we provide some Monte Carlo simulation examples to illustrate the performance of this procedure.

6.2 Monte Carlo Simulation Studies

We conduct four simulation examples to evaluate the finite-sample performance of the partial residual two-stage approach for ultrahigh dimensional partial linear models. Example 1-3 mainly focus on the first stage, and illustrate two desirable properties of PRSIS empirically: the sure screening property (Fan and Lv, 2008), which states that the procedure has a large probability to include all the truly relevant predictors; and the ranking consistency property (Zhu, Li, Li and Zhu, 2011), where the screening criterion ensures that all the relevant predictors tend to rank before all the irrelevant predictors with an overwhelming probability. Furthermore, we compare our screening procedure with SIS (Fan and Lv, 2008), and show that if the true model has a nonparametric component $\alpha(\cdot)$, our method performs consistently better than SIS, where the nonparametric pattern is ignored. Example 4 considers both the two stages, mainly the second, and compares the results from three penalty functions and different tuning parameter selection rules.

Simulation settings and evaluation criteria

For all the four examples below, we generate the random noise ε from normal distribution with mean 0 and variance that ensure the signal-to-noise ratio is not weak, specifically, choose $\text{var}(\varepsilon)$ such that the population $R^2 = \text{var}(\alpha(t) +$

$\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}(t))/\mathrm{var}(y(t))$ is approximately 75% for any fixed t. We consider the AR(1) correlation structure for the p-dim predictor $\mathbf{x}$ (Tibshirani, 1996), that is,

$$\mathrm{corr}(x_j,\ x_k) = \rho^{|j-k|},\ 1 \le j,\ k \le p,$$

where x_j and x_k are the jth and kth element of $\mathbf{x}$. We set $\rho = 0.5$ in all the examples. For the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, we assume $\beta_1 = 3$, $\beta_4 = 1.5$, $\beta_7 = 2$, and all the other β_j's are 0 (Fan and Li, 2001), so only x_1, x_4 and x_7 are truly relevant predictors. In addition, a total of m=100 simulation runs are conducted for each example.

The first three example focus on the screening stage, where the following criteria (Zhu, Li, Li and Zhu, 2011) are considered to assess the performance of PRSIS and to compare it with SIS.

- p_j: The proportion of the jth predictor being selected into the model with size d.
- p_a: The proportion that all relevant predictors are selected into the model. Both p_j and p_a are used to verify the sure screening property introduced above.
- R_j: The rank of $\widehat{\rho}_j$ in a decreasing order. This criterion illustrates the ranking consistency property that all important predictors should rank in the top.
- M: the minimum size of the model which contains all the true predictors, in other words, the largest rank of the true predictor. We report the 5%, 25%, 50%, 75% and 95% quantiles of M from m repetitions. We can assess both sure screening and ranking consistency property from M.

For each of the three examples, we consider three predictor dimensions ($p = 500,\ 1000,\ 2000$), and the submodel size is taken to be $d = \lfloor N/\log(N) \rfloor$. The experiment settings are described as follows.

Example 1. (Single observation) In this example, we consider the ordinary partial linear model without the longitudinal data structure, hence assume each subject has only one observation, i.e. $T_i = 1,\ i = 1, \ldots, n$ in model (6.3), and $N = n$. We consider two sample sizes $n = 100,\ 200$. The p-dim predictor $\mathbf{x}_i$ is generated

independently and identical-distributively (i.i.d.) from multivariate normal with mean 0, variance 1, and the aforementioned AR(1) correlation structure, i.e.

$$\mathbf{x}_i \sim MVN(\mathbf{0}, (\Sigma_{jk})_{p\times p}), \quad \text{where} \quad \Sigma_{jk} = 0.5^{|j-k|}, \quad j,k = 1,\ldots,p \tag{6.16}$$

The univariate index variable t_i is generated from Unif$(0,1)$, independent of $\mathbf{x}$. The nonparametric baseline function $\alpha(t) = \exp(3t)$ which is nonlinear and monotone.

Example 2. (Longitudinal observations) We study longitudinal data structure in model (6.3). Therefore, there exist within-subject correlation, hence we no longer have i.i.d. observations. We draw $T_i,\ i = 1,\ldots,n$, the number of observations for each subject in model (6.3), and the time points t_{ij} from the following discrete and continuous uniform distributions:

$$T_i \sim \text{Unif}\{2,3,\ldots,10\}, \quad \text{and} \quad t_{ij} \sim U(0,1).$$

To make the T_i observations for subject i correlated with each other, the T_i-dimensional random error $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_{i1}),\ldots,\varepsilon_i(t_{iT_i}))^{\mathrm{T}}$ is drawn from

$$\boldsymbol{\varepsilon}_i \sim MVN(\mathbf{0}, \Sigma_\varepsilon), \text{ where } (\Sigma_\varepsilon)_{jk} = \text{var}(\varepsilon)\cdot\rho_\varepsilon^{|j-k|}.$$

Here we set $\rho_\varepsilon = 0.5$, and $\text{var}(\epsilon)$ is chosen to satisfy the signal-to-noise ratio as aforementioned. Furthermore, we generate the predictors $\mathbf{x}_i$ in the following fashion to create the correlation among the T_i observations from the same subject. First we draw n i.i.d random vectors from (6.16), and treat the ith vector of the n samples as the first observation for the ith subject, then generate the remaining observations for the ith subject iteratively: Denote the current observation as $\mathbf{x}_1$, the corresponding time point as t_1, and the next observation as $\mathbf{x}_2$ with the time point t_2, then $\mathbf{x}_2$ is generated by

$$\mathbf{x}_2 = \mathbf{x}_1 + (t_2 - t_1)\boldsymbol{\delta}, \quad \text{where} \quad \boldsymbol{\delta} \sim MVN(\mathbf{0}, I).$$

Therefore, the observations within each subject are highly correlated. Therefore, we create two types of correlations: the correlation among predictors, and the within-subject correlation among observations. We take two sample size $n =$

50, 100, and the nonparametric component $\alpha(t) = 10\sin(2\pi t)$, which is nonlinear and non-monotonic.

Example 3. (Mimic SNP data) This example studies the longitudinal partial linear model with predictor being categorical, as in SNP data. Specifically, the predictors take values of 0, 1 or 2, and remain unchanged for different time points of the same subject. We first draw n i.i.d. random vectors $\mathbf{x}^*$'s by (6.16), and recode the x_{ij} as 0, 1, 2, $i = 1, \ldots, n$ and $j = 1, \ldots, T_i$, according to the 25% and 75% empirical quantiles of x^*_{ij}, q_1 and q_3:

$$x_{ij} = \begin{cases} 2, & \text{if } x^*_{ij} \le q_1; \\ 1, & \text{if } q_1 < x^*_{ij} \le q_3; \\ 0, & \text{if } x^*_{ij} > q_3. \end{cases}$$

By doing this we guarantee $P(x = 0) = P(x = 2) = 1/4$, and $P(x = 1) = 1/2$. Since the SNP values stay the same for different time points within each subject, we have $x_{ij} = x_{ij'},\ j, j' = 1, \ldots, T_i$. Moreover, the time points t_{ij} are generated from discrete uniform distribution to mimic the sparse observation time, i.e. $t_{ij} \sim \text{Unif}\{0.1,\ 0.2, \ldots, 1\}$. And the nonparametric function $\alpha(t) = \exp(3t)$. All the other quantities are generated in the same fashion as Example 2.

All the above three experiments focus the screening stage, while the following example 4 evaluate both stage 1 and 2.

Example 4. All the data generation techniques are identical with Example 3. But after the screening stage, we further conduct standard variable selection procedures to refine the screened submodel. We study three penalty functions: LASSO (Tibshirani, 1996), Adaptive LASSO (Zou, 2006) and SCAD (Fan and Li, 2001), and three tuning parameter selection criteria: cross-validation (CV, Picard and Cook, 1984), AIC (Akaike, 1974) and BIC (Schwarz, 1978) criterion.

For this example, denote $\widehat{\mathcal{M}}_{(k)}$ and $\widehat{\beta}_{j(k)},\ \ k = 1, \ldots, m$ to be the final chosen model and the jth estimated coefficient by the partial residual two-stage approach in the kth simulation run. We impose the following criteria to compare the performances of different penalty functions and tuning parameter selection rules (Liang, Wang and Tsai, 2011):

- The average size of the final chosen model: $m^{-1}\sum_{k=1}^{m} |\widehat{\mathcal{M}}_{(k)}|$.

- The coverage probability of the true model: $100\% \times m^{-1} \sum_{k=1}^{m} I(\widehat{\mathcal{M}}_{(k)} \supset \mathcal{M})$.
- The percentage of correct zero: $100\% \times \{(p-d_0)m\}^{-1} \sum_{k=1}^{m} \sum_{j=1}^{p} I(\widehat{\beta}_{j(k)} = 0) I(\beta_j = 0)$, where $d_0 = |\mathcal{M}|$. This quantity studies the probability of identifying the unimportant predictors.
- The average heritability: $m^{-1} \sum_{k=1}^{m} \widehat{\text{var}}(\mathbf{X}\widehat{\boldsymbol{\beta}}_{(k)}) / \widehat{\text{var}}(\mathbf{y})$, where $\widehat{\text{var}}(\mathbf{z})$ is the sample variance of vector $\mathbf{z}$. Notice that the heritability is expected to be relatively low because we do not take into account the baseline function $\alpha(t)$, which may explain the variation of y to a certain degree.

Simulation results

The detailed output of the four examples are reported in Tables 6.1-6.4. Table 6.1-6.3 illustrate empirically the performance of PRSIS and the comparison of PRSIS and SIS for Example 1-3. All these three tables show that PRSIS performs consistently better than SIS, especially when the sample size n is relatively small.For example, in Table 6.1, when $n = 100$, the probabilities p_a that PRSIS includes all the true predictors are 81%, 85% and 94% for $p = 2000,\ 1000,\ 500$, which are significantly larger than that of SIS (41%, 49% and 63%). Same conclusion can be drawn from other criteria and other examples. This is not surprising because SIS fails to capture the nonparametric component $\alpha(t)$.

Recall that Example 1 does not have longitudinal data structure and all the observations are i.i.d., while Example 2 consider multiple observations for each subject and contain within-subject correlation. However, the comparison between Table 6.2 and Table 6.1 show that although PRSIS ignores the longitudinal structure, we do not lose much information and still have large probability to include all important predictors in the screened submodel. In addition, the results in Table 6.2 is generally more thrilling than that in Table 6.3, which is because we gain less information by generating discrete time point t_{ij} and categorical $\mathbf{x}$. To save space, we mainly focus on PRSIS in example 2 in the discussion below. Results and conclusions are directly applicable for other examples.

In the PRSIS part of Table 6.2, p_1, p_4, p_7 and p_a depict the probability of including each important predictor x_1, x_4 x_7 and including all these three. The magnitude increases and tends to 1 as the sample size n increases from 50 to 100,

for every value of p, which demonstrates the sure screening property of PRSIS that as n goes to infinity, it has probability tending to 1 to include all the important predictors. The small values of R_1, R_4 and R_7 indicate that x_1, x_4 x_7 rank in the top, indicating the ranking consistency property. Moreover, they are consistent with the magnitude of true value of coefficients $\beta_1 = 3$, $\beta_4 = 1.5$ and $\beta_7 = 2$, i.e. x_1 tends to rank almost the first since β_1 is the largest, x_7 is ranking between x_1 and x_4, and x_4 ranks behind the others because β_4 is the smallest among the three. The minimum model size M states that generally (focus on the median of M based on 100 simulation runs)we only need small models to include all the important predictors. Furthermore, all the results improve substantially as n increases. The reason is that all the desirable properties (e.g. sure screening property and ranking consistency) hold theoretically only when $n \to \infty$. Therefore, when the sample size n is small, the results are not as good. Nevertheless, we can always take $d = \nu\lfloor N^{4/5}/\log(N^{4/5})\rfloor$, $\nu = 2, 3, \ldots$ instead of $d = \lfloor N^{4/5}/\log(N^{4/5})\rfloor$ used in all examples in this paper to enlarge the probability of including important predictors in practice.

Table 6.4 examine the second stage performance. To save space, we only report the output from PRSIS for $n = 100$, $p = 500,\ 1000$ in the first stage. From the table, SCAD+BIC tends to produce the sparsest model (Average Model Size is 8.51 for $p = 500$ and 9.29 for $p = 1000$), and hence highest probability of identifying unimportant predictors (98.89% for $p = 500$ and 99.36% for $p = 1000$). Furthermore, the average heritability of SCAD+BIC is higher than any other method (37.49% for $p = 500$ and 34.77% for $p = 1000$). Since all average model sizes are larger than 3, the true model size, we have large coverage probability and perfect percentage of incorrect zero for all the methods.

6.3 Real Data Analysis: Soybean Data

(Introduction to soybean data, which I don't have the information.. And why it's appropriate to use partial linear model: Biomass is controlled by time, but don't know in which pattern, and also by some SNPs.)

Here $p = 488$, $n = 184$, $T_i = 6, 7, 8$. The response y is the total biomass, t is the time point, $\mathbf{x}$ is the p-dim SNP information, coded as 0 and 1. (Do we need to

explain why it has only two genotypes? To impute the missing marker information, incorporate the rate of recombinant homozygotes.)

We only report the result from the partial residual two-stage approach with SCAD penalty and BIC tuning parameter selection rule. In the first stage, PRSIS reduces the dimension from 488 to 191; and in the second stage, the SCAD+BIC chooses 13 SNPs. The mean squared error MSE=0.122, and the total heritability is 11.01%. The information of chosen SNPs are listed in Table (6.5), where the individual heritability is the estimate of $\mathrm{var}(\beta_j x_j)/\mathrm{var}(y)$.

The tuning parameter selection plot and the estimated baseline function are in Figure 6.1. Therefore, $\alpha(t)$ is indeed a increasing function of t.

Table 6.1. Simulation results for Example 1.

									M				
p	n	p_1	p_4	p_7	p_a	R_1	R_4	R_7	5%	25%	50%	75%	95%
							PRSIS						
2000	200	1	1	1	1	1.00	3.25	2.70	3.00	3.00	3.00	4.00	6.00
	100	1	0.88	0.93	0.81	1.16	20.14	8.33	3.00	3.00	5.00	14.75	67.45
1000	200	1	1	1	1	1.01	3.18	2.80	3.00	3.00	3.00	4.00	6.00
	100	1	0.94	0.91	0.85	1.08	11.21	10.76	3.00	4.00	5.00	8.00	73.00
500	200	1	1	1	1	1.02	3.28	2.88	3.00	3.00	4.00	5.00	6.00
	100	1	0.96	0.98	0.94	1.07	7.56	4.49	3.00	3.00	5.00	6.00	21.85
							SIS						
2000	200	1	0.96	0.98	0.94	1.21	12.18	4.97	3.00	3.75	5.00	8.00	39.20
	100	0.97	0.58	0.70	0.41	4.76	71.38	36.81	3.95	10.75	28.00	88.00	412.30
1000	200	1	0.94	1	0.94	1.14	8.05	4.43	3.00	3.00	5.00	8.25	38.10
	100	0.99	0.75	0.66	0.49	2.39	38.71	40.88	3.00	7.00	21.00	56.75	267.35
500	200	1	0.98	0.99	0.97	1.14	5.50	4.39	3.00	3.00	4.00	7.00	15.05
	100	1	0.79	0.79	0.63	1.56	21.89	21.11	3.00	5.00	12.00	34.25	158.60

Table 6.2. Simulation results for Example 2.

										M		
p	n	p_1	p_4	p_7	p_a	R_1	R_4	R_7	5%	25%	50%	75%	95%
							PRSIS						
2000	100	1	0.98	1	0.98	1.00	5.65	2.56	3.00	3.00	3.00	4.00	7.35
	50	1	0.78	0.92	0.73	1.04	43.88	22.79	3.00	4.00	11.00	26.25	301.20
1000	100	1	1	1	1	1.01	2.96	2.46	3.00	3.00	3.00	3.00	6.00
	50	1	0.84	0.86	0.70	1.02	22.66	13.14	3.00	3.00	8.00	26.00	135.45
500	100	1	0.99	1	0.99	1.02	3.24	2.59	3.00	3.00	3.00	3.00	5.05
	50	1	0.85	0.98	0.83	1.01	11.18	5.19	3.00	3.00	5.00	12.25	56.55
							SIS						
2000	100	1	0.96	0.98	0.94	1.02	14.58	3.79	3.00	3.00	4.00	6.00	31.35
	50	1	0.54	0.76	0.41	1.34	81.44	63.51	4.00	11.00	33.50	117.50	573.00
1000	100	1	1	0.99	0.99	1.08	4.41	3.75	3.00	3.00	3.00	6.00	19.05
	50	1	0.73	0.80	0.57	1.38	50.33	27.24	3.00	5.00	15.00	80.50	269.90
500	100	1	0.98	0.98	0.96	1.06	4.53	3.36	3.00	3.00	3.00	4.00	9.60
	50	1	0.81	0.88	0.70	1.13	18.56	13.94	3.00	4.75	11.00	29.25	128.90

Table 6.3. Simulation results for Example 3.

										M		
p	n	p_1	p_4	p_7	p_a	R_1	R_4	R_7	5%	25%	50%	75%	95%
							PRSIS						
2000	100	1	0.96	1	0.96	1.04	9.02	3.05	3.00	3.00	4.00	6.00	27.00
	50	1	0.64	0.72	0.43	1.24	71.42	52.84	3.95	9.00	27.50	104.25	478.10
1000	100	1	0.96	0.98	0.94	1.02	7.65	4.82	3.00	3.00	4.00	6.00	54.15
	50	1	0.58	0.85	0.50	1.16	46.34	14.42	3.00	5.00	20.50	46.50	257.65
500	100	1	0.98	1	0.98	1.03	4.50	2.90	3.0	3.0	4.0	4.0	12.1
	50	1	0.76	0.88	0.66	1.14	27.71	12.13	3.00	5.00	9.00	28.25	166.05
							SIS						
2000	100	1	0.86	0.96	0.84	1.13	18.90	6.96	3.00	4.00	7.00	16.25	78.35
	50	0.98	0.43	0.52	0.21	3.27	120.49	99.42	5.00	26.00	66.50	226.25	829.80
1000	100	1	0.88	0.93	0.82	1.05	18.12	9.60	3.00	3.00	5.50	15.50	91.05
	50	1	0.45	0.64	0.27	1.59	92.78	42.94	4.00	18.50	52.00	168.25	422.95
500	100	1	0.95	0.98	0.94	1.07	7.45	4.35	3.0	3.0	4.0	8.0	33.0
	50	0.98	0.63	0.77	0.44	1.99	38.19	22.28	3.00	7.75	22.00	53.00	313.35

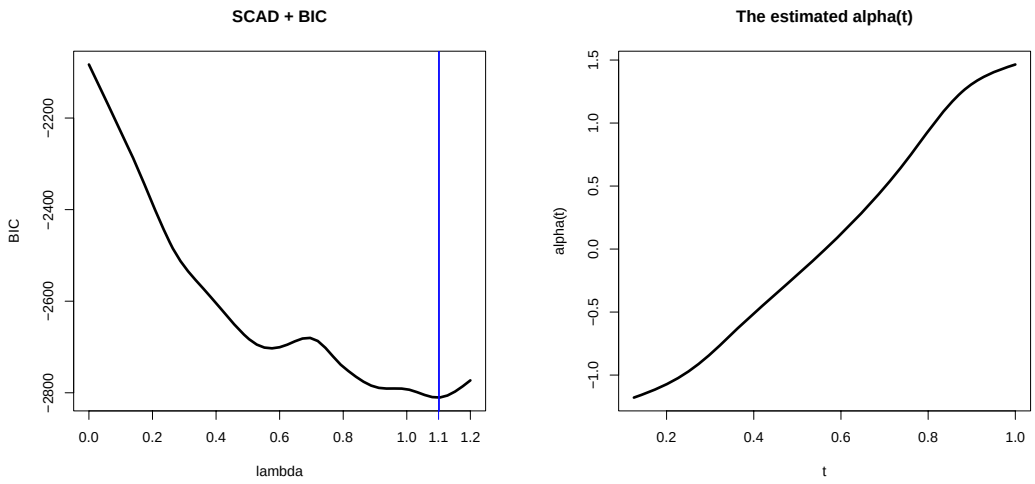
Table 6.4. Simulation results for Example 4.

Penalty and Tuning Parameter Selection Rule	Average Model Size	Coverage Probability (%)	% of Correct Zero	Average Heritability %
		$p = 500,\ n = 100$		
LASSO+CV	24.20	97	95.72	29.47
AdaptiveLASSO+AIC	18.84	97	96.81	33.82
AdaptiveLASSO+BIC	8.66	97	98.86	34.87
SCAD+AIC	16.72	97	97.23	35.38
SCAD+BIC	8.51	97	98.89	37.49
		$p = 1000,\ n = 100$		
LASSO+CV	24.88	99	97.80	26.43
AdapLASSO+AIC	19.39	98	98.35	30.38
AdapLASSO+BIC	9.89	97	99.30	32.39
SCAD+AIC	16.95	97	98.59	31.72
SCAD+BIC	9.29	97	99.36	34.77

Table 6.5. information and heritability of the SNPs chosen by SCAD+BIC.

SNP index	Chromosome	SNP names	Marker distance	Individual Heritability (%)
477		Satt592	32.6	2.718
463		$Sat_3$79	154.6	1.296
485		satt173	99.1	3.266
12		satt648	86.8	0.703
223		Satt397	123.2	0.975
489		satt608	105.3	0.305
308		A426T	185.4	0.148
486		satt188	100.9	1.045
470		satt237	50.4	0.233
144		$sat_2$52	93.5	0.125
225		$Sat_2$92	134.1	0.012
247		Satt204	54.2	0.166
246		Satt263	50.6	0.005

Figure 6.1. Tuning parameter selection and the estimated baseline function $\alpha(t)$.



Chapter 7

Conclusion and Future Research

7.1 Conclusion Remarks

Different two-stage approaches are proposed specifically for ultrahigh dimensional models with varying coefficient structure, longitudinal structure, and partially linear model setting. The screening score $\widehat{\rho}^*$ of the screening procedure for ultrahigh dimensional varying coefficient models, called conditional correlation independence screening (CCIS), is constructed based on the conditional correlation $\rho(x_j, y|u)$ between each predictor x_j and the response y given the index variable u, and we use kernel smoothing to estimate $\rho(x_j, y|u)$ via five conditional means. For future study, other smoothing techniques can be applied as long as they can guarantee the nonnegativity of the variance estimation. Iterative screening procedure CCIS for varying coefficient models is advocated to improve the finite sample performances, by which we can identify the variables that are jointly associated with the response but marginally not. This procedure can be extended to longitudinal data structure by ignoring the within-subject correlation in the screening stage. Simulation results show that we do not lose ranking consistency and sure screening property by doing this. For partially linear models, although they can be considered as a special case of time-varying coefficient models, we propose a different strategy PRSIS to reduce the dimensionality based on the partial residual method.

The ranking consistency and sure screening property of CCIS for varying coefficient models are proved in Chapter 4. Furthermore, all the two-stage approaches with the newly proposed screening methods are demonstrated in the genetic stud-

ies. To illustrate the two-stage approach for varying coefficient models with CCIS as the first stage, we analyze the GWAS data set from Framingham Heart Study (FHS). In the screening stage, CCIS is applied to take into account the effect of SNPs on BMI which may depend on the baseline age of subject, and in the post-screening variable selection stage, LASSO, Adaptive LASSO and SCAD penalized regression are modified for varying coefficient models. The report of prediction errors and model sizes indicates that CCIS+SCAD combination gives the sparsest model with smallest median squared prediction error (MSPE), and the generalized likelihood ratio test further convince us that CCIS+SCAD is sufficient for modeling the response. However, if the dynamic pattern of BMI is of more interest rather than the baseline value, we need to study the longitudinal structure. The same procedure as CCIS can be applied, by not taking into account the within-subject correlation. We select a different SNP set for explaining the dynamic pattern of BMI by the two-stage approach for longitudinal models, thus the SNP effect on baseline BMI is different from that on the whole BMI curve. To demonstrate the application of PRSIS for partially linear models, we study the SNP effect of soybean on the biomass of them. 13 SNPs are selected by the two-stage approach with PRSIS in the first stage and partial residual penalized regression with SCAD penalty in the second stage.

7.2 Future Research

In Chapter 6, we illustrated the sure screening property and ranking consistency of PRSIS for partially linear models through Monte Carlo simulations, but have not theoretically verified them as we did in Chapter 4 for CCIS. The regularity conditions and the theoretical framework are challenging but interesting for future work.

Furthermore, no matter which model setting we assume until now, we have only considered the continuous response case. In reality, however, discrete responses are often encountered with a certain distribution. For example, instead of BMI, suppose we are interested in studying the relationship between the smoking status y and SNPs x's, where y is coded as 1 for "yes" and 0 for "no". Then the response is now a binary variable from Bernoulli distribution with certain success probability.

Other possible distributions might be, but not limited to, Poisson distribution, Binomial distribution, etc. This motivates us to study the generalized varying coefficient model.

Specifically, we assume $(u_i, \mathbf{x}_i, y_i), i = 1, \ldots, n$, is a random sample from the population $(u, \mathbf{x}, y)$, where the random scalar y is from an exponential family with the canonical form of probability density function related to $\mathbf{x}$ and u,

$$f(y, \theta(u, \mathbf{x})) = \exp\{y\theta(u, \mathbf{x}) - b(\theta(u, \mathbf{x})) + c(y)\},$$

where $\theta(u, \mathbf{x})$ is an unknown function of u and $\mathbf{x}$, and $b(\cdot)$, $c(\cdot)$ are unknown functions. Note that we do not study the dispersion parameter here, and only model the mean function

$$E(y|u, \mathbf{x}) = b'(\theta(u, \mathbf{x})) \equiv m(u, \mathbf{x})$$

where b' is the first-order derivative of $b(\cdot)$. With the canonical link function $g = (b')^{-1}$, the generalized varying coefficient model can be represented as

$$g(m(u, \mathbf{u})) = \theta(u, \mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}(u), \tag{7.1}$$

and the density function now becomes

$$f(y|u, \mathbf{x}) = \exp\{y\mathbf{x}^T \boldsymbol{\beta}(u) - b(\mathbf{x}^T \boldsymbol{\beta}(u)) + c(y)\}. \tag{7.2}$$

In the above example, where the smoking status y is binary, the distribution of y can be characterized by

$$y|u, \mathbf{x} \sim Bernoulli(\pi(u)), \quad \text{logit}(\pi(u)) = \mathbf{x}^T \boldsymbol{\beta}(u),$$

where $\text{logit}(\pi(u)) = \log\{\pi(u)/(1-\pi(u))\}$. If y is the counting variable from Poisson distribution,

$$y|u, \mathbf{x} \sim Poisson(\lambda(u)), \quad \log(\lambda(u)) = \mathbf{x}^T \boldsymbol{\beta}(u).$$

Although the theoretical properties of the aforementioned ordinary feature screening method for varying coefficient models do not rely on the continuity of

response and the normality assumption, it is not straightforward to visualize the rationale of using it as the screening procedure for generalized varying coefficient models. Therefore, we develop a new screening procedure especially for generalized varying coefficient models.

Notice that conditional on u, the generalized varying coefficient model (7.1) becomes an ordinary generalized linear model. Motivated by Fan and Song (2010), we consider using the marginal maximum log-likelihood for each predictor x_j to represent its importance. Specifically, consider the p marginal models

$$g(m(u, \mathbf{u})) = \theta(u, \mathbf{x}) = \beta_{0j}(u) + \beta_j(u)x_j, \quad j = 1, \ldots, p.$$

According to the conditional density function (7.2), $\beta_{0j}(u)$ and $\beta_j(u)$ should maximize the population version of conditional log-likelihood $Q(\beta_{0j}(u),\ \beta_j(u)|u)$:

$$\begin{aligned} Q(\beta_{0j}(u),\ \beta_j(u)|u) &= E\left[\{y \cdot (\beta_{0j}(u) + \beta_j(u)x_j) - b(\beta_{0j}(u) + \beta_j(u)x_j)\}|u\right] \\ &= \beta_{0j}(u)E(y|u) + \beta_j(u)E(x_j y|u) - E\{b(\beta_{0j}(u) + \beta_j(u)x_j)|u\} \end{aligned}$$

Empirically, we apply kernel smoothing technique to estimate the three conditional expectations above, i.e.

$$\widehat{E}(y|u) = \sum_{i=1}^{n} \omega_i(u)y_i, \quad \widehat{E}(x_j y|u) = \sum_{i=1}^{n} \omega_i(u)x_{ij}y_i,$$
$$\widehat{E}(b(\beta_{0j}(u) + \beta_j(u)x_j)|u) = \sum_{i=1}^{n} \omega_i(u)b(\beta_{0j}(u) + \beta_j(u)x_{ij}),$$

where the weight $\{\omega_i(u),\ i = 1, \ldots, n\}$ is the normalized kernel function of u:

$$\omega_i(u) = \frac{K_h(u - u_i)}{\sum_{i=1}^{n} K_h(u - u_i)}.$$

Therefore, the estimated marginal coefficients $\hat{\beta}_{0j}(u)$ and $\hat{\beta}_j(u)$ are obtained by

$$(\hat{\beta}_{0j}(u), \hat{\beta}_j(u)) = \text{argmax}_{\beta_{0j}, \beta_j} Q(\beta_{0j}(u),\ \beta_j(u)|u),$$

where

$$Q(\beta_{0j}(u),\ \beta_j(u)|u) \ = \ \beta_{0j}(u)\widehat{E}(y|u) + \beta_j(u)\widehat{E}(x_j y|u) - \widehat{E}\{b(\beta_{0j}(u) + \beta_j(u)x_j)|u\}$$

In the optimization problem above, the maximizer $(\hat{\beta}_{0j}(u), \hat{\beta}_j(u))$ can be obtained by Newton-Raphson algorithm, where $b(\cdot)$ is determined by the conditional distribution of y given $\mathbf{x}$ and u. For example, if y is binary from Bernoulli distribution, $b(\theta) = \log(1 + e^{\theta})$, and if y is the counting variable from Poisson distribution, $b(\theta) = \exp(\theta)$.

After we obtain the estimates $(\hat{\beta}_{0j}(u), \hat{\beta}_j(u))$, we can compute the conditional maximum log-likelihood estimate given u:

$$\widehat{Q}(\hat{\beta}_{0j}(u),\ \hat{\beta}_j(u)|u) \ = \ \hat{\beta}_{0j}(u)\widehat{E}(y|u) + \hat{\beta}_j(u)\widehat{E}(x_j y|u) - \widehat{E}\{b(\hat{\beta}_{0j}(u) + \hat{\beta}_j(u)x_j)|u\}.$$

To average out the effect of u, we calculate $\widehat{Q}$ value for each u_i, and the final screening criterion is defined as

$$Q_j^* = \sum_{i=1}^{n} \widehat{Q}(\hat{\beta}_{0j}(u_i),\ \hat{\beta}_j(u_i)|u_i).$$

Then we sort Q_j^* in a decreasing order, and reduce the model size from ultrahigh p to moderate d by picking the top d predictors, that is, the submodel

$$M_\gamma = \{j :\ 1 \le j \le p,\ Q_j^* \text{ ranks among the top } \ d.\}.$$

To assess the performance of the feature screening procedure for the generalized varying coefficient model (7.1), we conduct the following Monte Carlo simulations. Same as before, the univariate index variable u_i and the covariate $\mathbf{x}_i$ are generated i.i.d in the following fashion:

$$\begin{aligned} &(u_i^*, \mathbf{x}_i) \sim MVN(\mathbf{0}, \Sigma), \quad \text{where } \ \Sigma_{jk} = \rho^{|j-k|}, \quad j, k = 1, \ldots, p+1, \\ &u_i = \Phi(u_i^*), \quad i = 1, \ldots, n. \end{aligned} \tag{7.3}$$

For all examples, we set $\rho = 0.8$, $p = 1000$, $n = 400$, and repeat the experiment 100 times. In each of the 100 simulations, we choose d to be d_0, $2d_0$ and $3d_0$, where

$d_0 = [n^{4/5}/\log(n^{4/5})]$. For evaluating the performances, we still use the following criteria:

- p_j: The proportion of the jth predictor being selected into the model with size d.
- p_a: The proportion that all active predictors are selected into the model.
- $rank_j$: The ranking of ρ_j^* in a decreasing order.
- M: the minimum size of the model which contains all the true predictors. We report the 5%, 25%, 50%, 75% and 95% quantiles of M from 100 simulations.

Example 1. This example is a linear regression model, where we generate the random noise $\{\varepsilon_i,\ i = 1, \ldots, n\}$ from i.i.d. $N(0, 1)$. The true sparse model is generated as

$$y = 1.8x_2 - 1.7x_{100} + 2x_{400} - 1.5x_{600} + 1.5x_{1000} + \varepsilon.$$

Table 7.1, 7.2, and 7.3 demonstrate the performances of all the four screening techniques, where SIS is the sure independence screening proposed by Fan and Lv (2008), MMLE is the marginal maximum likelihood method proposed by Fan and Song (2010), VCM is the screening technique especially for varying coefficient models proposed in chapter 3, and GVCM is the newly advocated method for generalized varying coefficient models.

The first table depicts the proportion of including each truly important variables into the submodel with size d_0, $2d_0$ and $3d_0$. A good screening procedure should be able to contain all the true predictors with large probability, called sure screening property (Fan and Lv, 2008). And The second table illustrates the ranking of each significant variable in terms of the corresponding screening scores. The small numbers of rankings for the important predictors indicate the validity of the screening method, proposed as ranking consistency property by Zhu, Li, Li and Zhu (2011). The third table, with the summary of minimum model size shows both sure screening property and ranking consistency property.

From the following three tables, we conclude that for linear models, which are special cases of varying coefficient models and generalized varying coefficient models, all the four methods work well and have an overwhelming probability to include all the truly important predictors, especially when the submodel size is $d = 2d_0$ or $d = 3d_0$, we can include all the important predictors with probability 1. Under this circumstance, SIS is preferable due to its computational efficiency. And since the screening method for generalized varying coefficient requires the iterative algorithm during the estimation procedure, we do not recommend it if the underlying model structure is linear.

Table 7.1. The proportions p_j and p_a for Example 1.

	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
			$d = d_0$			
SIS	1	1	1	1	0.98	0.98
MMLE	1	1	1	1	0.98	0.98
VCM	1	1	1	0.99	1	0.99
GVCM	1	1	1	1	0.99	0.99
			$d = 2d_0$			
SIS	1	1	1	1	1	1
MMLE	1	1	1	1	1	1
VCM	1	1	1	1	1	1
GVCM	1	1	1	1	1	1
			$d = 3d_0$			
SIS	1	1	1	1	1	1
MMLE	1	1	1	1	1	1
VCM	1	1	1	1	1	1
GVCM	1	1	1	1	1	1

Table 7.2. $rank_j$ of each true predictor x_j for Example 1.

	x_2	x_{100}	x_{400}	x_{600}	x_{1000}
SIS	2.8	4.09	1.36	7.2	7.9
MMLE	2.8	4.09	1.36	7.2	7.9
VCM	6.22	3.25	1.12	6.11	6.28
GVCM	6.28	3.39	1.14	6.36	6.66

Example 2. We now consider logistic regression. Suppose the response y is from

Table 7.3. The minimum model size M for Example 1.

	5%	25%	50%	75%	95%
SIS	5	7	10	12	16
MMLE	5	7	10	12	16
VCM	5.95	7	9	11	13
GVCM	5	7.75	9	11	14

$Bernoulli(\pi)$, where

$$\text{logit}(\pi) = 1.8x_2 - 1.7x_{100} + 2x_{400} - 1.5x_{600} + 1.5x_{1000}.$$

Table 7.4, 7.5 and 7.6 reported the performances of the four methods. In this example, the random error is no longer normally distributed, thus the theoretical assumption for SIS no longer hold. Nevertheless, we find that the SIS still performs well, and behaves approximately identically with MMLE. Also, the VCM and GVCM still have a large chance to include the significant variables, although they are less preferable for constant coefficient structures due to computation cost.

Table 7.4. The proportions p_j and p_a for Example 2.

	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
			$d = d_0$			
SIS	1	0.96	1	0.94	0.91	0.81
MMLE	1	0.96	1	0.94	0.91	0.81
VCM	0.97	0.98	1	0.95	0.93	0.83
GVCM	0.95	1	1	0.92	0.93	0.80
			$d = 2d_0$			
SIS	1	1	1	1	1	1
MMLE	1	1	1	1	1	1
VCM	1	1	1	1	1	1
GVCM	1	1	1	0.96	1	0.96
			$d = 3d_0$			
SIS	1	1	1	1	1	1
MMLE	1	1	1	1	1	1
VCM	1	1	1	1	1	1
GVCM	1	1	1	0.98	1	0.98

Table 7.5. $rank_j$ of each true predictor x_j for Example 2.

	x_2	x_{100}	x_{400}	x_{600}	x_{1000}
SIS	3.53	5.56	2	8.24	8.73
MMLE	3.52	5.6	2	8.32	8.75
VCM	6.85	4.72	1.53	7.24	7.47
GVCM	6.94	4.95	1.64	14.29	8.21

Table 7.6. The minimum model size M for Example 2.

	5%	25%	50%	75%	95%
SIS	7	9	12	16	24
MMLE	7	9	12	16	24
VCM	5	8	11	14	21
GVCM	7	10	12	16	32

Example 3. In this example we consider a varying coefficient model

$$y = \beta_2(u)x_2 + \beta_{100}(u)x_{100} + \beta_{400}(u)x_{400} + \beta_{600}(u)x_{600} + \beta_{1000}(u)x_{1000} + \varepsilon,$$

where the random noises are still from independent $N(0,1)$, and nonzero coefficients are defined by

$$\beta_2(u) = 3I(u > 0.4), \quad \beta_{100}(u) = 1 + u, \quad \beta_{400}(u) = (2 - 3u)^3$$
$$\beta_{600}(u) = 2\sin(2\pi u), \quad \beta_{1000}(u) = \exp\{u/(u+1)\}.$$

The results are reported in Table 7.7, 7.8 and 7.9. Since this example is designed for the varying coefficient screening procedure, we can see from the three table that VCM works the best in terms of both ranking consistency and sure screening property. Moreover, SIS and MMLE fail to detect x_{600} because $\beta_{600}(u) = 2\sin(2\pi u)$ has empirical mean close to 0, tending to be insignificant if treated as a constant coefficient. GVCM also works well in this example, since varying coefficient models are special cases of generalized varying coefficient models.

Example 4. We study the logistic regression with varying coefficients. Suppose

Table 7.7. The proportions p_j and p_a for Example 3.

	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
			$d = d_0$			
SIS	1	1	0.96	0	0.98	0
MMLE	1	1	0.96	0	0.98	0
VCM	1	1	1	1	0.97	0.97
GVCM	1	1	1	0.97	0.85	0.84
			$d = 2d_0$			
SIS	1	1	1	0.03	1	0.03
MMLE	1	1	1	0.03	1	0.03
VCM	1	1	1	1	1	1
GVCM	1	1	1	1	0.98	0.98
			$d = 3d_0$			
SIS	1	1	1	0.06	1	0.06
MMLE	1	1	1	0.06	1	0.06
VCM	1	1	1	1	1	1
GVCM	1	1	1	1	0.99	0.99

Table 7.8. $rank_j$ of each true predictor x_j for Example 3.

	x_2	x_{100}	x_{400}	x_{600}	x_{1000}
SIS	1.91	2.89	7.01	488.44	5.43
MMLE	1.91	2.89	7.01	488.44	5.43
VCM	1.15	3.96	2.57	8.13	6.76
GVCM	2.8	6.52	1.19	8.94	11.04

Table 7.9. The minimum model size M for Example 3.

	5%	25%	50%	75%	95%
SIS	36	195	519.5	722	951.4
MMLE	36	195	519.5	722	951.4
VCM	5	7.75	10	12	15
GVCM	7	10	12	15	24.1

y is a binary response from $Bernoulli(\pi(u))$, where

$$\text{logit}(\pi(u)) = \beta_2(u)x_2 + \beta_{100}(u)x_{100} + \beta_{400}(u)x_{400} + \beta_{600}(u)x_{600} + \beta_{1000}(u)x_{1000},$$

with the nonzero coefficients defined as Example 3.

Table 7.10, 7.11 and 7.12 indicate that both VCM and GVCM perform similarly and are able to detect significant variables with large probability, indicating that although VCM is not specifically designed for generalized varying coefficient models, it can still be applied without losing much information. GVCM also works well in terms of the three criteria. However, SIS and MMLE fail for the same reason as Example 3.

Table 7.10. The proportions p_j and p_a for Example 4.

	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
			$d = d_0$			
SIS	0.89	1	0.75	0.01	0.99	0.01
MMLE	0.87	1	0.76	0.01	0.99	0.01
VCM	0.93	1	1	0.93	0.98	0.84
GVCM	0.92	0.99	0.99	0.95	0.98	0.83
			$d = 2d_0$			
SIS	0.97	1	0.88	0.02	1	0.02
MMLE	0.97	1	0.89	0.02	1	0.02
VCM	0.99	1	1	1	1	0.99
GVCM	1	1	1	1	0.99	0.99
			$d = 3d_0$			
SIS	1	1	0.93	0.02	1	0.02
MMLE	1	1	0.93	0.02	1	0.02
VCM	1	1	1	1	1	1
GVCM	1	1	1	1	0.99	0.99

Table 7.11. $rank_j$ of each true predictor x_j for Example 4.

	x_2	x_{100}	x_{400}	x_{600}	x_{1000}
SIS	8.81	1.71	21.67	508.7	3.03
MMLE	8.82	1.69	21.64	508.63	3.06
VCM	8.53	2.35	3.53	7.44	4.56
GVCM	7.3	2.89	3.11	6.23	6.79

Example 5. In this example we consider the Poisson generalized varying coefficient model. Suppose y is a counting response from $Poisson(\lambda(u))$, where

$$\log(\lambda(u)) = \beta_2(u)x_2 + \beta_{100}(u)x_{100} + \beta_{400}(u)x_{400} + \beta_{600}(u)x_{600} + \beta_{1000}(u)x_{1000},$$

Table 7.12. The minimum model size M for Example 4.

	5%	25%	50%	75%	95%
SIS	70	256.25	530	749.75	939.7
MMLE	71	256.75	530	749.75	939.7
VCM	7	8	11	15	22
GVCM	6	9	10.5	14.25	27

with the nonzero coefficients

$$\beta_2(u) = I(u > 0.4), \quad \beta_{100}(u) = (1+u)/2, \quad \beta_{400}(u) = (2-3u)^2$$
$$\beta_{600}(u) = \sin(2\pi u), \quad \beta_{1000}(u = \exp\{u/(u+1)\}/2.$$

Since the mean value for Poisson variable is $\lambda(u) = \exp(\mathbf{x}^T\boldsymbol{\beta}(u))$, the right hand side of the above model cannot be too large, otherwise the exponential of it to be infinity. Therefore, instead of directly use the generated $\mathbf{x}$ in (7.3), we transform the x values by $2\mathbb{F}(x)-1 \in [-1,1]$, where $\mathbb{F}(x)$ is the empirical distribution function of x. Therefore, we guarantee $\lambda(u) = \exp(\mathbf{x}^T\boldsymbol{\beta}(u))$ within a reasonable range.

The results are demonstrated in the following three tables 7.13, 7.14, and 7.15. We discover the same pattern as the previous two examples, indicating that under generalized varying coefficient model setting, both VCM and GVCM are still valid, while SIS and MMLE are not necessarily.

We can also conduct real data analysis to illustrate the application of this screening procedure for generalized varying coefficient models, as we did for CCIS and PRSIS, and develop some theoretical results.

Table 7.13. The proportions p_j and p_a for Example 5.

	p_2	p_{100}	p_{400}	p_{600}	p_{1000}	p_a
			$d = d_0$			
SIS	0.87	1	0.99	0.12	1	0.11
MMLE	0.87	1	0.99	0.12	1	0.11
VCM	1	1	1	1	1	1
GVCM	0.96	0.99	1	1	0.99	0.94
			$d = 2d_0$			
SIS	0.94	1	1	0.27	1	0.24
MMLE	0.94	1	1	0.27	1	0.24
VCM	1	1	1	1	1	1
GVCM	0.97	1	1	1	0.99	0.96
			$d = 3d_0$			
SIS	0.97	1	1	0.38	1	0.36
MMLE	0.97	1	1	0.38	1	0.36
VCM	1	1	1	1	1	1
GVCM	0.99	1	1	1	0.99	0.98

Table 7.14. $rank_j$ of each true predictor x_j for Example 5.

	x_2	x_{100}	x_{400}	x_{600}	x_{1000}
SIS	39.57	1.68	7.16	413.59	3.54
MMLE	39.54	1.69	7.14	413.67	3.49
VCM	7.67	2.37	3.97	3.56	5.53
GVCM	16.85	5.34	2.72	4	14.94

Table 7.15. The minimum model size M for Example 5.

	5%	25%	50%	75%	95%
SIS	41.95	137.75	379	725	928.15
MMLE	41.8	137	379	725	928.15
VCM	5	7	9	11.25	18.05
GVCM	6	9	11	16	72.85

Bibliography

Akaike, H. (1974) "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, **19**, 716–723.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics,* **37**, 373–384.

Cai, Z., Fan, J., and Li, R. (2000), "Efficient estimation and inferences for varying coefficient models," *Journal of the American Statistical Association*, **95**, 888–902.

Chu, C. and Marron, J. (1991), "Choosing a Kernel Regression Estimator," *Statistical Science*, **6**, 404–419.

Efron, B., Hastie, T., Joghstone, I., and Tibshirani, R. (2004), "Least angle regression," *The Annals of Statistics*, **32**, 407–499.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. (2002),"Analysis of longitudinal data, 2nd edition," *Oxford University Press, New York.*

Donoho, D. L. (2000), "High dimensional data analysis: The curse and blessing of dimensionality," *Aide-Memoire of the lecture in AMS con- ference: Math challenges of 21st Centrury.* http://www.stat.stanford.edu/donoho/Lectures.

Fan, J. and Gijbels, I. (1992), "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics,* **20**, 2008–2036.

Fan, J. (1993), "Local linear regression smoothers and their minimax efficiencies," *The Annals of Statistics,* **21**, 196–216.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, Chapman and Hall, New York, NY.

Fan, J., and Zhang, W. (1999), "Statistical estimation in varying coefficient models," *The Annals of Statistics*, **27**, 1491–1518.

Fan, J., and Zhang, W. (2001), "Simultaneous confidence bands and hypotheses testing in varying-coefficient models," *Scandinavian Journal of Statistics*, **27**, 1491–1518.

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

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized likelihood ratio statistics and Wilks phenomenon," *The Annals of Statistics*, **29**, 153–193.

Fan, J., and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and it oracle properties," *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J., and Li, R. (2004), "New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis," *Journal of the American Statistical Association*, **99**, 710–723.

Fan, J., and Peng, H. (2004), "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, **32**, 928–961.

Fan, J. and Li, R. (2006), "An Overview on Nonparametric and Semiparametric Techniques for Longitudinal Data," *Frontiers in Statistics, (J. Fan and H. Koul eds., Imperial College Press, London),* 277–303

Fan, J., and Li, R. (2006), "Statistical challenges with high dimensionality: feature selection in knowlege discorvery," *Proceedings of the International Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.)*, **Vol.III**, 595–622.

Fan, J. and Zhang, W. (2008), "Statistical methods with varying coefficient models," *Statistics and its interface,* **1,** 179–195.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space (with discussion)," *Journal of the Royal Statistical Society, Series B*, **70**, 849–911.

Fan, J., Samworth, R. and Wu, Y. (2009), "Ultrahigh dimensional feature selection: beyond the linear model," *Journal of Machine Learning Research*, **10**, 1829–1853.

Fan, J. and Song, R. (2010), "Sure independence screening in generalized linear models with NP-dimensionality," *The Annals of Statistics*, **38**, 3567–3604.

Fan, J. and Lv, J. (2010) "A selective overview of variable selection in high dimensional feature space". *Statistica Sinica* **20** 101–148.

Foster, D. and George, E. (1994) "The risk inflation criterioin for multiple regression,". *The Annals of Statistics* **22** 1947–1975.

Fu, W. (1998) "Penalized regressions: The Bridge versus the LASSO,". *Journal of Computational and Graphical Statistics* **7(3)** 397–416.

Gasser, T. and Müller, H. G. (1984), "Estimating Regression Functions and Their Derivatives by the Kernel Method," *Scandinavian Journal of Statistics*, **11**, 171–185.

Hall, P. and Kang, K. (2001), "Bootstrapping nonparametric density estimators with empirically chosen bandwidths," *Annals of Statistics*, 1443–1468.

Hall, P. and Miller, H. (2009), "Using generalized correlation to effect variable selection in very high dimensional problems," *Journal of Computational and Graphical Statistics*, **18**, 533–550.

Hastie, T. J. and Tibshirani, R. J. (1993), "Varying-coefficient models," *Journal of the Royal Statistical Society, Series B*, **55**, 757–796.

Hoerl, A. and Kennard, R. (1970) "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics* **12** 55–67.

Hoover, D., Rice, J., Wu, C., and Yang, L. (1998), "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data," *Biometrika,* **85**, 809–822.

Huang, J., Ma, S. G. and Zhang, C. H. (2008), "Adaptive Lasso for sparse high-dimensional regression models," *Statistica Sinica*, **18**, 1603–1618

Hunter, D. R., and Li, R. (2005), "Variable selection using MM algorithms," *The Annals of Statistics*, **27**, 1491–1518.

Kim, Y., Choi, H. and Oh, H. S. (2008), "Smoothly clipped absolute deviation on high dimensions," *Journal of the American Statistical Association,* **103**, 1665–1673.

Konishi, S. and Kitagawa, G. (1996), "Generalized information criteria in model selection," *Biometrika,* **83**, 875–890.

Leng, C., Lin, Y., and Wahba, G. (2006), "A note on the LASSO and related procedures in model selection," *Statistica Sinica,* **16**.

Li, K. (1987), "Asymptotic optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: discrete index set," *The Annals of Statistics,* **15**, 958–975.

Li, R., Dziak, J., and Ma, Y. (2006), "Nonconvex penalized least squares: characterizations, algorithm and application," *Manuscript.*

Li, R., and Liang, H. (2007), "Variable selection in semiparametric regression modeling," *The Annals of Statistics,* **36**, 261–286.

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

Li, R., Zhong, W. and Zhu, L.P. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, **107**, 1129–1139.

Mallows, C. (1973), "Some comments on C_p," *Technometrics,* **15**, 661–675.

Miller, A. (2002), "Subset selection in regression, 2nd edition," *New York: Chapman and HALL/CRC.*

Müller, H. (1988), *Nonparametric regression analysis of longitudinal data,* Springer-Verlag, New York.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, **90**, 1257–1270.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003), "Regression approach for microarray data analysis," *Journal of Computational Biology,* **10**, 961–980.

Shao, J. (1997), "An asymptotic theory for linear model selection," *Statistica Sinica,* **7**, 221–264.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *The Annals of Statistics*, **5**, 595–645.

Tibshirani, R. (1996), "Regression shrinkage and selection via LASSO," *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Wand, M. and Jones, M. (1995), *Kernel Smoothing,* Chapman and Hall/CRC.

Wang, H., Li, R., and Tsai, C. L. (2007b) "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, **94**, 553–558.

Wang, H. (2009), " Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, **104**, 1512–1524.

Wang, H. and Xia. Y. (2009), "Shrinkage Estimation of the Varying Coefficient Model," *Journal of the American Statistical Association*, **104**, 747–757.

Xia, Y., Zhang, W., and Tong, H. (2004), "Efficient estimation for semivarying coefficient models," *Biometrika,* **91**, 661–681.

Yuan, M., and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.

Zhang, W., Lee, S., and Song, X. (2002), "Local polynomial fitting in semivarying coefficient model," *Journal of Multivariate Analysis*, **82**, 166–188.

Zhang, W., and Lee, S. Y. (2007), "Variable bandwidth selection in varying coefficient models," *Journal of Multivariate Analysis*, **19**, 116–134.

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.

Zhu, L.P., Li, L., Li, R., and Zhu, L.X. (2011), "Model-free feature screening for ultrahigh dimensional data," *Journal of the American Statistical Association*, **106**, 1464–1475.

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.

Jingyuan Liu

325 Thomas Building
The Pennsylvania State University
State College, PA, 16802

Phone: (814)-321-7556
Email:jul221@psu.edu

Education

Ph.D. Candidate in Statistics, GPA: 4.0/4.0 — 08/08 - 05/13
Department of Statistics, Pennsylvania State University (PSU)
Dissertation: Statistical Methods for Ultrahigh Dimensional Models.

B.S. in Statistics, GPA: 3.9/4.0 — 09/04 - 06/08
School of Mathematics, Shandong University

Publications/ Manuscripts

Conditioning-correlation screening for varying coefficient models, J. Liu, R. Li, R. Wu. To be submitted to *Journal of the American Statistical Association*

Ultrahigh dimensional partially linear model and its application, J. Liu, R. Li, R. Wu, In preparation for *Bioinformatics*

Considerations in Identifying Treatment Effects on Transient Event Driven Health Status Changes Measured by Patient Reported Outcomes, J. Liu, J. Legg, M. May. Ready to be submitted to *Statistics in Medicine*

Model and algorithm for linkage disequilibrium analysis in a nonequilibrium population, J. Liu, Z. Wang, Y. Wang, R. Li, R. Wu. *Frontiers in Statistical Genetics and Methodology* 2012; 3: 78

Functional mapping of developmental processes: theory, applications, and prospects, K. Das, J. Liu, G. Fu, J. Li, Y. Li, C. Tong, R. Wu. *Methods in Molecular Biology* 2012; 871:227-243

Dynamic modeling of genes controlling cancer stem cell proliferation, Z. Wang, J. Liu, J. Wang, Y. Wang, N. Wang, Y. Li, R. Li, R. Wu. *Frontiers in Statistical Genetics and Methodology* 2012; 3: 84

Statistical models for genetic mapping in polyploids: challenges and opportunities, J. Li, K. Das, J. Liu, G. Fu, Y. Li, C. Tobias, R. Wu. *Methods in Molecular Biology* 2012; 871:245-261

Teaching Experience

Instructor, Department of Statistics, PSU — 08/10 - 08/11
Courses: Probablity; Applied Regression

Teaching Assistant, Department of Statistics, PSU — 08/08 - 05/13
Courses: Applied Statistics; Design of Experiment; Biostatistics; ANOVA and Design of Experiments; Experimental Methods

Working Experience

Statistician (Intern), Amgen Inc. — 05/12 - 08/12

Statistical Consultant, Statistical Consulting Center, PSU — 08/10 - 05/11

Awards

2011 Harkness Excellent Teaching Award, Dept. of Statistics, PSU — 12/11

Student Travel Grant, Joint Statistical Meetings — 08/11

PhD qualifier exam pass with distinction, Dept. of Statistics, PSU — 01/10

Master qualifier exam pass with distinction,Dept. of Statistics, PSU — 05/09

Student of Distinction, Shandong Province — 01/08