DOCTORAL THESIS

Variable selection in high dimensional semi-varying coefficient models

Chen, Chi

Date of Award:
2013

Variable Selection in High Dimensional Semi-Varying Coefficient Models

CHEN Chi

A thesis submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Principal Supervisor: Dr. PENG Heng

Hong Kong Baptist University

Aug 2013

Declaration

I hereby declare that this thesis represents my own work which has been done after registration for the degree of PhD at Hong Kong Baptist University, and has not been previously included in a thesis, dissertation submitted to this or other institution for a degree, diploma or other qualification.

Signature: _____

Date: Aug 2013

Abstract

With the development of computing and sampling technologies, high dimensionality has become an important characteristic of commonly used science data, such as some data from bioinformatics, information engineering, and the social sciences. The varying coefficient model is a flexible and powerful statistical model for exploring dynamic patterns in many scientific areas. It is a natural extension of classical parametric models with good interpretability, and is becoming increasingly popular in data analysis. The main objective of thesis is to apply the varying coefficient model to analyze high dimensional data, and to investigate the properties of regularization methods for high-dimensional varying coefficient models.

We first discuss how to apply local polynomial smoothing and the smoothly clipped absolute deviation (SCAD) penalized methods to estimate varying coefficient models when the dimension of the model is diverging with the sample size. Based on the nonconcave penalized method and local polynomial smoothing, we suggest a regularization method to select significant variables from the model and estimate the corresponding coefficient functions simultaneously. Importantly, our proposed method can also identify constant coefficients at same time. We investigate the asymptotic properties of our proposed method and show that it has the so called "oracle property."

We apply the nonparametric independence Screening (NIS) method to varying coefficient models with ultra-high-dimensional data. Based on the marginal varying coefficient model estimation, we establish the sure independent screening property under some regular conditions for our proposed sure screening method. Combined with our proposed regularization method, we can systematically deal with high-dimensional or ultra-high-dimensional data using varying coefficient models.

The nonconcave penalized method is a very effective variable selection method. However, maximizing such a penalized likelihood function is computationally challenging, because the objective functions are nondifferentiable and nonconcave. The local linear approximation (LLA) and local quadratic approximation (LQA) are two popular algorithms for dealing with such optimal problems. In this thesis, we revisit these two algorithms. We investigate the convergence rate of LLA and show that the rate is linear. We also study the statistical properties of the one-step estimate based on LLA under a generalized statistical model with a diverging number of dimensions. We suggest a modified version of LQA to overcome its drawback under high dimensional models. Our proposed method avoids having to calculate the inverse of the Hessian matrix in the modified Newton Raphson algorithm based on LQA.

Our proposed methods are investigated by numerical studies and in a real case study in Chapter 5.

Acknowledgements

First, I would like to take this opportunity to express my great appreciation and thanks to my supervisor, Dr. Peng Heng. Without his generous support and guidance, this work would not have completed. I would also like to thank my co-supervisor Prof. Zhu Lixing, for his helpful comments on my work.

I wish to thank Dr. Tong Tiejun, Dr. Yuan Xiaoming from whom I learned asymptotic theory and optimization theory. And I would like to thank Prof. Liao Lizhi for offering me generously his idea of modified Newton-Raphson method.

I also wish to thank all my classmates and my friends for their enormous help.

Finally, I would like to express my gratitude to my parents for their support and encouragement.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 High Dimension Regression Models

Regression models are very important statistical models and have wide practical applications. Given a group of pairs of data $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, independently sampled from a population, where $\mathbf{X}_i$ is a p dimensional random vector and Y_i is a random variable, the main purpose of a regression model is to estimate the regression function $f(x) = E(Y_i|\mathbf{X}_i = x)$. In other words, we need to estimate $f(x)$ from the following model:

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i, \ i = 1, \ldots, n,$$

where ε_i is a random error with zero mean. If the form of $f(x)$ is assumed to be known and can be written as $f(\mathbf{X}_i, \boldsymbol{\theta})$, where θ is an unknown parameter vector, then it is referred to as a parametric regression model; otherwise, it is a nonparametric regression model. A classical example of the parametric regression model is the linear regression model,

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon, \ i = 1, \ldots, n,$$

where $f(\cdot)$ is a linear function and $\boldsymbol{\theta} = \boldsymbol{\beta}$ is the unknown parameter vector in the model. The linear regression model assumes that the response variable Y_i and the predictor vector $\mathbf{X}_i$ have a linear relationship.

In practice, even when the form of the regression function $f(\cdot)$ is unknown, in some circumstances we still have some prior information about $f(\cdot)$, such as the symmetry property or any other special properties. Hence there is a type of statistical model

named a semiparametric model between a parametric model and a nonparametric model. The partial linear regression model (1.1) and the model (1.2) proposed by Li (1991) are examples of such semiparametric models.

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + f(U_i) + \varepsilon_i,\ i = 1, \ldots, n, \tag{1.1}$$

where the function $f(\cdot)$ is an unknown function and U_i is a random variable.

$$Y_i = f(\mathbf{X}_i\boldsymbol{\beta}_1, \ldots, \mathbf{X}_i\boldsymbol{\beta}_K, \varepsilon_i), \quad i = 1, \ldots, n, \tag{1.2}$$

where function $f(\cdot)$ is an unknown function defined by R^{K+1} , $K \leq p$, $\boldsymbol{\beta}_j, 1 \leq j \leq K$ are unknown orthogonal $p \times 1$ vectors and ε_i and $\mathbf{X}_i$ are independent.

In a linear regression model, the unknown parameter vector $\boldsymbol{\beta}$ can be estimated by some closed equation forms. In semiparametric and nonparametric models, the unknown function $f(\cdot)$ can usually be estimated by the kernel regression technique, local polynomial technique or the base function expansion technique. For details, see Fan and Gijbels (1996). However the efficiency of these estimates, whether for parametric regression, semiparametric or nonparametric regression models, are seriously affected by the dimension of the predictor vector $\mathbf{X}_i$. In classical statistical regression models, we regard $(\mathbf{X}_i, Y)$ as a $p + 1$ dimension vector and the dimension p is fixed constant even when the sample size n might be very large. Hence the effect of p the dimension of $\mathbf{X}_i$ is always ignored in theoretical investigations and numerical studies using statistical regression models.

However, the development of science has given rise to many statistical problems in which the dimension of $\mathbf{X}_i$ is no longer small or fixed. For example, DNA microarray technology, which involves a collection of microscopic DNA spots attached to a solid surface, helps scientists to use DNA microarrays to measure the expression levels of large numbers of genes simultaneously. In the data collected by this brand new technology, the dimension p is the number of genes being studied. While, p is usually several hundreds or even thousands, the sample size n, which is the number of DNA samples, is quite small. We can say that n is far smaller than p. Even in classical statistical problems we might introduce more variables to reduce possible modeling biases, and this may also cause an increase in dimension p. Fan and Peng

(2004) studied the gender discrimination suit problem of the Fifth National Bank of Springfield. In their suggested linear and partial linear regression model for the data, the dimension p depends on the sample size. Such problems are so-called high dimensional regression modeling (HDRM) problems.

An early reference to this kind of problem is the seminal paper by Neyman and Scott (1948). In the early years, due to the problems in X-ray crystallography, where the typical values for the number of parameters p and sample size n were in the range of 10 to 500 and 100 to 10,000 respectively, Huber (1973) noticed that in a model selection context, the number of parameters is often large and should be modeled as p_n, which tends to infinity with the sample size. Donoho (2000), using Web term-document data, sensor array data, gene expression data, and consumer financial history data, demonstrate that large sample sizes with high dimensions are important characteristics of such datasets. Donoho also showed that even in a classical setting, such as the Framingham Heart study, the sample size is as large as $n = 25,000$ and dimension $p = 100$. In this example, $p = O(n^{1/2})$ or $p = O(n^{1/3})$.

It is clear that for many modern statistical problems, the assumption that dimension p is finite or a constant is no longer appropriate. Classical methods need to be re-investigated under new high dimensional model settings and new statistical methods may be needed to propose to solve HDRM problems. Solving HDRM problems has become a hot topic in modern statistical science.

1.2 Variable Selection via Penalty

HDRM is a challenging problem because we have to face the so-called "curse of dimensionality." Note that as the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. To obtain a statistically sound and reliable result, the amount of data need to support the result often grows exponentially with the dimensionality. In addition, organizing and searching data often rely on detecting areas where objects form groups with similar properties; however in high dimensional data all objects appear to be sparse and dissimilar in many ways,

which reduces the efficiency of common data organization strategies.

A natural way of avoiding the "curse of dimensionality" is to select significant variables when modeling the data. In the past 50 years, many variable selection methods have been developed by statisticians. These methods are useful in practice and have been applied in many kinds of statistical contexts. Miller (1990) provided a comprehensive summary of various variable selection procedures. One of the basic ideas for variable selection is based on penalization. For the linear regression model

$$Y_i = \mu_i + \varepsilon_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i,\ i = 1, \ldots, n,$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is an $n \times 1$ vector, $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$ is an $n \times p$ matrix and $\boldsymbol{\beta} = (\beta_1, \ldots \beta_p)$ is a $p \times 1$ parametric vector. Variable selection methods are used to select a subset $\mathbf{X}_M$ from $\mathbf{X}$, then estimate β_M based on the selected model $\mathbf{X}_M$. Usually, $\mathbf{X}_M$ is selected by minimizing the criterion in the form of

$$(\mathbf{Y} - \mathbf{X}_M\hat{\boldsymbol{\beta}}_M)^T(\mathbf{Y} - \mathbf{X}_M\hat{\boldsymbol{\beta}}_M) + \lambda|\mathbf{X}_M|\hat{\sigma}^2$$

subject to all possible subsets $\mathbf{X}_M$ of $\mathbf{X}$; λ is a positive penalized parameter, $|\mathbf{X}_M|$ is the dimension of $\mathbf{X}_M$, and $\hat{\sigma}^2$ is an estimate of the variance of the error.

Most of the early variable selection methods mostly used a fixed penalty based on the size of the model. For example, Akaike (1973) stated the Akaike information criterion (AIC) ($\lambda = 2$), Mallows (1973) defined the Mallows' C^p criterion ($\lambda = 2$), and Schwarz (1978) introduced the Bayesian information criterion (BIC) using $\lambda = \log(n)$. These methods often use the stepwise selection and deletion procedure and are practically useful. There are also many modifications of these criteria. For example, Foster and George (1994) proposed the risk inflation criterion (RIC) with $\lambda = 2\log(p)$, and Tibshirani and Knight (1999) suggested the covariance inflation criterion (CIC) with $\lambda = 4\sum_{j=1}^{|\mathbf{X}_M|}\log(n/j)/|\mathbf{X}_M|$. All of these methods perform well when the dimension p is not large, but can not be applied when the dimension p is large. Breiman (1996) and Fan and Li (2001) pointed out that these selection procedures ignore stochastic errors inherited during the variable selection stages. Further, stepwise selection and deletion procedures make these procedures computationally intensive. Therefore, their theoretical properties are relatively difficult to understand and the selection procedures are lack of stability.

To avoid these drawbacks, we need to select variables automatically and simultaneously. One way to achieve this is to penalize the coefficient β_j by the L_q penalty function where $0 < q$. The bridge regression proposed by Frank and Friedman (1993) and the nonnegative garrote proposed by Breiman (1995) are methods based on this idea. Tibshirani (1996) proposed an important method based on the L_1 penalized least squares, called the least absolute shrinkage and selection operator (LASSO). Suppose that we have data $(\mathbf{X}_i, Y_i)$; $i = 1, \ldots, n$, $\mathbf{X}_i$ are the predictor variables and Y_i are the response variable. Furthermore, we suppose that all observations are independent and all X_{ij} are standardized. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, the LASSO estimate $(\hat{\alpha}, \hat{\beta})$ is defined by

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin} \left\{ \sum_{i=1}^{N} (Y_i - \alpha - \sum_{j}^{p} \beta_j X_{ij})^2 \right\} \text{ subject to } \sum_{j}^{p} |\beta_j| \leq t, \qquad (1.3)$$

where $t \geq 0$ is the tuning parameter. As $\hat{\alpha} = \bar{y}$, without loss of generality we can omit α.

All of these convex penalty methods select variables by shrinking and zeroing coefficients. Meanwhile, those convex penalties, such as the quadratic penalties used in the ridge regression and the L_1 penalty function used in LASSO, produce shrinkage estimators of the parameters to make trade-offs between bias and variance. Fan and Peng (2004) stated that most convex penalties can create unnecessary biases when the true parameters are large and parsimonious models cannot be produced. Fan and Li(2001) pointed out that the penalty functions have to be singular at the origin to produce sparse solutions, they also have to satisfy certain conditions to produce continuous estimates, and need to be bounded by a constant to produce nearly unbiased estimates for large coefficients. For the bridge regression and the LASSO, their associated L_q or L_1 penalty functions do not satisfy all of the preceding three required properties. Zou (2006) proposed an adaptive LASSO estimator that satisfies these three properties. However, one problem with this method is that it requires an initial $n^{1/2}$ consistent estimator which is difficult to achieve in high-dimensional cases even though Huang, Ma and Zhang (2008) proposed a sufficient condition for constructing an initial $n^{1/2}$-consistent estimator for high dimensional cases.

Based on the above three properties, Fan and Li (2001) proposed a new unified approach using nonconcave penalized least squares to select variable automatically and simultaneously and estimate their coefficients. A good nonconcave penalty function is their proposed smoothly clipped absolute deviation (SCAD) penalty function.

Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be n predictor-response variable pairs that are assumed to be a random sample. We estimate the regression coefficient β by minimizing the penalized least square

$$C(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \sum_{i=1}^{p} p_\lambda(|\beta_j|) \tag{1.4}$$

where the penalty function p_λ is the SCAD penalty as follows

$$p_\lambda(|\beta|) = \lambda|\beta|\ I(0 \le |\beta| < \lambda)\ +\ (\frac{a\lambda(|\beta| - \lambda) - (|\beta|^2 - \lambda^2/2)}{(a-1)} + \lambda^2)$$

$$\times I(\lambda \le |\beta| \le a\lambda)\ +\ (\frac{(a-1)\lambda^2}{2} + \lambda^2)I(|\beta| \ge a\lambda)$$

If we let $p_\lambda = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$, we obtain the so-called hard thresholding estimator which is not a continuous estimate of $\boldsymbol{\beta}$ as shown by Fan and Li (2001).

The SCAD penalty estimator obtained by minimizing (1.4) satisfies all three properties of unbiasedness, sparsity, and continuity. Furthermore, Fan and Peng (2004) applied the SCAD penalty method successfully when dimension p is no longer fixed, and they also extended the asymptotic properties of the nonconcave penalized least-squares method to the nonconcave penalized likelihood method when the number of parameters in the model is diverging. In high dimensional cases, the SCAD method has the so-called oracle property, i.e. it selects the model consistently and estimates the parameters efficiently as the real parsimony model is known. This is impossible for LASSO (Meinshausen and Bühlmann, 2006; Zou, 2006). Though Zou (2006) proposed an adaptive LASSO estimator that also satisfies the three desirable properties, this procedure requires an initial $n^{1/2}$ consistent estimator, which is difficult to construct in high-dimensional cases. Huang, Ma, and Zhang (2008) proposed a sufficient condition for constructing an initial $n^{1/2}$-consistent estimator for high-dimensional cases, but their sufficient condition is difficult to check in practice. Kim, Choi, and Oh (2008) extended the results of Fan and Peng (2004) to high dimension circumstances when $p > n$. They applied the SCAD method to a regression problem of

gene microarrays in dimension $p = 3000, 1000, 500$ and sample size $n = 100$. The results showed that the SCAD penalized method had the best performance when $p = 1000, 3000$ while LASSO was best when $p = 500$. These results suggest that the SCAD penalized method is a promising method for high-dimensional modeling problems even when either the true model is not sparse or the signal variables are not strongly correlated. The SCAD estimator does not require any initial $n^{1/2}$-consistent estimator and it still has the oracle property in high-dimensional cases. The nonconcave penalized likelihood method has also been applied to other statistical regression models, such as Cox's proportional hazard model (Fan and Li, 2002). Kim, Choi and Oh (2008) and Fan and Lv (2011) showed the relationship between the SCAD estimators and the oracle estimators in ultra-high dimensional linear or generalized linear regression models. Variable selection in ultra-high dimensional models is becoming an important and challenging subject in statistical science.

1.3 Numerical Algorithms for Nonconcave Penalized Methods

As shown above, the SCAD penalized estimate has some very good statistical properties, such as unbiasedness, sparsity, and continuity. It has been shown that given some regular conditions, the nonconcave or SCAD penalized likelihood estimators perform as well as the oracle estimators. However, the nonconcave penalty functions also have some unfriendly drawbacks, such as singularity and nonconvexity. These drawbacks prompt statisticians to invent numerical algorithms that are capable of maximizing a nondifferentiable nonconcave function. To solve such challenging problem, Antoniadis and Fan (2001) proposed nonlinear regularized Sobolev interpolators (NRSI) and a regularized one-step estimator (ROSE) for wavelets smoothing with nonconvex penalized least squares. They also applied the graduated nonconvexity (GNC) algorithm for high-dimensional nonconvex penalized least squares problems.

However GNC is computationally intensive and its implementation depends on a sequence of tuning parameters, so Fan and Li (2001) proposed a new unified algorithm for minimizing the SCAD penalty problems via local quadratic approximations and

called it local quadratic approximation (LQA). The LQA algorithm has two parts: a local quadratic approximation and a modified Newton-Raphson algorithm. Without loss of generality, the target function can be written as

$$\ell(\boldsymbol{\beta}) \ + \ n\sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{1.5}$$

The functions p_λ are singular at the origin and they do not have continuous second order derivatives, but they can be locally approximated by a quadratic function. Suppose we have a nonzero initial value that is close to the true minimizer of (1.5), then p_λ can be locally approximated by a quadratic function as

$$p_\lambda(|\beta_j|) \ \approx \ p_\lambda(|\beta_{j0}|) \ + \ \frac{1}{2}\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}(\beta_j^2 \ - \ \beta_{j0}^2).$$

We assume that the log-likelihood function is smooth with respect to β so that its first two partial derivatives are continuous. Then $\ell(\beta)$ can also be locally approximated by a quadratic function and the Newton-Raphson algorithm can be used. Fan and Li (2001) stated that the estimators obtained by the aforementioned algorithm with a few iterations can always be regarded as one-step estimators, which is as efficient as the fully iterative method. Hunter and Li (2005) studied the convergence property of the LQA algorithm and they found that the LQA algorithm is one of the minimize-maximize (MM) algorithms, which are extensions of the well-known EM algorithm.

However the LQA has a serious drawback. Fan and Li (2001) and Hunter and Li (2005) both pointed out that once a coefficient is reduced to zero, it will stay at zero. One way of avoiding this problem is to using the one-step estimates from the iterative LQA algorithm with good starting estimators.

Kim, Choi, and Oh (2008) studied the SCAD method in high dimensional statistical models. They decomposed the SCAD penalty as the sum of the convex and concave functions, then used the CCCP algorithm (An and Tao, 1997; Yuille and Rangarajan, 2003) for optimization and to find piecewise-linear-regularized-solution-paths (Rosset and Zhu, 2007) during each iteration. The CCCP algorithm has been used in many learning problems, including those of Shen, Tseng, Zhang, and Wong (2003) and Collobert, Sinz, Weston, and Bottou (2006). The key idea of the CCCP algorithm is to update the solution using the minimizer of the tight convex upper

bound of the objective function obtained at the current solution. An important property of the CCCP algorithm is that after each iteration, the objective function always decreases. Thus, the solution eventually converges to a local minimum. Further, the coefficients that are initially zero can be nonzero in iterative steps of the CCCP-SCAD algorithm; that is, the CCCP-SCAD algorithm is less sensitive to the choice of initial solutions. Because the CCCP-SCAD algorithm treats all coefficients equally, computing time is not seriously affected by the initial solution.

However, once again, the problem with the one-step LQA and CCCP estimators is that they cannot have a sparse representation. It is most unfortunate to lose the most attractive and important property of the nonconcave penalized likelihood estimator: sparsity.

Zou and Li (2008) proposed a brand new algorithm for the SCAD penalty, based on local linear approximation (LLA). They stated three significant advantages of LLA: first, it is not necessary to delete any small coefficients or choose the size of perturbation to avoid numerical instability; second, LLA is proven to be the best convex minorization-maximization (MM) algorithm, and its convergence is the same as the MM algorithm (Lange, Hunter, and Yang, 2000); third, LLA produces a sparse estimate. They proved that when tuning appropriate parameter is chosen with a good initial value, the one-step LLA estimator has the oracle properties.

The idea of LLA is as follows

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j{}^{(0)}|) + p'_\lambda(|\beta_j{}^{(0)}|)(|\beta_j| - |\beta_j{}^{(0)}|) \tag{1.6}$$

Similar to the LQA algorithm, the maximization of the penalized likelihood can be carried out by an iterative algorithm and efficient LARS algorithm (Efron *et al.*, 2004). That is,

$$\boldsymbol{\beta}^{(k+1)} = \text{argmax}\{\sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) - n \sum_{j=1}^{p} p'_\lambda(|\beta_j{}^{(k)}|)|\beta_j|\}. \tag{1.7}$$

The iterations stop if the sequence converges. This is the LLA algorithm. The final estimate will have a sparse representation. Beyond this statistical advantage, the LLA algorithm has the the same computational efficiency as the the LASSO, and the approximation is numerically stable, and thus, the drawback of backward

variable selection can be avoided in LLA algorithm. Further, Zou and Li (2008) showed that the one-step LLA estimator has good properties, such as the oracle property, asymptotic normality, sparsity, and continuity. And to the computational efficiency the maximization can be solved by efficient algorithms, such as the least angle regression (LARS) algorithm (Efron, Hastie, Johnstone, and Tibshirani, 2004).

The LARS algorithm is a major breakthrough in the development of the LASSO-type methods. Zou and Hastie (2005) modified the LARS algorithm to compute the solution paths of the elastic net. Rosset and Zhu (2007) generalized the LARS type algorithm to a class of optimization problems with a LASSO penalty. The LARS algorithm was used to simplify the computations in an empirical Bayes model for LASSO (Yuan and Lin, 2006). Zou and Li (2008) implemented LARS in the LLA algorithm with $\ell_i(\boldsymbol{\beta})$ as a quadratic term.

The LQA algorithm depends on the Newton-Raphson method while the LLA algorithm depends on the LARS method. Normally, the Newton-Raphson method converges q-quadratically while the LARS method should only converge linearly and requires the statistical model to have a particular structure, such as a linear regression model structure. This means that the Newton-Raphson method should be faster and more effective, and in particular, it can be applied to more generalized statistical regression models. However, in the Newton-Raphson method, we have to compute an inverse matrix at each step, which is quite awkward if the model dimensionality is high. The LARS method, does not involve such issues. This inconvenience urged us to modify the Newton-Raphson method to avoid having to calculate the inverse of the Hessian matrix.

Without loss of generality, we use following target function to optimize

$$f(\boldsymbol{\beta}) \;=\; \frac{1}{2}(\mathbf{Y} \;-\; \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} \;-\; \mathbf{X}\boldsymbol{\beta}) \;+\; \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{1.8}$$

where the function p is the SCAD penalty function. Given a suitable initial value $\boldsymbol{\beta}^{(0)}$, the direct Newton-Raphson method will give

$$[\mathbf{X}^T\mathbf{X} \;+\; \mathrm{diag}(\ldots, {p'}_\lambda(|{\beta_j}^{(0)}|)/|\beta_j|, \ldots)]\boldsymbol{\beta} = \; \mathbf{X}^T\mathbf{Y}.$$

We rewrite the above equation as

$$[\mathbf{X}^T\mathbf{X} \;+\; D(\boldsymbol{\beta}^{(0)})]\boldsymbol{\beta} \;=\; d(\boldsymbol{\beta}^{(0)}). \tag{1.9}$$

If we need to solve $\boldsymbol{\beta}^{(1)}$ from (1.8), we have to compute the inverse matrix of $\mathbf{X}^T\mathbf{X} + D(\boldsymbol{\beta}^{(0)})$. When the model dimension is high, this task is impossible, so the question is whether we can find a way to avoid having to calculate the inverse matrix.

We split the matrix $\mathbf{X}^T\mathbf{X}$ into three parts: L, Λ, and L^T, where L is a strictly lower diagonal matrix, Λ is a diagonal matrix and $X^TX = L + L^T + \Lambda$. After some easy calculation we have an approximate system for (1.9):

$$(L + \Lambda + D(\boldsymbol{\beta}^{(0)}))\boldsymbol{\beta} = d(\boldsymbol{\beta}^{(0)}) - L\boldsymbol{\beta}^{(0)}. \tag{1.10}$$

It is clear that $L + \Lambda + D(\beta^{(0)})$ is an upper diagonal matrix which means that (1.5) can be solved easily by a backward substitution algorithm. If we could prove its convergence, this modified Newton-Raphson algorithm would no doubt be much more effective than the original Newton-Raphson algorithm and we could apply this new method to the LQA algorithm and many other statistical optimization problems.

1.4 Sure Independence Screening in Sparse Ultra-high-dimensional models

High-throughput data of unprecedented size and complexity are frequently seen in many contemporary statistical studies, such as genetic, microarray, proteomic, and functional magnetic resonance imaging studies, and in functional data and high-frequency financial data. In all of these examples, the number of variables p can grow much faster than the number of observations n, $\log(p) = O(n^a)$ where $0 < a < 1/2$. Fan and Lv (2010) referred to this as nonpolynomial (NP) dimensionality. Dimension reduction and feature selection play pivotal roles in these ultra-high-dimensional problems.

Classical methods for dimension reduction and feature selection such as LASSO (Tibshirani 1996), SCAD, and other folded-concave penalized methods (Fan, 1997; Fan and Li, 2001), the Dantzig selector (Candes and Tao, 2007), the Elastic net (Enet) (Zou and Hastie, 2005), and MCP (Zhang, 2010) have excellent performance when the model dimension of is not very large. However, due to the curse of dimensionality in terms of the simultaneous challenges of computational expediency,

statistical accuracy, and algorithmic stability, these methods are limited in their ability to handle ultra-high-dimensional problems. Kim, Choi, and Oh (2008) studied the SCAD method on high dimensions when $p > n$, but not on NP dimensionality. Motivated by these concerns, Fan and Lv (2008) introduced a new framework for variable screening via correlation learning in ultra-high dimensional statistical models. Hall, Titterington, and Xue (2008) used a different marginal utility, derived from an empirical likelihood point of view. Hall and Miller (2009) proposed a generalized correlation ranking, which allows nonlinear regression. Huang, Horowitz, and Ma (2008) also investigated the marginal of the bridge regression in the ordinary linear model. These methods focus on studying the marginal pseudo-likelihood and are fast but crude in terms of reducing NP dimensionality to a more moderate size. To overcome this problem, Fan and Lv (2008) and Fan, Samworth, and Wu (2009) introduced some methodological extensions to sure independence screening (SIS), including iterative SIS (ISIS) and multistage procedures such as SIS-SCAD and SIS-LASSO, to select variables and estimate parameters simultaneously.

However, these marginal screening methods have some methodological challenges. When the covariates are not jointly normal, even if the linear model holds in the joint regression, the marginal regression can be highly nonlinear. Thus, SIS based on nonparametric marginal regression becomes a natural candidate. In practice, there is often little prior information to indicate that the effects of the covariates take a linear form or belong to any other finite-dimensional parametric family. Thus using a more flexible class of nonparametric models, such as the additive model, would be an appropriate choice. The works of Koltchinskii and Yuan (2008) and Ravikumar *et al.*(2009) are closely related to the COSSO method proposed by Lin and Zhang (2006) with fixed minimal signals, which does not converge to 0. Huang, Horowitz, and Wei (2010) applied an adaptive LASSO to additive models with fixed minimal signals. Meier, Geer, and Bühlmann (2009) proposed a penalty that combines sparsity and smoothness with a fixed design for high-dimensional additive models. In ultra-high dimensional additive model settings, all of these methods still suffer from the aforementioned three challenges, because they can be viewed as extensions of penalized pseudo-likelihood approaches to nonparametric additive models. Backfit-

ting, commonly used algorithm in additive models, makes the situation even more challenging because of its great computational expense.

Fan, Feng, and Song (2011) applied the idea of sure independence screening to an ultra-high dimensional additive model by ranking the magnitudes of the marginal estimators, nonparametric marginal correlations, and the marginal residual sum of squares. Their work is a natural extension of the SIS procedures proposed by Fan and Lv (2008) and Fan and Song (2010). They approximated the nonparametric additive components by using a B-spline basis. The component selection in additive models can then be viewed as a functional version of the grouped variable selection which have been studied by Yuan and Lin (2006), Kim, Kim, and Kim (2006), Wei and Huang (2007), and Meier, Geer, and Bühlmann (2009). The method that Fan, Feng and Song (2011) proposed is called nonparametric independence screening(NIS). They also proposed an iterative NIS procedure in which variable selection and component function estimation can be achieved simultaneously. Their work can be readily adapted to other smoothing methods (Silverman, 1984; Horowitz, Klemelä, and Mammen 2006), and such as local polynomial regression (Fan and Jiang 2005), wavelet approximations (Antoniadis and Fan, 2001; Sardy and Tseng, 2004), and smoothing splines (Speckman, 1985), and can also be applied to other semiparametric models, such as varying coefficient models (Fan and Zhang 1999).

Using a multistage procedure, NIS-SCAD, we can apply the NIS method to other ultra high dimensional nonparametric or semiparametric models, such as varying coefficient models, to reduce the number of dimensions and select significant variables in the model.

1.5 Varying Coefficient Models

Linear regression is the oldest ancient, but still the most convenient, tool for parametric statistical inference, although linear regression models are unrealistic in many applications. Several useful data-analytic modeling techniques have been proposed to relax traditional parametric models and to exploit possible hidden nonlinear structures in the data. For example, additive models (Breiman and Friedman, 1985; Hastie

and Tibshirani, 1990), varying coefficient models (Hastie and Tibshirani, 1993; Fan and Zhang, 1999, 2000; Chiang, 2001), low-dimensional interaction models (Friedman, 1991; Gu and Wahba, 1992; Stone *et al.*, 1997), and partially linear models (Wahba, 1984; Green and Silverman, 1994), and their hybrids (Carroll *et al.*, 1997; Fan *et al.*, 1998; Heckman *et al.*, 1998; Fan *et al.*, 2003). Due to the special structure of these nonparametric and semiparametric models, the "curse of dimensionality" can be avoided when the number of predictor variables in the model is not large. As pointed out by Fan and Zhang (2008), these models are very useful tools for exploring dynamic patterns in many scientific areas. For example, Fan and Zhang (1999) used varying coefficient models to study dynamic changes in the Hong Kong environment over time and with different factors, and Cheng and Zhang (2007) also used the varying coefficient models to study changes in infant mortality in China over time.

The varying coefficient model (VCM) is a very important semiparametric model. Because of its flexibility and interpretability, the VCM has been applied to many scientific areas in the past 10 years. As a classical semiparameteric model, it has also underdone deep and exciting methodological and theoretical developments. The varying coefficient model was first introduced by Cleveland, Grosse, and Shyu (1991). To consider multivariate predictor variables containing a scalar U and a vector $\mathbf{X} = (X_1, \ldots, X_p)^T$, the varying-coefficient models assume the form of multivariate regression function,

$$m(U, \mathbf{X}) = \mathbf{X}^T \mathbf{a}(U) \tag{1.11}$$

for an unknown functional coefficient $\mathbf{a}(U) = (a_1(U), \ldots, a_p(U))^T$, where the function m is the regression function. (1.11) is the basic form of the varying coefficient models. It is a useful extension of the thresholding models in Tong (1990) and Chen and Tsay (1993) with the time series set-up. It also appears a natural choice for longitudinal data analyses in which one wishes to explore the extent to which covariates affect response changes over time (Hoover *et al.*, 1998; Wu *et al.*, 1998). The varying coefficient models are also useful for analysing functional types of data (See Ramsay and Silverman, 1997, and Brumback and Rice, 1998).

The smoothness of the coefficient function $a_j(U)$ is essential in the varying coef-

ficient models. Hastie and Tibshirani (1993) proposed an estimate for $a_j(U)$ via the dynamic linear model. The development in local polynomial modeling (See Fan and Gijbels, 1996; Fan and Zhang (1999) allowed to derive the asymptotic mean-square errors for the one-step and two-step estimation procedures of varying coefficient models. They found that the one-step estimator for a_p inherits nonnegligible approximation errors and is not optimal, but the two-step estimator of a_p achieves the optimal rate of convergence. Another important point is that the asymptotic conditional bias and the asymptotic conditional variance of the two-step estimator of a_p does not rely on the one-step estimation bandwidth h_1.

Fan and Zhang (2008) gave a detailed review of the estimation methods used in the the varying coefficient models. Usually, there are three approaches to estimate the $\mathbf{a}(U)$ in (3.1): kernel-local polynomial smoothing (Wu, Chiang, and Hoover, 1998; Fan and Zhang, 1999); polynomial spline (Huang and Shen, 2004; Huang, Wu, and Zhou, 2004) and smoothing spline (Hastie and Tibshirani, 1993; Hoover, Rice, Wu, and Yang, 1998). The kernel-local polynomial smoothing is a reasonable estimation method for the varying coefficient model, which adaptively avoids the boundary effect for the estimate of $\mathbf{a}(U)$. By Taylor's expansion of $\mathbf{a}(U)$, for each given u, the local linear estimator $\hat{\mathbf{a}}(U)$ is the part corresponding to $\mathbf{a}$ of the minimizer in the following optimization problem

$$L(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} \left\{ y_i - {X_i}^T \mathbf{a} - {X_i}^T \mathbf{b}(U_i - u) \right\}^2 K_h(U_i - u)$$

where K is a kernel function. Then the matrix form of $\hat{\mathbf{a}}(U)$ can be written as

$$\hat{\mathbf{a}}(U) = (I_p, \mathbf{0}_p) \left({\Gamma_u}^T W_u \Gamma_u \right)^{-1} {\Gamma_u}^T W_u \mathbf{Y}$$

Fan and Zhang (2000) got the asymptotic distribution of above $\hat{a}_1$. From this result they constructed simultaneous confidence bands of $\hat{a}_1$ and applied them to hypothesis testing inference. Bandwidth selection is essential to the estimation of varying coefficient functions. Basically, using an optimal bandwidth the estimator of varying coefficient functions should be able to minimize its mean squared error (MSE). Based on this idea and the pre-asymptotic substitution idea of Fan and Gijbels (1995), Fan and Zhang (2000) proposed an efficient bandwidth selection method for the local linear estimates of the varying coefficient model.

Constructing simultaneous confidence bands for varying coefficient functions is a very important issue for solving the inference problems of varying coefficient regression models. Wu *et al.* (1998) and Chiang *et al.* (2001) studied the pointwise confidence interval for the functional coefficients in varying coefficient models. In Wu (1998), he investigated the Bonferroni-type confidence bands. Huang *et al.* (2002, 2004) investigated the pointwise confidence intervals and confidence bands based on the polynomial spline approach and the Bonferroni adjustment.

A typical hypothesis testing problem for the varying coefficient model is to test

$$H_0 : a_j(U) = C_j \leftrightarrow H_1 : a_j(U) \neq C_j,$$

where C_j is a constant. Based on the maximum of the normalized deviations, Fan and Zhang (2000) suggested the following test statistics T_j for the above hypothesis, and H_0 is rejected if the value of the test statistic T_j exceeds the asymptotic critical value $C_\alpha = -log\{-0.5\ log(1-\alpha)\}$.

$$T_j = (-2\ logh)^{1/2}(\sup_{0\leq\mu\leq 1} |\{\widehat{var}(\hat{a}_j|D)\}^{1/2} \times (\hat{a}_j - \hat{C}_j - \widehat{bias}(\hat{a}_j|D))| - d_{v,n}), \quad (1.12)$$

where, $\hat{C}_j = n^{-1}\ \sum_{i=1}^{n} \hat{a}_j(U_i)$.

Cai, Fan and Yao (2000) developed a bootstrap-based test method for the hypothesis in which they used the idea of a generalized likelihood ratio (GLR) (Fan *et al.*, 2001).

Another important issue with the varying coefficient models is how to estimate $a_j(U), j = 1, \ldots, p$ when the components of $a_j(U), j = 1, \ldots, p$ have different degrees of smoothness. The main difficulty with this issue is that the smoother components need a larger bandwidth while the less smooth components need a smaller bandwidth. Fan and Zhang (1999) proved that (see Theorem 1 of Fan and Zhang, 1999) no matter how the bandwidth is chosen, one-step estimation cannot optimally estimate the smoother components [$a_p(U)$ in Theorem 1 of Fan and Zhang (1999)]. As mentioned above, they also proved that two-step estimation always outperforms one-step estimation when estimating the smoother components (see Theorem 2 of Fan and Zhang, 1999). When estimating the less smooth components, one-step and two-step estimators work equally well. Another advantage of two-step estimation is that the asymptotic conditional variance of the two-step estimator of a_p does not rely

on the one-step estimation bandwidth h_1, thus we can use a smaller bandwidth to detain the one-step estimator with large variance and small bias.

Without loss of generality, we assume that a_p is smoother than other functional coefficients. We rewrite model (1.11) as

$$y_i = \sum_{j=1}^{p-1} a_j(U_i)x_{ij} + a_p(U_i)x_{ip} + \varepsilon.$$

Using one-step estimation we have the estimates $\hat{\mathbf{a}}_j(U), j = 1, \ldots, p$ mentioned above. Then, we replace all $a_j(U_i), j = 1, \ldots, p-1$ by $\hat{a}_j(U_i)$, and we have the synthetic model

$$\hat{y}_i = y_i - \sum_{j=1}^{p-1} \hat{a}_j(U_i)x_{ij} = a_p(U_i)x_{ip} + \varepsilon.$$

Once again we use Taylor's expansion on $a_p(U_i)$ and following the same technique we have $\hat{a}_p(U_i)$. This is the two-step estimator of $a_p(U_i)$.

There are some extensions of varying coefficient models. For example, Fan *et al.* (2003) introduced a new varying coefficient model in which the variable U is unknown and unobservable, called the adaptive varying coefficient model

$$E(Y_i|\mathbf{X}) = \sum_{j=1}^{p} g_j(\boldsymbol{\beta}^T\mathbf{X}_i)X_j,$$

where $\boldsymbol{\beta}$ is an unknown direction, and $\mathbf{X} = (X_1, \ldots, X_p)^T$. Unlike the varying coefficient model, $U = \boldsymbol{\beta}^T\mathbf{X}$ is an unknown index. Fan *et al.* (2003) showed that the model is identifiable unless $E(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}\boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\gamma}^T\mathbf{X} + c$. If $\boldsymbol{\beta}$ is given, the model is really a varying-coefficient model and the functional coefficients can be estimated by using the method mentioned above, resulting in the estimates $\hat{g}(\cdot, \boldsymbol{\beta})$. Substituting this into the original adaptive varying coefficient model, we can use the least squares method to estimate $\boldsymbol{\beta}$. Fan *et al.* (2003) suggested an iterative algorithm to estimate $\boldsymbol{\beta}$ and $g(\cdot)$ simultaneously. They also extended such a model to a more generalized adaptive varying coefficient model with two index parameter vectors.

The hypothesis testing mentioned in Fan and Zhang (2000) tells us hat even in the varying coefficient model, some of the components of $a_j(U)$ may be constant. Then the model can be rewrite as

$$Y = {Z_1}^T\mathbf{a}_1(U) + {Z_2}^T\mathbf{a}_2 + \varepsilon. \tag{1.13}$$

Fan and Zhang (2008) claimed that the model (1.13) cannot be viewed as a special case of the varying coefficient model because we already know that $\mathbf{a}_2$ is a constant vector. Zhang, Lee and Song (2002) proposed a two-step estimation procedure. They showed that the two-step estimator of $\widehat{\mathbf{a}}_2$ has a convergence rate of $O_p(n^{-1/2})$. This result is important because we can replace $\mathbf{a}_2$ by $\widehat{\mathbf{a}}_2$ with little influence on the estimation of the functional coefficient $\mathbf{a}_1$ and model (1.13) will become an ordinary varying coefficient model.

The convergence rate of the estimator of $\mathbf{a}_2$ suggested by Zhang *et al.* (2002) is optimal; however, the asymptotic variance of the estimate does not reach the lower bound for the semiparametric model. To overcome this drawback, Fan and Huang (2005) proposed a so-called profile least-squares technique to estimate $\mathbf{a}_2$. The key point is that $\mathbf{a}_2$ is pretended to know, and model (1.13) is rewritten as follows:

$$\tilde{y}_i = y_i - {Z_{i2}}^T \mathbf{a}_2 = {Z_{i1}}^T \mathbf{a}_1(U_i) + \varepsilon_i \tag{1.14}$$

Then we apply the estimation procedure suggested by Fan and Zhang (1999) to get the estimator of $\mathbf{a}_1(U_i)$. The next step is to substitute $\mathbf{a}_1(U_i)$ by the estimate $\hat{\mathbf{a}}_1(U_i)$ to obtain the following synthetic model

$$Y_i \ - \ {Z_{i1}}^T \tilde{\mathbf{a}}_1(U_i) \ = \ {Z_{i2}}^T \mathbf{a}_2 \ + \ \varepsilon_i.$$

Then by the least squares estimation, the estimator of $\mathbf{a}_2$ can be obtained. Fan and Huang (2005) showed that the covariance matrix of the estimator of $\mathbf{a}_2$ reaches the lower bound for semiparametric models. Furthermore, Xia, Zhang and Tong (2004) proposed a cross-validation based model selection procedure to determine which components are constant and which are functional in practice.

Like generalized linear models (GLM), the generalized varying coefficient model (GVCM) is defined as follows

$$g(m(U, \mathbf{X})) = \theta(U, \mathbf{X}) = \mathbf{X}^T \mathbf{a}(U). \tag{1.15}$$

where $g(\cdot)$ is the known link function and $\mathrm{E}(Y|\mathbf{X}) = m(U, \mathbf{X})$. Cai, Fan and Yao (2000) have established the asymptotic normality of the local maximum likelihood estimator of $\mathbf{a}(u)$, and they also showed that the bias of $\hat{a}(U)$ is the same as that in standard varying coefficient models. Thus they concluded that the local maximum

likelihood estimation is efficient. In practice, however, the local maximum likelihood estimator does not usually have a closed form. To overcome this computational difficulty, Cai, Fan and Li (2000) proposed a one-step Newton-Raphson estimation for $\mathbf{a}(U)$, and showed that the one-step Newton-Raphson estimator can save computational cost by an order of tens with excellent performance. For bandwidth selection, Fan and Zhang (2008) suggested using cross-validation.

Like normal varying coefficient models, a typical hypothesis problem for generalized varying coefficient models is whether certain coefficients are really varying with u or whether certain coefficients are significantly different from 0. These are

$$H_0 \ : \ a_k(u) = a_k, \quad k \ = \ 1, \ldots, p, \quad \text{and} \quad H_0 \ : \ a_k(u) = 0, \quad \text{for certain } k.$$

Cai, Fan, and Yao (2000) used the generalized maximum likelihood ratio test developed by Fan, Zhang and Zhang (2001) to successfully construct the test statistics for these hypotheses. They also proved that the asymptotic distribution of the test statistic under the null hypothesis is distribution free and not related to the value of $\mathbf{a}(u)$ i.e. the Wilks phenomenon.

Fan and Zhang (2008) gave more details about the practical application of the varying coefficient models for analysing longitudinal and functional data, survival analysis, nonlinear time series, and time-varying diffusion models. All of these applications show that the varying coefficient models are developing rapidly and their applications are becoming wider. High dimensionality is the most important characteristic of recent datasets, and hence through deep investigations and research on high dimensional or ultra-high dimensional varying coefficient models are becoming increasingly necessary and important.

1.6 Organization of the Dissertation

In this dissertation, three inter-related topics related to high-dimensional semi-varying coefficient modeling are studied. In Chapter 2, the oracle properties and asymptotic properties of the non-concave penalized least squares in high-dimensional semi-varying coefficient models are considered. In Chapter 3, nonparametric independence screening(NIS) is applied to ultra-high-dimensional semi-varying coefficient models.

In Chapter 4, a modified Newton-Raphson method is considered to avoid calculating the inverse of the Hessian matrix. In Chapter 5, we discuss some practical issues about the statistical methods investigated in this dissertation, and then present all of our numerical study results. In Chapter 6, we discuss the work of this dissertation and some related directions for further investigation.

Chapter 2

Model Selection in Semivarying Coefficient Models with Diverging Number of Variables

2.1 Introduction

High dimensionality is an important characteristic of much recently collected data. Most popular statistical models have been investigated under high-dimensional model settings. Regularization or penalization methods are very useful tools for selecting and estimating these high dimensional statistical models. With the development of group penalized methods and to reduce the model bias, many semiparametric or nonparametric models have been applied to model high dimensional data. As an important tool for exploring the dynamic patterns in many scientific areas and to test the efficiency or validity of newly developed statistical methods, the varying coefficient model has also been investigated under high dimensional model settings in some studies.

Consider the following linear regression model

$$\mathbf{Y} = \sum_{i=1}^{K} \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is an $n \times 1$ vector, $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \mathbf{I})$, $\mathbf{X}_i$ is an $n \times p_i$ matrix corresponding to the ith factor and $\boldsymbol{\beta}_i$ is a coefficient vector of size $p_i, i = 1, \ldots, K$. To select groups of

variables $\boldsymbol{\beta}_i, i = 1, \ldots, K$ from the above model, Yuan and Lin (2006) proposed the group LASSO to estimate and select group variables "all in or all out" simultaneously by minimizing

$$\frac{1}{2}\left\|\mathbf{Y} - \sum_{i=1}^{K} \mathbf{X}_i \boldsymbol{\beta}_i\right\| + \lambda \sum_{i=1}^{K} \|\boldsymbol{\beta}_i\|_2$$

where $\|\boldsymbol{\beta}_i\|_2 = (\sum_{j=1}^{p_i} \beta_{ij})^{1/2}$. The group LASSO uses an ℓ_2 norm of the coefficients associated with a group of variables in the penalty function and is a natural extension of the LASSO (Tibshirani, 1996). Yuan and Lin (2006) suggested an efficient revised LARS algorithm to find the solution path for the group LASSO.

As claimed by Yuan and Lin (2006), the group variable selection method has broad applications, such as the multi-factor analysis of variance (ANOVA) problem and the nonparametric or semiparametric regression model selection problem. Such methodology has been widely studied. Antoniadis and Fan (2001) studied a class of block-wise shrinkage approaches for regularized wavelet estimation in nonparametric regression problems. Meier, Geer, and Bühlmann (2009) studied the group LASSO for logistic regression. Zhao, Rocha, and Yu (2009) proposed a quite general composite absolute penalty for group selection, which includes the group LASSO as a special case. Huang, Ma, Xie, and Zhang (2009) proposed a group bridge method to simultaneously select group variable and individual variables. Breheny and Huang (2009) investigated the group bridge method under a generalized linear regression model.

An important application of group LASSO is in nonparametric statistical modeling, including nonparametric additive model and varying coefficient models. Let $(Y_i, \mathbf{X}_i), i = 1, 2, \ldots, n$ be random vectors that are independently and identically distributed as $(Y, \mathbf{X})$, where Y is a response variable, and $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ is a p-dimensional covariate vector. The nonparametric additive model posits that

$$Y_i = \mu + \sum_{i=1}^{p} f_j(X_{ij}) + \varepsilon_i,\ 1 \le i \le n,$$

where μ is an intercept term, X_{ij} is the jth component of $\mathbf{X}_i$, the f_j's are unknown functions and ε_i is an unobserved random variable with a mean of zero and finite variance of σ^2. Suppose some components, f_j, are zero. One objective of nonparametric additive modeling is to select nonzero components and estimate them. For

fixed p, Lin and Zhang (2006) proposed a component selection and smoothing operator (COSSO) to simultaneously select and estimate nonzero component functions in the nonparametric additive model. The COSSO can be regarded as a group LASSO procedure in a reproducing kernel Hilbert space. Combining the smoothing spline and ℓ_2 norm of component functions, Meier, van de Geer and Bühlmann (2009) proposed a variable selection method for the ultra-high-dimensional nonparametric additive model. Their method is some closely related to the group LASSO. Huang, Horowitz and Wei (2010) studied the group LASSO for the nonparametric additive model based on B-spline approximation to the nonparametric components. Cui, Peng, Wen and Zhu (2013) suggested using power spline to approximate the nonparametric component and using the group bridge selection method to select the nonzero component function, and spline knots simultaneously.

For the varying coefficient model

$$Y_i = \sum_{k=1}^{p} X_{ik}\beta_k(U_i) + \varepsilon_i,\ \ i = 1, \ldots, n,$$

Wang, Chen and Li (2007) and Wang and Xia (2009) considered the use of group LASSO and SCAD methods to select significant variables and estimate and $\beta_i(U), i = 1, \ldots, p$ simultaneously. Xue, Qu and Zhou (2010) applied the ℓ_2 norm SCAD methods for variable selection in generalized linear varying coefficient models and considered its selection and estimation properties. When $p \gg n$, Wei, Huang and Li (2011) studied the group LASSO for the above varying coefficient model based on B-spline approximation of the varying coefficient functions. Xue and Qu (2012) considered difference convex programming to approximate the ℓ_0 penalty under high-dimensional varying coefficient models. They showed that their method can select significant coefficient functions consistently and features the oracle properties. By applying B-spline approximation to the varying coefficient functions, Lian (2012) applied the group LASSO to high dimensional generalized varying coefficient models.

The above group selection methods for nonparametric additive models and varying coefficient model cannot provide more subtle estimations when the real model is a partial nonparametric additive model or a so-called semivarying-coefficient model. These group selection methods cannot simultaneously identify which component functions in the partial nonparametric additive models are linear functions, or which co-

efficient functions are constants in the semivarying coefficient models. For a partial nonparametric additive model, based on power spline approximation, Cui *et al.* (2013) recently proposed a group selection method to determine the zero, linear and nonlinear component functions in the partial nonparametric additive model. Zhang, Cheng and Liu (2011) suggested a similar selection method using the COSSO technique. Huang, Wei and Ma (2012) proposed a semiparametric regression pursuit method to develop a partial nonparametric additive model and determine the zero, linear and nonlinear components in the model.

However, there has been little work on the structure estimation and selection procedure for semivarying coefficient models, especially when the dimension of the model diverges with the sample size. In this chapter, based on the local polynomial regression technique, nonconcave SCAD penalized group method, we suggest a model selection procedure for semivarying coefficient models when the dimension of the model diverges with the sample size. Our method can determine the zero, constant, and non-constant coefficient functions simultaneously. Then, from the model selection results, we propose a simple and efficient method to estimate constant coefficients in the model. In Section 2.2, we describe our method clearly. In Section 2.3, we investigate the theoretical properties of our proposed method, and some practical issues are discussed in Section 2.4. The numerical studies for our proposed methods are present in Chapter 5. The proofs of the theorems presented shown in Section 2.3 are relegated to the Appendix of this chapter, Section 2.5.

2.2 Regularized Estimation for Semivarying Coefficient Models By Local Linear Fitting

Suppose we have a sample $(U_i,\ \mathbf{X}_i^{\mathrm{T}},\ Y_i)$, $i = 1,\ \cdots,\ n$, from $(U,\ \mathbf{X}^{\mathrm{T}},\ Y)$ where $\mathbf{X} = (X_1,\ \cdots,\ X_{p_n})^{\mathrm{T}}$ and

$$Y = \sum_{j=1}^{p_n} a_j(U_i)X_j + \varepsilon. \tag{2.1}$$

Without loss of generality, we assume the support set of U is $[0,\ 1]$. Here, we consider p_n, the dimension of the model or the number of predictor variables in the model to

be diverging with the sample size. Hence when p_n is some large, it is reasonable to assume that some of $a_j(\cdot)$, $j = 1, \cdots, p_n$, in (2.1) may be zero, or some nonzero constant. Hence there are two objectives for the model estimation and selection. One is to determine the zero coefficients from $a_j(\cdot)$, $j = 1, \cdots, p_n$, and the other is to find those nonzero constant coefficients and estimate them efficiently. Although kernel smoothing (Wang and Xia, 2009) and B-spline approximation (Wei, Huang, and Li, 2011) are good tools as modeling varying coefficient models, these two objectives cannot be achieved simultaneously. In this chapter, we propose a regularized method whereby the local linear fitting is used to approximate the varying coefficient functions in the above model. As shown by Fan and Zhang (1999), the local linear fitting can estimates the varying coefficient function and its derivative simultaneously. Hence combined with the idea of regularization, it is possible to determine the zero, nonzero constant and nonconstant coefficient functions simultaneously by the local linear fitting.

Throughout this paper, for any vector $\mathbf{u} \in R^p$, we define $\|\mathbf{u}\| = \left(\mathbf{u}^{\mathrm{T}}\mathbf{u}/p\right)^{1/2}$.

By the Taylor's expansion, $a_j(U_i)$ can be approximated by $a_j(U)$ as

$$a_j(U_i) \approx a_j(U) + h \cdot a_j^{(1)}(U)\frac{U_i - U}{h}$$

when U_i is in a small neighbourhood of U. Hence to estimate $a_j(U)$ and $a_j'(U)$, as shown by Fan and Zhang (1999), $\hat{a}_j$ and $\hat{b}_j, j = 1, \ldots, p_n$, which minimize the objective function

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p_n}\left\{a_j + b_j\frac{U_i - U}{h}\right\}X_{ij}\right]^2 K_h(U_i - U),$$

are just the estimates of $a_j(U)$ and $a_j'(U)$. If $a_j(U) \equiv 0$ for any U, then $\hat{a}_j(U) \approx 0$ and $\hat{a}_j'(U) \approx 0$. If $a_j(U)$ is a nonzero constant, although $\hat{a}_j(U)$ would not be close to zero, $\hat{a}_j'(U)$ should be still close to zero. Hence by letting $U = U_1, \ldots, U_n$ and applying the idea of penalization to the norms of $\hat{a}_j(U)$ and $\hat{a}_j'(U)$ in the equation

above leads to the following objective function:

$$\sum_{k=1}^{n}\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\sum_{j=1}^{p_n}\left\{a_{jk}+b_{jk}\frac{U_i-U_k}{h}\right\}X_{ij}\right]^2 K_h(U_i-U_k)$$
$$+\quad n\sum_{j=1}^{p_n}\left\{p_{\lambda_j}(\|\mathbf{b}_j\|)+p_{\lambda_j}(\|\mathbf{a}_j\|)\right\}, \tag{2.2}$$

where $\mathbf{b}_j=(b_{j1},\ \cdots,\ b_{jn})^{\mathrm{T}}$, $\mathbf{a}_j=(a_{j1},\ \cdots,\ a_{jn})^{\mathrm{T}}$. $p_{\lambda_j}(\cdot)$ is a penalty function and λ_j is a tuning parameter.

By minimizing (2.2) with respect to $(\mathbf{a}_j,\ \mathbf{b}_j)$, $j=1,\ \cdots,\ p_n$, we can get the estimates of the curves $a_j(U)$ and $a_j'(U), j=1,2,\ldots,p_n$. Let the minimizer be $(\hat{\mathbf{a}}_j,\ \hat{\mathbf{b}}_j)$,

$$\hat{\mathbf{a}}_j=(\hat{a}_{j1},\ \cdots,\ \hat{a}_{jn})^{\mathrm{T}},\quad \hat{\mathbf{b}}_j=\left(\hat{b}_{j1},\ \cdots,\ \hat{b}_{jn}\right)^{\mathrm{T}}.$$

We expect $\|\hat{\mathbf{a}}_j\|=0$ when when $a_j(\cdot)\equiv 0$ and $\|\hat{\mathbf{b}}_j\|=0$ when $a_j(\cdot)$ is a nonzero constant, C_j. If $\|\hat{\mathbf{b}}_j\|=0$ but $\|\hat{\mathbf{a}}_j\|\neq 0$, then C_j can be simply estimated as

$$\hat{C}_j=n^{-1}\sum_{i=1}^{n}\hat{a}_{ji}. \tag{2.3}$$

We expect that this estimate also has some optimal properties with $\sqrt{n}$ convergence rate as shown by Fan and Zhang (2000).

We believe that the L_1 penalty function can be used in (2.2) to obtain a reasonable estimate as in Wang and Xia (2009), when the dimension of the varying coefficient model is diverging with the sample size. The L_1 penalty will produce biased estimates and those biases will be accumulate to affect the final model estimation and selection. Hence we suggest using the nonconcave penalty function to replace the L_1 penalty function in (2.2), Specially we apply the SCAD penalty function in the following numerical and theoretical studies, although other nonconcave penalty functions, such as MCP (Zhang, 2010), can be also applied.

2.3 Asymptotical Properties of the Proposed Estimation

In this section, we investigate the theoretical properties of our proposed estimates of varying coefficient functions based on (2.2) and (2.3) when the dimension of the

model (2.1) is diverging with the sample size.

Without loss of generality, we assume $a_j(\cdot)$, $j = 1, \cdots, p_{0n}$, are nonconstant functional, $a_j(\cdot) = C_j$, $j = p_{0n}+1, \cdots, p_{1n}$, are nonzero constant, and $a_j(\cdot) = 0$, $j = p_{1n}+1, \cdots, p_n$. Define

$$\boldsymbol{\beta}_k^{(1)} = (\beta_{k1}, \ldots, \beta_{kp_{1n}+p_{0n}})^T = (a_{1k}, \ldots, a_{p_{1n}k}, b_{1k}, \ldots, b_{p_{0n}k})^T,$$

$$\boldsymbol{\beta}_k^{(2)} = (\beta_{kp_{1n}+p_{0n}+1}, \ldots, \beta_{k2p_n})^T = (a_{p_{1n}+1k}, \ldots, a_{p_nk}, b_{p_{0n}+1k}, \ldots, b_{p_nk})^T,$$

$$\boldsymbol{\beta}_k^* = (\boldsymbol{\beta}_k^{(1)T}, \boldsymbol{\beta}_k^{(2)T})^T$$

and

$$\boldsymbol{\beta}_j = (\beta_{1j}^*, \ldots, \beta_{nj}^*)^T.$$

Under the model assumption, we know that the real $\boldsymbol{\beta}_k^{(2)T} = 0$ and $\boldsymbol{\beta}_k^{(1)T}$ are related to those significant predictor variables.

By the definitions above, (2.1) is written as

$$\sum_{k=1}^{n} \frac{1}{n} \sum_{i=1}^{n} \left[Y_j - X_{ik}^{(1)} \boldsymbol{\beta}_k^{(1)} - X_{ik}^{(2)} \boldsymbol{\beta}_k^{(2)} \right]^2 K_h(U_i - U_k) + n \sum_{j=1}^{p_n} \left\{ p_{\lambda_j}(\|\boldsymbol{\beta}_j\|) \right\}, \qquad (2.4)$$

where

$$X_{ik}^{(1)} = \left(X_{i1}^{(1)}, \ldots, X_{ip_{n0+n1}}^{(1)} \right) = \left(X_{i1}, \ldots, X_{ip_{1n}}, X_{i1} \frac{U_i - U_k}{h}, \ldots, X_{ip_{0n}} \frac{U_i - U_k}{h} \right)$$

and

$$X_{ik}^{(2)} = \left(X_{ip_{n0+n1}+1}^{(1)}, \ldots, X_{i2p_n}^{(1)} \right) = \left(X_{ip_{n1}+1}, \ldots, X_{ip_n}, X_{ip_{0n}+1} \frac{U_i - U_k}{h}, \ldots \right).$$

As shown by Fan and Li (2001), the good property of the nonconcave penalized estimate is that the estimate has the so called "oracle property." This means that under generalized regular conditions, the nonconcave penalized method can estimate the model just as the nonsignificant variables in the model are known advance. Hence, we expect that our proposed estimation method for the varying coefficient model with the dimension diverging with the sample size also has such an oracle property. The oracle estimator for the varying coefficient model shown above is defined by $\hat{\boldsymbol{\beta}}_k^o = (\boldsymbol{\beta}_k^{(1)o}, \mathbf{0}^{(2)}), k = 1, \ldots, n$, where $\boldsymbol{\beta}_k^{(1)o}$ is the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} \left[y_j - X_{ik}^{(1)} \boldsymbol{\beta}_k^{(1)} \right]^2 K_h(U_i - U_k).$$

To prove the oracle estimate is also a minimizer of (2.2), or even the asymptotical global minimum of (2.2), the following regularity conditions are needed.

A1. For $k = 1, \ldots, n, j = 1, \ldots, p_n$, $|X_{kj}|$ is bounded by a constant M_1.

A2. The density function of U, $f(u)$, is continuous and positive on the interval $[a, b]$. The second derivative of $f(u)$ is continuous and bounded.

A3. For any u, the matrix $\Omega(u) = \mathrm{E}(\mathbf{X}^T\mathbf{X}|U = u)$ is non-singular and its eigenvalues are bounded by the constants M_2 and M_2, and $r''_{ij}(u)$ is continuous for $i, j = 1, \ldots, p_n$ where $r''_{ij}(u) = \mathrm{E}(X_{ki}X_{kj}|U = u)$.

A4. The function $K(t)$ is a symmetric density function with a compact support.

A5. $a_j(\cdot), j = 1, \ldots, p_{0n}$ have a continuous four order derivative.

A6. $p_{n1} = O(n^{c_1})$ for some $O < c_1 < \frac{1}{2}$ and $p_n/n \to 0$.

A7. There exist positive constants c_2 and M_3 such that

$$n^{(1-c_2/)/2} \min_{j=1,\ldots,p_{n1}} \|a_j(u)\| > M_3, \quad \text{and} \quad n^{(1-c_2/)/2} \min_{j=1,\ldots,p_{n0}} \|ha'_j(u)\| > M_3$$

where $\|a_j(u)\| = (\mathrm{E}a_j^2(u))^{1/2}$.

A8. ε is independent of X, and $\mathrm{E}|\varepsilon|^n \leq \sigma^2 n! a^{n-2}/2$ to any $n \geq 2$, where $\sigma^2 = \mathrm{E}\varepsilon^2$.

Remark: Condition (A1) is a reasonable condition and can be easily satisfied by most applications. Conditions (A2)–(A5) are often used in the theoretical investigation of many statistical methods applied in varying coefficient models. Those conditions will be useful for studying the uniform properties of the proposed estimate. Condition (A6) is much generalized, compared to the results of Fan and Peng (2004) and Wei, Huang and Li (2011). This means that under other regular conditions and if the total number of predictor variables increases more slowly than the sample size, as long as the number of significant variables increases no faster than $\sqrt{n}$, our proposed estimate method will be able to estimate the varying coefficient function model efficiently and simultaneously select the model consistently. Condition (A7) is to show how strong a signal our proposed regularization method can be detected. Condition (A8) is

imposed on the error in the model. It requires to have an the error has approximated exponential tail. This is a generalized condition for the theoretical investigation of high dimensional statistical modeling methods.

Theorem 2.1 *(Sparsity). Let $\hat{\boldsymbol{\beta}}$ be the global minimum of (2.2) with the SCAD penalty and a regularization parameter λ_n. Then under the regular conditions A1—A8, we have*

$$\lim_{n\to\infty} \mathrm{P}(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^o) = 1$$

provided that $(h^4+h^2\sqrt{\log n/(nh)})p_{1n}n^{(1-c_2)/2} \to 0$, $\lambda_n = o(n^{-(1-(c_2-c_1))/2})$, $\lambda_n/(h^2 p_{1n}) \to \infty$, *and* $\lambda_n\sqrt{nh/\log n} \to \infty$ *as* $n \to \infty$.

Remark: The conditions in the above theorem concerns the bandwidth and the tuning parameter are not very strong. When p_{n1}, the number of the significant variables in the model is a constant, and the signals of those significant varying coefficient functions are strong, i.e. $c_1 = 0$ and $c_2 = 1$, the conditions for h and the tuning parameter λ_n are easily satisfied, and then with a probability tending to one, the oracle estimate should be the global minimum of (2.2). Especially, under such a situation, if we let $h = O(n^{-1/5})$ and $\lambda_n = o(\sqrt{\log n/nh})$, the estimates of nonconstant varying coefficient functions have an optimal convergence rate. Conversely, when the number of the significant variables in the model is also diverging with the sample size, and the signals of the significant variables are not too strong, by appropriately selecting λ_n and a small bandwidth h, the conditions of theorem 1 can be still satisfied. Hence, it is possible to select the significant variables and identify the constant coefficients consistently with a probability tending to one. For example, if $p_{1n} = O(n^{1/4})$, i.e. $c_1 = 1/4$, and $c_2 = 2/3$, then we just let $h = n^{-1/3}$ and $\lambda_n = n^{-7/24}/\log n$, and the conditions of Theorem 1 are satisfied. The proposed method can select the significant variables and determine the constant coefficients simultaneously and consistently.

Let $\hat{C}_j^o$ be the estimator of C_j obtained by assuming that we know $a_j(\cdot) = 0$, $j = p_{1n}+1, \cdots, p_n$, in model (2.1), and by applying the above penalized estimation procedure with $a_j(\cdot)$, $j = p_{1n}+1, \cdots, p_n$, is replaced by 0 and no penalty is imposed on $\boldsymbol{\alpha}_j$, $j = 1, \cdots, p_n$, and then we define $\hat{\mathbf{C}}^o$ as

$$\hat{\mathbf{C}}^o = \left(\hat{C}^o_{p_{0n}+1}, \cdots, \hat{C}^o_{p_{1n}}\right)^{\mathrm{T}}.$$

Let $\hat{a}_j^o(u_0)$ be the estimator of $a_j(u_0)$ obtained by the standard estimation for varying coefficient models (see Fan and Zhang, 1999). Under the assumption that we know $a_j(\cdot) = 0,\ j = p_{1n}+1,\ \cdots,\ p_n$, and $a_j(\cdot) = C_j,\ j = p_{0n}+1,\ \cdots,\ p_{1n}$, are constant in model (2.1), we also know the true value of C_j. Then we have the following two theorems about the oracle estimates of $\hat{a}_j^o(U)$ and $\hat{C}_j^o$.

Theorem 2.2 *(Asymptotic properties of $\hat{\alpha}_j^o(U)$). Under under the regular conditions A1—A8, when $h^2 p_{1n} \to 0$, $nh^9 p_{1n} \to 0$ and $p_{1n}\sqrt{\log n/nh} \to 0$ when $n \to \infty$, for any p_{0n} by 1 dimensional vector A_n with $\|A_n\| = 1$ we have*

$$\begin{aligned} & \sqrt{nh} A_n^T [\{\Omega^{(1)}(u)\}_{11}^{-1}]^{-1/2} \left\{ \hat{a}_1^o(u) - a_1(u) - \frac{1}{2}h^2\mu_2 a_1^{''}(u), \ldots, \right. \\ & \left. \hat{a}_{p0n}^o(u) - a_{p0n}(u) - \frac{1}{2}h^2\mu_2 a_{p0n}^{''}(u) \right\}^T \\ \sim\ & \mathcal{N}\left(0, \frac{\sigma^2\nu_0}{f(u)}\right) \end{aligned}$$

as $n \to \infty$, where $\{\Omega^{(1)}(U)\}_{11}^{-1}$ is the first p_{0n} by p_{0n} prime submatrix of $\{\Omega^{(1)}(U)\}^{-1}$.

Theorem 2.3 *(Asymptotic properties of $\hat{C}_j^o$). Under the regular conditions A1—A8, when $h^2 p_{1n} \to 0$, $nh^8 p_{1n} \to 0$ and $p_{1n}\sqrt{\log n/nh} \to 0$ when $n \to \infty$, then for any $p_{1n} - p_{0n}$ by 1 dimensional vector B_n with $\|B_n\| = 1$, we have*

$$\sqrt{n} B_n^T [\{\Omega^{(1)}(U)\}_{22}^{-1}]^{-1/2} (\hat{\mathbf{C}}^o - \mathbf{C}) \sim \mathcal{N}\left(0, \sigma^2\right)$$

as $n \to \infty$, where $\{\Omega^{(1)}(U)\}_{22}^{-1}$ is the last $p_{1n} - p_{0n}$ by $p_{1n} - p_{0n}$ prime submatrix of $\{\Omega^{(1)}(U)\}^{-1}$.

Remark: From the remark about Theorem 1, we know that the regular conditions in these two theorems can be easily satisfied. Hence, from these two theorems we know that in a generalized situation our proposed estimates have good asymptotic properties, just as the nonsignificant variables and constant coefficients in the model are known in advance when those regular conditions are satisfied.

2.4 Practical Issues

To apply the proposed method in practical applications, there are two important issues need to consider. The first one is how to find the solution for the optimal problem (2.2), and the second one is how to select an appropriate tuning parameter in (2.2).

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $X_j = (X_{1j}, X_{2j}, \ldots, X_{nj})^T$ and

$$\mathbf{X}_k = (X_1, \ldots, X_{p_n}, \mathbf{U}_k X_1, \ldots, \mathbf{U}_k X_{pn}),$$

where $\mathbf{U}_k = \text{diag}\left\{\frac{U_1 - U_k}{h}, \ldots, \frac{U_n - U_k}{h}\right\}$.

Although many efficient algorithms as shown in Chapter 1 have been proposed to solve the nonconcave penalized problem, most are only useful for high-dimensional linear regression models. As shown in Chapter 4, the quasi-Newton-Raphson algorithm and quadratic approximation proposed by Fan and Li (2001) is an efficient algorithm for many statistical models as long as those models have enough smoothing parameters in the models. Hence, given an initial value of $\mathbf{a}_k = \{a^o_{1k}, \ldots, a^o_{p_n k}\}^T$, and $\mathbf{b}_k = \{b^o_{1k}, \ldots, b^o_{p_n k}\}^T$ and $k = 1, \ldots, n$. we follow Fan and Li (2001) and update our estimates for those varying coefficient functions. First we use quadratic approximation to approximate the penalty function $p_{\lambda_n}(\|\mathbf{a}_j\|)$ and $p_{\lambda_n}(\|\mathbf{b}_j\|)$ by

$$p_{\lambda_n}(\|\mathbf{a}_j\|) = p_{\lambda_n}(\|\mathbf{a}^o_j\|) + p'_{\lambda_n}(\|\mathbf{a}^o_j\|) \cdot \frac{\|\mathbf{a}_j\|^2 - \|\mathbf{a}^o_j\|^2}{2\|\mathbf{a}^o_j\|}$$

and

$$p_{\lambda_n}(\|\mathbf{b}_j\|) = p_{\lambda_n}(\|\mathbf{b}^o_j\|) + p'_{\lambda_n}(\|\mathbf{b}^o_j\|) \cdot \frac{\|\mathbf{b}_j\|^2 - \|\mathbf{b}^o_j\|^2}{2\|\mathbf{b}^o_j\|}.$$

Next, following the procedure of the Newton Raphson algorithm, the estimates of $\mathbf{a}_k$ and $\mathbf{b}_k$ are updated by the following equation.

$$\begin{aligned}(\mathbf{a}_k^T, \mathbf{b}_k^T)^T = \quad & (\mathbf{X}_k^T \mathbf{W}_k \mathbf{X}_k)^{-1} \Big\{ \mathbf{X}_k \mathbf{W}_k \mathbf{Y} \\ + \quad & n \cdot \text{diag}\left(p'_{\lambda_n}(\|\mathbf{a}^o_1\|) \frac{a^o_{1k}}{\|\mathbf{a}^o_1\|}, \ldots, p'_{\lambda_n}(\|\mathbf{a}_{p_n}\|) \frac{a^o_{p_n k}}{\|\mathbf{a}^o_{p_n}\|}, p'_{\lambda_n}(\|\mathbf{b}^o_1\|) \frac{b_{1k}}{\|\mathbf{b}^o_1\|}, \ldots, \right. \\ & \left. p'_{\lambda_n}(\|\mathbf{b}^o_{p_n}\|) \frac{a_{p_n k}}{\|\mathbf{b}^o_{p_n}\|} \right) \Big\}.\end{aligned}$$

Note that when the dimension of the model is high, even larger than the sample size, calculating the inverse matrix of $\mathbf{X}_k^T \mathbf{W}_k \mathbf{X}_k$ would be quite difficult. The revised

Newton Raphson algorithm shown in Chapter 1 can be applied here to simplify some of the calculation. The other problem is how to select an initial estimate when the dimension of the model is no larger than the sample size. Empirically the weighted least squares estimate would be a good initial estimate for the above algorithm. If the dimension of the model is larger than the sample size, to improve the stability of the model selection, the sure screening step will be needed as shown in Chapter 4.

Fan and Li (2001) suggested using the GCV criteria to select the tuning parameter for nonconcave penalized least squares under a linear regression model. Wang, Li, and Tsai (2007) showed that the GCV criteria, which is similar to the AIC criteria, would not select a consistent tuning parameter. Hence they suggested using a BIC criteria to select the tuning parameter for nonconcave penalized methods with a fixed dimension. Chen and Chen (2008) re-examined the BIC criteria in model selection when the dimension of the model is quite large and suggested an extended BIC for the model selection with large model spaces. Following the idea of the BIC and extended BIC, we use the following modified BIC criteria to select the tuning parameter λ_n when the dimension of the model is diverging with the sample size n.

First, given λ_n, we can get the estimates of $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$. The residual can be calculated as

$$\mathrm{SSE}_{\lambda_n} = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p_n} X_{ij} \hat{a}_j(U_i) \right)^2 .$$

Based on SSE_{λ_n}, the BIC value for λ_n is defined as

$$\mathrm{BIC}_{\lambda_n} = \log \mathrm{SSE}_{\lambda_n} + \gamma \mathrm{DF}_{\lambda_n} \cdot \frac{\log n}{n}$$

where DF_{λ_n} is the approximated degrees of freedom. To simplify our numerical studies, we used the number of nonzero components of $\hat{\mathbf{a}}$. More accurate approximation for the degrees of freedom for the model estimate can be calculated based on the idea of generalized degrees of freedom as shown in Fan and Li (2001). However, as shown by our numerical studies in Chapter 5, our approximated degrees of freedom is reasonable good when the signals of the significant predictor variables are quite strong. γ in the above equation is the so-called "inflated factor," to take into account both the number of unknown varying coefficients and the complexity of the model space. Based on the experience of Cui, *et al.* (2013) and the discussions of Luo and

Wahba (1997) and Friedman and Silverman (1989), we suggest selecting γ within the interval $(1.2, 3)$. In our numerical study, the inflated factor is taken as 1.5.

2.5 Appendix: Proofs

Lemma 2.1 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random vectors, where Y_i's are scalar variables. Assume further that Y_i is bounded by a constant M, or independent of X_i and satisfies $\mathrm{E}Y_i = 0$ and $\mathrm{E}|Y_i|^k \leq c^{k-2}k!EY_i^2$, $i = 1, \ldots, n$ and where c is a constant. Let K be a bounded positive function with a bounded support, satisfying a Lipschitz condition. When n is large enough, then as $nh \to \infty$, we have*

$$\mathrm{P}\left(\sup_{x\in[a,b]}\left|n^{-1}\sum_{i=1}^{n}\{K_h(X_i-x)Y_i - \mathrm{E}[K_h(X_i-x)Y_i]\}\right| > C\sqrt{\frac{\log n}{nh}}\right) < 2\exp(-C_2\log n)$$

where C_2 is a positive constant that depends on C, M and the Lipschitz constant. If C is large enough, C_2 can be larger than a given positive value.

Proof: Let $Q_h(x) = n^{-1}\sum_{i=1}^{n} K_h(X_i - x)Y_i$. Partition the interval $[a, b]$ into $N = [n/h]^{\frac{1}{2}}$ subintervals I_j of equal length. Let x_j be the center of I_j. Then

$$\sup_{x\in[a,b]}|Q_h(x) - \mathrm{E}Q_h(x)| \leq \max_{1\leq j\leq N}|Q_h(x_j) - \mathrm{E}Q_h(x_j)| + C_1(nh)^{-1/2}$$

where C_1 is related to the length of $[a, b]$ and the Lipschitz condition. By the inequality above, then

$$\begin{aligned}
&\mathrm{P}\left(\sup_{x\in[a,b]}\left|n^{-1}\sum_{i=1}^{n}\{K_h(X_i-x)Y_i - \mathrm{E}[K_h(X_i-x)Y_i]\}\right| > C\sqrt{\frac{\log n}{nh}}\right)\\
\leq\ &\mathrm{P}\left(\max_{1\leq j\leq N}|Q_h(x_j) - \mathrm{E}Q_h(x_j)| + C_1(nh)^{-1/2} > C\sqrt{\frac{\log n}{nh}}\right)\\
\leq\ &\mathrm{P}\left(\max_{1\leq j\leq N}|Q_h(x_j) - \mathrm{E}Q_h(x_j)| > (C - C_1)\sqrt{\frac{\log n}{nh}}\right).
\end{aligned}$$

First, we consider that X_i and Y_i are independent, and Y_i satisfies $\mathrm{E}(Y_i) = 0$ and $\mathrm{E}|Y_i|^k \leq c^{k-2}k!EY_i^2$, $i = 1, \ldots, n$. It is easy to see that $\mathrm{E}(K_h(X_i - x)Y_i)^2 = \frac{\sigma_Y^2}{h}$. Note that there exists a constant L such that $K_h(\cdot) < L/h$, so

$$\mathrm{E}|K_h(X_i - x)Y_i|^k \leq \frac{L^{k-2}}{h^{k-2}}c^{k-2}k!\frac{\sigma_Y^2}{h}.$$

Define $c^* = \frac{cL}{h}$ and note $\mathrm{E}(K_h(X_i - x)Y_i)^2 = \sigma_Y^2/h$, by Bernstein's inequality (see Bosq, 1998), then we have

$$\begin{aligned}
& \mathrm{P}\left(\left|n^{-1}\sum_{i=1}^{n}\{K_h(X_i - x)Y_i - \mathrm{E}[K_h(X_i - x)Y_i]\}\right| > (C - C_1)\sqrt{\frac{\log n}{nh}}\right) \\
\leq\ & 2\exp\left(-\frac{(C - C_1)^2\sqrt{n}\log n/h}{2\sqrt{n}\mathrm{E}(K_h(X_i - x)Y_i)^2 + 2(C - C_1)cL/h\sqrt{\log n/h}}\right) \\
\leq\ & 2\exp\left(-\frac{(C - C_1)^2 n\log n/h}{4n\sigma_Y^2/h}\right) \\
=\ & 2\exp(-C^*\log n).
\end{aligned}$$

Because $nh \to \infty$, we get

$$\begin{aligned}
& \mathrm{P}\left(\max_{1\leq j\leq N}|Q_h(x_j) - \mathrm{E}Q_h(x_j)| > (C - C_1)\sqrt{\frac{\log n}{nh}}\right) \\
\leq\ & \sqrt{n/h}\cdot 2\exp(-C^*\log n) \leq 2\exp(-(C^* - 1)\log n) = \exp(-C_2\log n).
\end{aligned}$$

For any given positive N, as C is large enough, we always have $C_2 > N$.

Next, we consider the situation that Y_i is bounded. By Bernstein's inequality(see Bosq, 1998), we have

$$\begin{aligned}
& \mathrm{P}\left(\left|n^{-1}\sum_{i=1}^{n}\{K_h(X_i - x)Y_i - \mathrm{E}[K_h(X_i - x)Y_i]\}\right| > (C - C_1)\sqrt{\frac{\log n}{nh}}\right) \\
\leq\ & 2\exp\left(-\frac{(C - C_1)^2 n\log n/h}{4n\mathrm{E}(K_h(X_i - x)Y_i)^2 + 2MC\sqrt{n\log n/h}}\right) \\
\leq\ & 2\exp\left(-\frac{(C - C_1)^2 n\log n/h}{4nM^2/h + 2MC\sqrt{n\log n/h}}\right) \\
=\ & 2\exp(-C^*\log n).
\end{aligned}$$

Because $nh \to \infty$, we get

$$\begin{aligned}
& \mathrm{P}\left(\max_{1\leq j\leq N}|Q_h(x_j) - \mathrm{E}Q_h(x_j)| > (C - C_1)\sqrt{\frac{\log n}{nh}}\right) \\
\leq\ & \sqrt{n/h}\cdot 2\exp(-C^*\log n) \leq 2\exp(-(C^* - 1)\log n).
\end{aligned}$$

Then, in a similar way, $C_2 = C^* - 1$, for any given positive N, and as C is large enough, we always have $C_2 > N$. □

Proof of Theorem 2.1

For notational simplicity, we drop the subscript n. Let $\mathcal{P} = (1, 2, \ldots, p_{n0+n1})$ and $\mathcal{N} = (p_{n0+n1+1}, \ldots, p_{2n})$. Define

$$\mathbf{W}_k^{\frac{1}{2}} = \text{diag}\{K_h^{\frac{1}{2}}(U_1 - U_k), \ldots, K_h^{\frac{1}{2}}(U_n - U_k)\},$$

$$\mathbf{Y}_k = \mathbf{W}_k^{\frac{1}{2}} \cdot (Y_1, \ldots, Y_n)^T,$$

and

$$\mathbf{X}_k^{(1)} = \mathbf{W}_k^{\frac{1}{2}} \cdot (X_{1k}^{(1)T}, \ldots X_{nk}^{(1)T})^T, \quad \mathbf{X}_k^{(2)} = \mathbf{W}_k^{\frac{1}{2}} \cdot (X_{1k}^{(2)T}, \ldots X_{nk}^{(2)T})^T.$$

Hence (2.2) can be written as

$$C(\boldsymbol{\beta}) = \sum_{k=1}^{n} \|\mathbf{Y}_k - \mathbf{X}_k^{(1)}\boldsymbol{\beta}_k^{(1)} - \mathbf{X}_k^{(2)}\boldsymbol{\beta}_k^{(2)}\|^2 + n\sum_{j=1}^{2p_n} p_{\lambda_n}(\|\boldsymbol{\beta}_j\|).$$

Let $\hat{\mathbf{X}}_k^{(2)}$ be the columnwise projection of the jth column of $\mathbf{X}_k^{(2)}$ onto the linear space spanned by the column vectors of $\mathbf{X}_k^{(1)}$. Let $\tilde{\mathbf{X}}_k^{(2)} = \mathbf{X}_k^{(2)} - \hat{\mathbf{X}}_k^{(2)}$. By condition A.3, it is easy to show that $\tilde{\mathbf{C}}_k^{(2,2)} = \tilde{\mathbf{X}}_k^{(2)T}\tilde{\mathbf{X}}_k^{(2)}/n$ also has a small eigenvalue larger than a certain number $\tilde{\rho}$.

Let $\hat{\mathbf{Y}}_k = \mathbf{X}_k^{(1)}\hat{\boldsymbol{\beta}}^{(1)o}, k = 1, \ldots, n$, then for a given $\boldsymbol{\beta} \in R^{2p_n \times n}$ the equation above can be written as

$$\begin{aligned} C(\boldsymbol{\beta}) &= \sum_{k=1}^{n} \|\hat{\mathbf{Y}}_k - \mathbf{X}_k^{(1)}\boldsymbol{\beta}_k^{(1)} - \hat{\mathbf{X}}_k^{(2)}\boldsymbol{\beta}_k^{(2)}\|^2 \\ &\quad + \sum_{k=1}^{n} \|\mathbf{Y}_k - \hat{\mathbf{Y}}_k - \tilde{\mathbf{X}}_k^{(2)}\boldsymbol{\beta}_k^{(2)}\|^2 + n\sum_{j=1}^{2p_n} p_{\lambda_n}(\|\boldsymbol{\beta}_j\|) \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} \sum_{k=1}^{n} \|\mathbf{Y}_k - \hat{\mathbf{Y}}_k - \tilde{\mathbf{X}}_k^{(2)}\boldsymbol{\beta}_k^{(2)}\|^2 &= \sum_{k=1}^{n} \|\mathbf{Y}_k - \hat{\mathbf{Y}}_k\|^2 \\ &\quad + \sum_{k=1}^{n} \boldsymbol{\beta}_k^{(2)T}\tilde{\mathbf{C}}_k^{(2,2)}\boldsymbol{\beta}_k^{(2)} \\ &\quad - \sum_{k=1}^{n} 2\sum_{j\in\mathcal{N}} \beta_{jk} < \mathbf{Y}_k - \hat{\mathbf{Y}}_k, \tilde{X}_k^{(j)} > . \end{aligned}$$

To prove the theorem, it is sufficient to show that

$$\text{P}(C(\boldsymbol{\beta}) \geq C(\hat{\boldsymbol{\beta}}^o) \text{ for all } \boldsymbol{\beta} \in R^{2p_n \times n}) \to 1,$$

as $n \to \infty$.

For a given $\boldsymbol{\beta} \in R^{2p_n \times n}$, let $\mathcal{P}^+ = \{j \in \mathcal{P} : \|\boldsymbol{\beta}_j\| > a\lambda_n\}$, and $\mathcal{N}^+ = \{j \in \mathcal{N} : \|\boldsymbol{\beta}_j\| > \lambda_n\}$, and let $\mathcal{P}^- = \mathcal{P} - \mathcal{P}^+$ and $\mathcal{N}^- = \mathcal{N} - \mathcal{N}^+$. Consider the linear space $\mathcal{A}_k$ spanned by $\mathbf{X}_k^{\mathcal{A}} = \{\hat{X}_k^{(j)}, j \in \mathcal{P}^+ \cup \mathcal{N}\}$ where we let $\hat{X}_k^{(j)} = X_k^{(j)}$ for $j \in \mathcal{P}$, and $\boldsymbol{\beta}^{\mathcal{A}} = \{\boldsymbol{\beta}_j, j \in \mathcal{P}^+ \cup \mathcal{N}\}$. Similarly, we let $\mathbf{X}_k^{\mathcal{A}^c} = \{\hat{X}_k^{(j)}, j \in \mathcal{P}^-\}$ and $\boldsymbol{\beta}^{\mathcal{A}^c} = \{\boldsymbol{\beta}_j, j \in \mathcal{P}^-\}$.

First, by Condition A7 and Lemma 2.2, note that we have

$$\min_{j \in \mathcal{P}} \|\hat{\boldsymbol{\beta}}_j^o\| = O_p^+(n^{-(1-c_2)/2}) \geq a\lambda_n, \tag{A.2}$$

where $O_p^+(n^a)$ is defined as a sequence of positive random variables such that there exists a positive number τ with $\mathrm{P}(n^{-a}O_p^+(n^a) \geq \tau) \to 1$ as $n \to \infty$. Then let $\hat{\mathbf{Y}}_k^{\mathcal{A}}$ be the projection of $\hat{\mathbf{Y}}$ onto $\mathcal{A}$ and it is easy to show that

$$\sum_{k=1}^{n} \|\hat{\mathbf{Y}}_k - \hat{\mathbf{Y}}_k^{\mathcal{A}}\|^2 = nrO_p^+(n^{-1+c_2}), \tag{A.3}$$

where r is the cardinality of $\mathcal{P}^-$.

By Lemma 2.3, under the regular conditions, for any $\boldsymbol{\beta} \in R^{2p_n \times n}$ and $j \in \mathcal{N}$, we have

$$\sum_{k=1}^{n} \beta_{kj} < \mathbf{Y}_k - \hat{\mathbf{Y}}_k, \tilde{X}_k^{(j)} >= o_p(n\lambda_n)\|\boldsymbol{\beta}_j\|. \tag{A.4}$$

Next, it is easy to show that

$$\sum_{k=1}^{n} \boldsymbol{\beta}_k^{(2)'} \tilde{\mathbf{C}}_k^{(2,2)} \boldsymbol{\beta}_k^{(2)} \geq \tilde{\rho} \sum_{k=1}^{n} \sum_{j \in \mathcal{N}} \beta_{kj}^2 \geq \tilde{\rho} \sum_{j \in \mathcal{N}^+} \sum_{k=1}^{n} \beta_{kj}^2 \geq n\tilde{\rho}\lambda_n \sum_{j \in \mathcal{N}^+} \|\boldsymbol{\beta}_j\| \tag{A.5}$$

and

$$\begin{aligned} &\mathrm{P}\left(\sum_{j=1}^{p_n} \left(p_{\lambda_n}(\|\boldsymbol{\beta}_j\|) - p_{\lambda_n}(\|\hat{\boldsymbol{\beta}}_j^o\|) \right) \geq \lambda_n \sum_{j \in \mathcal{N}^-} \|\boldsymbol{\beta}_j\| - rO(\lambda_n^2), \forall\ \boldsymbol{\beta} \in R^{n \times p_n} \right) \\ \to\ & 1. \end{aligned} \tag{A.6}$$

Then (A.1)—(A.7) imply

$$\begin{aligned} C(\boldsymbol{\beta}) - C(\hat{\boldsymbol{\beta}}^o) \ =\ & \sum_{k=1}^{n} \|\hat{\mathbf{Y}}_k - \hat{\mathbf{Y}}_k^{\mathcal{A}} + \hat{\mathbf{Y}}_k^{\mathcal{A}} - \mathbf{X}_k^{\mathcal{A}}\boldsymbol{\beta}_k^{\mathcal{A}} - \mathbf{X}_k^{\mathcal{A}^c}\boldsymbol{\beta}_k^{\mathcal{A}^c}\|^2 + \sum_{k=1}^{n} \boldsymbol{\beta}_k^{(2)'} \tilde{C}_k^{(2,2)} \boldsymbol{\beta}_k^{(2)} \\ & -2\sum_{k=1}^{n} \sum_{j \in \mathcal{N}} \beta_{kj} < \mathbf{Y}_k - \hat{\mathbf{Y}}_k, \tilde{X}_k^{(j)} > \\ & +n \sum_{j=1}^{2p_n} \left\{ p_{\lambda_n}(\|\boldsymbol{\beta}_j\|) - p_{\lambda_n}(\|\hat{\boldsymbol{\beta}}_j^o\|) \right\}, \end{aligned}$$

and then that the probability of

$$C(\boldsymbol{\beta}) - C(\hat{\boldsymbol{\beta}}^o) \geq \sum_{k=1}^{n} \|\hat{\mathbf{Y}}_k - \hat{\mathbf{Y}}_k^{\mathcal{A}}\|^2 + \sum_{k=1}^{n} \|\hat{\mathbf{Y}}_k^{\mathcal{A}} - \mathbf{X}_k^{\mathcal{A}}\boldsymbol{\beta}_k^{\mathcal{A}}\|^2 - \sum_{k=1}^{n} \|\hat{\mathbf{X}}_k^{\mathcal{A}^c}\boldsymbol{\beta}_k^{\mathcal{A}^c}\|^2$$
$$+n\lambda_n \sum_{j\in\mathcal{N}^-} \|\boldsymbol{\beta}_j\| + n\tilde{\rho}\lambda_n \sum_{j\in\mathcal{N}^+} \|\boldsymbol{\beta}_j\| - o_p(n\lambda_n)\sum_{j\in\mathcal{N}} \|\boldsymbol{\beta}_j\| - nrO(\lambda_n^2)$$

for all $\boldsymbol{\beta} \in R^{2p_n\times n}$ converges to 1 as $n \to \infty$. Note that largest eigenvalue of $\mathbf{X}_k^{\mathcal{A}^c T}\mathbf{X}_k^{\mathcal{A}^c}/n$ is bounded by a certain positive constant M_4 because $\mathbf{C}_k$ is bounded. So we have

$$\sum_{k=1}^{n} \|\hat{\mathbf{X}}_k^{\mathcal{A}^c}\boldsymbol{\beta}_k^{\mathcal{A}^c}\|^2 \leq \sum_{k=1}^{n} M_4 \sum_{j\in\mathcal{P}^-} \beta_{jk}^2 \leq M_4 nra^2\lambda_n^2.$$

Thus the probability of

$$C(\boldsymbol{\beta}) - C(\hat{\boldsymbol{\beta}}^o) \geq nr\left(O_p^+(n^{-1+c_2}) - O(\lambda_n^2)\right) + n(\min(1,\tilde{\rho})\lambda_n - o_p(\lambda_n))\sum_{j\in\mathcal{N}} \|\boldsymbol{\beta}_j\|$$

for all $\boldsymbol{\beta} \in R^{2p_n\times n}$ converges to 1. Because $O(\lambda_n^2) = o(n^{-1+c_2})$, the left side of the equation above becomes nonnegative in probability, and the proof is complete. □

Lemma 2.2 *Under the regular conditions for Theorem 2.1, we have*

$$\min_{j\in\mathcal{P}} \|\hat{\boldsymbol{\beta}}_j^o\| = O_p^+(n^{-(1-c_2)/2}).$$

Proof: Following the proof steps of Theorem 2.2, and by Lemma 2.1, we know that

$$\sup_{j\in\mathcal{P}} \sup_{u\in[a,b]} |\hat{\beta}_j(u) - \beta_j(u)| = O_p((h^4 + h^2\sqrt{\log n/(nh)})(p_{1n} + p_{0n})).$$

Hence by the condition A.7 and the following proof of Theorem 2.2 , we have

$$\|\hat{\boldsymbol{\beta}}_j^o\| \geq \|\boldsymbol{\beta}_j\| - O_p((h^4 + h^2\sqrt{\log n/(nh)})(p_{1n} + p_{0n})) = O_p^+(n^{-(1-c_2)/2}).$$

□

Lemma 2.3 *Under the regular conditions for Theorem 2.1, for any $\boldsymbol{\beta} \in R^{2p_n\times n}$, we have*

$$\max_{j\in\mathcal{N}} \sum_{k=1}^{n} \beta_{kj} < \mathbf{Y}_k - \hat{\mathbf{Y}}_k, \tilde{X}_k^{(j)} >= o_p(n\lambda_n)\|\boldsymbol{\beta}_j\|.$$

Proof: Because

$$\sum_{k=1}^{n}\beta_{kj}<\mathbf{Y}_k-\hat{\mathbf{Y}}_k,\tilde{X}_k^{(j)}>\leq n\|\boldsymbol{\beta}_j\|\left(\frac{1}{n}\sum_{k=1}^{n}<\mathbf{Y}_k-\hat{\mathbf{Y}}_k,\tilde{X}_k^{(j)}>^2\right)^{\frac{1}{2}},$$

we only need to show that

$$\left(\frac{1}{n}\sum_{k=1}^{n}<\mathbf{Y}_k-\hat{\mathbf{Y}}_k,\tilde{X}_k^{(j)}>^2\right)^{\frac{1}{2}}=o_p(\lambda_n).$$

Similar to the proof of Theorem 2.2,

$$\begin{aligned}
<\mathbf{Y}_k-\hat{\mathbf{Y}}_k,\tilde{X}_k^{(j)}> &= \frac{1}{n}\tilde{X}_k^{(j)T}\{\mathbf{I}-\mathbf{X}_k(\mathbf{X}_k^T\mathbf{X}_k)^{-1}\mathbf{X}_k^T\}\mathbf{Y}_k \\
&= \frac{1}{n}\tilde{X}_k^{(j)T}\mathbf{W}_k^{\frac{1}{2}}\begin{pmatrix}\frac{1}{2}\sum_{j=1}^{p_{0n}}X_{1j}a_j''(\xi_1)(U_1-U_k)^2\\ \vdots \\ \frac{1}{2}\sum_{j=1}^{p_{0n}}X_{nj}a_j''(\xi_n)(U_n-U_k)^2\end{pmatrix} \\
&\quad +\frac{1}{n}\tilde{X}_k^{(j)T}\{\mathbf{I}-\mathbf{X}_k(\mathbf{X}_k^T\mathbf{X}_k)^{-1}\mathbf{X}_k^T\}\boldsymbol{\varepsilon}_k \\
&\triangleq I_1+I_2.
\end{aligned}$$

By Lemma 2.1, when n is large enough, uniformly for j,k and U_k we have

$$I_1=O_p(h^2p_{0n})\quad I_2=O_p\left(h^2+\sqrt{\frac{\log n}{nh}}\right).$$

Hence

$$\begin{aligned}
\sum_{k=1}^{n}\beta_{kj}<\mathbf{Y}_k-\hat{\mathbf{Y}}_k,\tilde{X}_k^{(j)}> &\leq n\|\boldsymbol{\beta}_j\|\left(\frac{1}{n}\sum_{k=1}^{n}\left(\frac{1}{n}<\mathbf{Y}_k-\hat{\mathbf{Y}}_k,\tilde{X}_k^{(j)}>^2\right)^{\frac{1}{2}}\right) \\
&= n\|\boldsymbol{\beta}_j\|O_p\left(h^4p_{0n}^2+\frac{\log n}{nh}\right)^{\frac{1}{2}} \\
&= o_p(n\lambda_n\|\boldsymbol{\beta}_j\|).
\end{aligned}$$

□

Proof of Theorem 2.2

For u_0, define

$$\mathbf{W}_0=\mathrm{diag}\{K_h(U_1-u_0),\ldots,K_h^{\frac{1}{2}}(U_n-u_0)\},$$

$$\mathbf{X}_0^{(1)}=(X_{1k}^{(1)T},\ldots X_{nk}^{(1)})^T,\quad ,\mathbf{Y}=(Y_1,\ldots,Y_n)^T,$$

and

$$\boldsymbol{\beta}_0^{(1)} = (a_1(u_0), \ldots, a_{p_{0n}}(u_0), C_1, \ldots, C_{p_{1n}-p_{0n}}, a_1'(u_0), \ldots, a_{p_{0n}}'(u_0))^T.$$

Then the oracle estimate of $\boldsymbol{\beta}_0^{(1)}$ can be expressed as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_0^{(1)} &= (\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)})^{-1}\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{Y} \\
&= (\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)})^{-1}\mathbf{X}_0^{(1)T}\mathbf{W}_0 \begin{pmatrix} \sum_{j=1}^{p_{0n}} X_{1j}a_j(U_1) + \sum_{j=p_{0n}}^{p_{1n}} X_{1j}C_j \\ \vdots \\ \sum_{j=1}^{p_{0n}} X_{nj}a_j(U_n) + \sum_{j=p_{0n}}^{p_{1n}} X_{nj}C_j \end{pmatrix} \\
&\quad +(\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)})^{-1}\mathbf{X}_0^{(1)T}\mathbf{W}_0\boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}_0^{(1)} + (\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)})^{-1}\mathbf{X}_0^{(1)T}\mathbf{W}_0 \begin{pmatrix} \frac{1}{2}\sum_{j=1}^{p_{0n}} X_{1j}a_j''(\xi_1)(U_1-u_0)^2 \\ \vdots \\ \frac{1}{2}\sum_{j=1}^{p_{0n}} X_{nj}a_j''(\xi_n)(U_1-u_0)^2 \end{pmatrix} \\
&\quad +(\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)})^{-1}\mathbf{X}_0^{(1)T}\mathbf{W}_0\boldsymbol{\varepsilon} \\
&\triangleq \boldsymbol{\beta}_0^{(1)} + I_1 + I_2.
\end{aligned}$$

By Lemma 2.1, when n is large enough and because $p_{1n}/n \to 0$, we have

$$\begin{aligned}
&\sup_{i,j, \text{ and } u_0\in[a,b]} \left\| \frac{1}{n}\sum_{i=1}^{n}\{X_{ki}X_{kj}K_h(U_i-u_0) - \mathrm{E}X_{ki}X_{kj}K_h(U_i-u_0)\} \right\| \\
= \; & O_p\left(\sqrt{\frac{\log n}{nh}}\right).
\end{aligned}$$

Hence

$$\frac{1}{n}\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)} = \frac{1}{n}\mathrm{E}\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)} + R = A + V$$

where V denotes a matrix whose elements are uniformly bounded by $O_p\left(\sqrt{\frac{\log n}{nh}}\right)$. For A, by the basic calculation, we know than

$$A = \mathrm{diag}(\Omega^{(1)}, \mu_2\Omega^{(0)})f(u_0) + B$$

where B denotes a matrix whose elements are uniformly bounded by $O_p(h^2)$. By the equations above, we have

$$(\frac{1}{n}\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)})^{-1} = \mathrm{diag}(\Omega^{(1)}(u_0), \mu_2\Omega^{(0)}(u_0))^{-1}f^{-1}(u_0) + O_p\left(h^2 + \sqrt{\frac{\log n}{nh}}\right).$$

From the conditions on h and p_{1n} we know that the maximum eigenvalue of the second term on the right side is bounded by

$$O_p\left((p_{0n}+p_{1n})\left(h^2+\sqrt{\frac{\log n}{nh}}\right)\right)=o_p(1).$$

In a similar way and by Lemma 2.1,

$$\begin{aligned}
&\frac{1}{n}\mathbf{X}_0^{(1)T}\mathbf{W}_0\begin{pmatrix}\frac{1}{2}\sum_{j=1}^{p_{0n}}X_{1j}a_j''(\xi_1)(U_1-u_0)^2\\ \vdots\\ \frac{1}{2}\sum_{j=1}^{p_{0n}}X_{nj}a_j''(\xi_n)(U_1-u_0)^2\end{pmatrix}\\
=\;&\frac{1}{2}f(u_0)\mu_2h^2\mathrm{diag}(\Omega^{(1)}(u_0),\mathbf{0})\begin{pmatrix}a_1''(u_0)\\ \vdots\\ a_{p_{0n}}''(u_0)\\ \mathbf{0}\end{pmatrix}\\
&+O_p\left(h^4+h^2\sqrt{\frac{\log n}{nh}}\right)
\end{aligned}$$

where $O_p(\cdot)$ denotes a vector whose elements are uniformly bounded by the order of $o_p(h^4+h^2\sqrt{\frac{\log n}{nh}})$. Hence, by the equations above,

$$\begin{aligned}
&\mathrm{bias}(A_n^T(a_1(u_0),\ldots,a_{p_{0n}}(u_0))^T)\\
=\;&\left\{\frac{1}{2}\mu_2h^2[A_n^T,\mathbf{0}](a_1''(u_0),\ldots,a_{p_{0n}}''(u_0),\mathbf{0})^T+O_p\left(\left\{h^4+\sqrt{\frac{h^4\log n}{n}}\right\}\sqrt{p_{p_{0n}+p_{1n}}}\right)\right\}\\
=\;&[A_n^T,\mathbf{0}]\left\{\frac{1}{2}\mu_2h^2a_1''(u_0),\ldots,\frac{1}{2}\mu_2h^2a_1''(u_0),\mathbf{0}\right\}^T+o_p(1/\sqrt{nh}),
\end{aligned}$$

where $\mathbf{0}$ is a 1 by $p_{1n}-p_{0n}$ vector.

For I_2, it is easy to see that $\mathrm{E}I_2=0$ and

$$\mathrm{Var}([A_n^T,0]I_2)=[A_n^T,\mathbf{0}](\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)})^{-1}\mathbf{X}_0^{(1)T}\mathbf{W}_0^2\mathbf{X}_0^{(1)}(\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)})^{-1}[A_n^T,\mathbf{0}]^T$$

where $\mathbf{0}$ is a 1 by $p_{1n}+p_{0n}$ vector.

Following the same steps as before and by Lemma 1, we have

$$\begin{aligned}
\mathrm{Var}(\sqrt{n}[A_n^T, \mathbf{0}]^T I_2) &= [A_n^T, \mathbf{0}]\mathrm{diag}(\Omega^{(1)}(u_0), \mu_2\Omega^{(0)}(u_0))^{-1} f^{-1}(u_0) \\
&\quad \times \frac{\sigma^2}{h}\mathrm{diag}(\nu_0\Omega^{(1)}(u_0), \nu_2\Omega^{(0)}(u_0)) f(u_0) \\
&\quad \times \mathrm{diag}(\Omega^{(1)}(u_0), \mu_2\Omega^{(0)}(u_0))^{-1} f^{-1}(u_0)[A_n^T, \mathbf{0}]^T \\
&\quad + O_p\left(h^2 + \sqrt{\frac{\log n}{nh}}\right) \\
&= \frac{\sigma^2\nu_0}{hf(u_0)}[A_n^T, \mathbf{0}]\{\Omega^{(1)}(u_0)\}^{-1}[A_n^T, \mathbf{0}]^T + o(1) \\
&= \frac{\sigma^2\nu_0}{hf(u_0)} A_n^T \{\Omega^{(1)}(u_0)\}_{11}^{-1} A_n + o(1),
\end{aligned}$$

where $\mathbf{0}$ is a 1 by $p_{1n} - p_{0n}$ vector and $\{\Omega^{(1)}(u_0)\}_{11}^{-1}$ is the p_{0n} by p_{0n} first prime submatrix of $\{\Omega^{(1)}(u_0)\}^{-1}$

Finally, by checking the Lindeberg-Feller condition, we have

$$\begin{aligned}
&\sqrt{nh}[A, \mathbf{0}]_n^T \{\Omega^{(1)}(u)\}^{1/2} \Bigg\{ \hat{a}_1^o(u) \\
&\quad - a_1(u) - \frac{1}{2}h^2\mu_2 a_1''(u), \ldots, \hat{a}_{p0n}^o(u) - a_{p0n}(u) - \frac{1}{2}h^2\mu_2 a_{p0n}''(u) \Bigg\}^T \\
\sim\ & \mathcal{N}\left(0, \frac{\sigma^2\nu_0}{f(u)}\right)
\end{aligned}$$

as $n \to \infty$ and $1 \le j \le p_{0n}$, where $\mathbf{0}$ is a $p_{1n} - p_{0n}$ by 1 zero vector. Theorem 2.2 has been proved. □

Proof of Theorem 2.3

Following the proof of Theorem 2.2 and by Lemma 2.1, we know that

$$\begin{aligned}
&\sup_{u_0 \in [a,b]} \Big| [\mathbf{0}_{1\times p_{0n}}, B_n^T]^T \left\{ [\mathbf{0}_{1\times p_{0n}}, \hat{a}_{p_{n0}+1}(u_0), \ldots, \hat{a}_{p_{n1}(u_0)}]^T - [0_{1\times p_{0n}}, \mathbf{C}^T]^T \right\} \\
- \ & [\mathbf{0}_{1\times p_{0n}}, B_n^T, \mathbf{0}_{1\times p_{0n}}]^T (\mathbf{X}_0^{(1)T}\mathbf{W}_0\mathbf{X}_0^{(1)})^{-1}\mathbf{X}_0^{(1)T}\mathbf{W}_0\boldsymbol{\varepsilon} \Big| \\
=\ & O_p(\{h^4 + h^2(nh)^{-1/2}\log^{\frac{1}{2}} n\}\sqrt{p_{1n}}).
\end{aligned}$$

Then

$$\begin{aligned}
&\sqrt{n}\left|\frac{B_n}{n}\sum_{i=1}^{n}(\hat{a}_{p_{0n}+1}(U_i),\ldots,\hat{a}_{p_{1n}})^T-\mathbf{C}\right.\\
&\left.-\frac{[\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}]^T}{n}\sum_{i=1}^{n}(\mathbf{X}_i^{(1)T}\mathbf{W}_i\mathbf{X}_i^{(1)})^{-1}\mathbf{X}_i^{(1)T}\mathbf{W}_i\boldsymbol{\varepsilon}\right|\\
=\;&\sqrt{n}\cdot O_p(\{h^4+h^2(nh)^{-1/2}\log^{1/2}n\}\sqrt{p_{1n}})\\
=\;&o_p(1).
\end{aligned}$$

Therefore, $1/\sqrt{n}B_n^T(\sum_{i=1}^{n}(\hat{a}_{p_{0n}+1}(U_i),\ldots,\hat{a}_{p_{1n}}(U_i))^T-\mathbf{C})=\sqrt{n}B_n(\hat{\mathbf{C}}-\mathbf{C})$ has the same distribution with

$$\frac{[\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}]^T}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{X}_i^{(1)T}\mathbf{W}_i\mathbf{X}_i^{(1)})^{-1}\mathbf{X}_i^{(1)T}\mathbf{W}_i\boldsymbol{\varepsilon}.$$

Hence its variance is

$$\begin{aligned}
&\mathrm{Var}\left(\frac{[\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}]}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{X}_i^{(1)T}\mathbf{W}_i\mathbf{X}_i^{(1)})^{-1}\mathbf{X}_i^{(1)T}\mathbf{W}_i\boldsymbol{\varepsilon}\right)\\
=\;&\frac{\sigma^2}{n}(1,\ldots,1)\begin{pmatrix}[\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}](\mathbf{X}_1^{(1)T}\mathbf{W}_1\mathbf{X}_1^{(1)})^{-1}\mathbf{X}_1^{(1)T}\mathbf{W}_1\\ \vdots\\ [\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}](\mathbf{X}_n^{(1)T}\mathbf{W}_n\mathbf{X}_n^{(1)})^{-1}\mathbf{X}_n^{(1)T}\mathbf{W}_n\end{pmatrix}\\
&\times\begin{pmatrix}[\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}](\mathbf{X}_1^{(1)T}\mathbf{W}_1\mathbf{X}_1^{(1)})^{-1}\mathbf{X}_1^{(1)T}\mathbf{W}_1\\ \vdots\\ [\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}](\mathbf{X}_n^{(1)T}\mathbf{W}_n\mathbf{X}_n^{(1)})^{-1}\mathbf{X}_n^{(1)T}\mathbf{W}_n\end{pmatrix}^T\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}\\
=\;&\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}[\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}](\mathbf{X}_i^{(1)T}\mathbf{W}_i\mathbf{X}_i^{(1)})^{-1}\mathbf{X}_i^{(1)T}\mathbf{W}_i\\
&\times\mathbf{W}_j\mathbf{X}_j^{(1)}(\mathbf{X}_j^{(1)T}\mathbf{W}_i\mathbf{X}_j^{(1)})^{-1}[\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}]^T.
\end{aligned}$$

Then by Lemma 2.1, the regular condition A.3 and the properties of the Kronecker product, then after some calculation, we have

$$\begin{aligned}
&\mathrm{Var}\left(\frac{[\mathbf{0}_{1\times p_{0n}},B_n^T,\mathbf{0}_{1\times p_{0n}}]}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{X}_i^{(1)T}\mathbf{W}_i\mathbf{X}_i^{(1)})^{-1}\mathbf{X}_i^{(1)T}\mathbf{W}_i\boldsymbol{\varepsilon}\right)\\
=\;&\sigma^2\mathrm{E}[\mathbf{0}_{1\times p_{0n}},B_n^T]\{\Omega^{(1)}(U)\}^{-1}[\mathbf{0}_{1\times p_{0n}},B_n^T]^T](1+o(1))\\
=\;&\sigma^2\mathrm{E}B_n^T\{\Omega^{(1)}(U)\}_{22}^{-1}B_n(1+o(1)),
\end{aligned}$$

where $\{\Omega^{(1)}(U)\}_{22}^{-1}$ is the $p_{1n}-p0n$ by $p_{1n}-p0n$ last prime submatrix of $\{\Omega^{(1)}(U)\}^{-1}$. Similar to the proof of Theorem 2.2, by checking the Lindeberg-Feller condition and some calculation, we finish the proof of Theorem 2.3. □

Chapter 3

Independent Sure Screening in Ultra-High-Dimensional Semi-varying Coefficient Models

3.1 Introduction

High dimensionality is an important characteristic of most popular data sets. Especially for some important data, such as image data, genetical or microarray data, the dimension of the data is considerably higher than the sample size, i.e $\log p = O(n^a)$ where $a \in (0, 1/2)$. Following Fan and Lv (2010), we call this nonpolynomial (NP) dimensionality or ultra-high dimensionality. Analyzing such ultra-high-dimensional data, presents simultaneous challenges to computational expediency, statistical accuracy, and algorithmic stability. Classical statistical methods are limited in their ability to handle such problems due to the "curse of dimensionality." Many studies have shown that regularization methods such as the LASSO (Tibshirani,1996), SCAD (Fan and Li, 2001), Dantzig selector (Canades and Tao, 2007), and MCP (Zhang, 2010), are appropriate for dealing with ultra-high dimensional linear regression models or other ultra-high dimensional parametric models. However, to reduce the model bias and allow it to be more flexible, ultra-high-dimensional nonparametric or semiparametric models are a good choice for modeling ultra-high-dimensional data when there is little prior information to indicate whether the effects of the covariates take a

linear form or belong to any other parametric family. The varying coefficient model is an important semiparametric model because of its flexibility and interpretability, but an appropriate estimation method is needed so that it can be used to model ultra-dimensional data and investigate the properties of those methods.

In Chapter 2, we investigated the regularization methods for varying coefficient models when the dimension of the model is diverging with the sample size, i.e. $p = o(n)$. However as Fan and Lv (2010) claimed, although such regularization methods can be applied to deal with ultra-high dimensional modeling problems, in such setting, these methods have many limitations, and are computationally expense. Fan and Lv (2008) introduced a sure independent screening method for ultra-high linear regression model. They suggested using the marginal correlations between the predictor variables and the response variable to screen and reduce the dimensionality of the model to relatively large, and then using the regularization method to select the final model. Fan, Feng and Song (2011) extended this sure screening idea to ultra-high dimensional additive models, and proposed a nonparametric independence screening method. Based on the ideas in these two papers, in this chapter, we consider ultra-high-dimensional varying coefficient models and suggest a nonparametric sure independent screening method. We investigate the sure screening properties and discuss some important practical implemental issues for our proposed sure screening method.

This chapter is organized as follows: In Section 3.2 we introduce the nonparametric independence screening (NIS) procedure in varying coefficient models. The theoretical properties are discussed in Section 3.3. in Section 3.4, we discuss some practical issues for implementation of the proposed method. The numerical studies for our proposed methods are presented in Chapter 5. The proofs of the theorems shown in Section 3.3 are relegated to the Appendix of this Chapter, Section 3.5.

3.2 NIS in Varying Coefficient Models

Suppose we have a random sample $\{(\mathbf{X}_i, U_i, Y_i)\}^n{}_{i=1}$, from the population

$$Y = m(\mathbf{X}, U) + \varepsilon = \beta_0(U) + \sum_{j=1}^{p_n} \beta_j(U) X_j + \varepsilon, \tag{3.1}$$

where $\mathbf{X} = (X_1, \cdots, X_{p_n})^T$, U is a random variable, and ε is the random error with conditional mean

$$\mathrm{E}(\varepsilon|U, X_1, \ldots, X_{p_n}) = 0$$

and conditional variance

$$\mathrm{var}(\varepsilon|U, X_1, \ldots, X_{p_n}) = \sigma^2(U).$$

When p_n is ultra high, the most important task for the above model is to expeditiously identify important variables. Similar to Fan, Feng, and Song (2011), to avoid the curse of dimensionality we consider the following p_n marginal varying coefficient regression problems:

$$\min_{a_{0j}, a_j} \mathrm{E}(Y - a_{0j}(U) - a_j(U)X_j)^2. \tag{3.2}$$

Let the minimizer of (3.2) be $a_{0j}(U)$ and $a_j(U)$. By simple calculation, it is easy to show that

$$a_j(U) = \frac{\mathrm{Cov}(X_j, Y|U)}{\mathrm{Var}(X_j|U)}, \quad \text{and} \quad a_{0j}(U) = \mathrm{E}(Y|U) - a_j(U)\mathrm{E}(X_j|U).$$

We then rank the utility of the covariates in Model (3.1) according to $\mathrm{E}a_j^2(U)$, and select a small group of significant covariates by thresholding.

To obtain a sample version of the marginal varying coefficient regression, we use a B-spline basis approximation on a_{0j} and a_j. Let S_n be the space of polynomial splines of degree $l > 1$ and let $\psi_{jk}(u), k = 1, \cdots, d_n$ denote a normalized B-spline basis with $\|\psi_{jk}(u)\|_\infty \leq 1$. For any a_{n0j} and a_{nj} belonging to S_n, we have

$$a_{n0j}(u) = \sum_{k=1}^{d_n} \beta_{0jk}\psi_{jk}(u) \quad \text{and} \quad a_{nj}(u) = \sum_{k=1}^{d_n} \beta_{jk}\psi_{jk}(u)$$

for some coefficients β_{0jk} and β_{jk}. Under some smoothness conditions, the functional coefficients a_{0j} and a_j can be well approximated by functions in S_n. The sample version of the marginal regression problem can be expressed as

$$\begin{aligned} & \min_{a_{n0j}, a_{nj}} \mathrm{P}_n(Y - a_{n0j}(u) - a_{nj}(u)X_j)^2 \\ = & \min_{\beta_{0j}, \beta_j} \mathrm{P}_n(Y - \boldsymbol{\psi}_j^T(u)\boldsymbol{\beta}_{0j} - \boldsymbol{\psi}_j^T(u)\boldsymbol{\beta}_j X_j)^2, \end{aligned} \tag{3.3}$$

where $\boldsymbol{\psi}_j(u)$ denotes the d_n dimensional basis functions, and P_n is the expectation with respect to the empirical measure P_n. This univariate nonparametric smoothing

can be computed rapidly, even for NP-dimensional problems. We correspondingly define the minimizer of the component-wise least squares regression (Huang, Wu, Zhou, 2004),

$$\begin{aligned} & \hat{a}_{nj}(u_i) \\ = \; & (\mathbf{0}, \boldsymbol{\psi}_j^T(u_i)) \; \{\mathrm{P}_n[(\boldsymbol{\psi}_j^T(u), \boldsymbol{\psi}_j^T(u)X_j)^T(\boldsymbol{\psi}_j^T(u), \boldsymbol{\psi}_j^T(u)X_j)]\}^{-1} \\ & \times \mathrm{P}_n(\boldsymbol{\psi}_j^T(u), \boldsymbol{\psi}_j^T(u)X_j)^T\mathbf{Y}, \end{aligned}$$

where $\mathbf{0}$ is d_n by 1 vectors with all elements 0. We also define the expectation version of the above minimizer as

$$\begin{aligned} & a_{nj}(u_i) \\ = \; & (\mathbf{0}, \boldsymbol{\psi}_j^T(u_i)) \; \{\mathrm{E}[(\boldsymbol{\psi}_j^T(u), \boldsymbol{\psi}_j^T(u)X_j)^T(\boldsymbol{\psi}_j^T(u), \boldsymbol{\psi}_j^T(u)X_j)]\}^{-1} \\ & \times \mathrm{E}(\boldsymbol{\psi}_j^T(u), \boldsymbol{\psi}_j^T(u)X_j)^T Y. \end{aligned}$$

Next we select a set of variables:

$$\hat{M}_{v_n} = \{1 \leq j \leq p : v_n \leq \|\hat{a}_{nj}\|_n^2\}, \tag{3.4}$$

where $\|\hat{a}_{nj}\|_n^2 = n^{-1}\sum_{i=1}^{n} \hat{a}_{nj}(U_i)^2$ and v_n is a predefined threshold value. This independence screening ranks the importance according to the marginal strength of the marginal varying coefficient function. As shown above, the marginal varying coefficient function is related to the correlation between the response variable and the predictor variables conditional on the random variable U. By selecting the appropriate threshold value v_n, the NIS procedure is able to reduce the dimensionality from p_n to a possibly much smaller dimension $|M_{v_n}|$, according to the following theoretical results.

3.3 Asymptotic Sure Screening Properties of NIS

In this section we investigate the sure screening properties of varying coefficient models. Firstly, we assume that the true regression function (3.1) admits the structure

$$m(\mathbf{X}, U) = \beta_0(u) + \sum_{j=1}^{p_n} \beta_j(u)X_j. \tag{3.5}$$

Let $M = \{j : \mathrm{E}\beta_j^2(U) > 0\}$ be the true sparse model with nonsparsity size $s_n = |M|$. The theoretical basis of the sure screening is that the marginal signal of the active varying coefficient functions $(\|a_j(u)\|, j \in M)$ does not vanish. The following are regular conditions are needed:

A1. All X_j are normalized, bounded on $[a, b]$ and the marginal density function g_j of X_j satisfies $0 < K_1 < g_j(X_j) < K_2$, and $L_1 < \mathrm{Var}(X_j|U) < L_2$ for all j and some constants K_1, K_2, L_1 and L_2.

A2. The density of the random variable U is continuous and positive on the interval $[0, 1]$.

A3. All $a_j(u)$ belong to a class of functions F, whose rth derivative exists and is Lipschitz constant of the order α, that is

$$|a_j^{(r)}(s) - a_j^{(r)}(t)| \leq K|s - t|^{\alpha} \quad for\ s,\ t \in [0, 1]$$

for some positive K, nonnegative integer r, $0 < \alpha < 1$, $d = r + \alpha > 0.5$.

A4. $c_1 d_n n^{-2\kappa} \leq \min_{(j \in M)} \mathrm{E}a_j^2(U)$.

A5. All $a_j(u)$ have bounded sup norm: $\|a_j(u)\|_{\infty} \leq B_1$ for some positive number B_1.

A6. The random errors are i.i.d with a conditional mean 0, and for any positive number B_2, there exists a positive constant B_3 such that $\mathrm{E}[\exp(B_2|\epsilon_i|)|X_i, U_i] < B_3$.

A7. There exist positive constants c_1 and $0 < \gamma < 1$ such that ${d_n}^{-2d-1} \leq c_1(1 - \gamma)n^{-2\kappa}/C_1$.

Remark: Conditions A1 and A2 are reasonable conditions for nonparametric or semiparametric regression models. To reduce the effect of scale, according to the condition A1, the samples $X_{ij}, i = 1, \ldots, n$ must be normalized before the model estimation. The condition on the conditional variance between X_j and U is similar to the condition imposed by Stone (1985) for additive models. This condition simplifies some of the theoretical derivation in the proof. In fact, we can consider a more

subtle nonparametric sure screening method for the ultra high dimensional varying coefficient model to remove this condition on the conditional variance between X_j and U. Conditions A3 and A5 are smoothing conditions on the varying coefficient functions. Those conditions are very useful when we investigate the effect of spline approximation on those varying coefficient functions. Condition A4 just requires that the negative significant signal of the significant varying coefficient functions should be sufficiently strong for it to be identified from noise. Condition A6 assumes that the error in the model has a light tail or follows a sub-gaussian distribution. Such a condition has often been used in ultra-high-dimensional investigations. Condition A7 shows the relationship between the number of knots used for the spline approximation and the signal strength of the significant varying coefficient functions in the model. When those signals are very strong, i.e. $\kappa = 0$, d_n can be a finite constant.

Let $\tilde{\boldsymbol{\psi}}_j(u) = (\boldsymbol{\psi}_j^T(u), \boldsymbol{\psi}_j^T(u)X_j)^T$ and let the new d_n be two times the old d_n. With regular calculations, similar to Fan, Feng, and Song (2011), we can identify the following three properties of $\tilde{\boldsymbol{\psi}}$:

Fact 1. As shown by Stone (1985) for additive models, there exists a positive constant C_1 such that

$$\|a_{0j} - a_{n0j}\|^2 \leq C_1 {d_n}^{-2d} \quad \text{and} \quad \|a_j - a_{nj}\|^2 \leq C_1 {d_n}^{-2d}. \tag{3.6}$$

Fact 2. Because X_j is bounded, there exists a positive constant C_2 (Stone, 1985; Huang, Horowitz, and Wei, 2010), such that

$$\mathrm{E}\tilde{\psi}_{jk}^2(U_i) \leq C_2 {d_n}^{-1}. \tag{3.7}$$

Fact 3. Because X_j is a bounded variable, by Condition A1 and as shown in the proof of Lemma 3.1, there exist some positive constants D_1 and D_2 (Zhou, Shen, and Wolfe, 1998), such that

$$D_1 {d_n}^{-1} \leq \lambda_{min}(\mathrm{E}\tilde{\boldsymbol{\psi}}_j \tilde{\boldsymbol{\psi}}_j^T) \leq \lambda_{max}(\mathrm{E}\tilde{\boldsymbol{\psi}}_j \tilde{\boldsymbol{\psi}}_j^T) \leq D_2 {d_n}^{-1}. \tag{3.8}$$

By the regular Conditions A1-A7 and Facts 1-3, for the sure screening properties of our proposed NIS for ultra-high-dimensional varying coefficient models, we have the following theorems.

Theorem 3.1 *Under regular conditions A1-A7,*

(i) For any $c_2 > 0$, there exits some positive constants c_3 and c_4 such that

$$\begin{aligned} & \mathrm{P}(\max_j \left| \|\hat{a}_{nj}\|_n^{\ 2} - \|a_{nj}\|^2 \right|) \geq c_2 d_n n^{-2\kappa}) \\ \leq \ & p_n d_n \{(8+2d_n)\exp(-c_3 n^{1-4\kappa} {d_n}^{-3}) + 6d_n \exp(c_4 n {d_n}^{-3})\}. \end{aligned}$$

(ii) Taking $v_n = c_5 d_n n^{-2\kappa}$, $c_5 \leq c_1\gamma/2$, we have

$$\begin{aligned} & \mathrm{P}(M \subset \hat{M}_{v_n}) \\ \geq \ & 1 - s_n d_n \{(8+2d_n)\exp(-c_3 n^{1-4\kappa} {d_n}^{-3}) + 6d_n \exp(-c_4 n {d_n}^{-3})\}. \end{aligned}$$

Remark: If we take

$$\log p_n = o(n^{1-4\kappa} {d_n}^{-3} + n {d_n}^{-3}), \tag{3.9}$$

then the second part of the above theorem can be rewritten as

$$P(M \subset \hat{M}_{v_n}) \to 1.$$

The above theorem also tells us that to ensure the sure screening properties, an upper bound on the basis number should be $d_n = o(n^{1/3})$, while the regular condition A7 gives a lower bound on the number of basis, that is $d_n \geq B_4 n^{2\kappa/(2d+1)}$ while $B_4 = (c_1(1-\gamma)/C_1)^{-1/(2d+1)}$.

Similar to the discussion of Fan, Feng, and Song (2011), as for the size of the B-spline basis, the smaller d_n is, the larger the dimensionality we can handle. However, the size of d_n cannot be too small because the approximation error cannot be too large, which is why we need a lower bound on d_n. As the lower bound is defined by d in Condition A3, we can see that the smoother a_j is, the smaller d_n we can take and the higher the dimensionality we can handle.

The following theorem shows the correlation between the size of basis and the size of the selected models.

Theorem 3.2 *Under regular conditions A1-A7, if Var(Y) is bounded, then for any $v_n = c_5 d_n n^{-2\kappa}$, there exist positive constants c_3 and c_4 such that*

$$\begin{aligned} & \mathrm{P}[|\hat{M}_{v_n}| \leq O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}] \\ \geq \ & 1 - p_n d_n \{(8+2d_n)\exp(-c_3 n^{1-4\kappa} {d_n}^{-3}) + 6d_n \exp(-c_4 n {d_n}^{-3})\}, \end{aligned}$$

where $\Sigma = \mathrm{E}\tilde{\Psi}\tilde{\Psi}^T$, $\tilde{\Psi} = (\tilde{\psi}_1^T, \cdots, \tilde{\psi}_{p_n}^T)^T$.

As $\kappa < 1/2$, and $\lambda_{max}(\Sigma) = O(1)$ we can see that the size of the selected models should satisfy $|\hat{M}_{v_n}| = o(n)$. There is an important special case of the theorem, in that if all of the covariates are independent, then Σ is block diagonal. Then, from Fact 3, we have

$$\lambda_{max}(\Sigma) = O({d_n}^{-1}).$$

The selected model size should then be $|\hat{M}_{v_n}| = O(n^{2/3})$.

From the discussion above, Theorem 3.2 means that the proposed method has sure screening properties. It can reduce the dimension of the ultra high dimensional varying coefficient model to a relatively high-dimensional varying coefficient model, but keep the strong signal varying coefficient functions or significant predictor variables in the model. After the coarse selection in the first step using the proposed nonparametric sure screening method, a regularization method, such as the method investigated in Chapter 2, can be applied to a relatively large model to refine the selection and estimate the final model.

3.4 Practical Issues

We can apply the following two-step procedure to ultra high dimensional varying coefficient models to obtain get the final estimation of the model.

1. Applying the proposed nonparametric sure independent screening to roughly select significant variable set M.

2. Based on set M, apply the proposed regularization method in Chapter 2 to select the final model, and estimate the significant varying coefficient functions and determine the constant coefficients simultaneously.

To apply the above procedure in practice, several important issues concerning the nonparametric sure independent screening need to be considered.

- The first one is how to select an appropriate thresholding value v_n. In fact, the objective of NIS is to reduce the dimension of the model from ultra-high

to relatively high. Hence, the simplest way to determine the value of v_n is to set it as the $(n-1)$th or $(n/\log n)$th largest $\|\hat{a}_{nj}\|^2$. This means that at the first step, NIS will first select $n-1$ or $n/\log n$ variables with the largest norm of varying coefficient functions. According to our experience, such a method is quite simple and stable. To obtain a more subtle value of v_n, similar to Fan, Feng and Song, we can use the random permutation of Zhao and Li (2010). The detailed procedure can be found in Fan, Feng, and Song (2011). In our numerical study, we use the first method to determine the threshold value v_n and select significant variables by NIS

- The second issue is how to select the number of knots for marginal spline approximation. There is some discussion about this issue in Fan, Feng, and Song (2011). Similarly, according to our Theorem 3.1 and Theorem 3.2, when the marginal varying coefficient functions are quite smooth, a constant number of knots with even spacing should be sufficient for the nonparametric sure independent screening. It would not be sensible to the screening results. To improve the accuracy of the estimate for the marginal varying coefficient function, an adaptive spline smoothing technique, such as the penalized regression spline, can be used to estimate the marginal varying coefficient functions although the computational burden will increase and the sure screening properties of the adaptive estimates need further investigation.

- As shown in Fan and Lv (2008), when the covariates are jointly highly correlated, those marginal sure independent screening methods have some methodological challenges. The iterative procedure would be a good method to improve the accuracy of the final model selection and to dealt with some of the complex covariance structures among the predictor variables and response variables. The algorithm works as follows:

  1. For every $j \in \{1, \ldots, p_n\}$, as shown in Section 3.2, compute $\hat{a}_{nj}$, and select the following variables

     $$\mathcal{A}_1 = \{j : |\hat{a}_{nj}|^2 \geq v_n\}$$

     where v_n is $(n/\log n)$th largest $|\hat{a}_{nj}|^2$.

2. Apply the penalized method shown in Chapter 2 on the set $\mathcal{A}_1$ to select a subset $\mathcal{M}_1$ and the estimate of the varying coefficient functions $\hat{a}^*_{nj}$ corresponding to the variables X_j selected in $\mathcal{M}_1$.

3. For every $j \in \mathcal{M}_1^c = \{1, \ldots, p_n\} \setminus \mathcal{M}_1$, minimize

$$\mathrm{P}_n \left(Y - \hat{a}^*_{n0}(U) - \sum_{i \in \mathcal{M}_1} \hat{a}^*_{ni}(U) X_i - a_{nj}(U) X_j \right)^2,$$

with respect to $a_{nj} \in S_n$ for $j \in \mathcal{M}_1^c$. After marginal screening, as in Step 1, choose a set of indices, $\mathcal{A}_2$, then apply the regularization method on the set $\mathcal{M}_1 \bigcup \mathcal{A}_2$ to select a subset $\mathcal{M}_2$.

4. Iterate the process until $\mathcal{M}_l > s_0$ or $\mathcal{M}_l = \mathcal{M}_{l-1}$, where s_0 is a predefined number.

3.5 Appendix: Proofs

Lemma 3.1 *Under regular conditions, we have*

$$\min_{j \in M} \|a^2_{nj}(u)\| \geq c_1 \gamma d_n n^{-2\kappa}$$

Proof. By the property of the least squares estimator, we have

$$\mathrm{E}(Y - a_{n0j}(u) - a_{nj}(u) X_j) a_{nj}(u) X_j = 0,$$

$$\mathrm{E}(Y - a_{n0j}(u) - a_{nj}(u) X_j) a_{n0j}(u) = 0,$$

and

$$\mathrm{E}(Y - a_{0j}(u) - a_j(u) X_j) a_{nj}(u) X_j = 0,$$

$$\mathrm{E}(Y - a_{0j}(u) - a_j(u) X_j) a_{n0j}(u) = 0.$$

Then we can easily get,

$$\mathrm{E} a_{nj}(u) X_j (a_{0j}(u) + a_j(u) X_j - a_{n0j}(u) - a_{nj}(u) X_j) = 0$$

and

$$\mathrm{E} a_{n0j}(u) (a_{0j}(u) + a_j(u) X_j - a_{n0j}(u) - a_{nj}(u) X_j) = 0$$

As $a_j = a_{nj} + (a_j - a_{nj})$ and $a_{0j} = a_{n0j} + a_{0j} - a_{n0j}$, we have

$$\|a_{n0j}(u) + a_{nj}(u)X_j\|^2 = \|a_{0j}(u) + a_j(u)X_j\|^2 - \|a_{0j} + a_j(u)X_j - a_{n0j}(u) - a_{nj}(u)X_j\|^2.$$

Next, note that the eigenvalues of

$$\begin{bmatrix} 1 & \mathrm{E}(X_j|U) \\ \mathrm{E}(X_j|U) & \mathrm{E}(X_j^2|U) \end{bmatrix}$$

are bounded between two positive constants by Condition A1. Thus, this lemma follows under Condition A4 and Fact 1. □

Lemma 3.2 *(Bernsteins inequality I) (van der Vaart and Wellner, 1996) For independent random variables $Y_1, \cdots, Y_n$ with bounded ranges $[-M, M]$ and 0 means, we have*

$$P(|Y_1 + \cdots + Y_n| > x) \leq 2exp(-x^2/2(v + Mx/3))$$

For $v \geq Var(Y_1 + \cdots + Y_n)$

Lemma 3.3 *(Bernsteins inequality II; van der Vaart and Wellner, 1996) For independent random variables $Y_1, \cdots, Y_n$ 0 means, $E|Y_i|^m \leq m!M^{m-2}v_i/2$ for every $m \geq 2$ and i and some constants M, v_i, then*

$$P(|Y_1 + \cdots + Y_n| > x) \leq 2exp(-x^2/2(v + Mx))$$

For $v \geq Y_1 + \cdots + Y_n$

Lemma 3.4 *Under regular conditions, for any $\delta > 0$, there exist some positive constants c_6 and c_7 such that,*

$$\mathrm{P}(|(\mathrm{P}_n - \mathrm{E})\tilde{\psi}_{jk}Y| \geq \delta n^{-1}) \leq 4exp(-\delta^2/2(c_6 n {d_n}^{-1} + c_7\delta)).$$

Proof. We set $Y_i = m(\mathbf{X}_i, U_i) + \epsilon$ for every $i = 1, \cdots, n$, and let.

$$T_{jki} = \tilde{\psi}_{jk}(U_i)Y_i - \mathrm{E}\tilde{\psi}_{jk}(U_i)Y_i.$$

Then we have $T_{jki} = T_{jki1} + T_{jki2}$ where

$$T_{jki1} = \tilde{\psi}_{jk}(U_i)m(X_i, U_i) - \mathrm{E}\tilde{\psi}_{jk}(U_i)m(X_i, U_i)$$

and $T_{jki2} = \tilde{\psi}_{jk}(U_i)\varepsilon_i$.

Under regular conditions A1, A2, A3 A5 and Fact 2, it can be shown ,

$$|T_{jki1}| \leq 2B_1$$

and

$$\mathrm{Var}(T_{jki1}) \leq \mathrm{E}\tilde{\psi}^2_{jk}(U_i)m_i^2(X_{ij}, U_i) \leq B_1^2 C_2 d_n^{-1}.$$

By Lemma 3.2, for any $\delta_1 > 0$,

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} T_{jki1}\right| > \delta_1\right) \leq 2\exp\left(-\frac{1}{2}\frac{\delta_1^2}{nB_1^2C_2d_n^{-1} + 2B_1\delta_1/3}\right). \tag{A.1}$$

For T_{jki2}, when $r \geq 2$ and under regular condition A6 and Fact 2, we have

$$\begin{aligned}
\mathrm{E}|T_{jki2}|^r &\leq \mathrm{E}|\tilde{\psi}_{jk}(X_{ij})|^2\mathrm{E}(|\epsilon_i|^r|\mathbf{X}_i) \\
&\leq r!B_2^{-r}\mathrm{E}|\tilde{\psi}_{jk}(X_{ij})|^2\mathrm{E}\exp(B_2|\epsilon_i||\mathbf{X}_i) \leq B_3C_2d_n^{-1}r!B_2^{-r}.
\end{aligned}$$

Then using Lemma 3.3, for any $\delta_2 > 0$,

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} T_{jki2}\right| > \delta_2\right) \leq 2\exp\left(-\frac{1}{2}\frac{\delta_2^2}{2nB_2^{-2}B_3C_2d_n^{-1} + B_2^{-1}\delta_2}\right). \tag{A.2}$$

Hence, the lemma is proved by combining (A.1) and (A.2). □

Lemma 3.5 *Under regular conditions, for any $\delta > 0$, we have,*

$$\mathrm{P}(|\lambda_{\min}(\mathrm{P}_n(\tilde{\psi}_j\tilde{\psi}_j^T)) - \lambda_{\min}(\mathrm{E}(\tilde{\psi}_j\tilde{\psi}_j^T))| \geq d_n\delta/n) \leq 2d_n^2\exp\left\{-\frac{1}{2}\frac{\delta^2}{C_2nd_n{}^{-1} + \delta/3}\right\}$$

Proof. For any symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, and any $\|\mathbf{x}\| = \mathbf{1}$, by linear algebra knowledge, we have

$$\lambda_{min}(\mathbf{A} + \mathbf{B}) \geq \lambda_{\min}(\mathbf{A}) + \lambda_{\min}(\mathbf{B}).$$

Then it is easy to show that

$$\lambda_{\min}(\mathbf{A} - \mathbf{B}) \leq \lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B}),$$

and

$$\lambda_{\min}(\mathbf{B} - \mathbf{A}) \geq \lambda_{\min}(\mathbf{B}) - \lambda_{\min}(\mathbf{A}).$$

By these two inequalities, we have

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \leq \max\{|\lambda_{\min}(\mathbf{A}-\mathbf{B})|, |\lambda_{\min}(\mathbf{B}-\mathbf{A})|\}.$$

Next let $\mathbf{A} = \mathrm{P}_n(\tilde{\psi}_j\tilde{\psi}_j^T)$, $\mathbf{B} = \mathrm{E}(\tilde{\psi}_j\tilde{\psi}_j^T)$, and $\mathbf{D}_j = \mathbf{A} - \mathbf{B}$, then

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \leq \max\{|\lambda_{\min}(\mathbf{D}_j)|, |\lambda_{\min}(-\mathbf{D}_j)|\}. \tag{A.3}$$

To bound the right side of (A.3)), for any $\|\mathbf{x}\| = 1$ we have

$$|\mathbf{x}^T\mathbf{D}_j\mathbf{x}| \leq \|\mathbf{D}_j\|_\infty \left(\sum_{i=1}^{d_n} |x_i|\right)^2 \leq d_n\|\mathbf{D}_j\|_\infty.$$

Thus,

$$\lambda_{\min}(\mathbf{D}_j) = \min_{\|\mathbf{X}\|=1} (\mathbf{x}^T\mathbf{D}_j\mathbf{x}) \leq d_n\|\mathbf{D}_j\|_\infty,$$
$$\lambda_{\min}(\mathbf{D}_j) = -\max_{\|\mathbf{X}\|=1} (-\mathbf{x}^T\mathbf{D}_j\mathbf{x}) \geq -d_n\|\mathbf{D}_j\|_\infty.$$

It is equivalent to

$$|\lambda_{\min}(\mathbf{D}_j)| \leq d_n\|\mathbf{D}_j\|_\infty.$$

Thus, from (A.3), we can get,

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{min}(\mathbf{B})| \leq d_n\|\mathbf{D}_j\|_\infty. \tag{A.4}$$

Next we need to bound the right side of (A.4). As $\|\psi_{jk}\|_\infty \leq 1$ and X_{ij} are bounded, by Fact 2 we have,

$$\mathrm{Var}\{\tilde{\psi}_{jk}(U)\tilde{\psi}_{jl}(U)\} \leq \mathrm{E}\tilde{\psi}_{jk}^2(U)\tilde{\psi}_{jl}^2(U) \leq M \cdot \mathrm{E}\tilde{\psi}_{jk}^2(U) \leq C_2 d_n^{-1}.$$

Hence by Lemma 3.2, for any $\delta > 0$,

$$\mathrm{P}(|(\mathrm{P}_n - \mathrm{E})\tilde{\psi}_{jk}(U)\tilde{\psi}_{jl}(U)| > \delta/n) \leq 2\exp\left\{-\frac{\delta^2}{2(C_2 n d_n^{-1} + \delta/3)}\right\}. \tag{A.5}$$

Thus the lemma is proved by (A.5) together with (A.4). □

Lemma 3.6 *Under regular conditions, for any $\delta > 0$, and for any given constant c_4, there exist some positive constant c_8 such that*

$$\begin{aligned} &\mathrm{P}\left\{\lambda_{\max}\{(\mathrm{P}_n(\tilde{\psi}_j\tilde{\psi}_j^T))^{-1}\} - \lambda_{\max}\{(\mathrm{E}(\tilde{\psi}_j\tilde{\psi}_j^T))^{-1}\}\right. \\ \geq\ &\left. c_8\lambda_{\max}\{(\mathrm{E}(\tilde{\psi}_j\tilde{\psi}_j^T))^{-1}\}\right\} \leq 2{d_n}^2\exp(-c_4 n d_n^{-3}) \end{aligned}$$

Proof. First we set $\delta = c_9 D_1 n d_n^{-2}$, $c_9 \in (0,1)$. From Fact 3 and (A.5), for some positive constant c_4, we have,

$$\begin{aligned} & \mathrm{P}(|\lambda_{\min}(\mathrm{P}_n(\tilde{\psi}_j\tilde{\psi}_j^T)) - \lambda_{\min}(\mathrm{E}(\tilde{\psi}_j\tilde{\psi}_j^T))|) \\ \geq \;& c_9\lambda_{\min}(\mathrm{E}(\tilde{\psi}_j\tilde{\psi}_j^T))) \leq 2d_n^2\exp(-c_4 n d_n^{-3}). \end{aligned} \tag{A.6}$$

By linear algebra knowledge we know that $(\lambda_{\min}(H))^{-1} = \lambda_{\max}(H^{-1})$. Set $A = \lambda_{\min}\{P_n(\tilde{\psi}_j\tilde{\psi}_j^T)\}$, and $B = \lambda_{\min}\{\mathrm{E}(\tilde{\psi}_j\tilde{\psi}_j^T)\}$. The left side of (A.6) can be rewritten as

$$|A - B| \geq aB$$

for some $a \in (0,1)$. Alternatively,

$$A^{-1} - B^{-1} \leq -(1 - 1/(1+a))B^{-1}$$

and

$$A^{-1} - B^{-1} \geq (1/(1-a) - 1)B^{-1}.$$

By $1 - 1/(1+a) < 1/(1-a) - 1$, we have

$$|A^{-1} - B^{-1}| \geq (1/(1-a) - 1)B^{-1}.$$

Then from (A.6), we have proved this lemma. □

Proof of Theorem 3.1

Notice that

$$\|\hat{a}_{n0j}(U) + \hat{a}_{nj}(U)X_j\|_n^2 = (\mathrm{P}_n\tilde{\psi}_j Y)^T(\mathrm{P}_n\tilde{\psi}_j\tilde{\psi}_j^T)^{-1}\mathrm{P}_n\tilde{\psi}_j Y,$$

and

$$\|a_{n0j}(U) + a_{nj}(U)X_j\|^2 = (\mathrm{E}\tilde{\psi}_j Y)^T(\mathrm{E}\tilde{\psi}_j\tilde{\psi}_j^T)^{-1}\mathrm{E}\tilde{\psi}_j Y.$$

Let $a_n = \mathrm{P}_n\tilde{\psi}_j Y$, $B_n = (\mathrm{P}_n\tilde{\psi}_j\tilde{\psi}_j^T)^{-1}$, $a = \mathrm{E}\tilde{\psi}_j Y$ and $B = (\mathrm{E}\tilde{\psi}_j\tilde{\psi}_j^T)^{-1}$. Then we have,

$$\begin{aligned} & \|\hat{a}_{n0j} + \hat{a}_{nj}X_j\|_n^2 - \|a_{n0j} + a_{nj}X_j\|^2 \\ = \;& a_n^T B_n a_n - a^T B a = (a_n - a)^T B_n(a_n - a) + 2(a_n - a)^T B_n a + a^T(B_n - B)a \\ = \;& I + II + III. \end{aligned} \tag{A.7}$$

Firstly, consider I. By the Cauchy-Schwartz inequality, we have

$$I \leq \lambda_{\max}(B_n)\|\|a_n - a\|^2. \tag{A.8}$$

By Lemma 3.4, and using union bound probability,

$$\mathrm{P}(\|a_n - a\|^2 \geq d_n\delta^2 n^{-2}) \leq 4d_n \exp(-\delta^2/2(c_6 n {d_n}^{-1} + c_7\delta)). \tag{A.9}$$

Next by Lemma 3.6, we have

$$\mathrm{P}(|\lambda_{\max}(B_n) - \lambda_{\max}(B) \geq c_8\lambda_{\max}(B)) \leq 2d_n^2 \exp(-c_4 n d_n^{-3}).$$

Hence from Fact 3 that $\lambda_{\max}(B) \leq D_1^{-1}d_n$, so we have

$$\mathrm{P}(\lambda_{\max}(B_n) \geq (c_8+1)D_1^{-1}d_n) \leq 2d_n^2 \exp(-c_4 n d_n^{-3}). \tag{A.10}$$

Combining (A.7), (A.8), (A.9) and (A.10), we bound I as follows,

$$\begin{aligned} & \mathrm{P}(I \geq (c_8+1)D_1^{-1}d_n^2\delta^2/n^2) \\ \leq\ & 4d_n \exp(-\delta^2/2(c_6 n d_n^{-1} + c_7\delta)) + 2d_n^2 \exp(-c_4 n d_n^{-3}). \end{aligned} \tag{A.11}$$

Next, consider II. Using the Cauchy-Schwartz inequality again, we can easily see that

$$|II| \leq 2\|a_n - a\|\|B_n a\| \leq \|a_n - a\| \cdot \lambda_{\max}(B_n) \cdot \|a\|. \tag{A.12}$$

By Condition A5 and Fact 2, we have

$$\begin{aligned} \|a\|^2 &= \sum_{k=1}^{d_n}(\mathrm{E}\tilde{\psi}_{jk}Y)^2 = \sum_{k=1}^{d_n}(\mathrm{E}\psi_{jk}(U)m(X_j,U))^2 \\ &\leq \sum_{k=1}^{d_n} B_1^2 \mathrm{E}\psi_{jk}^2(U) \\ &\leq B_1^2 C_2. \end{aligned} \tag{A.13}$$

Using the same bounding method as for I above, (A.9) and (A.10), we have

$$\begin{aligned} & P(|II| \geq 2(c_8+1)D_1^{-1}C_2^{1/2}B_1 d_n^{3/2}\delta/n) \\ \leq\ & 4d_n \exp(-\delta^2(C - 6n d_n^{-1} + c_7\delta)) + 2d_n^2 \exp(-c_4 n d_n^{-3}). \end{aligned} \tag{A.14}$$

Finally, for III, just like I and II, we have,

$$III = a^T B_n((\mathrm{E} - \mathrm{P}_n)\tilde{\psi}_j\tilde{\psi}_j^T)Ba.$$

Again by the Cauchy-Schwartz inequality, we have

$$|III| \leq \lambda_{\max}(\mathbf{D_j})\lambda_{\max}(B_n)\lambda_{\max}(B)\|a\|^2,$$

where $\mathbf{D_j} = (\mathrm{E} - \mathrm{P}_n)\tilde{\psi}_j\tilde{\psi}_j^T$ After some simple calculations, for any $\|\mathbf{x}\| = 1$,

$$\mathbf{x}^T\mathbf{D}^T\mathbf{D}\mathbf{x} = \sum_i (\sum_j d_{ij}x_j)^2 \leq \|\mathbf{D}\|_\infty^2 d_n(\sum_{j=1}^{d_n} |x_i|^2) \leq d_n^2\|\mathbf{D}\|_\infty^2.$$

That is $\lambda_{\max}(\mathbf{D}) \leq d_n\|\mathbf{D}\|_\infty$. Hence we have

$$\|\mathbf{D}_j\| \leq d_n\|\mathbf{D}_j\|_\infty.$$

For the remaining part $\|B_n\|\|B\|\|a\|^2$, using the same bounding techniques (A.5), (A.9), (A.10), and Fact 3, then by the union bound of probability we have

$$\begin{aligned} & \mathrm{P}(|III| \geq (c_8+1)D_1^{-2}B_1^2C_2D_n^3\delta/n) \\ \leq\ & 2d_n^2\exp(-\delta^2/2(c_6nd_n^{-1} + c_7\delta)) + 2d_n^2\exp(-c_4nd_n^{-3}). \end{aligned} \tag{A.15}$$

Finally, combining the bounding of I, II, III, (A.11), (A.14) (A.15), the union bound of probability we have

$$\begin{aligned} & \mathrm{P}\left(\left|\|\hat{a}_{n0j}(U) + \hat{a}_{nj}(U)X_j\|_n^2 - \|a_{n0j}(U) + a_{nj}(U)X_j\|^2\right|\right. \\ & \left. \geq c_{10}d_n^2\delta^2/n^2 + c_1 1 d_n^{3/2}\delta/n + c_1 2 d_n^3\delta/n\right) \\ \leq\ & (8d_n + 2d_n^2)\exp(-\delta^2/2(c_6nd_n^{-1} + c_7\delta)) + 6\delta^2\exp(-c_4nd_n^{-3}) \end{aligned} \tag{A.16}$$

Next notice that the eigenvalues of

$$\begin{bmatrix} 1 & \mathrm{E}(X_j|U) \\ \mathrm{E}(X_j|U) & \mathrm{E}(X_j^2|U) \end{bmatrix}$$

are bounded between two positive constants by Condition A1, and let $c_2d_nn^{-2\kappa} = c_{10}d_n^2\delta^2/n^2 + c_1 1 d_n^{3/2}\delta/n + c_1 2 d_n^3\delta/n$, we then proved the first conclusion of Theorem 3.1.

For the second conclusion, we let,

$$W_n = \{\max_{j\in M} | \|\hat{a}_{nj}\|_n^2 - \|a_{nj}\|^2| \le c_1\gamma d_n n^{-2\kappa}\}$$

Using Lemma 1 we know that by the choice of v_n, we can have $M \subset \hat{M}_{v_n}$. From the first conclusion we have,

$$\mathrm{P}(W_n^c) \le s_n((8d_n + 2d_n^2)\exp(-c_3 n^{1-4\kappa} d_n^{-3}) + 6d_n^2 \exp(-c_4 n d_n^{-3})).$$

Then the second conclusion is proved. □

Proof of Theorem 3.2

First we prove that

$$\|\mathrm{E}\tilde{\Psi}Y\|^2 = O(\lambda_{max}(\Sigma)) \tag{A.17}$$

Let $\boldsymbol{\alpha}_n = \mathrm{argmin}\mathrm{E}(Y - \tilde{\Psi}^T\boldsymbol{\alpha})^2$. Then we have $E\tilde{\Psi}(Y - \tilde{\Psi}^T\alpha_n) = 0$.

so,

$$\|\mathrm{E}\tilde{\Psi}Y\|^2 = \boldsymbol{\alpha}_n^T \mathrm{E}\tilde{\Psi}\tilde{\Psi}^T \mathrm{E}\tilde{\Psi}\tilde{\Psi}^T\boldsymbol{\alpha}_n \le \lambda_{\max}(\Sigma)\boldsymbol{\alpha}_n^T \mathrm{E}\tilde{\Psi}\tilde{\Psi}^T\boldsymbol{\alpha}_n. \tag{A.18}$$

By orthogonal decomposition,

$$\mathrm{Var}(Y) = \mathrm{Var}(\tilde{\Psi}^T\alpha_n) + \mathrm{Var}(Y - \tilde{\Psi}^T\alpha_n)$$

Since $\mathrm{Var}(Y) = O(1)$, so it is clear that $\mathrm{Var}(\tilde{\Psi}^T\boldsymbol{\alpha}_n) = O(1)$. Hence we have

$$\boldsymbol{\alpha}_n^T E\tilde{\Psi}\tilde{\Psi}^T\boldsymbol{\alpha}_n = O(1).$$

Then by (A.18) we have (A.17).

Next, using Fact 3 and because all samples are bounded, we then have

$$\sum_{j=1}^{p_n} \|a_{nj}\|^2 \le \max_{1\le j\le p_n} \lambda_{\max}\{(\mathrm{E}\tilde{\psi}_j\tilde{\psi}_j^T)^{-1}\}\|\mathrm{E}\tilde{\Psi}Y\|^2 = O(d_n\lambda_{max}(\Sigma)).$$

From the above equation we can see that the number of $\{j : \|a_{nj}\|^2 > \epsilon d_n n^{-2\kappa}\}$ can not exceed $O(n^{2\kappa}\lambda_{max}(\Sigma))$ for any positive ϵ.

Define a set

$$H_n = \left\{ \max_{1\le j\le p_n} \left| \|\hat{a}_{nj}\|_n^2 - \|a_{nj}\|^2 \right| \le \epsilon d_n n^{-2\kappa} \right\}.$$

On H_n the number of $\{j : \|\hat{a}_{nj}\|_n^2 > 2\epsilon d_n n^{-2\kappa}\}$ cannot exceed the number of $\{j : \|a_{nj}\|^2 > \epsilon d_n n^{-2\kappa}\}$. If we let $\epsilon = c_5/2$, we have

$$P\left\{|\widehat{M}_{v_n}| \leq O\{n^{-2\kappa}\lambda_{max}(\Sigma)\}\right\} \geq P(H_n)$$

By the first conclusion of Theorem 3.1, we complete the proof of Theorem 3.2. $\square$

Chapter 4

Revisit Local Linear and Quadratic Approximation for Nonconcave Penalized Methods

4.1 Introduction

In Chapter 1, we reviewed the development of variable selection techniques. As shown by Fan and Li (2001) and Fan and Peng (2004), the nonconcave penalized method has some very good statistical properties such as unbiasedness, sparsity, and continuity. Particularly, under some regular conditions, the nonconcave penalized likelihood estimators perform as well as the oracle estimators; namely, they work as well as if the correct submodel were known. In Chapter 2, we also showed that the group nonconcave penalized method for high-dimensional semivarying coefficient models also has similar good properties. We have just proved that the oracle estimate is the global minimizing point of group nonconcave penalized weighted least squares with a probability tending to one. These good properties of the nonconcave penalized estimates are why we like to use the SCAD penalized method for high dimensional semivarying coefficient models. However, compared with the LASSO, the nonconcave penalty function has no convex properties and hence there exist multiple extreme points for the nonconcave penalized least squares or nonconcave penalized likelihood functions. However, similar to the LASSO, the nonconcave penalty function is also singular at

zero. These drawbacks hinder the development of nonconcave penalized methods, and urge statisticians to invent efficient numerical algorithms that are capable of maximizing or minimizing the nondifferentiable nonconcave function.

Antoniadis and Fan (2001) proposed nonlinear regularized Sobolev interpolators (NRSI) and a regularized one-step estimator (ROSE) for nonconvex penalized least squares problems under wavelets settings. They also applied the graduated nonconvexity (GNC) algorithm to minimize the high-dimensional nonconvex penalized least squares problem when the design matrix of the regression model is an orthogonal matrix. Fan and Li (2001) proposed a new unified revised Newton-Raphson algorithm for the minimization of SCAD penalty problems via local quadratic approximations and called it a local quadratic approximation (LQA). Hunter and Li (2005) showed the convergence of such local quadratic approximation algorithm when the dimension of the model is a fixed constant. However, in a sense the solution of the revised Newton-Raphson algorithm does not have sparse properties, and therefore, based on the so-called local linear approximation (LLA), Zou and Li (2008) proposed a new algorithm for nonconcave penalized methods. The efficient LARS algorithm can be used to update the estimate of LLA in its iterative process and retain sparsity for the final solution.

The LLA and LQA algorithms are the most effective algorithms for nonconcave penalized methods. However, these two algorithms have their drawbacks. The LLA algorithm only uses the linear term and the first derivative of the nonconcave penalty function to make the approximation. This means, as shown by the following theorem, that the convergence rate of the LLA estimator to the minimum or maximum is linear. Although the LQA also uses only the first derivative of the nonconcave penalty function to make the approximation, it uses a quadratic form to approximate the second derivative of the nonconcave penalized function and based on the idea of the efficient Newton Raphson algorithm, it would in some sense have the so-called superlinear convergence rate. However it does not have the sparse property as it will slow its convergence rate. Nevertheless, as shown by Zou and Li (2008), applying the LLA algorithm depends on the quadratic approximation to the full statistical model. If the model cannot be simply approximated by the quadratic form, the LLA algorithm will

not be efficient and will not be able to converge to an appropriate solution. In this sense, the LQA algorithm has more generalized applications, as long as the second derivative of the likelihood function can be approximated accurately. However, an important drawback of the LQA algorithm is that it has to calculate the inverse of the Hessian matrix of the penalized likelihood or penalized least squares. When the dimension of the model is even larger than the sample size, such an inverse matrix does not even exist. When the dimension of the model is diverging with the sample size, calculating the inverse matrix will also considerably increase the computational burden and make the results unstable.

In this chapter, we revisit the local linear approximation (LLA) and local quadratic approximation algorithms (LQA) for nonconcave penalized likelihood functions. We merely show that the convergence rate of the LLA is linear, but the idea can be applied to the generalized likelihood function with a diverging number of dimensions as long as the LASSO has efficient algorithms for it. As shown in Chapter 1, based on the idea of local quadratic approximation, we suggest a new revised Newton Raphson algorithm for nonconcave penalized methods without calculating the inverse of the Hessian matrix. This chapter is organized as follows: in Section 4.2, we revisit the LLA algorithm and investigate its convergence rate. We also study the statistical properties of its one-step estimate when the dimension of the model is diverging with the sample size. In Section 4.3 we propose our new local quadratic approximation procedure without calculating the inverse of the Hessian matrix. We also show its convergence in this section. All of the theoretical proofs are relegated to the appendix of this chapter, Section 4.4.

4.2 Revisiting the LLA for Nonconcave Penalized High Dimensional Likelihoods

Suppose that $\{X_i, Y_i\}$ are n identically and independently distributed samples, where X_i denotes the p_n-dimension predictor and Y_i is the response variable, and the conditional log-likelihood is given by $l(X_i, Y_i, \boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is the parameter vector in the model. In the linear regression model, or generalized linear regression model, Y_i de-

pends on X_i through a linear combination of $X_i^T\boldsymbol{\beta}$, and the conditional log-likelihood is given by $l(X_i^T\boldsymbol{\beta}, Y_i)$. When the dimension p of X_i is quite high, the dimension of $\boldsymbol{\beta}$ would also be quite large. Hence we assume that the dimension of $\boldsymbol{\beta}$ is diverging with the sample size n. To simplify, we denote the conditional likelihood function as $l_i(\boldsymbol{\beta})$.

In practice, some components of X_i have no significant correlation with Y_i, and hence some component of $\boldsymbol{\beta}$ should be zero. Selecting an appropriate model to fit real data is, therefore, very important in many applications. Regularization or penalized methods are examples of such popular statistical methods. The goal of these methods is to identify significant variables and estimate their corresponding coefficients efficiently. The generalized penalized likelihood function is given in the following form:

$$Q(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) - n\sum_{j=1}^{p_n} p_{\lambda_j}(|\beta_j|) \tag{4.1}$$

where $p_\lambda(\cdot)$ is the penalty function and λ_j is a predefined tuning parameter. Similarly, the penalized least squares can be defined as

$$Q(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n}(Y_i - X_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^{p_n} p_{\lambda_j}(|\beta_j|). \tag{4.2}$$

By maximizing the penalized likelihood function we can then get the estimate of $\boldsymbol{\beta}$. The LASSO and SCAD are two popular penalty functions used for penalized likelihood functions because their solutions have the so called "sparsity" properties. The SCAD penalty (Fan and Li (2001)) is a nonconcave function defined by $p_\lambda(0) = 0$ and for $|\beta| > 0$,

$$p'_\lambda = \lambda I(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{a-1} I(|\beta| > \lambda) \tag{4.3}$$

Note that the SCAD function is a nonconcave function, and maximizing the nonconcave penalized likelihood is a challenging problem. The LLA algorithm proposed by Zou and Li (2008) for the SCAD is a popular algorithm used to maximize nonconcave penalized likelihood functions. When the dimension of $\boldsymbol{\beta}$ is a fixed constant, Zou and Li (2008) showed that the LLA is the best convex minorization-maximization (MM) algorithm and if the regularization parameter is appropriately chosen, the one-step LLA estimates enjoy the oracle properties, provided that the initial estimates are good enough.

The LLA Algorithm and its Convergence Rate

Given an initial $\beta_j^{(0)}$, Zou and Li (2008) suggested using a local linear approximation to approximate the penalty function $p_\lambda(|\beta_j|)$:

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + p'_\lambda(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|) \tag{4.4}$$

Then given the initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$, the estimate of the maximum penalized likelihood and penalized least squares can be updated as

$$\hat{\boldsymbol{\beta}}^{(1)} = \arg\max_\beta \Big[\sum_{i=1}^{n} l_i(\boldsymbol{\beta}) - n \sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^{(0)}|)|\beta_j| \Big] \tag{4.5}$$

and

$$\hat{\boldsymbol{\beta}}^{(1)} = \arg\min_\beta \Big[\sum_{i=1}^{n} (Y_i - X_i^T\boldsymbol{\beta})^2 + n \sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^{(0)}|)|\beta_j| \Big]. \tag{4.6}$$

Then $\hat{\boldsymbol{\beta}}^{(1)}$ can be regarded as the new initial estimate, and following the procedure shown as above, the estimate of $\boldsymbol{\beta}$ is updated again, and stops when the estimate's sequence converges. The LLA algorithm is distinguished from the LQA algorithm because the LLA estimator naturally adopts a sparse representation. The LLA algorithm inherits the good features of the LASSO. The LASSO's efficient algorithm can be used to update the estimate in the above equation, hence in some sense the LLA is efficient in computation.

However, the LLA only uses the first derivative of the penalty function to make the approximation. If the initial estimate is not so good, the convergence of speed will be rather slow. The following theorem and a simple example show that the convergence speed of LLA is linear, thus when applying the LLA algorithm, the initial estimate of $\boldsymbol{\beta}$ is very important to improve the computational efficiency.

Theorem 4.1. Denote $\boldsymbol{\beta}^{(k)}$ as the series of $\boldsymbol{\beta}$ obtained (4.6) using the LLA with the SCAD penalty function, and let $\boldsymbol{\beta}^*$ be the true minimizer of (4.2). If the eigenvalues of $\mathrm{E}X_iX_i^T$ are bounded between two constant positive values, then by an appropriate choice of a in the SCAD penalty function, we have

$$|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^*| \leq C|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^*|,$$

where C is a real number satisfying $0 < C < 1$.

The above theorem just shows that the convergence rate of the LLA algorithm is linear. Such results can not be improved. For example, consider the following simple penalized least squares example:

$$\frac{1}{2}(y-\theta)^2 + p_\lambda(|\theta|),$$

where $p_\lambda(\cdot)$ is the SCAD penalty function.

Then, given the initial estimate of $\theta^{(0)}$ between 2λ and $a\lambda$, and a positive y, by the iterative LLA algorithm, we have

$$\begin{aligned}\theta^{(1)} &= y - p'_\lambda(|\theta^{(0)}|)_+, \\ \theta^{(2)} &= y - p'_\lambda(|\theta^{(1)}|)_+.\end{aligned}$$

Hence

$$\theta^{(2)} - \theta^{(1)} = p'_\lambda(|\theta^{(0)}|)_+ - p'_\lambda(|\theta^{(1)}|)_+ = \frac{\theta^{(0)} - \theta^{(1)}}{a-1},$$

and then

$$\|\theta^{(2)} - \theta^{(1)}\| = \frac{1}{a-1}\|\theta^{(1)} - \theta^{(0)}\|.$$

This means that the convergence rate of the LLA algorithm in this simple example cannot be faster than the linear rate, so the result of the above theorem can not be improved. However as shown by Zou and Li (2008), if the initial estimate of $\boldsymbol{\beta}$ is close enough to the real value of $\boldsymbol{\beta}$ when the dimension of the model is fixed, a one-step estimate is good enough and has the so called "oracle property." Next, we investigate the properties of such a one-step estimate when the dimension of the model is diverging with the sample size.

One-Step Estimation by LLA when the Dimension of the Model is Diverging

As demonstrated by Fan and Chen (1999) and Cai, Fan and Li (2000), both empirically and theoretically, the one-step method is as efficient as the fully iterative method, provided that the initial estimators are reasonably good. Although one may further define a k-step estimator, it is generally unnecessary. Under generalized regular conditions, Zou and Li (2008) showed that a one-step estimate by LLA also has

the "oracle property" when the dimension of the model is fixed constant. Here, we also aim to show that even when the dimension of the model is diverging with the sample size, a one-step estimate still retains the oracle property.

Suppose that the log-likelihood function is smooth and has the first three derivatives with respect to $\boldsymbol{\beta}$. Given a certain initial $\boldsymbol{\beta}^{(0)}$, the one-step estimate is defined as

$$\boldsymbol{\beta}^{(1)} \;=\; \arg\max_{\beta}\Big[\sum_{i=1}^{n} l_i(\boldsymbol{\beta}) - n\sum_{j=1}^{p_n} p'_{\lambda}(|\beta_j^{(0)}|)|\beta_j|\Big]. \tag{4.7}$$

It is clear that (4.1) reduces to the one-step estimates in linear regression models, if we are willing to assume that the error ϵ satisfies a normal distribution. Now we show that in the general likelihood setting, $\boldsymbol{\beta}^{(1)}$ is the desired one step estimates with the oracle property. We denoted $\boldsymbol{\beta}^{(1)}$ by $\hat{\boldsymbol{\beta}}^{(ose)}$. Let $\boldsymbol{\beta}_{n0} = \{\boldsymbol{\beta}_{n01}, \mathbf{0}\}$ be a p_n dimensional vector and its first s_n elements $\boldsymbol{\beta}_{n01}$ are nonzero and the others are zero. Let $I_n(\boldsymbol{\beta}_{n0})$ be the Fisher information matrix and $I_n(\boldsymbol{\beta}_{n01}) = I_n(\boldsymbol{\beta}_{n01},\ 0)$ be the known Fisher information $\boldsymbol{\beta}_{n02} = 0$.

Regular Conditions for Likelihood Functions:

Let $a_n = \max\{p'_{\lambda_n}(|\beta_j^{(0)}|), j = 1, \ldots, p_n\}$ and $b_n = \max\{p''_{\lambda_n}(|\beta_j^{(0)}|), j = 1, \ldots, p_n\}$. The following regular conditions are needed for the penalty function and initial value $\boldsymbol{\beta}^{(0)}$.

A1. $\liminf_{n\to\infty} \liminf_{\theta\to 0_+} p'_{\lambda_n}(\theta)/\lambda_n > 0$,

A2. $a_n = o_p(\frac{1}{(np_n)^{1/2}})$,

A3. $b_n = o_p\Big(\frac{1}{p_n^{1/2}}\Big)$,

A4. there are constants C and D when $\theta_1,\ \theta_2\ >\ C\lambda_n$, we have $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \le |\theta_1 - \theta_2|$.

Condition [A1] makes the penalty function singular at the origin so that the penalized likelihood estimators possess the sparsity property. If the initial estimate of $\boldsymbol{\beta}^{(0)}$ is a consistent estimate for the real $\boldsymbol{\beta}$, conditions [A2] and [A3] are easy to satisfy. [A4] is a smoothness condition that is imposed on the nonconcave penalty functions.

Unlike finite dimensional problems, we are unable to assume that likelihood functions are invariant in regression models with a diverging number of parameters. In addition to regular conditions (A1)-(A4), we have the following conditions for likelihood functions and samples.

A5. For every n, the observations (X_{ni}, Y_i) are independent and identically distributed with the probability density f_n, which has a common support. The first and second derivatives of the likelihood function satisfy the equations

$$\mathrm{E}_{\beta_n}\Big\{\frac{\partial \log f_n}{\partial \beta_{nj}}\Big\} = 0$$

and

$$\mathrm{E}_{\beta_n}\Big\{\frac{\partial \log f_n}{\partial \beta_{nj}}\frac{\partial \log f_n}{\partial \beta_{nk}}\Big\} = -\mathrm{E}_{\beta_n}\Big\{\frac{\partial^2 \log f_n}{\partial \beta_{nj}\partial \beta_{nk}}\Big\}.$$

A6. The Fisher information matrix $I_n(\beta_n) = \mathrm{E}\Big[\Big\{\frac{\partial \log f_n}{\partial \beta_n}\Big\}\Big\{\frac{\partial \log f_n}{\partial \beta_n}\Big\}^T\Big]$ satisfies the conditions

$$0 < C_1 < \min\{I_n(\beta_n)\} \le \max\{I_n(\beta_n)\} < C_2 < \infty$$

and

$$E_{\beta_n}\Big\{\frac{\partial \log f_n}{\partial \beta_{nj}}\frac{\partial \log f_n}{\partial \beta_{nk}}\Big\}^2 < C_3 < \infty,$$

$$E_{\beta_n}\Big\{\frac{\partial^2 \log f_n}{\partial \beta_{nj}\partial \beta_{nk}}\Big\}^2 < C_4 < \infty.$$

A7. There is an open subset ω_n of $\Omega \in R^{p_n}$ that contains the true parameter point β_n. For all $\boldsymbol{\beta}_n \in \omega_n$, the density admits all third derivatives which also satisfies

$$\mathrm{E}\Big|\frac{\partial^3 \log f_n}{\partial \beta_{nj}\beta_{nk}\beta_{nl}}\Big|^2 < C_5 < \infty.$$

A8. Let the values of $\beta_{n01}, \cdots \beta_{n0s_n}$ be non-zero while others are zero. For these nonzero β_{n0j}, we have

$$\min|\beta_{n0j}|/\lambda_n \to \infty.$$

Under conditions [A6] and [A7], the second and fourth moments of the likelihood function are imposed. The information matrix of the likelihood function is assumed to be positive definite, and its eigenvalues are uniformly bounded. [A8] shows the

rate at which the penalized likelihood can distinguish non-vanishing parameters from 0.

We show that in the general likelihood setting, $\boldsymbol{\beta}^{(1)}$ is the desired one-step estimate, denoted by $\hat{\boldsymbol{\beta}}^{(ose)}$.

Theorem 4.2. Let p_{λ_n} be a nonconcave penalty function. If $(np_n)^{\frac{1}{2}}\lambda_n \to \infty$, $p_n^5/n \to 0$ and $\lambda_n \to 0$, then under regular conditions [A1]-[A8], the one-step LLA estimates $\boldsymbol{\beta}^{(1)} = \hat{\boldsymbol{\beta}}^{(ose)}$ must satisfy

a. sparsity: with probability tending to one, $\hat{\boldsymbol{\beta}}_{n2}^{(ose)} = 0$, and

b. asymptotic normality: $n^{\frac{1}{2}} A_n I^{1/2}(\hat{\boldsymbol{\beta}}_{n1}^{(ose)} - \boldsymbol{\beta}_{n01}) \ \to \ N(0,\ G)$.

where A_n is a $q \times s_n$ matrix such that $A_n A_n^T \to G$ and G is a $q \times q$ non-negative symmetrical matrix.

Choosing of the initial value $\boldsymbol{\beta}^{(0)}$ is very important in LLA. Initial values that do not satisfy regular conditions [A2] and [A3] might not generate a one-step estimator that shares above sparsity and asymptotic properties. When the model dimension p_n is not too large, as for these conditions in Theorem 2, we can still use the unpenalized least square estimator as the initial value. It is clear that under linear regression models, the least squares estimator obviously satisfies [A2] and [A3] (see Portnoy, 1988). When dimension p_n is large or larger than the sample size n, we can use other methods such as sure independence screening or the LASSO to generate an initial estimator to satisfy [A2] and [A3].

4.3 Revisiting the LQA Algorithm

Singularity and nonconcavity make it difficult to maximize the nonconcave penalized likelihood functions. Given an initial value $\boldsymbol{\beta}^{(0)}$ that is close to the true value of $\boldsymbol{\beta}$, Fan and Li (2001) proposed a local approximation of the firs-order derivative of the penalty function using the linear function

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\text{sign}(\beta_j) \approx [p'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|]\beta_j.$$

Thus, they use an LQA to approximate the penalty function:

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}[p'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|](\beta_j^2 - \beta_j^{(0)2}). \tag{4.8}$$

With such an approximation, we can calculate the second derivative of the penalty function and hence the Newton-Raphson algorithm can be applied to update the estimate and find the maximizer of the penalized likelihood. Specifically, given an initial value of $\hat{\boldsymbol{\beta}}^{(0)}$, we repeatedly solve

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \arg\max_\beta \Big[\sum_{i=1}^n l_i(\hat{\boldsymbol{\beta}}^{(k)}) - n \sum_{j=1}^p \frac{p'_\lambda(|\hat{\beta}_j^{(k)}|)}{2|\hat{\beta}_j^{(k)}|}\beta_j^2 \Big], \tag{4.9}$$

or

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} - \{\nabla\ell(\hat{\boldsymbol{\beta}}^{(k)}) + n\Sigma_\lambda(\hat{\boldsymbol{\beta}}^{(k)})^{-1}\}^{-1}\{\nabla\ell(\hat{\boldsymbol{\beta}}^{(k)}) + nU_\lambda(\hat{\boldsymbol{\beta}}^{(k)})\},$$

where

$$\nabla\ell(\hat{\boldsymbol{\beta}}^{(k)}) = \sum_{i=1}^n \frac{\partial l_i(\hat{\boldsymbol{\beta}}^{(k)})}{\partial\boldsymbol{\beta}},$$

$$\nabla^2\ell(\hat{\boldsymbol{\beta}}^{(k)}) = \sum_{i=1}^n \frac{\partial^2 l_i(\hat{\boldsymbol{\beta}}^{(k)})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},$$

$$\Sigma_\lambda(\hat{\boldsymbol{\beta}}^{(k)}) = \mathrm{diag}\{p'_\lambda(|\hat{\beta}_1^{(k)}|)/|\hat{\beta}_1^{(k)}|, \ldots, p'_\lambda(|\hat{\beta}_{p_n}^{(k)}|)/|\hat{\beta}_{p_n}^{(k)}|\},$$

and

$$U_\lambda(\hat{\boldsymbol{\beta}}^{(k)}) = \Sigma_\lambda(\hat{\boldsymbol{\beta}}^{(k)})\hat{\boldsymbol{\beta}}^{(k)}.$$

Finally, we stop the iteration process if the estimate sequence $\hat{\boldsymbol{\beta}}^{(k)}$ converges.

When the dimension of the model is quite high, the LQA has a drawback. For every step, it needs to calculate the inverse of the Hessian matrix which increases the computational burden and makes the results unstable. Nevertheless, similar to the Newton Raphson algorithm, the LQA has a fast convergence rate, is easy to implement in a wide range of applications, and can deal with complex likelihood functions or statistical models. Hence how to apply LQA in high dimensional statistical models is an interesting problem. The key point is how to calculate the inverse of the Hessian matrix, and when such a matrix is singular, how to revise the Newton Raphson algorithm to update the estimate.

Construction of a Modified Newton-Raphson method

To simplify the problem, we just consider the LQA algorithm under the penalized least squares. The following discussion can be also be extended to the generalized penalized likelihood function without any modification.

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\beta} Q(\boldsymbol{\beta}) = \arg\min_{\beta} \Big[\frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\boldsymbol{\beta})^2 + n\sum_{j=1}^{p_n}\frac{p'_\lambda(|\beta_j^{(k)}|)}{2|\beta_j^{(k)}|}\beta_j^2\Big] \quad (4.10)$$

Using the Newton-Raphson algorithm to solve (4.8) we have

$$\left[\mathbf{X}^T\mathbf{X} + n\cdot\mathrm{diag}\left(\frac{p'_\lambda(|\beta_1^{(k)}|)}{|\beta_1^{(k)}|},\cdots,\frac{p'_\lambda(|\beta_{p_n}^{(k)}|)}{|\beta_{p_n}^{(k)}|}\right)\right]\boldsymbol{\beta}^{(k+1)} = \mathbf{X}^T\mathbf{Y}, \quad (4.11)$$

and

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \left[\mathbf{X}^T\mathbf{X} + n\cdot\mathrm{diag}\left(\frac{p'_\lambda(|\beta_1^{(k)}|)}{|\beta_1^{(k)}|},\cdots,\frac{p'_\lambda(|\beta_{p_n}^{(k)}|)}{|\beta_{p_n}^{(k)}|}\right)\right]^{-1}\mathbf{X}^T\mathbf{Y}, \quad (4.12)$$

where $\mathbf{Y} = (Y_1,\ldots,Y_n)^T$ and $\mathbf{X} = (X_1^T,\ldots,X_n^T)^T$.

Note that the computation for the inverse matrix of $\mathbf{X}^T\mathbf{X}$ might be very time-consuming, and the inverse matrix does not exist when the dimension of X_j is larger than the sample size. This motivates us to suggest the following modification to the Newton-Raphson method.

First, because $\mathbf{X}^T\mathbf{X}$ is a symmetrical matrix, we can split it into three parts, L, Λ, and L^T. L is a strictly lower diagonal matrix of $\mathbf{X}^T\mathbf{X}$ while Λ is a diagonal matrix of $\mathbf{X}^T\mathbf{X}$, and $\mathbf{X}^T\mathbf{X} = L + L^T + \Lambda$. Denote $D(\beta^{(k)}) = \mathrm{diag}(\frac{p'_\lambda(|\beta_1^{(k)}|)}{|\beta_1^{(k)}|},\cdots,\frac{p'_\lambda(|\beta_{p_n}^{(k)}|)}{|\beta_{p_n}^{(k)}|})$.

Then we can approximate (4.9) as follows:

$$(L^T + \Lambda + D(\boldsymbol{\beta}^{(k)}))\boldsymbol{\beta}^{(k+1)} = \mathbf{X}^T\mathbf{Y} + L\boldsymbol{\beta}^{(k)}. \quad (4.13)$$

Combining the local quadratic approximation method and the modified Newton-Raphson method, we can update the estimate by the following equation

$$\boldsymbol{\beta}^{(k+1)} = (L^T + \Lambda + D(\boldsymbol{\beta}^{(k)}))^{-1}(\mathbf{X}^T\mathbf{Y} + L\boldsymbol{\beta}^{(k)}). \quad (4.14)$$

Note that the matrix $L^T + \Lambda + D(\boldsymbol{\beta}^{(k)})$ is an upper diagonal matrix, so (4.13) and (4.14) can easily be solved by backward substitutions with no need to calculate

$(L^T + \Lambda + D(\boldsymbol{\beta}^{(k)}))^{-1}$. The complexity of this modified Newton-Raphson method is $O(p^2)$, which is much smaller than the complexity of the inverse matrix computation $(O(p^3))$. When dealing with ultra-high-dimensional cases, our newly modified Newton method will be much more effective if an appropriate initial estimate can be chosen for the nonconcave penalized penalty function.

4.4 Proofs

Proof of Theorem 4.1.

For simplicity, we denote $g(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\boldsymbol{\beta})^2$ and $g'(\boldsymbol{\beta}) = A\boldsymbol{\beta} - b$ where $A = \sum_{i=1}^{n} X_i^T X_i$ and $b = 2\sum_{i=1}^{n} X_i Y_i$. Next let $h(\boldsymbol{\beta}) = n\sum_{j=1}^{p_n} p_\lambda(\beta_j)$, then the original penalized least squares (4.2) can be rewritten as

$$Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + h(\boldsymbol{\beta}).$$

By the LLA, we have

$$g'(\boldsymbol{\beta}^{(k+1)}) = A\boldsymbol{\beta}^{(k+1)} - b = -h'(|\boldsymbol{\beta}^{(k)}|)\mathrm{sgn}(\boldsymbol{\beta}^{(\mathrm{k+1})}).$$

Hence $\beta^{(k+1)}$ can be shown as

$$\boldsymbol{\beta}^{(k+1)} = A^{-1}(-h'(|\boldsymbol{\beta}^{(k)}|)\mathrm{sgn}(\boldsymbol{\beta}^{(\mathrm{k+1})}) + \mathrm{b}),$$

so we have

$$\begin{aligned}\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)} &= A^{-1}\{-h'(|\boldsymbol{\beta}^{(k)}|)\mathrm{sgn}(\boldsymbol{\beta}^{(\mathrm{k+1})}) + \mathrm{b} - \mathrm{A}\boldsymbol{\beta}^{(\mathrm{k})}\} \\ &= (-A^{-1})\{h'(|\boldsymbol{\beta}^{(k)}|)\mathrm{sgn}(\boldsymbol{\beta}^{(\mathrm{k+1})}) - \mathrm{h}'(|\boldsymbol{\beta}^{(\mathrm{k-1})}|)\mathrm{sgn}(\boldsymbol{\beta}^{(\mathrm{k})})\}.\end{aligned}$$

For the SCAD penalty function and $\beta_j \geq 0$, we have

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

For simplicity, we use λ for λ_n because the sample size n has no effect on this algorithm. As $a > 2$, we have $0 < \frac{1}{a-1} < 1$, and we choose a large enough a such that the positive number $C = \frac{1}{a-1}\|nA^{-1}\|$ satisfies $0 < C < 1$.

We rewrite $h'(|\boldsymbol{\beta}^{(k)}|)\text{sgn}(\boldsymbol{\beta}^{(k+1)}) - h'(|\boldsymbol{\beta}^{(k-1)}|)\text{sgn}(\boldsymbol{\beta}^{(k)})$ as follows,

$$\begin{aligned} h'(|\boldsymbol{\beta}^{(k)}|) - h'(|\boldsymbol{\beta}^{(k-1)}|) &= n\sum_{j=1}^{p_n}\lambda\Big\{I(|\beta_j^{(k)}| \le \lambda)\text{sgn}(\beta_j^{(k+1)}) - I(|\beta_j^{(k-1)}| \le \lambda)\text{sgn}(\beta_j^{(k)}) \\ &\quad + \frac{(a\lambda - |\beta_j^{(k)}|)_+}{(a-1)\lambda} I(|\beta_j^{(k)}| > \lambda)\text{sgn}(\beta_j^{(k+1)}) \\ &\quad - \frac{(a\lambda - |\beta_j^{(k-1)}|)_+}{(a-1)\lambda} I(|\beta_j^{(k-1)}| > \lambda)\text{sgn}(\beta_j^{(k)})\Big\} \\ &\hat{=} n\sum_{j=1}^{p_n} I_j. \end{aligned} \tag{A.2}$$

First we consider the situation where $\beta_j^{(k+1)}$ and $\beta_j^{(k)}$ are both negative or positive. There are four relationships between $\beta_j^{(k)}$ and $\beta_j^{(k-1)}$.

- If $\beta_j^{(k)} \le \lambda < \beta_j^{(k-1)}$, (A.2) has the form $\frac{\beta_j^{(k-1)} - \lambda}{a-1}$. Since $\beta_j^{(k)} \le \lambda$, then we have that $|I_j| \le |(\beta_j^{(k-1)} - \lambda)| \le \frac{1}{a-1}|\beta_j^{(k)} - \beta_j^{(k-1)}|$.
- This result also holds if $\beta^{(k-1)} \le \lambda < \beta^{(k)}$.
- If $\beta^{(k-1)} \le \lambda$ and $\beta^{(k)} \le \lambda$, then we have $|I_j| = 0 \le \frac{1}{a-1}|\beta_j^{(k)} - \beta_j^{(k-1)}|$.
- If $\beta^{(k-1)} > \lambda$ and $\beta^{(k)} > \lambda$, then we have $|\beta_j^{(k+1)} - \beta_j^{(k)}| = \frac{1}{(a-1)}|\beta_j^{(k)} - \beta_j^{(k-1)}|$.

We can obtain the same result when $\beta_j^{(k+1)}\beta_j^{(k)} < 0$. From the above results, we have

$$\|\beta^{(k+1)} - \beta^{(k)}\| \le \frac{1}{a-1}\|nA^{-1}\|\|\beta^{(k)} - \beta^{(k-1)}\| = C\|\beta^{(k)} - \beta^{(k-1)}\|.$$

Hence the theorem is proved. □

Proof of Theorem 4.2

Let $L(\boldsymbol{\beta}) = \sum_{i=1}^n l_i(\boldsymbol{\beta})$. Now we define a new function R,

$$\boldsymbol{\beta}^{(1)} = \arg\max_{\beta}\ R(\boldsymbol{\beta}) = \arg\max_{\beta}\Big[L(\boldsymbol{\beta}) - n\sum_{j=1}^{p} p'_\lambda(|\beta_j^{(0)}|)|\beta_{nj}|\Big]. \tag{A.3}$$

Let $\alpha_n = p_n^{1/2}(n^{-1/2} + a_n)$ and set $\|u\| = C$, where C is a large enough constant. By Taylor expansion of the real parameter $\boldsymbol{\beta}_0$, we have

$$D(u) = R(\boldsymbol{\beta}_0 + \alpha_n u) - R(\boldsymbol{\beta}_0) \le L(\boldsymbol{\beta}_0 + \alpha_n u) - L(\boldsymbol{\beta}_0) + (II) = (I) + (II). \tag{A.4}$$

Here (II) can be rewritten as

$$(II) = -n\sum_{j=1}^{p_n} p'_\lambda(|\beta_j^{(0)}|)(|\beta_{0j} + \alpha_n u_j| - |\beta_{0j}|).$$

For (I), we have

$$\begin{aligned}(I) &= \alpha_n \nabla^T L(\boldsymbol{\beta}_0)u + \frac{1}{2}u^T\nabla^2 L(\boldsymbol{\beta}_0)u\alpha_n^2 + \frac{1}{6}\nabla^T\{u^T\nabla^2(\boldsymbol{\beta}^*)u\}u\alpha_n^3 \\ &= I_1 + I_2 + I_3,\end{aligned}$$

where $\boldsymbol{\beta}^*$ is a vector between $\boldsymbol{\beta}_0 + \alpha_n u$ and $\boldsymbol{\beta}_0$.

For I_1, we have,

$$|I_1| = |\alpha_n \nabla^T L(\boldsymbol{\beta}_0 u)| \leq \alpha_n \|\nabla^T L(\boldsymbol{\beta}_0)\|\|u\| = O_p(\alpha_n^2 n)\|u\|. \tag{A.5}$$

For I_2,

$$\begin{aligned}I_2 &= \frac{1}{2}u^T\big[\frac{1}{n}\nabla^2 L(\boldsymbol{\beta}_0) - E\nabla^2 L(\boldsymbol{\beta}_0)\big]un\alpha_n^2 - \frac{1}{2}u^T I(\boldsymbol{\beta}_0)un\alpha_n^2 \\ &= -\frac{n\alpha_n^2}{2}u^T I(\boldsymbol{\beta}_0)u + op(1)n\alpha_n^2\|u\|^2.\end{aligned} \tag{A.6}$$

For I_3, using the Cauchy-Schwarz inequality we obtain,

$$|I_3| = \Big|\frac{1}{6}\sum_{i,j,k=1}^{p}\frac{\partial L(\boldsymbol{\beta}^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}u_i u_j u_k\Big| \leq \frac{1}{6}\sum_{l=1}^{n}(p^3 C_5)^{1/2}\|u\|^3\alpha_n^3$$

As $p_n^4/n \to 0$ and $p_n^2\alpha_n \to 0$, we have,

$$I_3 = o_p(n\alpha_n^2)\|u\|^2. \tag{A.7}$$

For (II), using the Cauchy-Schwarz inequality and regular condition [A1], we have

$$\begin{aligned}|(II)| &\leq n\alpha_n\Big[\sum_{j=1}^{s_n}\Big|p'(|\beta_{nj}^{(0)}|)\mathrm{sgn}(\beta_{n0j})u_j\Big| + \sum_{s_n+1}^{p_n}\Big|p'(|\beta_{nj}^{(0)}|)\mathrm{sgn}(\beta_{n0j})u_j|I(\beta_{n0j}=0)\Big|\Big] \\ &\leq s_n^{1/2}n\alpha_n a_n\|u\| + (p_n - s_n)^{1/2}n\alpha_n a_n\|u\| \leq n\alpha_n^2\|u\|.\end{aligned} \tag{A.8}$$

Combining (A.5)-(A.8), when $\|u\|$ is large enough, all terms I_1, I_3, and (II) are dominated by I_2, which is negative. Thus, we know that $\boldsymbol{\beta}^{(1)}$ is a root-p/n-consistent estimator for $\boldsymbol{\beta}_0$.

Now, using the following lemma, we aim to prove that the one-step estimator possesses the sparsity property.

Lemma 4.1. Under regular conditions, if $\lambda_n \to 0$, $(n/p_n)^{1/2}\lambda_n \to 0$ and $p^5/n \to 0$, then for any given maximizer $\boldsymbol{\beta}^{(1)}$ satisfying $\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_0\| = O_p((p/n)^{1/2})$ and any constant C, we have

$$\max_{\beta} R\{\boldsymbol{\beta}\} = R\{(\boldsymbol{\beta}_1^{(1)T},\ \boldsymbol{\beta}_2^{(1)T})^T\} = R\{(\boldsymbol{\beta}_1^{(1)T},\ 0)^T\}$$

with a probability tending to 1.

Proof: To prove this lemma, we set $\epsilon_n = C(p_n/n)^{1/2}$. We only need to show that for any given maximizer $\boldsymbol{\beta}^{(1)}$ satisfying $\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_0\| = O_p((p/n)^{1/2})$, for $j = s_n+1, \cdots, p_n$ we have

$$\frac{\partial R(\boldsymbol{\beta}^{(1)})}{\partial \beta_j} < 0 \qquad \text{for } 0 < \beta_j < \epsilon_n, \tag{A.9}$$

and

$$\frac{\partial R(\boldsymbol{\beta}^{(1)})}{\partial \beta_j} > 0 \qquad \text{for } -\epsilon_n < \beta_j < 0. \tag{A.10}$$

By Taylor expansion of point β_{0j}, we have

$$\frac{\partial R(\beta_n)}{\partial \beta_{nj}} = I_1 + I_2 + I_3 + I_4.$$

For I_1, we have

$$I_1 = \frac{\partial L(\boldsymbol{\beta}^{(1)})}{\partial \beta_j} = O_p(n^{1/2}) = O_p(n{p_n}^{1/2}). \tag{A.11}$$

We also have

$$I_4 = -np'_\lambda(|\beta_j^{(0)}|)\mathrm{sgn}(\beta_j^{(1)}). \tag{A.12}$$

For I_2, we have

$$\begin{aligned} I_2 \ &= \ \sum_{l=1}^{p_n} \frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_l}(\beta_l^{(1)} - \beta_{0l}) = \sum_{l=1}^{p_n} \Big\{ \frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_l} - \mathrm{E}\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_l} \Big\}(\beta_l^{(1)} - \beta_{0l}) \\ &\quad + \sum_{l=1}^{p_n} \mathrm{E}\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_l}(\beta_l^{(1)} - \beta_{0l}^{(1)}) = K_1 + K_2. \end{aligned}$$

Using the Cauchy-Schwarz inequality and by $\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_0\| = O_p(p_n/n^{1/2})$, we have

$$|K_2| = \Big| n\sum_{l=1}^{p_n} I_n(\beta_{n0})(j,l)(\beta_l^{(1)} - \beta_{0l}) \Big| \le nO_p\Big((\frac{p_n}{n})^{1/2}\Big)\Big\{ \sum_{l=1}^{p_n} I_n^2 \Big\}^{1/2}.$$

Because the eigenvalues of the information matrix are bounded according to regular conditions, we have,

$$\sum_{l=1}^{p_n} I_n^2 = O_p(1),$$

and so

$$K_2 = O_p((np_n)^{1/2}).$$

Using the same method we can have

$$K_1 = O_p((np_n)^{1/2}).$$

Thus, we conclude that

$$I_2 = O_p((np_n)^{1/2}). \tag{A.13}$$

Again, split I_3 as follows:

$$\begin{aligned} I_3 &= \sum_{l,k=1}^{p_n} \frac{\partial^3 L(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_l\partial\beta_k}(\beta_l^{(1)} - \beta_{0l})(\beta_k^{(1)} - \beta_{0k}) \\ &= \sum_{l,k=1}^{p_n} \Big\{ \frac{\partial^3 L(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_l\partial\beta_k} - \mathrm{E}\frac{\partial^3 L(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_l\partial\beta_k} \Big\}(\beta_l^{(1)} - \beta_{0l})(\beta_k^{(1)} - \beta_{0k}) \\ &\quad + \sum_{l=1}^{p_n} \mathrm{E}\frac{\partial^3 L(\boldsymbol{\beta}^*)}{\partial\beta_j\partial\beta_l\partial\beta_k}(\beta_l^{(1)} - \beta_{0l})(\beta_k^{(1)} - \beta_{0k}) \\ &= K_3 + K_4. \end{aligned}$$

For K_4, under regular conditions we have

$$|K_4| \le C^{1/2} n p_n \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_0\|^2 = O_p(p_n^2) = o_p((np_n)^{1/2}). \tag{A.14}$$

By the Cauchy-Schwarz inequality, and following the same technique as above, under regular conditions we have

$$K_3 = O_p\Big\{ \Big(n p_n^2 \frac{p_n^2}{n} \Big)^{1/2} \Big\} = o_p((np_n)^{1/2}). \tag{A.15}$$

Combining (A.11), (A.13), (A.14), and (A.15), we know that

$$I_1 + I_2 + I_3 = O_p((np_n)^{1/2}).$$

However, as we know that $(p_n/n)^{1/2}/\lambda_n \to 0$, and $p'(\theta)/\lambda_n > 0$, combined with the regular condition [A1]–[A3] we have,

$$\frac{\partial R(\boldsymbol{\beta}^{(1)})}{\partial\beta_j} = n\lambda_n \Big\{ - \frac{p'(|\beta_j^{(0)}|)}{\lambda_n} + O_p\Big((\frac{p_n}{n})^{1/2}/\lambda_n \Big) \Big\} + O(1).$$

It is easy to see that the sign of $\beta_j^{(0)}$ determines the sign of $\frac{\partial R(\boldsymbol{\beta}^{(1)})}{\partial\beta_j}$. By the regular conditions [A1]-[A3] for the initial estimate $\boldsymbol{\beta}^{(0)}$, we know that (A.9) and (A.10) are satisfied. Thus, we prove the above lemma. □

As shown above we know that there is a root-(n/p_n)-consistent local maximizer $\boldsymbol{\beta}^{(1)}$ for $R(\boldsymbol{\beta})$. Thus from the above lemma we know that the first result of Theorem 4 has been proved. Now we prove the second result, the asymptotic normality of the one-step estimate.

We let $R(\boldsymbol{\beta}^{(1)}) = R(\boldsymbol{\beta}_1^{(1)}, 0)$. As $\boldsymbol{\beta}^{(1)}$ is a local maximizer of R, we have $\nabla R(\boldsymbol{\beta}_1^{(1)}) = 0$, so by the Taylor expansion of $\boldsymbol{\beta}_{01}$, we can get

$$\begin{aligned} & \frac{1}{n}\Big[\{\nabla^2 L(\boldsymbol{\beta}_{01})\}(\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_{01}) - n\sum_{j=1}^{p_n} p'_\lambda(|\beta_j^{(0)}|)\mathrm{sgn}(\beta_{\mathrm{j}}^{(1)})\Big] \\ = \; & \frac{1}{n}\Big[\nabla L(\boldsymbol{\beta}_{01}) + \frac{1}{2}(\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_{01})^T\nabla^2\{\nabla L(\boldsymbol{\beta}_1^*)\}(\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_{01})\Big]. \end{aligned}$$

Here, $\boldsymbol{\beta}_1^*$ lies between $\boldsymbol{\beta}_1^{(1)}$ and $\boldsymbol{\beta}_{01}$. Let $L_1 = \nabla^2 L(\boldsymbol{\beta}_{01})$ and $L_2 = \frac{1}{2}(\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_{01})^T\nabla^2\{\nabla L(\boldsymbol{\beta}_1^*)\}(\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_{01})$. Under regular conditions and by the Cauchy-Schwarz inequality, we get

$$\Big\|\frac{1}{n}L_2\Big\|^2 \le \frac{1}{n^2}\sum_{i=1}^{n} n^2\|\boldsymbol{\beta}_{n1}^{(1)} - \boldsymbol{\beta}_{n01}\|^4 \sum_{j,k,l=1}^{s_n} C_5 = O_p\Big(\frac{p_n^2}{n}\Big)O_p(p_n^3) = o_p(\frac{1}{n}). \quad \text{(A.16)}$$

In a similar way, for any ϵ we have

$$\begin{aligned} P\Big(\Big\|\frac{1}{n}\nabla^2 L(\boldsymbol{\beta}_{01}) + I(\boldsymbol{\beta}_{01})\Big\| \ge \frac{\epsilon}{p_n}\Big) \; & \le \; \frac{p_n^2}{n^2\epsilon^2}E\sum_{i,j=1}^{s_n}\Big\{\frac{\partial L(\boldsymbol{\beta}_{01})}{\partial\beta_i\partial\beta_j} - rE\frac{\partial L(\boldsymbol{\beta}_{01})}{\partial\beta_i\partial\beta_j}\Big\}^2 \\ & = \; \frac{p_n^4}{n} = o(1). \end{aligned}$$

Hence we get

$$\Big\|\frac{1}{n}\nabla^2 L(\boldsymbol{\beta}_{01}) + I(\boldsymbol{\beta}_{01})\Big\| = o_p(1/p_n). \quad \text{(A.17)}$$

Under regular conditions, we know that

$$\lambda_i\Big\{\frac{1}{n}L_1 + I(\boldsymbol{\beta}_{01})\Big\} = o_p\Big(\frac{1}{p_n^{1/2}}\Big),$$

where λ_i is the ith eigenvalue of a symmetrical matrix A_n. Because $\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_{01} = O_p((p_n/n)^{1/2})$,

$$\Big\{\frac{1}{n}L_1 + I(\beta_{01})\Big\}(\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_{01}) = O_p\Big(\frac{1}{n^{1/2}}\Big). \quad \text{(A.18)}$$

Then by combining (A.17) and (A.18), we know that

$$I(\boldsymbol{\beta}_{01})(\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_{01}) = \frac{1}{n}\nabla L(\boldsymbol{\beta}_{01}) + o_p\Big(\frac{1}{n^{1/2}}\Big). \quad \text{(A.19)}$$

Then let $Y_{ni} = \frac{1}{n^{1/2}} A_n I^{-1/2}(\boldsymbol{\beta}_{01}) \nabla L(\boldsymbol{\beta}_{01})$. Under regular conditions and some routine calculations, we know that Y_{ni} satisfies the conditions of the Lindeberg-Feller central limit theorem. Thus we finish the proof of Theorem 4.2. $\square$

Chapter 5

Numerical Study of High-dimensional Varying Coefficient Models

5.1 Implementation of Group-SCAD Method

The biggest problem with the SCAD penalty is the non-differentiability of the penalty function p_{λ_n}. To overcome this problem, in this section we apply the LQA algorithm proposed by Fan and Li (2001). As shown in Chapter 2, we use the following iteration,

$$\begin{aligned}\beta^{(k+1)} &= \arg\min_{\beta}\Big[\sum_{k=1}^{n}\frac{1}{n}\sum_{i=1}^{n}\Big[y_j - X_{ik}^{(1)}\boldsymbol{\beta}_k^{(1)} - X_{ik}^{(2)}\boldsymbol{\beta}_k^{(2)}\Big]^2 K_h(U_i - U_k) \\ &\quad + n\sum_{j=1}^{p_n}\frac{p_\lambda'(|\beta_j^{(k)}|)}{2|\beta_j^{(k)}|}\beta_j^2\Big]. \qquad (5.1)\end{aligned}$$

We outline the algorithm as follows:

- Step 1: initialize $\boldsymbol{\beta}^{(1)}$ by using (5.1) with $\lambda = 0$, setting $k = 0$.
- Step 2: solve $\boldsymbol{\beta}_{k+1}$ from $\boldsymbol{\beta}^{(k)}$ by using (5.1) with the predefined λ; and
- Step 3: iterate Step 2 until convergence of β.

If we apply the group-SCAD algorithm, another important problem is the selection of the tuning parameter λ. The procedure described above will give an estimator $\hat{\boldsymbol{\beta}}$.

The number of varying coefficients identified by $\boldsymbol{\beta}$ is denoted by df_λ, while $p_n - df_\lambda$ is the number of non-varying coefficients. Thus we have the corresponding RSS_λ as follows:

$$\mathrm{RSS}_\lambda \;=\; \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_j - X_{ijk}^{(1)}\boldsymbol{\beta}_k^{(1)} - X_{ijk}^{(2)}\boldsymbol{\beta}_k^{(2)}\right]^2 K_h(U_i - U_k). \tag{5.2}$$

Then we use a BIC-type criterion to select λ

$$\mathrm{BIC}_\lambda \;=\; \log(\mathrm{RSS}_\lambda) + df_\lambda \times \frac{\log(nh)}{nh} + (p_n - df_\lambda) \times \frac{\log(n)}{n}. \tag{5.3}$$

Finally, the tuning parameter can be chosen as follows:

$$\hat{\lambda} \;=\; \arg\min_{\lambda} \mathrm{BIC}_\lambda.$$

As shown by Fan and Li (2001), 3.7 is a good choice for the parameter a used in the SCAD. Hence, we mainly use the above method to select the penalized parameter λ for the SCAD penalized regression. In all the following simulations and discussions, we always set the value of a as 3.7.

Because we use a local linear approximation for those varying coefficient $\beta(u)$, we need to choose the bandwidth h. Fan and Zhang (1999) pointed out that the optimal rate cannot be achieved using the one-step estimator, thus we should use two-step estimation. Another advantage of the two-step estimation is that it is not very sensitive to the choice of initial bandwidth h_0, if h_0 is small enough so that the bias in the first-step smoothing is negligible. In the second step, the problem is really a univariate smoothing problem. Therefore, we can apply univariate bandwidth selection procedures such as CV or GCV to select the smoothing parameter, which makes the bandwidth selection problem easier. In our study, we use this two-step method for choosing bandwidth h. In the first step, we get a one-step estimator $\hat{\boldsymbol{\beta}}_{\mathrm{os}}$ by minimizing the non-penalized function

$$\sum_{k=1}^{n}\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{p_n}\left\{a_{jk} + b_{jk}\frac{U_i - U_k}{h}\right\}x_{ij}\right]^2 K_h(U_i - U_k).$$

Based on the result, we use use cross-validation(CV) or generalized cross-validation(GCV) to select the bandwidth for the one-step estimator. Finally we use $h_0 = \frac{1}{2}\hat{h}$ as the initial bandwidth.

5.1.1 Simulation Examples

Example 1. Consider the following three varying coefficient models:

$$(I)\ Y_i = 2\sin(2\pi Z_i)X_{i1} + 4Z_i(1-Z_i)X_{i2} + \sigma\varepsilon_i,$$
$$(II)\ Y_i = \exp(2Z_i - 1)X_{i1} + 8Z_i(1-Z_i)X_{i2} + 2\cos^2(2\pi Z_i)X_{i3} + \sigma\varepsilon_i,$$
$$(III)\ Y_i = 4Z_iX_{i1} + 2\sin(2\pi Z_i)X_{i2} + X_{i3} + \sigma\varepsilon_i,$$

where $X_{i1} = 1$ and $(X_{i2}, \ldots, X_{ip})^T$ is generated from a multivariate normal distribution with $\text{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$ for any $2 \leq j_1, j_2 \leq p$. ε_i is simulated from $N(0,1)$; the index variable is simulated from uniform $[0,1]$. $n = 100, 200, 400$ and $p = 7, 10$.

For Model (I), only the first two variables are relevant and for both Models (II) and (III), the first three are relevant. The value of σ is given by 1.5. A total of 1,000 simulation replications are conducted for each model setup. For every simulated dataset, we first fit an unpenalized varying coefficient estimate using the two-step estimation method proposed in Fan and Zhang (1999). We use the Epanechnikov kernel in all of our approximations of varying coefficients. We compute the bandwidth h_0 by the CV method. This bandwidth is then used in the SCAD, where the optimal tuning parameter is determined by the BIC criterion mentioned above.

To summarize the simulation results, whenever the estimated model misses at least one relevant predictor, we classify it as an underfitted model. and whenever the estimated model includes at least one irrelevant predictor but does not miss any relevant predictors, we classify it as an overfitted model. Whenever the resulting model is exactly the same as the true model, we classify it as a correctly fitted model. Then, the percentage of experiments with correctly (under, over) fitted models is summarized. We also report the average number of correctly and incorrectly identified 0 coefficients. The average number correctly identified is restricted to the true zero coefficients. The incorrect number depicts the average number of coefficients erroneously set to 0. To evaluate the estimation accuracy we use the relative estimation error (REE):

$$REE \ = \ 100 \ \times \ \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}|\hat{\beta}_j(u_i) \ - \ \beta_{0j}(u_i)|}{\sum_{i=1}^{n}\sum_{j=1}^{p}|\bar{\beta}_j(u_i) \ - \ \beta_{0j}(u_i)|}.$$

where $\bar{\beta}$ is either the unpenalized estimator or the oracle estimator while $\hat{\beta}$ is our SCAD penalized estimator. The corresponding REE value measures the estimation accuracy of $\hat{\beta}$ to that of $\bar{\beta}$. For each model and parameter setting, the median of REE values (denoted as MREE) are summarized. If we take $\hat{\beta}$ as our SCAD estimator and $\bar{\beta}$ is the unpenalized estimator, then if the MREE value is smaller than 1, it means our SCAD method is more accurate than the unpenalized estimator. The smaller this percentage is, the higher the accuracy of our estimator.

Figure 1-3 shows a typical result that under the dimensional condition $p = 10$. The true curve and the estimated curve for the varying coefficients of $(I), (II)$, and (III) are depicted by a dot-dash line for the true function and a solid line for the estimated function. Here, $\hat{a}_{j0}$ and $\hat{a}_{j1}$ are chosen as, the median in all of our simulation results. We can see from these figures that the true function and the estimated function are very close to each other, implying that our SCAD estimator has a satisfactory estimation error.

All of the relevant results are presented in Tables 5.1 and 5.2. Clearly, the MREE ratio of the penalized estimator to the unpenalized estimator is much less than 1 in all cases, thus indicating that the SCAD estimator is much more accurate than the unpenalized estimates. As the dimension p increases, especially when $p = 10$ and $n = 100$, the value of the MREE increases; however, as p is less than n, our penalized estimators are still much more accurate than the unpenalized estimates. As we can see from Table 5.2 when $n = 200$ and $n = 400$ it is clear that our SCAD estimator still has a very satisfactory model selection ability. For every fixed p, the percentage of correctly fitted models steadily increases as the sample size increases, and quickly approaches 100%, thus confirming that our SCAD estimator can indeed identify the true model consistently. The MREE ratio of the penalized estimator to the oracle estimator quickly approaches 100%, which corroborates the oracle properties of the SCAD penalized estimator.

n	Number of estimated zeros		Percentage of models			MREE (%)	
	Correct	Incorrect	Under	Correct	Over	PE/UPE	PE/OE
			Model I				
100	4.81	0.08	0.07	0.78	0.15	40.22	120.86
200	4.98	0.01	0.01	0.95	0.03	36.87	114.13
400	5.00	0	0	1.00	0	34.59	109.95
			Model II				
100	3.82	0.02	0.03	0.86	0.11	56.03	108.39
200	3.99	0	0.01	0.98	0.01	52.38	107.52
400	4.00	0	0	1.00	0	49.61	106.82
			Model III				
100	3.87	0.01	0.03	0.89	0.07	53.77	114.95
200	3.99	0	0.01	0.97	0.01	49.38	110.03
400	4.00	0	0	1.00	0	47.62	106.28

Table 5.1: The simulation results of Example 1 with $p = 7$.

n	Number of estimated zeros		Percentage of models			MREE (%)	
	Correct	Incorrect	Under	Correct	Over	PE/UPE	PE/OE
			Model I				
100	7.08	0.09	0.15	0.63	0.22	76.57	137.15
200	7.82	0	0.02	0.90	0.07	42.26	124.38
400	7.95	0	0	0.95	0.05	37.49	119.57
			Model II				
100	6.29	0.07	0.12	0.70	0.18	63.36	129.86
200	6.81	0.02	0.02	0.91	0.06	44.32	117.38
400	6.96	0	0	0.96	0.04	39.38	113.74
			Model III				
100	6.21	0.06	0.11	0.73	0.16	51.92	128.53
200	6.85	0.02	0.03	0.90	0.07	40.87	116.62
400	6.97	0	0.01	0.97	0.02	36.21	112.26

Table 5.2: The simulation results of Example 1 with $p = 10$.

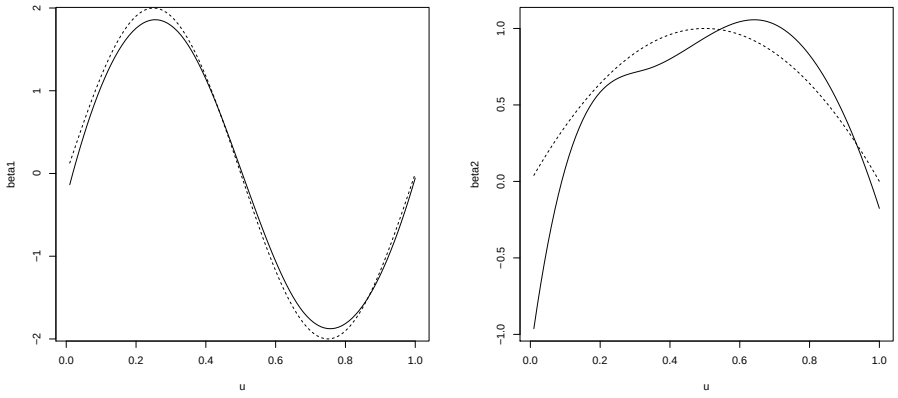

Figure 1: The varying coefficients of (I)

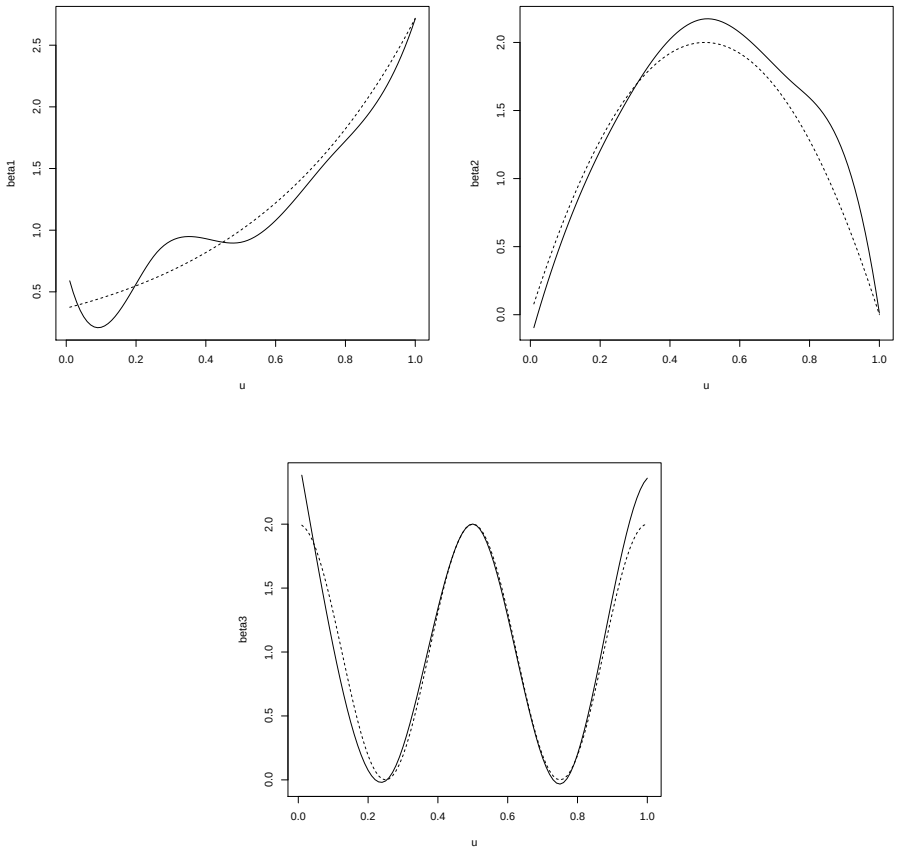

Figure 2: The varying coefficients of (II)

Example 2. We generate random samples with $p = 7, 10$ and $n = 100, 200, 400$ from

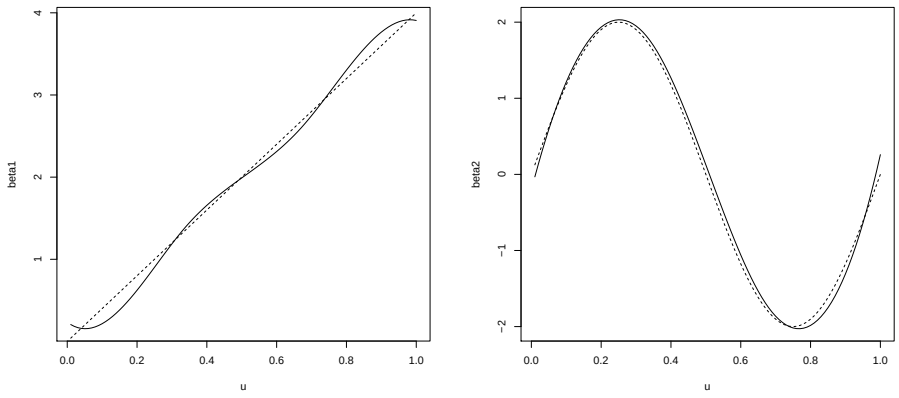

Figure 3: The varying coefficients of (III)

the following three models:

$$\begin{aligned}
(I)\ Y_i &= 2\sin(2\pi Z_i)X_{i1} + 4Z_i(1-Z_i)X_{i2} \\
&\quad +0.5X_{i4} + 0.5X_{i5} + X_{i6} + 0.1X_{i7} + \sigma\varepsilon_i, \\
(II)\ Y_i &= 3\sin(2\pi Z_i)X_{i1} + 8Z_i(1-Z_i)X_{i2} + \cos^2(2\pi Z_i)X_{i3} \\
&\quad +X_{i4} + 0.5X_{i5} + X_{i6} - 0.5X_{i7} + \sigma\varepsilon_i, \\
(III)\ Y_i &= 3Z_iX_{i1} + 2\sin(2\pi Z_i)X_{i2} + 15Z_i(1-Z_i)X_{i3} \\
&\quad +X_{i4} - X_{i5} + X_{i6} + \sigma\varepsilon_i,
\end{aligned}$$

where $X_{i1} = 1$ and $(X_{i2}, \ldots, X_{ip})^T$ is generated from a multivariate normal distribution with $\mathrm{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$ for any $2 \le j_1, j_2 \le p$; and ε_i is simulated from $N(0,1)$. The value of σ is given by 0.5. The index variable is simulated from either uniform $[0,1]$. A total of 200 simulation replications are conducted for each model setup.

We use the same algorithm as in Example 1. To evaluate the estimation accuracy of the proposed procedure, we again consider the REE mentioned above. We also collect determine the number of experiments with correct model identifications, i.e., the identified model has the same varying coefficient terms and invariant coefficient terms as the true model. As all three models have constant coefficients, we separately summarize the ratio of correctly estimated constants. Again, we use the REE for the constant parts, and we list the MREE ratio of constant penalized estimators (CPE) to constant oracle estimators (COE).

Figure 1-3 shows a typical result under the dimensional condition $p = 10$. The true curve and the estimated curve for the varying coefficients of $(I), (II)$, and (III) are depicted by a dot-dash line for the true function and a solid line for the estimated function. Here, $\hat{a}_{j0}$ and $\hat{a}_{j1}$ are chosen as the median in all of our simulation results. Again, we can see from these figures that the true function and the estimated function are very close to each other, thus implying that our SCAD estimator has a satisfactory estimation error.

All of the relevant results are summarized in Tables 5.3, 5.4, and 5.5. It is clear that as the dimension p increases, all of MREE ratios also increase. However, as long as our dimension p_n satisfies certain limitations, most of these MREE ratios of the penalized estimator to the unpenalized estimator are still much less than 100%. This means that the proposed SCAD estimates are much more efficient than the unpenalized estimates. The correct frequency and MREE value seems to fall significantly when $p = 10,\ n = 100$, this is due to the small sample size. When n increases, our SCAD estimator still works efficiently. As before, the number of experiments with correct model identifications steadily increases as the sample size increases and quickly approaches 100%, which confirms that our SCAD method can indeed identify the true model consistently. The ratio of correctly estimated constant coefficients is also increases to 100% quickly. From this result, we can conclude that our local polynomial approximation and SCAD penalty method can estimate the varying part and the constant part simultaneously. Moreover, we find that the MREE ratio of the penalized estimator to the oracle estimator quickly approaches 100%, which corroborates the oracle properties of the proposed SCAD estimator.

For the constant part, we list the median of the estimators for $\beta_j,\ j = 4, 5, 6, 7$ when $n = 200$ and $n = 400$ in Table 5.5. As not all of the estimators for these β_j are constant, we let the value of those estimators that are not constants be positive ∞. Then, we collect these constant estimations and list the median of the estimators and the median of the standard errors from these 200 simulation results of $\hat{\beta}_j,\ j = 4, 5, 6, 7$. The results show that our SCAD penalty method has a good estimation performance.

5.2 Numerical Study for Ultra-high-dimensional Varying Coefficient Models

5.2.1 Implementation of NIS-Group-SCAD Method

It is not a good choice to apply the NIS method directly, because as Fan and Lv (2008) point out, in practice the NIS will still suffer from false negatives (i.e., miss some

n	Model	Correct identification frequency		MREE (%)		
		Overall	Constant	PE/UPE	PE/OE	CPE/COE
100	Model I	0.63	0.61	52.1	113.7	112.9
200	Model I	0.94	0.92	48.7	101.8	102.1
400	Model I	0.99	0.99	45.9	100.4	100.5
100	Model II	0.69	0.65	60.2	115.2	115.7
200	Model II	0.96	0.94	56.6	109.1	108.6
400	Model II	0.99	0.99	52.4	106.3	105.1
100	Model III	0.76	0.73	64.9	124.6	122.5
200	Model III	0.97	0.95	61.1	112.8	113.2
400	Model III	1.00	0.99	55.8	102.5	103.1

Table 5.3: The simulation results of Example 2 with $p = 7$.

n	Model	Correct identification frequency		MREE (%)		
		Overall	Constant	PE/UPE	PE/OE	CPE/COE
100	Model I	0.37	0.33	63.4	128.5	137.2
200	Model I	0.68	0.67	51.2	110.4	112.8
400	Model I	0.92	0.89	48.3	104.7	105.3
100	Model II	0.41	0.36	69.7	131.8	133.4
200	Model II	0.71	0.69	58.3	114.1	115.6
400	Model II	0.95	0.93	53.5	108.9	106.9
100	Model III	0.52	0.48	73.2	135.2	139.1
200	Model III	0.80	0.78	62.8	118.3	116.9
400	Model III	0.97	0.97	56.9	109.7	107.5

Table 5.4: The simulation results of Example 2 with $p = 10$.

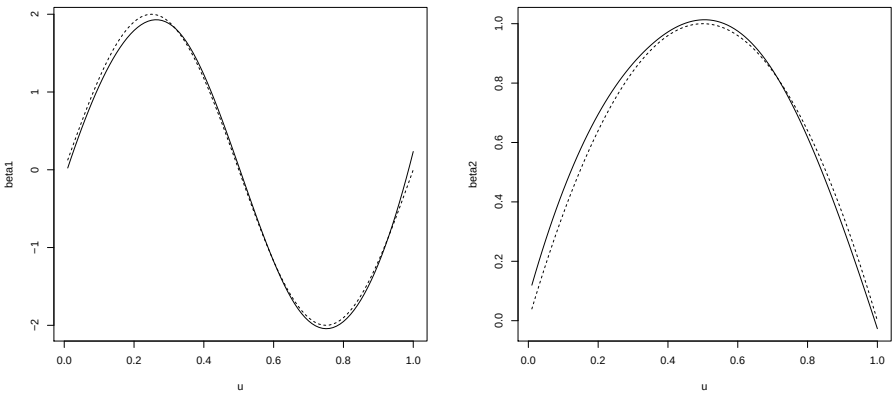

Figure 4: The varying coefficients of (I)

important predictors that are marginally weakly correlated but jointly correlated with the response), and false positives (i.e., select some unimportant predictors that are highly correlated with the important ones). Thus, a natural way to solve these disadvantages in practice is to introduce a two-step iterative algorithm. The first step involves large-scale variable screening (NIS) and the second step involves moderate-

p	n	Model	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$
7	200	Model I	0.45(56)	0.44(61)	0.94(75)	0.07(51)
7	400	Model I	0.49(33)	0.48(38)	0.98(48)	0.09(29)
7	200	Model II	0.93(63)	0.45(57)	0.92(70)	-0.43(64)
7	400	Model II	0.97(41)	0.49(32)	0.97(47)	-0.48(38)
7	200	Model III	0.92(62)	-0.91(68)	0.93(59)	-
7	400	Model III	0.98(39)	-0.96(44)	0.97(41)	-
10	200	Model I	0.38(77)	0.41(80)	0.86(94)	0.04(65)
10	400	Model I	0.46(52)	0.47(56)	0.94(71)	0.08(44)
10	200	Model II	0.85(84)	0.38(76)	0.87(87)	-0.39(71)
10	400	Model II	0.94(61)	0.45(58)	0.95(65)	-0.46(52)
10	200	Model III	0.84(85)	-0.83(93)	0.86(89)	-
10	400	Model III	0.93(57)	-0.92(65)	0.94(62)	-

Table 5.5: Median values and median standard deviations (multiplied by 1000) of constant coefficient estimators

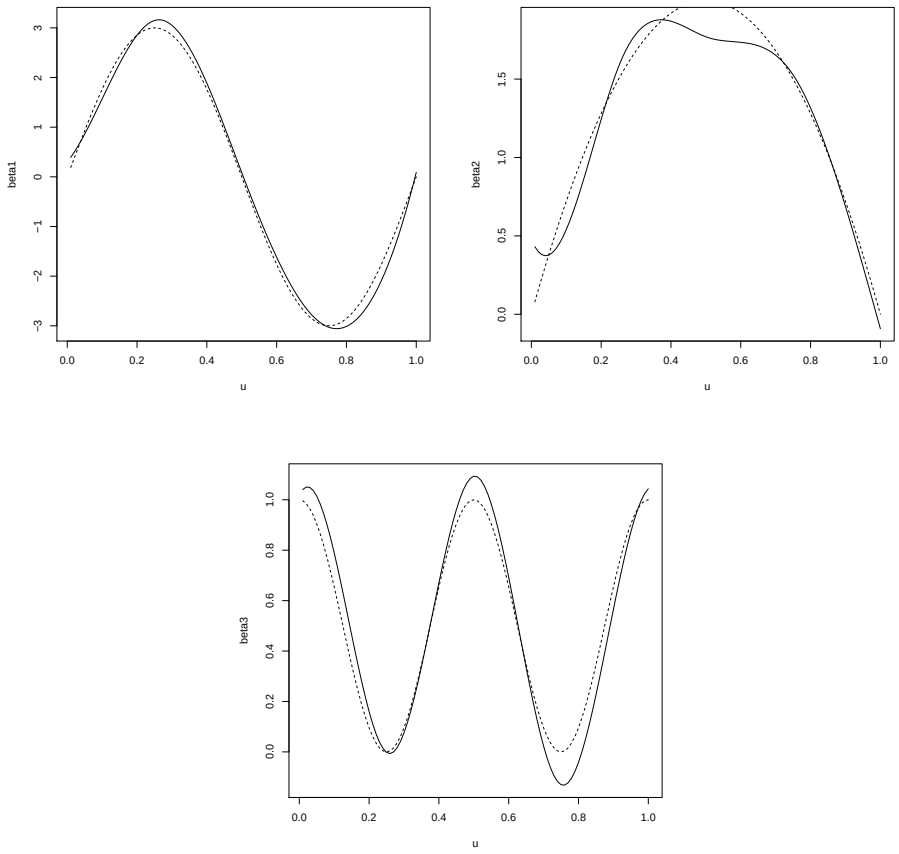

Figure 5: The varying coefficients of (II)

scale variable selection in which we use the above mentioned group-SCAD method.

Fan, Feng, and Song (2011) constructed the INIS algorithm for the NIS method. In step 2 they used a penGAM (Meier, Geer, and Bhlmann, 2009) to select a subset. In their paper they commented that in step 2, any variable selection method for additive models, such as SpAM (Ravikumar et al. 2009) and the adaptive group LASSO for the additive models of Huang, Horowitz, and Wei (2010), would work. Hence we can use the group-SCAD in step 2 rather than the penGAM method. We choose the same truncation term $d_n = O(n^{1/5})$. Given data (X_i, Y_i), we have the following algorithm:

First, we need to calculate

$$\hat{a}_{nj}(u_i) = (\psi_j(u_i))^T \; E[(\psi_j(u_i)X_j)(\psi(u_i)_j X_j)^T]^{-1} \; E((\psi_j(u_i))X_j)Y.$$

We randomly permute the rows of X to get $\tilde{X}$ and let $\omega(q)$ be the qth quantile of

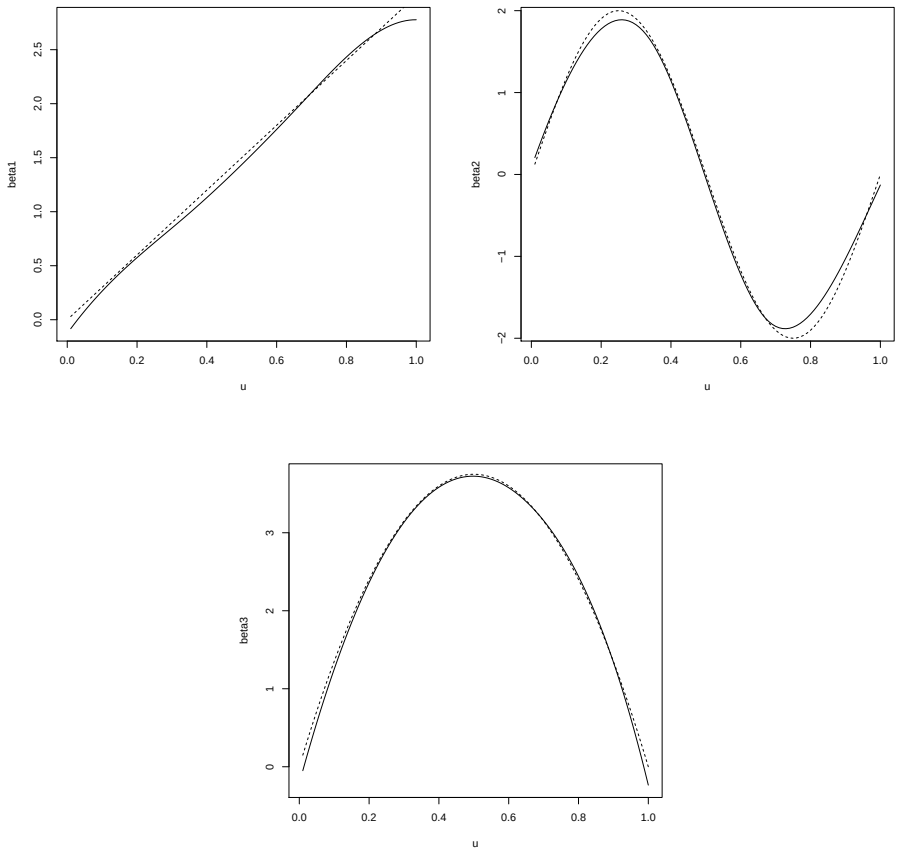

Figure 6: The varying coefficients of (III)

$\|\hat{a}^*_{nj}(u_i)\|^2_n$, where

$$\hat{a}^*_{nj}(u_i) = \text{argmin } \mathrm{P}_n(Y - a_{nj}(\tilde{X}))^2.$$

Then, we use NIS to select the following variables,

$$A_1 \;=\; \Big\{j : \; \|\hat{a}^*_{nj}\|^2_n \geq \omega(q)\Big\}.$$

Here, $q = 1$.

Second, we apply the group-SCAD method to the set A_1 to select a subset M_1. We use the LQA-SCAD method for this optimization problem.

Third, for every $j \in M_1^c$, minimize

$$\mathrm{P}_n\Big(Y - \sum_{i \in M_1} a_{ni}(X_1) - a_{nj}(X_j)\Big)^2,$$

with respect to a_{ni}. This regression reflects the additional contribution of the jth components, conditional on the existence of the variable set M_1. After marginally

screening, as in the first step, we choose a set of indices as A_2. The determination of size is the same as in Step 1, except that only the variables not in M_1 are randomly permuted. Now, we use the group-SCAD on the set $M_1 \cup A_2$.

Finally, iterate the process until $|M_l| > s_0$ or $|M_l = M_{l+1}|$.

After this INIS step, we can use group-SCAD method on the finally selected model. Then following the same algorithm in section 5.1, we can complete the variable selection in ultra-high dimensional varying coefficient models.

5.2.2 Simulation Example

Consider the following varying-coefficient model

$$Y_i = \sin(2\pi\mu)X_{i1} + ((2\mu-1)^2 + 0.5)X_{i2} + (\exp(2\mu-1)-1)X_{i3} + \varepsilon_i, \quad , i = 1, \ldots, 200.$$

where $(X_{i1}, \ldots, X_{ip})^T$ is generated from a multivariate normal distribution with $\text{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$, the index variables U_i are generalized from a uniform [0,1], and ε_i is simulated from $N(0, 1)$.

It is clear that only the first three covariates are relevant for predicting the response variable, and the rest are null variables and do not contribute to the model prediction. Now we consider the model with $p = 400, 800, 1200$ to examine the model selection performance and estimation when p exceeds or p is far larger than the sample size. To assess the estimation accuracy of the penalized methods, we also consider the oracle model. The oracle model only contains the first three relevant variables and is only available in simulation studies where the true information is known.

Figure 7 shows a typical result under the dimensional condition $p = 1200$, with the estimated curve and the true curve for the varying coefficients of the above model. Here, $\hat{a}_{j0}$ and $\hat{a}_{j1}$ are chosen as the median of all of our simulation results.

Table 5.6 summarizes the simulation results. It gives the relative total averaged integrated squared errors (RTAISEs) of the NIS-SCAD estimator to the oracle estimator. It also reports the percentage of correct fitting (C), underfitting (U) and overfitting (O) over 200 simulation runs for the penalized methods.

From these results, we can conclude that when dimension p is very large, we are dealing with ultra high dimensional problems, and our NIS-SCAD method can still

select the variables we need with satisfactory accuracy as the RTAISEs is not far from 1. Comparing our results with those of Xue and Qu (2012) when dimension p equals 400, we can see that our NIS-SCAD method is more accurate. Further, the result with $n = 400$ and $n = 200$, confirms that the false selection rate converges to 0 as n increases.

p	RTAISE	C	U	O
	n=200			
400	1.61	0.85	0.01	0.13
800	1.68	0.81	0.01	0.17
1200	1.77	0.75	0.03	0.21
	n=400			
400	1.49	0.92	0.01	0.06
800	1.57	0.87	0.03	0.10
1200	1.66	0.79	0.04	0.15

Table 5.6: The simulation results with $p = 400,\ 800,\ 1200$ and $n = 200,\ 400$.

5.3 Real Case study

The Fifth National Bank of Springfield faced a gender discrimination suit*. The charge was that its female employees received substantially lower salaries than its male employees. The banks employee database (based on 1995 data) is listed in Albright, Winston and Zappe (1999). For each of its 208 employees the data set includes the following variables:

- EduLev: education level, a categorical variable with categories 1 (finished high school), 2 (finished some college courses), 3 (obtained a bachelors degree), 4 (took some graduate courses), 5 (obtained a graduate degree).
- JobGrade: a categorical variable indicating the current job level, with IC6 possible levels (6 highest).

*This example and the accompanying data set are based on a real case. Only the banks name has been changed, according to Example 11.3 of Albright, Winston and Zappe (1999).

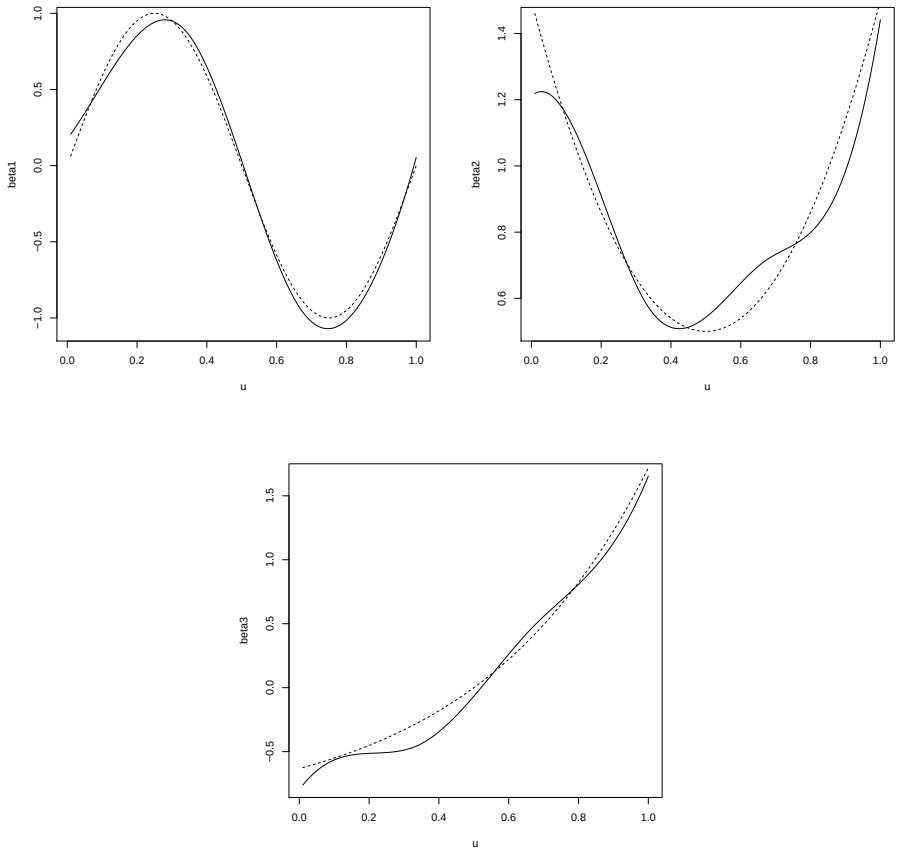

Figure 7: The varying coefficients of an ultra-high dimensional model

- YrHired: year that an employee was hired.
- YrBorn: year that an employee was born.
- Gender: a categorical variable with values Female and Male.
- YrsPrior: number of years of work experience at another bank prior to working at the Fifth National Bank.
- PCJob: a dummy variable that takes the value 1 if the empolyees current job is computer related and 0 otherwise.
- Salary: current (1995) annual salary in thousands of dollars.

A naive comparison of the average salaries of males and females will not work, as there are many confounding factors that affect salary. Our main interest is to provide, after adjusting the contributions from confounding factors, a good estimate

for the average salary difference between male and female employees. Hence it is very reasonable to build a large statistical varying coefficient model to reduce possible modeling biases. In building such a model the estimability of parameters is a factor in choosing the number of parameters, which depends on the sample size.

We set

$$\begin{aligned} \text{Salary} &= \beta_0(\text{YrsExp}) + \beta_1(\text{YrsExp})(\text{Female}) + \beta_2(\text{YrsExp})\text{PCJob} \\ &\quad + \sum_{i=1}^{4} \beta_{2+i}(\text{YrsExp})\text{Edu}_i + \sum_{i=1}^{5} \beta_{6+i}(\text{YrsExp})\text{JobGrd}_i \\ &\quad + \beta_{12}(\text{YrsExp})(\text{Age}) + \varepsilon. \end{aligned} \tag{5.4}$$

We rewrite the above model as

$$\begin{aligned} \text{Salary} &= \beta_0(u) + \beta_1(u)X_1 + \beta_2(u)X_2 + \sum_{i=1}^{4} \beta_{2+i}(u)X_i \\ &\quad + \sum_{i=1}^{5} \beta_{6+i}(u)X_i + \beta_{12}(u)X_{12} + \varepsilon, \end{aligned} \tag{5.5}$$

where the variable YrsExp is the years of working experience, computed from the variables YrHired and YrsPrior.

In this case, we apply the group-SCAD described in Section 5.1 for the varying coefficient model regression. The bandwidth and tuning parameter selection follow the same procedure as in section 5.1. To apply the penalized likelihood method, first the index variable YrsExp $=$ u is transformed so that its marginal distribution is $[0, 1]$.

Our method shows that the variables "Edu1", "Edu4" and "Age" have coefficients of 0. Thus the model is actually a 9-dimensional model. For the remaining variables, "PCJob", "JobGrd1", "JobGrd3" have constant coefficients, while "Female", "Edu2", "Edu3", "JobGrd2", "JobGrd4", and "JobGrd4" have varying coefficients. Figure 8-10 shows the estimations of these varying coefficients.

The coefficient of the variable "Female" is a varying coefficient and for most of the time its value is negative. This means that there is indeed an ongoing gender discrimination situation. From the original data, we can see that with similar work experience and education levels, female workers usually earn less than male workers.

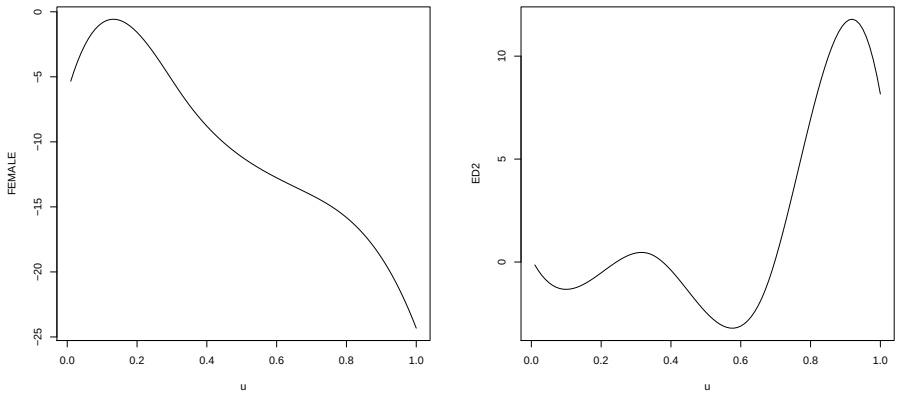

Figure 8: Female and Edu2 coefficients

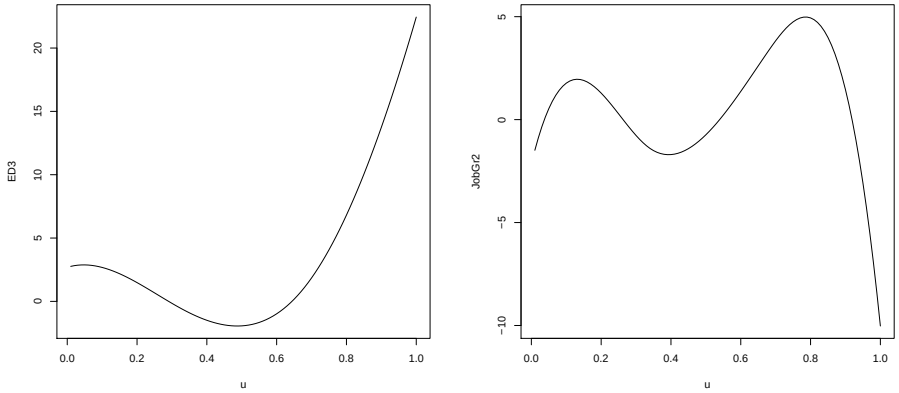

Figure 9: Edu3 and JobGrd2 coefficients

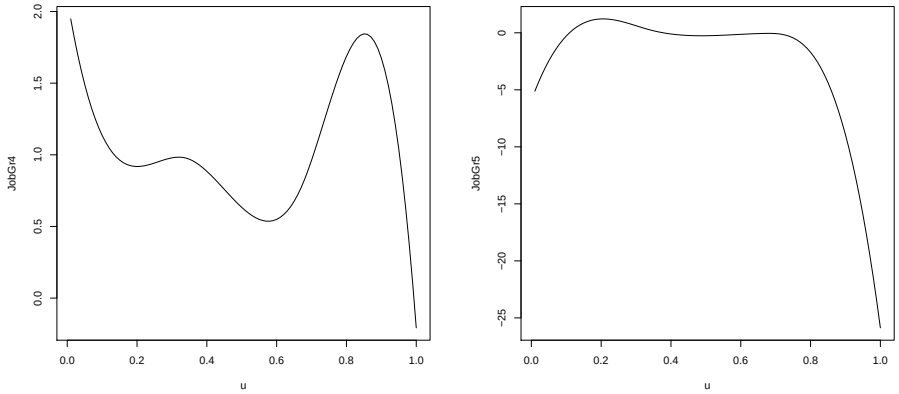

Figure 10: JobGrd4 and JobGrd5

5.4 Numerical Study of Penalized Spline

5.4.1 Implementation of some algorithms

In this section, we apply the one-step LLA-SCAD method, the LQA-SCAD method and the modified LQA-SCAD method in our simulation study.

For the penalized spline regression, first we let the initial number of knots be proportional to the sample size. A number of large initial knots can reduce the approximation error. The nonconcave penalized regression spline can automatically select knots and estimate the nonparametric regression function simultaneously. Here we follow the formula proposed by Friedman and Silverman (1989) to select the number of initial knots k,

$$k = [n/M(n,\alpha)] + 1,$$

where $0.05 \leq \alpha \leq 0.1$, the sample size $n \geq 15$, and $M \approx L_{\max}(\alpha)/3$ denotes a minimum span between two placed knots, $L_{\max}(\alpha) \approx -\log\{-(1/n)\ln(1-\alpha)\}$.

In a nonparametric regression setting, the matrix $n^{-1}X_n^T X_n$ might be singular, resulting in a large variance in our initial spline estimator. Thus, we have to weigh the initial estimate in penalty terms such that the variance of the weighted estimate is of the order of $O(n^{-1/2})$, which is the same order of λ_n. We take

$$\omega_j = \Big[(\frac{1}{n}X_n^T X_n)_{jj}^{-1}\Big]^{-1/2}.$$

Here, in practice, we replace the inverse operation by the generalized inverse operation because the matrix might be singular. This replacement is reasonable because even when $\hat{\boldsymbol{\beta}} = (X_n^T X_n)^{-1} X_n Y$ is not a consistent estimate of $\boldsymbol{\beta}$, $X_n\hat{\boldsymbol{\beta}}$ is still a consistent estimate of $X_n\boldsymbol{\beta}$.

Our aim is to minimize the following objective function:

$$\min_{\beta} \ \sum_{j=1}^{n}\{y - f(x_j,\boldsymbol{\beta})\}^2 \ + \ n\sum_{j=1}^{k} p_{\lambda_n}(|\omega_{p+i-1}\beta_{p+i-1}|), \tag{5.6}$$

where ω_j are the weights, and

$$f(x_j,\boldsymbol{\beta}) \ = \ \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{p-1}x^{p-1} + \sum_{i=1}^{k}\beta_{p+i-1}(x-t_i)_+^{p-1}.$$

Generally, the under-smoothing of the penalized regression spline is caused by an excessive number of knots, and the problem is attenuated by convex penalties to produce shrinkage estimates of the basis function coefficients, such as the rough penalty used in smoothing splines. Here the thresholding rule provides an attractive alternative to reduce the problem of under smoothing. We may reduce the number of knots adaptively by a thresholding rule. However, the properties of unbiasedness and continuity retain the smoothing and stability of the penalized regression spline when we reduce the number of knots.

The SCAD penalty function is singular at the origin. Thus we need to use some approximation method to solve (5.6). For the LLA algorithm it is obvious that that solving $\boldsymbol{\beta}^{(1)}$ is not much different from solving the LASSO. There are many ways to solve LASSO-type problems, and here we choose LARS. The LARS algorithm (Eforn, Hasttie, Johnstone, and Tibshirani, 2004) is a major breakthrough in the development of LASSO-type methods. There are also many modified LARS methods for different kinds of situations. For the LQA method, we apply the algorithm in Fan and Li (2001). For the modified LQA-SCAD method, we apply our algorithm stated in Chapter 4. Finally we compare the results of these three methods.

Similar to Fan and Li (2001), a modified GCV is applied to choose the regularization parameter λ for all three methods. For example, when we use the LQA or modified LQA method, we just find λ, which minimizes

$$\text{GCV}(\lambda) \;=\; \frac{1}{n}\frac{\|y - X\hat{\beta}(\lambda)\|^2}{[1 - \gamma e(\lambda)]^2},$$

where $\hat{\beta}(\lambda)$ is the penalized least-squares estimate for a given λ, $e(\lambda) = \text{trace}[P_X(\hat{\beta}(\lambda))]$, $P_X(\beta) = X(X^TX + \Sigma(\beta))^{-1}X^T$, and

$$\Sigma_{\lambda_n} = \{\text{diag}p'_{\lambda_n}(|\omega_{j1}\beta_{j10}|)/|\omega_{j1}\beta_{j10}|, \cdots, p'_{\lambda_n}(|\omega_{jd}\beta_{jd0}|)/|\omega_{j1}\beta_{j10}|\}.$$

The inflation factor is used here because a lot of basis functions are used in the model. In this simulation section, we use $\gamma = 2.5$. As in the above simulations, we set $a = 3.7$.

5.4.2 Simulation Examples

1. $Y = \sin(2(4X-2) + 2\exp(-16X^2) + \sigma\varepsilon$,
   where X is equally sampled from [0,1], $\sigma = 0.3$, $n = 256$, with 400 repetitions.

2. $Y = 2.2(4\sin 4\pi X - \text{sgn}(x-0.3) - sgn(0.72-X)) + \sigma\varepsilon$,
   where X is equally sampled from [0,1]. $\sigma = 1$, $n = 2048$ with 31 repetitions.

We use simulation examples to examine the performance of the LLA, LQA, and modified LQA methods.

The initial number of knots can be set by following the equation:

$$\Big[\frac{3n}{-\log_2\{(-1/n)\ln(1-0.1)\}}\Big].$$

For the first example, we set the initial knots number of knots as $k = 60$, and $k = 432$ for the second example. λ is selected by the MGCV above.

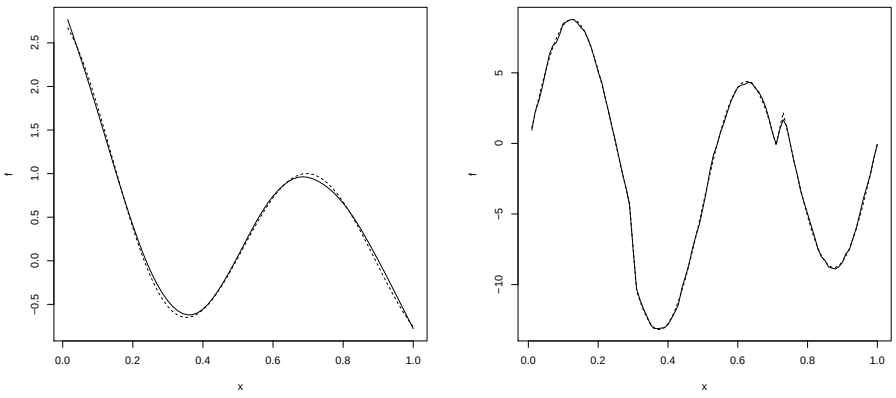

Figure 11: The smoothing spline regression

Example	LQA-SCAD	LLA-SCAD	mLQA-SCAD
1	5.3(3.2)	5.4(3.4)	5.6(3.7)
2	52 (8)	53 (8)	55(9)

Table 5.7: The simulation results with three algorithm for SCAD.

In Table 8 we present the medians of the MSEs, and we can see that all three methods have a fairly good performance. The original LQA method is slightly better

than the other two with a smaller MSE. In practice, the original LQA converges to the limit point more rapidly. The fits of the SCAD penalized regression spline with median performance are shown in Figure 11.

Chapter 6

Conclusion and Discussion

In Chapter 2, we apply the local polynomial approximation method and the group SCAD penalty method to semi-varying coefficient models with a diverging number of variables. Our proposed method can select significant variables and identify the constant coefficient simultaneously. Our method extends the other regularization methods for varying coefficient models. In particular, we consider the situation in which the dimension of the varying coefficient model is diverging with the sample size. We prove that under some regular conditions, our proposed SCAD estimator retains the sparsity and asymptotic normality properties in high-dimensional models. We suggest using a BIC-type criterion to select the tuning parameter in the penalty function. Our numerical results reported in Chapter 5 indicate that our proposed methods have better accuracy in finding the true regression model and the constant coefficients. In Chapter 3, an independent sure screening method is applied to ultra-high dimensional semi-varying coefficient models. Our work is based on the B-spline approximation for varying coefficient functions. We prove that under some regular conditions, the sure screening property of the proposed sure screening method can be established. We also show that in our numerical studies, the sure screening method with regularization methods can select the model consistently, even when the model dimension is ultra-high. These results can be viewed as an extension of the nonparametric independent screening (NIS) for additive models proposed by Fan, Feng, and Song (2011). Together with the results in Chapters 2 and 3, we establish a complete theory for variable selection whether for high- or ultra-high-dimensional semi-varying

coefficient models. In Chapter 4, we show that in high-dimensional situations with a suitable initial value, the one-step LLA estimator also has the oracle properties. We also construct a brand new approximation algorithm for the modified Newton-Raphson algorithm using the local quadratic approximation. This new algorithm considerably reduce the computational burden because it does not need to calculate the inverse of the Hessian matrix at every iteration step. In simulations, we show that this new method has a good performance compared to the LLA and original LQA.

Because of the high dimensionality, some asymptotic properties, such as the convergence rates, may not be as good as those in a finite dimension setting. However, the classic local polynomial approximation method is still a powerful tool in varying coefficient models. With a few modifications, this method can be applied to high-dimensional modeling problems, as proven by the theoretical results in Chapter 2. Due to the similarity between varying coefficient models and additive models, in Chapter 3 we naturally apply the NIS method based on B-spline approximation to varying coefficient models. The SCAD method has many advantages which make this method an excellent choice for many variable selection problems. However, the non-concave property of this penalty function makes the optimization problem complex. The LQA and LLA are the two best algorithms to solve this calculation disadvantage of SCAD. The modified Newton-Raphson method we establish in Chapter 4 is a powerful tool that can ease the computational burden in many statistical models, such as the additive models. The fast convergence rate of local quadratic approximation is still very attractive.

In Chapter 2, the number of parameters truly grows to infinity with the sample size. However, we still pose some strict limitations on the dimension size. It poses a big theoretical and methodological challenge for further study of the traditional data analysis methods in situations where the number of parameters tends to infinity at a faster rate. After model estimation, we need to discuss testing methods for making inferences. How to apply classical inference methods to high-dimensional varying coefficient models is a good issue for further research. Another important topic is how to extend our proposed methods to high-dimensional generalized varying coefficient

models. Although this problem seems to be a natural extension of the original varying coefficient models, the high-dimensional generalized varying coefficient models have more wider applications in practice. In Chapters 2 and 3, we only use the local linear Taylor expansion. Could we use a more complicated expansion, such as the quadratic Taylor expansion, to determine which coefficient function is a linear function so that the model is more easy to interpret? In Chapter 4, a big problem still remains. When the dimension of the model diverges with the sample size more quickly, will these approximation methods still be available? As the choice of initial value is crucial for the LLA, can we find a good initial value under more universal conditions? Fan, Xue and Zou (2012) studied this problem and suggested using the LASSO estimator as the initial value $\boldsymbol{\beta}_{(0)}$ when the dimension p is larger than the sample size n. Their work provide an interesting point for our further research on group penalized methods for semiparametric models.

Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Edited by B. N. Petrov and F. Csaki, 267 - 281. Akademiai Kiado, Budapest.

[2] An, L., T, H. and Tao, P. D. (1997). Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. *Journal of Global Optimization*, **11**, 253 - 285.

[3] Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussion). *Journal of the American Statistical Assocation*, **96**, 939 - 967.

[4] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373 - 384.

[5] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**, 2350 - 2383.

[6] Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, **80**, 580 - 619.

[7] Breheny, P. and Huang, J. (2009). Penalized methods for bilevel variable selection. *Statistics and Its interface*, **2**, 369 - 380.

[8] Brumback, B. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistiscal Assocation*, **93**, 961 - 994.

[9] Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models.*Journal of the American Statistical Assocation*, **95**, 888 - 902.

[10] Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, **35**, 2313 - 2404.

[11] Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal otf the American Statistical Assocation*, **92** 477 - 489.

[12] Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika*, **95**, 759 - 771.

[13] Chen, R. and Tsay, R. J. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Assocation*, **88**, 298 - 308.

[14] Cheng, M. Y. and Zhang, W. (2007). Statistical estimation in generalized multiparameter likelihood models, manuscript.

[15] Chiang, C. T., Rice, J. A. and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Assocation*, **96**, 605 - 619.

[16] Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). Local regression models. In *Statistical Models in S*, Chambers, J. M. and Hastie, T. J., eds, 309 - 376. Wadsworth & Brooks, Pacific Grove.

[17] Collobert, R., Sinz, F., Weston, J. and Bottou, L. (2006). Large-scale transductive SVMs, *Journal of Machine Learning Research*, **7**, 1687 - 1712.

[18] Cui, X., Peng, H., Wen, S. Q. and Zhu, L. X. (2013). Component selection in an additive models. *Scandinavian Journal of Statistics*, to appear.

[19] Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century.*

[20] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics*, **32**, 407 - 499.

[21] Fan, J. (1997). Comments on Wavelets in Statistics: A Review by A. Antoniadis, *Journal of the Italian Statistical Association*, **6**, 131 - 138.

[22] Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation. *Journal of Royal Statistical Society*, B, **61**, 927 - 943.

[23] Fan, J., Farmen, M. and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society*, B, **60**, 591 - 608.

[24] Fan, J. and Huang, T. (2005). Profile Likelihood Inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**, 1031 - 1057.

[25] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models, *Journal of American Statistical Association*, **116**, 544 - 557.

[26] Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society*, B, **57**, 371 - 394.

[27] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

[28] Fan, J. and Jiang, J. (2005). Nonparametric inferences for additive models. *Journal of the American Statistical Association*, **100**, 890 - 907.

[29] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348 - 1360.

[30] Fan, J. and Li, R. (2002). Variable selection for Coxs proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74C99.

[31] Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion). *Journal of the Royal Statistical Society*, B, **70**, 849 - 911.

[32] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101 - 148.

[33] Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with NP-Dimensionality. *IEEE – Information Theory*, **57**, 5467-5484.

[34] Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *The Annals of Statistics*, **32**, 928 - 961.

[35] Fan, J., Samworth, R. and Wu, Y. (2009). Ultra-dimensional variable selection via independent learning: beyond the linear model. *Journal of Machine Learning Research*, **10**, 1829 - 1853.

[36] Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality, *The Annals of Statistics*, 38, 3567 - 3604.

[37] Fan, J., Xue, L. and Zou, H. (2012). Strong oracle optimality of folded concave penalized estimation. Manuscript.

[38] Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society*, B, **65**, 57 - 80.

[39] Fan, J., Zhang, C. M. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, **29**, 153 - 193.

[40] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**, 1491 - 1518.

[41] Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, **27**, 715 - 731.

[42] Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models, *Statistics and Its Inference*, **1**, 179 - 195.

[43] Foster, D. and George, E. (1994). The risk infation criterion for multiple regression. *The Annals of Statistics*, **22**, 1947 - 1975.

[44] Frank, I. E. and Frideman, J. H. (1993). A statistical view of some chemometric regression tools (with discussion). *Technometrics*, **35** 109 - 148.

[45] Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, **19**, 1 - 141.

[46] Fridman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additiving modeling, *Technometrics*, **31**, 3-21.

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

[48] Gu, C. and Wahba, G. (1993). Smoothing spline ANOVA with component-wise Bayesian confidence intervals. *Journal of Computational and Graphical Statistics*, **2**, 97 - 117.

[49] Hall, P. and Miller, H. (2009). Using generalised correlation to effect variable selection in very high dimensional problems, *Journal of Computational and Graphical Statistics*, **18**, 533 - 550.

[50] Hall, P., Titterington, D. M. and Xue, J. H. (2008). Discussion of Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society*, B, **70**, 889 - 890.

[51] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall, London.

[52] Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of Royal Statistical Society*, B, **55**, 757 - 796.

[53] Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998). Charac- terizing selection bias using experimental data, *Econometrica*, **66**, 1017 - 1098

[54] Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809 - 822.

[55] Horowitz, J., Klemelä, J. and Mammen, E. (2006). Optimal Estimation in Additive Regression Models, *Bernoulli*, **12**, 271 - 298.

[56] Hu, T. and Xia, Y. C. (2012). Adaptive semi-varying coefficient model selection, *Statistica Sinica*, to appear.

[57] Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-Dimensional regression models. *The Annals of Statistics*, **36**, 587 - 613.

[58] Huang, J., Horowitz, J. and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, **38**, 2282 - 2313.

[59] Huang, J., Ma, S., Xie, H. and Zhang, C. H. (2009). A group bridge approach for variable selection. *Biometrika*, **96**, 339C355.

[60] Huang, J., Wei, F. and Ma, S. G. (2012). Semiparametric regression pursuit, *Statistica Sinica*, **22**, 1403-1426.

[61] Huang, J., Ma, S. G. and Zhang, C. H. (2008). Adaptvie Lasso for Sparse High Dimensional Regression Models, *Statistica Sinica*, **18**, 1603 - 1618.

[62] Huang, J. Z. and Shen, H. (2004). Functional coefficient regression models for nonlinear time series: A polynomial spline approach. *Scandinavian Journal of Statistics*, **31**, 515 - 534.

[63] Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements, *Biometrika*, **89**, 111 - 128.

[64] Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, **14**, 763 - 788.

[65] Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo, *The Annals of Statistics*, **1**, 799 - 821.

[66] Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, **33**, 1617 - 1642.

[67] Kim, Y., Choi, H. and Oh, H.S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of American Statistical Assocation*, **103**, 1665 - 1673.

[68] Kim, Y., Kim, J. and Kim, Y. (2006). Blockwise Sparse Regression, *Statistica Sinica* , **16**, 375 - 390.

[69] Koltchinskii, V. and Yuan, M. (2008). Sparse Recovery in Large Ensembles of Kernel Machines, in *21st Annual Conference on Learning Theory - COLT 2008*, Helsinki, Finland, July 9-12, 2008, eds. R. A. Servedio and T. Zhang, Omnipress, pp. 229 - 238.

[70] Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational n and Graphical Statististics*, **9**, 1 - 59.

[71] Li, K. C. (1991). Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association*, **86**, 316 - 327.

[72] Lian, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica*, **22**, 1563 - 1588.

[73] Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression, *The Annals of Statistics*, **34**, 2272 - 2297.

[74] Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Assocation*, **92**, 107-114.

[75] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **12**, 661 - 675.

[76] Meier, L., Geer, V. and Bühlmann, P. (2009). High-Dimensional Additive Modeling. *The Annals of Statistics*, 37, 3779 - 3821.

[77] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with Lasso, *The Annals of Statistics*, **34**, 1436 - 1462.

[78] Miller, A. (1990). *Subset Selection in Regression*, Chapman and Hall/CRC, London.

[79] Neyman, J. and Scott, E. L. (1948). Consistent estimation based on partially consistent observations, *Econometrica*, **16**, 1 - 32.

[80] Peng, H. (2004). Nonconcave penalized spline, manuscript.

[81] Portnoy, S. (1988). Asymptotic behavior of likeli hood methods for exponen tial families when the number of parameters tends to infinity. *The Annals of Statistics*, **16**, 356C366.

[82] Ramsay, J. O. and Silverman, B. W. (1997). *The Analysis of Functional Data*. Springer-Verlag, Berlin.

[83] Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2009). Spam: sparse additive models. *Journal of the Royal Statistical Society*, B , **71**, 1009 - 1030.

[84] Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, **35**, 1012 - 1030.

[85] Sardy, S. and Tseng, P. (2004). AMlet, RAMlet, and GAMlet: Automatic nonlinear fitting of additive models, obust and generalized, with Wavelets, *Journal of Computational and Graphical Statistics*, 13, 283 - 309.

[86] Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461 - 464.

[87] Shen, X., Tseng, G. C., Zhang, X. and Wong, W. H. (2003). On ψ-Learning. *Journal of the American Statistical Association*, **98**, 724 - 734.

[88] Silverman, B. (1984). Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics*, **12**, 898 - 916.

[89] Speckman, P. (1985). Spline smoothing and optimal Rates of convergence in nonparametric regression models. *The Annals of Statistics*, **13**, 970 - 983.

[90] Stone, C. (1985). Additive Regression and Other Nonparametric Models. *The Annals of Statistics*, **13**, 689 - 705.

[91] Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, **25**, 1371 - 1470.

[92] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, B, **58**, 267 - 288.

[93] Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society*, B, **61**, 529-546.

[94] Tong, H. (1990). *NonLinear Time Series: A Dynamical System Approach*, Oxford University Press, Oxford.

[95] Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*, New York: Springer.

[96] Wahba, G. (1984). Partial spline models for semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series* , Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319 - 329. Institute of Statistical Mathematics, Tokyo.

[97] Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553 - 568.

[98] Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, **104**, 747 - 757.

[99] Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, **23**, 1486 - 1494.

[100] Wei, F. and Huang, J. (2007). Consistent group selection in high- dimensional linear regression, Technical Report 387, University of Iowa.

[101] Wei, F., Huang, J. and Li, H. Z. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statitica Sinica*, **21**, 1515-1540.

[102] Wu, C. O., Chiang, C. T. and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of American Statistical Aossciation*, **93**, 1388 - 1402.

10, 433 - 456.

[103] Xia, Y., Zhang, W. and Tong, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika*, **91**, 661 - 681.

[104] Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying ce-officient models with grobal optimality, *Journal of Machine Learning Research*, **13**, 1973-1998.

[105] Xue, L., Qu, A. and Zhou. J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association*, **105**, 1518-1530.

[106] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, B, **68**, 49 - 67.

[107] Yuille, A. and Rangarajan, A. (2003). The concaveCconvex procedure. *Neural Computation*, **15**, 915 - 936.

[108] Zhang, C. H. (2010). Nearly unbiased variable section under minimax concave penalty, *The Annals of Statistics*, **38**, 894 - 942.

[109] Zhang, H., Cheng, G. and Liu, Y. (2011). Linear or Nonlinear? automatic discovery for partially linear models. *Journal of the American Statistical Association*, **106**, 1099 - 1112.

[110] Zhang, W., Lee, S. Y. and Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *Journal of Multivariate Analysis*, **82**, 166 - 188.

[111] Zhao, D. S., and Li, Y. (2010). Principled sure independence screening for cox models with ultra-high dimensional covariates. manuscript, Harvard University.

[112] Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, **37**, 3468 - 3497.

[113] Zhou, S., Shen, X. and Wolfe, D. A. (1998), Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, **26**, 1760 - 1782.

[114] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Assocation*, **101**, 1418 - 1429.

[115] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, B, **67**, 301 - 320.

[116] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, **36**, 1509 - 1533.

Curriculum Vitae

Academic qualifications of the thesis author, Mr. Chen Chi:

- Received the degree of Bachelor of Science (Mathematics) from Fudan University (China), July 2006.
- Received the degree of Master of Science (Mathematics) from Fudan University (China), Jan 2010.

Aug 2013