VARIABLE SELECTION FOR FUNCTIONAL INDEX COEFFICIENT MODELS AND ITS APPLICATIONS IN FINANCE AND ENGINEERING

by

Bingduo Yang

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2012

Approved by:

____
Dr. Zongwu Cai

____
Dr. Jiancheng Jiang

____
Dr. Weihua Zhou

____
Dr. Sheng-Guo Wang

ABSTRACT

BINGDUO YANG. Variable selection for functional index coefficient models and its applications in finance and engineering. (Under the direction of DR. ZONGWU CAI)

Variable selection with a non-concave penalty function has become popular in recent years, since it has ability to select significant variables and to estimate unknown regression coefficients simultaneously. In this dissertation, firstly, I consider variable selection in a functional index coefficient model under strong mixing context. Due to the fact that the model is in a semiparametric form so that the convergence rate of parametric estimator is faster than nonparametric estimator, my selection procedures with smoothly clipped absolute deviation penalty function consist of two steps. The first is to select significant covariates with functional coefficients and it is then to do variable selection for local significant variables with parametric coefficients. The asymptotic properties such as consistency, sparsity and the oracle property of these two step estimators are established, whereas easy computational algorithms are suggested to highlight the implementation of the proposed procedures. Finally, Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed estimators and selection procedures. Two financial examples including functional index coefficient autoregressive models and functional index coefficient models for the stock return predictability are extensively studied.

In the second part of this dissertation, I consider the estimation and variable selection for the local annual average daily traffic (AADT) using different groups of variables. It is well documented that in transportation networks, AADT estimation is very important to decision making, planning, air quality analysis, etc and a regression method may be one of the most popular methods for estimating AADT on non-counters roads. The existing literatures focus on how to collect different groups of predicting variables, and to select significant variables by t-test and F-test. However, there is no theory on the validity of these multiple selecting steps. Furthermore, variables collected for high functional class roads maybe not suitable for the estimation of local AADT because of lacking counters. The variable selection by smoothly clipped absolute deviation penalty (SCAD) procedure is proposed and it can select significant variables and estimate unknown regression coefficients simultaneously at one step. The estimation algorithm and the tuning parameters selection are also presented. To demonstrate the usefulness of the proposed procedure, I use the real data observed from Mecklenburg County of North Carolina in 2007 for illustration. The analysis result shows that the selection procedure is indeed valid and it further improves the local AADT estimation by incorporating satellite information. The proposed method outperforms some other regression methods when it is applied to local AADT estimation.

The third part of this dissertation is to consider how to calculate seasonal factors in annual average daily traffic (AADT) and vehicle miles traveled (VMT). It is well known that seasonal factors are very important to the estimation of AADT and VMT and they are used to transfer one or two days measured traffic data at portable traffic monitoring sites to the AADT. Most literatures focus on taking the average of seasonal factors within groups of roads. Factor grouping including three techniques to calculate seasonal factors has been recommended by the Federal Highway Administration (FHWA). However, as recognized, it is difficult to select a representative group sample of roads. In this part, to calculate seasonal factors, I propose a new nonparametric approach by introducing the distance kernel and by using the local weights. The nonparametric seasonal factors estimation and test procedure are presented. Moreover, the proposed approach can be extended to grouping cases if prior information of grouping is available. Finally, the real example demonstrates the new approach by using the data observed in the North Carolina.

ACKNOWLEDGMENTS

I would like to gratefully and sincerely thank my supervisor Dr. Zongwu Cai for his guidance, understanding, patience, and most importantly, his friendship during my graduate studies at UNC Charlotte. It is because of him that I came to the USA and pursued my Ph.D. degree. He encouraged me to not only grow as a statistician and an econometrician, but also as an instructor and an independent thinker. He asked me to explore different new topics and let me do what I am interested in. He gave me the opportunity to develop my own individuality and self-sufficiency with independent thinking. His wisdom, knowledge and commitment to the highest standards inspired and motivated me. He always made himself available to clarify my doubts despite his busy schedules. Meanwhile, he revised the draft of my dissertation so carefully and made it the way as it is now. Furthermore, he often treated us for lunch or dinner. His wife and he hosted our Ph.D. students including myself many times throughout years. Many thanks to Dr. Zongwu Cai for everything he has done for me.

I would like to thank Dr. Jiancheng Jiang for his guidance of study in my graduate life, suggestions in the dissertation and useful discussions in the seminars. I usually ask Dr. Jiang for help if I have any questions in the research. He is really sweet and patient to give me academic and professional commitments. In particular, I should thank him for being a member of my oral exam and doctoral dissertation defense committee. My special thanks also go to Dr. Weihua Zhou for being a member of my oral exam and doctoral defense committee.

I also want to convey my appreciation to Dr. Sheng-Guo Wang for his unending encouragement and support. I did projects under his supervision and he taught me how to

write and revise our papers more than ten times face to face and step by step. I would like to thank him for his valuable guidance, scholarly inputs and consistent encouragement I received throughout the research work.

I would like to thank the Department of Mathematics and Statistics for its supports of my Ph.D. study. In particular, I would like to thank Dr. Avrin, Dr. Kazemi and Chairman Dr. Alan Dow for their hard work and patient service. I would like to take this opportunity to thank the China Scholarship Council for the two-year financial support during my visit as an exchange Ph.D. student in the USA.

Most importantly, I would like to thank my parents, for their financial support when needed, their unwavering faith and allowing me to be as ambitious as I wanted and to do what I thought.

Finally, I am very grateful for the friendship from all the members of our seminar. In particular, I should thank Dr. Linman Sun for helpful discussion in my graduate life. My colleagues and friends, Yonggang Wang and Li Wu have all extended their support in a very special way. I learned a lot from them, through their personal and scholarly interactions, their suggestions at various points of my research program. Meanwhile, I would like to express my deep appreciation to those who love me and help me and those who provided so much support and encouragement throughout this process. In particular, I should thank Director Marian Beane for her input and accessibility. The University Hills Baptist Church, through its global coffee program, helps me know more about the Christ and American culture. Finally, I should thank Dr. Jianping Fan and his wife Hailan Zhong for their kind friendship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

This dissertation covers three topics, including variable selection for functional index coefficient models and their applications in finance, efficient local annual average daily traffic (AADT) estimation via smoothly clipped absolute deviation (SCAD) variable selection based on regression models and nonparametric approach to calculate seasonal factors for AADT estimation. In this chapter, I will introduce the main results I have done to these tree topics, respectively.

1.1 Variable Selection for Functional Index Coefficient Models

Varying-coefficient model proposed by Hastie and Tibshirani (1993) has gained more and more attention in recent years due to desirable properties such as its flexibility and dimension reduction in nonparametric sense. To incorporate more variables in the functional coefficients and to overcome the difficulty of the curse of dimensionality, Fan, Yao and Cai (2003) proposed the following functional index coefficient model (FIM)

$$y_i = g^T(\beta^T Z_i)X_i + \varepsilon_i, \qquad 1 \le i \le n, \tag{1.1}$$

where y_i is a dependent variable, $X_i = (X_{1i}, X_{2i}, \ldots X_{pi})^T$ is a $p\times1$ vector of covariates, Z_i is a $d\times1$ vector of covariates, ε_i are independently identically distributed (i.i.d) with mean 0 and standard deviation σ, $\beta \in R^d$ is a $d\times1$ vector of unknown parameters and $g(\cdot) = (g_1(\cdot)\ldots g_p(\cdot))^T$ is a vector of $p-$dimensional unknown functional coefficients. Assume that $\parallel \beta \parallel= 1$ or the first element of β is positive for identification.

Due to the efficiency of estimation and the accuracy of prediction, it is very important to select significant variables and exclude insignificant variables in equation

(1.1). Meanwhile, almost all existed variable selection procedures are based on the assumption that the observations are identically independently distributed (i.i.d). In this chapter, I consider variable selection in functional index coefficient models under strong mixing conditions.

It is clear that model (1.1) is a semiparametric model. Therefore, to estimate $g(\cdot)$, the initial estimators of $\hat{\beta}$ are needed and they might not have huge effects on the final estimation of $g(\cdot)$ if the sample size n is large enough, since the convergence rate of the parametric estimators $\hat{\beta}$ is faster than the nonparametric function estimators $\hat{g}(\cdot)$. Thus, to estimate $g(\cdot)$ and β, a two-stage procedure is needed. Here, I propose variable selection and estimation in two steps as follows. Firstly, I select the significant covariates with functional coefficients, and then variable selection is applied for choosing local significant variables with parametric coefficients.

Step One: Given an initial estimator $\hat{\beta}$ such that $\| \hat{\beta} - \beta \| = O_p(1/\sqrt{n})$, minimize the penalized local least squares $Q(\widehat{g}, \hat{\beta}, h)$ to obtain $\widehat{g}(\cdot)$, where (or maximize the penalized local likelihood),

$$Q(\widehat{g}, \hat{\beta}, h) = \sum_{j=1}^{n} \sum_{i=1}^{n} \left\{ y_i - \widehat{g}^T \left(\hat{\beta}^T Z_j \right) X_i \right\}^2 K_h \left(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j \right) + n \sum_{k=1}^{p} P_{\lambda_k} \left(\| \widehat{g}_{\cdot k} \| \right) \tag{1.2}$$

with $K(\cdot)$ being the kernel function, $K_h(z) = K(z/h)/h$ and $P_{\lambda_k}(\cdot)$ being the penalty function, specified later.

Step Two: Given the estimator of function $\hat{g}(\cdot)$, minimize the penalized global least squares $Q(\beta, \hat{g})$ (or maximize the penalized global likelihood), where

$$Q(\beta, \hat{g}) = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \hat{g}^T(\beta^T Z_i) X_i \right)^2 + n \sum_{k=1}^{d} \Psi_{\lambda_n}(|\beta_k|) \tag{1.3}$$

with $\Psi(\cdot)$ being a penalty function, specified later..

In Chapter 2, I study the large sample behavior of these estimators such as consistency, sparsity and the oracle property, meanwhile, computational algorithms are outlined. Finally, Monte Carlo simulations are conducted to examine the finite sample performance and two financial examples including functional index coefficient autoregressive models and functional index coefficient models for the stock return predictability are extensively studied.

The first real example I consider is to use functional index coefficient autoregressive models (FIAR) to analyze asset return data. The FIAR model for this real example is defined as

$$r_t = \sum_{j=1}^{p} g_j(\beta^T \mathbf{r}_t) r_{t-j} + \varepsilon_t,$$

where r_t is the asset return and $g_j(\cdot)$'s are unknown functions in R^d for $j \leq p$ and $\mathbf{r}_t = (r_{t-1}, \cdots, r_{t-d})^T$ is a vector of lagged returns. Clearly, it is an extension of functional coefficient autoregressive (FAR) model, which was proposed by Chen and Tsay (1993). To explore the performance of functional index coefficient autoregressive models for asset returns, by taking $p = 6$, I simply assume my working model as follows.

$$r_t = \sum_{j=1}^{6} g_j(\mathbf{z}_t) r_{t-j} + \varepsilon_t,$$

where $\mathbf{z}_t = \beta_1 r_{t-1,t} + \beta_2 r_{t-2,t} + \beta_3 r_{t-3,t}$ and I assume $\beta_1^2 + \beta_2^2 + \beta_3^2 = 1$ in order to satisfy the identification assumption.

The data for asset returns consist of daily, weekly and monthly returns on the Dow Jones Industrial Average, NASDAQ Composite and $S\&P$ 500 INDEX. I use two step variable selection procedures to select significant variables and to estimate unknown coefficients simultaneously. Firstly, I do variable selection on the regressors based on penalized local maximum likelihood, then I do variable selection on the local variables based on penalized global maximum likelihood. When two step estimations and variable selections are employed in my model, the estimated coefficients of local variables and the norms of covariates are reported in Table 2.4. An interesting finding is that, all local variables perform the same for one day return of three index with similar parameter coefficients. However, we cannot find this phenomenon with the return of one week horizon and one month horizon.

Another interesting example I consider is the predictability for the stock return, which is very important in empirical finance since it is the center issue to the asset allocation for practitioners in finance markets. Many literatures have revealed that the coefficients of predictors may depend on other financial variables. In this section, I consider the predictability for the stock return with the functional index coefficient models of Fan, Yao and Cai (2003), which can incorporate multiple variables in the coefficients. I specify two type models as below.

$$\text{Model 1:} \quad r_t = g_1(\mathbf{z}_t)z_{1,t} + g_2(\mathbf{z}_t)z_{2,t} + g_3(\mathbf{z}_t)z_{3,t} + g_4(\mathbf{z}_t)z_{4,t} + \varepsilon_t,$$

where $\mathbf{z}_t = \beta_1 z_{1,t} + \beta_2 z_{2,t} + \beta_3 z_{3,t} + \beta_4 z_{4,t}$ and z_{jt} is a financial variable described below, and

$$\text{Model 2:} \quad r_t = \sum_{j=1}^{6} g_j(\mathbf{z}_t) r_{t-j} + \varepsilon_t.$$

In Model 1, the covariates and local variables $\{z_{j,t}\}$ include “BamAa”, the spread between Moody’s Baa corporate bond yield and Moody’s Aaa corporate bond yield, “Bam3m”, the spread between Moody’s Baa corporate bond yield and a three-month Treasury bill, “term1year”, the term spread between the one year and three-month Treasury yields, and “term10year”, the term spread between the ten year and three-month Treasury yields. In Model 2, I let the lagged returns to be covariates. To match the predictors, I let the lagged data as covariates and local variables. The dependent variables include monthly returns on the Dow Jones Industrial Average, NASDAQ Composite and $S\&P$ 500 INDEX in both two models. Finally, the detailed analysis results are summarized in Table 2.6 and Table 2.7 or Figure 2.11$\sim$ Figure2.13.

1.2 Efficient Local AADT Estimation via SCAD Variable Selection Based on Regression Models

In transportation networks, annual average daily traffic estimation is very important to decision making, planning, air quality analysis, etc. Regression method may be one of the most popular approaches used for estimating AADT on non-counters roads. Most literatures focus on how to collect different groups of predicting variables, and to select significant variables by t-test and F-test. However, there is no theory on the validity of these multiple selecting steps. This Chapter focuses on the estimation and variable selection for the local AADT using different groups of variables.

To illustrate the proposed method is practically useful, I consider a real data set observed in Mecklenburg County of North Carolina in 2007. I consider four groups of 19 variables including general driving behavior, characteristics of the roads, information from satellite and socioeconomic variables. The incorporated satellite information has a great improvement in our model, and it makes R-square to go up from 0.48 to 0.65. According to the R-square and the prediction error, our method produces a better result to estimate AADT in the local functional class roads.

1.3 Nonparametric Approach to Calculate Seasonal Factors for AADT Estimation

Seasonal factors are very important to the estimation of annual average daily traffic (AADT) and Vehicle Miles Traveled (VMT) and they are used to transfer one or two days measured traffic data at portable traffic monitoring sites to the AADT. Most literatures focus on taking the average of seasonal factors within groups of roads.

In this chapter, to calculate the seasonal factors, I propose a nonlinear regression model based on the nonparametric method by introducing the distance kernel and by using the local weights. The factors utilize the similarity of seasonal variability and traffic characteristics at the count sites in a nearby area. They are decomposed into monthly factors and weekly factors. Then, I introduce a nonlinear distance weighting kernel to estimate the weekly factors. It puts more weight on the observation points which are much closer to the interested point, and puts less weight on the far away observation points. Thus, it makes the seasonal factor estimation more reasonable and accurate to be close to the true value.

Firstly, I can calculate seasonal factors as follows

$$F_{mw} = F_m \cdot F_w, \tag{1.4}$$

where F_{mw} is the seasonal factor for the m-th month and the w-th week, F_m is the monthly factor for the m-th month, and F_w is the weekly factor for the w-th day in a week.

Secondly, I can obtain the Nadaraya-Watson estimator of $F_w(x_0, y_0)$ by

$$\hat{g}_w(x_0, y_0) = \sum_{i=1}^{n} w_i F_w(x_i, y_i), \tag{1.5}$$

where w_i is defined in (4.9), defined later.

To test whether there exists location effect or not, I construct the hypothesis testing by generalized likelihood ratio test (GLR test). At last, the detail estimation procedures for the seasonal factors and AADT are clearly presented by an example.

CHAPTER 2: VARIABLE SELECTION FOR FUNCTIONAL INDEX COEFFICIENT MODELS

2.1 Introduction

Varying-coefficient model proposed by Hastie and Tibshirani (1993) has gained more and more attention during the recent years. Many extensions (Xia and Li, 1999; Fan and Zhang, 1999; Cai, Fan and Li, 2000; Fan, Zhang, and Zhang, 2001; Huang, Wu, and Zhou, 2002; Fan, Yao and Cai, 2003; Fan and Huang, 2005; Li and Liang, 2008) have been considered on the estimation of parameters and functionals and hypotheses testing. Specially, to overcome the difficulty of the curse of dimensionality, Fan, Yao and Cai (2003) proposed the following functional index coefficient model (FIM)

$$y_i = g^T(\beta^T Z_i)X_i + \varepsilon_i, \qquad 1 \leq i \leq n, \tag{2.1}$$

where y_i is a dependent variable, $X_i = (X_{1i}, X_{2i}, \ldots X_{pi})^T$ is a $p \times 1$ vector of covariates, Z_i is a $d \times 1$ vector of covariates, ε_i are independently, identically distributed (i.i.d) with mean 0 and standard deviation σ, $\beta \in R^d$ is a $d \times 1$ vector of unknown parameters and $g(\cdot) = (g_1(\cdot) \ldots g_p(\cdot))^T$ is a vector of $p-$dimensional unknown functional coefficients. We assume that $\parallel \beta \parallel = 1$ or the first element of β is positive for identification.

Xia and Li (1999) studied the asymptotic properties of model (2.1) under mixing conditions when the index part of the above model is not constraint to linear combination of Z_i. However, due to the efficiency of estimation and the accuracy of prediction, it is very important to select significant variables in Z_i and exclude insignificant variables in equation (2.1). Fan, Yao and Cai (2003) proposed a combination of the t-statistic and the Akaike information criterion (AIC) to select significant variables of Z_i and they deleted the least significant variables in a given model according to t-value, and selected the best model according to the AIC. However there is no theoretical foundation for their work. As mentioned in Fan and Li (2001), a stepwise deletion procedure may suffer stochastic errors inherited in the multiple stages. Thus, it is very critical to develop a variable selection procedure which can simultaneously select significant variables and estimate unknown regression coefficients for the above model.

In fact, the motivation of this study comes from functional coefficient autoregressive (FAR) model proposed by Chen and Tsay (1993). The coefficients in FAR model are unknown functional form and depend on lagged terms and it satisfies

$$r_t = g_1(\mathbf{r}^*_{t-1})r_{t-1} + \cdots + g_p(\mathbf{r}^*_{t-1})r_{t-p} + \varepsilon_t, \tag{2.2}$$

where $\mathbf{r}^*_{t-1} = (r_{t-i_1}, r_{t-i_2}, \cdots, r_{t-i_d})'$ for $j = 1, \cdots, d$. Due to the curse of dimensionality, Chen and Tsay (1993) just considered one single threshold variable case $\mathbf{r}^*_{t-1} = r_{t-k}$ for some lagged term r_{t-k}. To overcome the curse of dimensionality and incorporate more variables in the functional coefficients β's, we assume that $\mathbf{r}^*_{t-1}$ is a linear combination of r_{t-i_j}'s, e.g. $\mathbf{r}^*_{t-1} = \beta^T \mathbf{r}_t$, where $\mathbf{r}_t = (r_{t-1}, \cdots, r_{t-d})^T$. The FAR model can be reduced as a special case of FIM of Fan, Yao and Cai (2003). We name it as functional index coefficient autoregressive models (FIAR).

$$r_t = g_1(\beta^T \mathbf{r}_t)r_{t-1} + \cdots + g_p(\beta^T \mathbf{r}_t)r_{t-p} + \varepsilon_t. \tag{2.3}$$

As mentioned above, there is no theory on the variable selection procedure for model (2.1) and so, neither is for model (2.3). Also, Fan, Yao and Cai (2003) did not address how to select the covariates r_{t-j} in model in (2.3). This motivates us to do variable selection with local variables $\mathbf{r}_t$ and covariates r_{t-i} in model (2.3).

Variable selection methods and their algorithms can be tracked back to four decades ago. Pioneer criterions include the Akaike information criterion (Akaike, 1973) and the Bayesian information criterion (Schwarz, 1978). Various shrinkage type methods have been developed recently, including but not limited to the nonnegative garrotte (Breiman, 1995; Yuan and Lin, 2006), bridge regression (Fu, 1998), least absolute shrinkage and selection operator (Tibshirani, 1996; Knight and Fu, 2000), smoothly clipped absolute deviation (Fan and Li, 2001), adaptive LASSO (Zou, 2006), and so on. The reader is referred to the review paper by Fan and Lv (2010) for details. Here we prefer the SCAD of Fan and Li (2001) since it merits three properties of unbiasedness, sparsity and continuity. Furthermore, it has oracle property. Namely, the resulting procedures perform as well as if the subset of significant variables were known in advance.

The shrinkage method has been successfully extended to semiparametric models; for example, variable selection in partially linear models (Liang and Li, 2009), partially linear models in longitudinal data (Fan and Li, 2004), single-index models (Kong and Xia, 2007), semiparametric regression models (Brent et al., 2008; Li and Liang, 2008), varying coefficient partially linear models with errors-in-variables (Zhao and Xue, 2010), and partially linear single-index models (Liang et al., 2010), and the references therein.

However, the aforementioned papers focused mainly on the variable selection of significant variables with parametric coefficients. Also, the shrinkage method was extended to select significant variables with functional coefficients. Lin and Zhang (2006) proposed component selection and smoothing operator (COSSO) for model selection and model fitting in multivariate nonparametric regression models in the framework of smoothing spline analysis of variance (ANOVA), meanwhile, Zhang and Lin (2006) extended the COSSO to the exponential families. Wang, Li and Huang (2008) proposed the variable selection procedures with basis function approximations and SCAD, which is very similar to the COSSO, and they argued that their procedures can select significant variables with time-varying effects and estimate the nonzero smooth coefficient functions simultaneously. Huang, Joel and Wei (2010) proposed to use the adaptive group LASSO for variable selection in nonparametric additive models based on a spline approximation, in which the number of variables and additive components may be larger than the sample size. Adopted the idea of grouping method in Yuan and Lin (2006), Wang and Xia (2009) used kernel LASSO (KLASSO) to shrinkage functional coefficient in the varying coefficient models. Their pure nonparametric shrinkage procedure is different from spline and basis functions (Lin and Zhang, 2006; Wang, Li and Huang, 2008; Huang, Joel and Wei, 2010).

Almost all the variable selection procedures mentioned above are based on the assumption that the observations are identically independently distributed (i.i.d). To the best of our knowledge, there are few papers to consider variable selections under non i.i.d settings. It might not be appropriate if it is applied in to analyze financial and economic data, since most of the financial/economic data are week dependent. To address this issue, Wang, Li and Tsai (2007) extended to the regression model with autoregressive errors via LASSO. In this Chapter, we consider variable selection in functional index coefficient models under very general dependent structure – strong mixing conditions. Our variable selection procedures include two steps. Firstly, we select the significant covariates with functional coefficients, and then do variable selection for local significant variables with parametric coefficients.

The rest of this paper is organized as follows. In Section 2.2, we present the conditions for identification in functional index coefficient models, two step estimation procedures and some properties of SCAD penalty functions. In Section 2.3, we propose variable selection procedures for covariates with functional coefficients, and establish consistency, sparsity and the oracle property of the estimators. In Section 2.4, variable selection procedures for local variables with parametric coefficients are provided together with their asymptotic properties. A simple bandwidth selection method is introduced, and two computational algorithms for these two procedures are also developed in Section 2.5. Monte Carlo simulation results for two cases are reported in Section 2.6. In Section 2.7, two financial examples are extensively studied. They include functional index coefficient autoregressive models (FIAR) and functional index coefficient models (FIM) for the stock return predictability. Section 2.8 concludes the chapter and all the regularity conditions and technical proofs are gathered in the appendix.

2.2 Identification, Estimation and Penalty Function

2.2.1 Identification

The identification problem in single index model was first investigated by Ichimura (1993), and extensively studied by Li (2007) and Horowitz (2009). Meanwhile, partial conditions for identification in functional index coefficient models were showed in Fan, Yao and Cai (2003). Here we present the conditions for identification below.

Theorem 1. (Identification in functional index coefficient models) Assume that dependent variable Y is generated by equation (2.1), $X_i = (X_{1i}, X_{2i}, \dots X_{pi})^T$ are $p-$dimensional vector variables and Z are $d-$ dimensional vector variables. $\beta \in R^d$ are $d-$dimensional unknown parameters and $g(\cdot) = (g_1(\cdot) \dots g_p(\cdot))^T$ are $p-$dimensional unknown vector functional coefficients. Then β and $g(\cdot)$ are identified if and only if the following conditions hold:

(I1) The vector functions $g(\cdot)$ are continuous, bounded, and not constant everywhere.

(I2) The components of Z are continuously distributed random variables.

(I3) There exists no perfect multi-collinearity within each components of Z and none of the components of Z is constant.

(I4) There exists no perfect multi-collinearity within each components of X.

(I5) The first element of β is positive, and $\parallel \beta \parallel = 1$, where $\parallel \cdot \parallel$ is the Euclidean norm (L_2 norm) and $\parallel \beta \parallel = \sqrt{\beta_1^2 + \beta_2^2 + \cdots + \beta_d^2}$.

(I6) $E(Y|X, Z)$ can not be the form as below

$$E(Y|X, Z) = \alpha^T X \beta^T X + \gamma^T X + c,$$

where $X = Z$, and α, $\gamma \in R^d$, $c \in R$ are constant and α and β are not parallel to each other.

Remark 1: Assumption I1 holds true since continuous and bounded functions are commonly assumed in nonparametric estimation, and it is obvious that β can not be identified if $g(\cdot)$ is a constant. we can relax Assumption I2 with some components of Z being discrete random variables. But two more conditions should be imposed, see Ichimura (1993) and Horowitz (2009) for detail. The perfect multi-collinearity problem in Assumptions I3 and I4 is similar to that for the classic linear models. In fact, it is also hard to get an exact estimator of β if high correlation of components exists in Z and X, respectively. It is not identified if any one component of Z is constant. For example, if Z_1=1, $E(Y|X, Z) = g^T(\beta_1 + \beta_2 Z_2 + \cdots + \beta_d Z_d)X = f^T(\beta_2 Z_2 + \cdots + \beta_d Z_d)X$. An alternative of Assumption I5 is to let the first coefficient be 1, e.g. $\beta_1 = 1$. However, it is infeasible for estimation and variable selection simultaneously, since we do not have any prior information that whether the coefficient β_1 of Z_1 is zero or not. Assumption I6 is also imposed by Fan, Yao and Cai (2003).

2.2.2 Estimation Procedures

Model (2.1) can be regarded as a semiparametric model. Therefore, to estimate both the functionals $g(\cdot)$ and parameters β, it is common to use a two-stage approach. To estimate $g(\cdot)$, it needs a initial estimator of $\hat{\beta}$ which might have little effects on the final estimation of $g(\cdot)$ if the sample size n is large enough, since the convergence rate of the parametric estimators $\hat{\beta}$ is faster than the nonparametric function estimators $\hat{g}(\cdot)$. Here we propose variable selection and estimation in two steps:

Step One: Given an initial estimator $\hat{\beta}$ such that $\parallel \hat{\beta} - \beta \parallel = O_p(1/\sqrt{n})$, minimize the penalized local least squares $Q(\widehat{g}, \hat{\beta}, h)$ to obtain $\widehat{g}(\cdot)$, where (or maximize the penalized local likelihood),

$$Q(\widehat{g}, \hat{\beta}, h) = \sum_{j=1}^{n} \sum_{i=1}^{n} \left\{ y_i - \widehat{g}^T \left(\hat{\beta}^T Z_j \right) X_i \right\}^2 K_h \left(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j \right) + n \sum_{k=1}^{p} P_{\lambda_k} \left(\parallel \widehat{g}_{\cdot k} \parallel \right), \tag{2.4}$$

with $K(\cdot)$ is the kernel function, $K_h(z) = K(z/h)/h$ and $P_{\lambda_k}(\cdot)$ is the penalty function. As recommended, an initial estimator $\hat{\beta}$ can be obtained by various algorithms such as the method in Fan, Yao and Cai (2003). As long as the initial estimator satisfies $\parallel \hat{\beta} - \beta \parallel = O_p(1/\sqrt{n})$, as expected, the parameter estimator $\hat{\beta}$ does not have any effect on the shrinkage estimation of functional coefficients $\hat{g}(\cdot)$ in the above equation. We choose penalty term $P_{\lambda_k}(\cdot)$ as SCAD function, which is described in Section 2.2.3, and the L_2 functional norm $\parallel \widehat{g}_{\cdot k} \parallel$ is defined in Section 2.3.1. The purse of using the penalized locally weighted least squares is to select significant variable X_i in model (2.1).

Note that when the penalty term $P_{\lambda_k}(z) = \lambda_k |z|$, the penalized local likelihood becomes the Lasso type, so that the above penalized local least squares in (2.4) is reduced to the case in the paper by Wang and Xia (2009).

Step Two: Given the estimator of function $\hat{g}(\cdot)$, minimize the penalized global least squares $Q(\beta, \hat{g})$ (or maximize the penalized global likelihood), where

$$Q(\beta, \hat{g}) = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \hat{g}^T(\beta^T Z_i) X_i\right)^2 + n \sum_{k=1}^{d} \Psi_{\lambda_n}(|\beta_k|) \tag{2.5}$$

with $\Psi(\cdot)$ being a penalty function.

Clearly, the above general setting may cover several other existing variable selection procedures as a special case. For example, when $p = 1$ and the regressor $X = 1$, the above procedure becomes variable selection for the single-index model in Kong and Xia (2007), which provided an alternative variable selection method called separated cross validation to do variable selection in the single-index model. When $p = 2$ and the only one regressor is market return, then the above model reduces to the case in the paper by Cai and Ren (2011) for an application in finance. In particular, they considered a nonparametric estimate of time-varying betas and alpha in the conditional capital asset pricing model (CAPM) with a variable selection. However, Cai and Ren (2011) did not provide any theory for the variable selection procedures. Furthermore, the model includes a special case of variable selection in partially linear single-index models as addressed in Liang et al. (2010), if only the first functional coefficient $g(\cdot)$ is nonlinear and all others are constant. Finally, it includes variable selection in semiparametric regression modeling by Li and Liang (2008), if the dimension of local variables $d = 1$ and some of the functional coefficients $g(\cdot)$ are constant and others are not.

2.2.3 Penalty Functions

As pointed out by Fan and Li (2001), a good penalty function should enjoy the following three nice properties.

(a) Unbiasedness, the estimator should be unbiased when the true unknown parameter is large.

(b) Sparsity, the estimator is a threshold rule, it can set small estimator to be

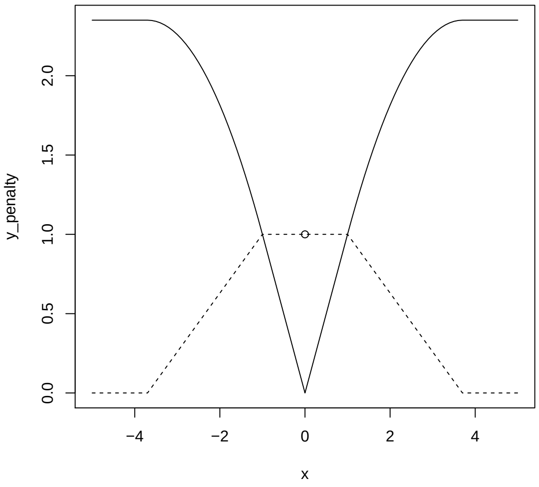

Figure 2.1: SCAD penalty function (solid line) and its derivative (dotted line) $a = 3.7$ and $\lambda = 1$

zero automatically.

(c) Continuity, the estimator is continuous to avoid instability in model prediction.

To achieve all the aforementioned three properties, Fan and Li (2001) proposed the following so called SCAD penalty function,

$$P_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & |\beta| \le \lambda, \\ -(|\beta|^2 - 2a\lambda|\beta| + \lambda^2)/[2(a-1)], & \lambda < |\beta| \le a\lambda, \\ (a+1)\lambda^2/2, & |\beta| > a\lambda, \end{cases} \tag{2.6}$$

The important property for the SCAD penalty function is that it has the following first derivative,

$$P'_\lambda(|\beta|) = \begin{cases} \lambda, & |\beta| \le \lambda, \\ (a\lambda - |\beta|)/(a-1), & \lambda < |\beta| \le a\lambda, \\ 0, & |\beta| > a\lambda, \end{cases} \text{ for some } a > 2. \tag{2.7}$$

so that it makes the computational implementation easily. The plots of penalty function and its derivative are displayed in Figure 2.2.3. It can be clearly seen that $P_\lambda(|\beta|)$ is not differentiable at 0 with respect to β. Thus it is not easy to minimize the penalized least squares functions due to its singularity. Fan and Li (2001) suggested to approximate the penalty function by a quadratic function as

$$P_\lambda(|\beta_j|) \approx P_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\{P'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)2}) \quad \text{for } \beta_j \approx \beta_j^{(0)}. \tag{2.8}$$

Alternatively, Zou and Li (2008) proposed local linear approximation (LLA) for non-concave penalty functions as

$$[P_\lambda(|\beta_j|)]' = P'_\lambda(|\beta_j|)\text{sign}(\beta_j) \approx \{P'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}\beta_j, \tag{2.9}$$

which can reduce the computational cost without losing any statistical efficiency. Recently, other algorithms such as mimorize-maximize (MM) algorithm are proposed by Hunter and Li (2005).

Given a good initial value $\beta^{(0)}$ such as maximal likelihood estimator (MLE) without the penalty term, in view of (2.8), we can find the one-step estimator as follow

$$\beta^{(1)} = \text{argmin}\frac{1}{2}(\beta - \beta^{(0)})^T[-\nabla^2\ell(\beta^{(0)})](\beta - \beta^{(0)}) + n\sum_{k=1}^{d}\frac{P'_\lambda(|\beta_k^{(0)}|)}{2|\beta_k^{(0)}|}\beta_k^2, \tag{2.10}$$

where $\nabla^2\ell(\beta^{(0)}) = \partial^2\ell(\beta_0)/\partial\beta\partial\beta^T$. As argued in Fan and Li (2001), there is no need to iterate until it converges as long as the initial estimator is reasonable. Also, the MLE estimator from the full models without penalty term can be regard as the reasonable estimator. For using the local linear approximation in Zou and Li (2008) and (2.9), the sparse one-step estimator given in (2.10) becomes to

$$\beta^{(1)} = \text{argmin}\frac{1}{2}(\beta - \beta^{(0)})^T[-\nabla^2\ell(\beta^{(0)})](\beta - \beta^{(0)}) + n\sum_{k=1}^{d}\frac{P'_\lambda(|\beta_k^{(0)}|)}{|\beta_k^{(0)}|}\beta_k. \tag{2.11}$$

As demonstrated in Zou and Li (2008), this one step estimator is as efficient as the fully iterative estimator, provided that the initial estimator is good enough. For example, we let $\beta^{(0)}$ be the maximal likelihood estimator without the penalty term.

2.3 Variable Selection for Covariates with Functional Coefficients

2.3.1 Notations and Technical Conditions

Let $\{(X_i, Z_i, y_i)\}$ be a strictly stationary and strong mixing sequence, and $f(z, \beta)$ be the density function of $z = \beta^T Z$, where β is an interior point of the compact set $\mathbb{B}$. Define $\mathcal{A}_z = \{Z | f(Z, \beta) \geq \varepsilon, \forall \beta \in \mathbb{B} \quad \text{and} \quad \exists a, b, \beta^T Z \in [a, b]\}$ as the domain of Z, such that $\beta^T Z$ is bounded and the density of $f(Z, \beta)$ is bounded away from 0. Also, define the domain of bandwidth h, $\mathcal{H}_n = \{h| \quad \exists C_1 \text{and} C_2, C_2 n^{-1/5} < h < C_1 n^{-1/5}\}$. For $Z \in \mathcal{A}_z, \beta \in \mathbb{B}$, and $h \in \mathcal{H}_n$, define n by p matrix penalized estimator as

$$\widehat{G}\left(\widehat{\beta}\right) = \left[\widehat{g}\left(\widehat{\beta}^T Z_1\right), \cdots, \widehat{g}\left(\widehat{\beta}^T Z_n\right)\right]^T = \left[\widehat{g}_{\cdot 1}, \cdots, \widehat{g}_{\cdot p}\right],$$

where

$$\widehat{g}\left(\widehat{\beta}^T Z\right) = \left[\widehat{g}_1\left(\widehat{\beta}^T Z\right), \cdots, \widehat{g}_p\left(\widehat{\beta}^T Z\right)\right]^T \in R^p,$$

and

$$\widehat{g}_{\cdot k} = \left[\widehat{g}_k\left(\hat{\beta}^T Z_1\right), \cdots, \widehat{g}_k\left(\widehat{\beta}^T Z_n\right)\right]^T \in R^n.$$

Similarly, we can define true value $G_0(\beta), g_0\left(\beta^T Z\right)$ and $g_{0 \cdot k}$. Without loss of generality, we assume that first p_0 functional coefficients are non-zero, and other $p - p_0$ functional coefficients are zero, e.g. $\| g_{\cdot k} \| \neq 0$ and $g_{\cdot k}$ are not constant everywhere for $1 \leq k \leq p_0$, $\| g_{\cdot k} \| = 0$ for $p_0 < k \leq p$. Let $\alpha_n = \max\{P'_\lambda(\| g_{\cdot k} \|) : 1 \leq k \leq p_0\}$. Then $\alpha_n = 0$ as $n \to \infty$. For an arbitrary matrix $X = (X_{ij})$, we define L_2 norm as $\| X \| = \sqrt{\sum_{i,j} X_{i,j}^2}$ and the kernel function $K_h = K(x/h)/h$. Also we define object function

$$Q(\widehat{G}, \hat{\beta}, h) = \sum_{j=1}^{n} \sum_{i=1}^{n} \{y_i - \widehat{g}^T\left(\hat{\beta}^T Z_j\right) X_i\}^2 K_h\left(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j\right) + n \sum_{k=1}^{p} P_{\lambda_k}(\| \widehat{g}_{\cdot k} \|). \tag{2.12}$$

By minimizing the above object function with respect to $\hat{g}(\cdot)$, one can obtain the penalized local least squares estimator for $g(\cdot)$.

To study the asymptotic distribution of the penalized local least squares estimator, we impose some technical conditions as follows.

(A1) The vector functions $g(\cdot)$ are bounded, not constant everywhere and have continuous second order derivatives with respect to the support of $\mathcal{A}_z$.

(A2) For any $\beta \in \mathbb{B}$ and $Z \in \mathcal{A}_z$, the density function $f\left(\beta^T Z\right)$ is continuous and there exists a small positive ε such that $f\left(\beta^T Z\right) > \varepsilon$.

(A3) The kernel function $K(z)$ is twice continuously differentiable on the support $(-1, 1)$, Let $\int z^2 K(z)dz = \mu_2$, and $\int K^2(z)dz = \nu_0$.

(A4) $\lim_{n\to\infty} \inf_{\theta\to 0^+} P'_{\lambda_n}(\theta)/\lambda_n > 0$, $n^{-1/10}\lambda_n \to 0$, $h \propto n^{-1/5}$ and $\| \hat{\beta} - \beta \| = O_p(1/\sqrt{n})$.

(A5) Define $\Omega\left(\beta^T Z\right) = E\left(X_i X_i^T | \beta^T Z\right)$, $\Omega\left(\beta^T Z\right)$ is nonsingular and has bounded second order derivative on $\mathcal{A}_z$.

(A6) $\{(X_i, Z_i, y_i)\}$ is a strictly stationary and strongly mixing sequence with mixing coefficient $\alpha\left(m\right) = O\left(\rho^m\right)$ for some $0 < \rho < 1$.

(A7) Let $z = \beta^T Z$, the conditional density $f\left(z_i, z_s | z_j\right)$ is continuous and has bounded second order derivative.

(A8) Let $\Omega\left(z_i, z_s, z_j\right) = E\left(X_i X_i^T X_s X_s^T | z_i, z_s, z_j\right)$ be continuous and has bounded second order derivative. Define $\Omega\left(z_j, z_j, z_j\right) = \Omega\left(z_i, z_s, z_j\right)|_{z_i = z_j, z_s = z_j}$, $\Omega_1\left(z_i, z_s, z_j\right) = \partial\Omega\left(z_i, z_s, z_j\right)/\partial z_i$ and $\Omega_2\left(z_i, z_s, z_j\right) = \partial\Omega\left(z_i, z_s, z_j\right)/\partial z_s$.

Remark 2: The conditions in A2 imply that the distances between two ranked values $\beta^T Z_{(i)}$ are at most order of $O_p(\log n/n)$ (Janson 1987). For any value $\Lambda \in \mathcal{A}_z$, we can find a closest value $\beta^T Z_j$ of Λ such that $\mid \beta^T Z_j - \Lambda \mid = O_p(\log n/n)$. With the conditions in A1, $\parallel g(\beta^T Z_j) - g(\Lambda) \parallel = O_p(\log n/n)$, which is smaller order of nonparametric convergence rate $n^{-2/5}$. This implies that we only need to estimate $\hat{g}(\beta^T Z_i)$ for $i = 1, 2, \cdots, n$ rather than $\hat{g}(\Lambda)$ for all the domain $\Lambda \in \mathcal{A}_z$. For the detailed argument, we refer to the paper by Wang and Xia (2009). A3 is the common assumption in nonparametric estimation. $\parallel \hat{\beta} - \beta \parallel = O_p(1/\sqrt{n})$ in A4 implies that the estimators of $\hat{\beta}$ have little effect in the estimation of $\hat{g}(\cdot)$ if the sample size n is large, since the convergence rate of the local parametric estimators $\hat{\beta}$ is faster than the nonparametric function estimators $\hat{g}(\cdot)$. The assumptions in A5 - A8 are very standard and used for the proof under mixing conditions; see Cai, Fan and Yao (2000).

2.3.2 Asymptotic Properties

To obtain the oracle property of the estimator, firstly, we present Lemma 1 - Lemma 3 below.

Lemma 1: Let $\{(X_i, Z_i, y_i\}$ be a strong mixing and strictly stationary sequence with mixing coefficient as in (A6), the conditional densities $f_{z_1|x_1}(z|X)$ and $f_{z_1,z_\ell|x_1,x_\ell}(z_1, z_\ell| X_1, X_\ell)$ are bounded for all $\ell > 1, h \propto n^{-1/5}$, then

$$\begin{aligned}\sup_{Z\in\mathcal{A}_z,\beta\in\mathbb{B},h\in\mathcal{H}_n} & \left|\frac{1}{n}\sum_{i=1}^{n}[K_h(\beta^T Z_i - \beta^T Z)X_i X_i^T - E(K_h(\beta^T Z_i - \beta^T Z)X_i X_i^T)]\right| \\ = & \quad O\left(\frac{(\log n)^{1/2}}{n^{3/5}h}\right).\end{aligned}$$

The above lemma directly follows from Lemma A.2 of Xia and Li (1999). Let $\hat{\Sigma}(\beta^T Z) = \frac{1}{n}\sum_{i=1}^{n} K_h(\beta^T Z_i - \beta^T Z)X_i X_i^T$. By Assumption A5, it is not hard to derive

$E(K_h(\beta^T Z_i - \beta^T Z)X_i X_i^T) = f(\beta^T Z)\Omega(\beta^T Z) + O(h^2)$. Then, we have

$$\begin{aligned} & \left|\hat{\Sigma}(\beta^T Z) - f(\beta^T Z)\Omega(\beta^T Z)\right| \\ \leq \quad & \left|\hat{\Sigma}(\beta^T Z) - E(K_h(\beta^T Z_i - \beta^T Z)X_i X_i^T)\right| \\ & + \left|E(K_h(\beta^T Z_i - \beta^T Z)X_i X_i^T) - f(\beta^T Z)\Omega(\beta^T Z)\right| \\ \leq \quad & O\left(\frac{(\log n)^{1/2}}{n^{3/5}h}\right) + O(h^2). \end{aligned} \tag{2.13}$$

Lemma 2: Let $\{(X_i, Z_i, y_i\}$ be a strong mixing and strictly stationary sequence. Under Assumptions A1−A8. Assume that $h \propto n^{-1/5}$, $n^{-1/10}\alpha_n \to 0$ and $\| \hat{\beta} - \beta \| = O_p(1/\sqrt{n})$, we have

$$n^{-1}\sum_{i=1}^{n} \| \hat{g}\left(\hat{\beta}^T Z_i\right) - g_0\left(\beta_0^T Z_i\right) \|^2 = O_p(n^{-4/5}).$$

Lemma 3: Let $\{(X_i, Z_i, y_i\}$ be a strong mixing and strictly stationary sequence, $h \propto n^{-1/5}$, $\lim_{n\to\infty} \inf_{\theta\to 0^+} P'_{\lambda_n}(\theta)/\lambda_n > 0$, and $n^{-1/10}\lambda_n \to 0$. Then, $\| \hat{g}_{.k} \| = 0$ as $n \to \infty$ for $k > d_0$.

The above lemma shows the sparsity of the estimator $\hat{g}_{.k}$ for $k > d_0$.

Theorem 2 (Oracle Property): Let (X_i, Z_i) be a strong mixing and strictly stationary sequence. Under Assumptions (A1)−(A8), $\lim_{n\to\infty} \inf_{\theta\to 0^+} P'_{\lambda_n}(\theta)/\lambda_n > 0$, $h \propto n^{-1/5}$ and $n^{-1/10}\lambda_n \to 0$ as $n \to \infty$, then

(a) Sparsity: $\| \hat{g}_b(\hat{\beta}^T Z_j) \| = 0 \quad j = 1, \cdots, n$, where

$$\hat{g}_b(\hat{\beta}^T Z_j) = [\hat{g}_{p_0+1}(\hat{\beta}^T Z_j), \hat{g}_{p_0+2}(\hat{\beta}^T Z_j), \cdots, \hat{g}_p(\hat{\beta}^T Z_j)]^T.$$

(b) Asymptotic Normality:

$$\sqrt{nh}\left(\hat{g}_a(\hat{\beta}^T Z_j) - g_{0a}(\beta_0^T Z_j) - B(\beta_0^T Z_j)\right) \sim N(0, V_{\beta_0^T Z_j}),$$

where $V_{\beta_0^T Z_j} = \nu_0 M^{-1}_{\beta_0^T Z_j}\sigma^2$, and

$$B(\beta_0^T Z_j) = h^2\mu_2 M^{-1}_{\beta_0^T Z_j}\{\int X_a X_a^T \dot{g}(\beta_0^T Z_j) f_z(X_a, \beta_0^T Z_j) dX_a + \frac{1}{2}\ddot{g}(\beta_0^T Z_j) M_{\beta_0^T Z_j}\}$$

with $M_{\beta_0^T Z_j} = f(\beta_0^T Z_j)\Omega(\beta_0^T Z_j)$ and $f_z(X_a, z) = \partial f(X_a, z)/\partial z$.

Theorem 2 indicates that the estimator merits the oracle property by our variable selection procedures. Note that our result can be extended to entire domain $Z \in \mathcal{A}_z$ from Remark 2.

2.4 Variable Selection for Local Significant Variables with Parametric Coefficients

2.4.1 Notations and Technical Conditions

Let $\{(X_i, Z_i, y_i)\}$ be a strictly stationary and strong mixing sequence,$f(z, \beta)$ be the density function of $z = \beta^T Z$, and β be an interior point of the compact set $\mathbb{B}$. Define $\mathcal{A}'_z = \{Z : f(Z, b) \geq \delta, \forall b \in \mathbb{B}\}$, where δ is a small positive constant. Also, define penalized least squares object function

$$Q(\beta, \hat{g}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{g}^T(\beta^T Z_i)X_i)^2 + n\sum_{k=1}^{d} P_{\lambda_n}(|\beta_k|). \tag{2.14}$$

We assume the first d_1 coefficients of β are nonzero, and all rest of parameters are zero, e.g., $\beta_0 = (\beta_{10}^T, \beta_{20}^T)^T$, all elements of β_{10} with dimension d_1 are nonzero, and $d - d_1$ dimensional coefficients $\beta_{20} = 0$. Finally, define $V_n = \sum_{i=1}^{n} g_z^T(\beta_0^T Z_i)X_i(Z_i - E(Z_i|\beta_0^T Z_i))\varepsilon_i$, where vector $g_z(\cdot)$ is the first derivative of function $g(\cdot)$ vector, and ε_i are independently, identically distributed (i.i.d) with mean 0 and standard deviation σ. Let $\widetilde{V}_0 = \frac{1}{n}Var(V_n)/\sigma^2$, and define ε be an asymptotically standard normal random $d-$dimensional vector such that $V_n = n^{1/2}\sigma\widetilde{V}_0^{1/2}\varepsilon$. $V_{1n} = \sum_{i=1}^{n} g_z^T(\beta_{10}^T Z_{1i})X_i(Z_{1i} - E(Z_{1i}|\beta_{10}^T Z_{1i}))\varepsilon_{1i}$, where ε_{1i} is the same as ε_i since $\beta_{20} = 0$. Similarly, we define $\widetilde{V}_{10} = \frac{1}{n}Var(V_{1n})/\sigma^2$ and ε_1 be an asymptotically standard normal random d_1-dimensional vector such that $V_{1n} = n^{1/2}\sigma\widetilde{V}_{10}^{1/2}\varepsilon_1$.

To study the asymptotic distribution of the penalized least squares estimator $\hat{\beta}$, we impose some technique conditions as below.

(B1) The vector functions $g(\cdot)$ are bounded, not constant everywhere and have continuous second order derivatives with respect to the support of $\mathcal{A}'_z$.

(B2) The components of Z are continuously distributed random variables.

(B3) The kernel function $K(z)$ is twice continuously differentiable on the support $(-1, 1)$, Let $\int z^2 K(z)dz = \mu_2$, and $\int K^2(z)dz = \nu_0$.

(B4) $\lim_{n\to\infty} \inf_{\theta\to 0^+} P'_{\lambda_n}(\theta)/\lambda_n > 0$, $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ and $h \propto n^{-1/5}$.

(B5) $\{(X_i, Z_i, y_i)\}$ is a strictly stationary and strong mixing sequence with mixing coefficient $\alpha(m) = O(\rho^m)$ for some $0 < \rho < 1$.

(B6) $E(\varepsilon_i|X_i, Z_i) = 0, E(\varepsilon_i^2|X_i, Z_i) = \sigma^2, E|X_i|^m < \infty$ and $E|y_i|^m < \infty$ for all $m > 0$.

Remark 3: The conditions in B1 and B2 and Section 2.2.1 ensure identification of the models. The second order differentiability of vector functions $g(\cdot)$ in B1 and kernel function $K(z)$ in B3 leads to that the order of bias term for nonparametric estimator is $O_p(h^2)$. This assumption is standard for a nonparametric method. The assumptions in B4 indicate the oracle property in Theorem 4. An alternative condition for bandwidth in Ichimura (1993) is $nh^8 \to 0$. However, the condition $nh^8 \to 0$ is still satisfied with our condition $h \propto n^{-1/5}$ in B4. Assumptions in B5 are the common conditions with week dependent data. Most financial models satisfy this conditions, such as ARCH and GARCH models; see Cai (2002). For Assumption B6, it is not hard to extend to the heteroscedasticity case, $E(\varepsilon_i^2|X_i, Z_i) = \sigma^2(X_i, Z_i)$, while Assumption B6 requires the moment conditions of X and y so that the Chebyshev inequality can be applied.

2.4.2 Asymptotic Properties

It follows from Theorem 1 in Xia and Li (1999) that

$$\hat{Q}_1(\beta, h) = \tilde{S}(\beta) + T(h) + R_1(\beta, h) + R_2(h),$$

where $\hat{Q}_1(\beta, h) = \sum_{i=1}^{n}(y_i - \hat{g}^T(\beta^T Z_i)X_i)^2$, $T(h)$ and $R_2(h)$ do not depend on β, and $R_1(\beta, h)$ is an ignorable term. Furthermore,

$$\tilde{S}(\beta) = n[\tilde{V}_0^{1/2}(\beta - \beta_0) - n^{-1/2}\sigma\varepsilon]^T[\tilde{V}_0^{1/2}(\beta - \beta_0) - n^{-1/2}\sigma\varepsilon] + R_3 + R_4(\beta),$$

where R_3 does not depend on β and h, and $R_4(\beta)$ is an ignorable term.

Theorem 3: Let $\{(X_i, Z_i, y_i)\}$ be a strictly stationary and strong mixing sequence. Let $a_n = max\{P'_{\lambda_n}(\beta_k) : \beta_k \neq 0\}$, and $\hat{\beta} = \arg min_{\beta \in \mathbb{B}} Q(\beta, \hat{g})$. Under Assumptions B1−B6 and if $max\{P''_{\lambda_n}(\beta_k) : \beta_k \neq 0\} \to 0$, then the order of $\parallel \hat{\beta} - \beta_0 \parallel$ is $O_p(n^{-1/2} + a_n)$.

If the penalty function is SCAD function, $a_n = 0$ as sample size $n \to 0$, and $\parallel \hat{\beta} - \beta_0 \parallel = O_p(n^{-1/2})$.

Theorem 4 (Oracle Property). Let $\{(X_i, Z_i, y_i)\}$ be a strictly stationary and strong mixing sequence. Under Assumptions B1−B6, by assuming $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then

(a) **Sparsity:**

$$\widehat{\beta}_2 = 0.$$

(b) **Asymptotic Normality:**

$$\sqrt{n}(\widehat{\beta}_1 - \beta_{10}) \to N(0, \sigma^2 V_{10}^{-1}),$$

where $V_{10} = E\left[\left(Z - E\left(Z|\beta_{10}^T Z\right)\right) g_z^T\left(\beta_{10}^T Z\right) X\right]\left[\left(Z - E\left(Z|\beta_{10}^T Z\right)\right) g_z^T\left(\beta_{10}^T Z\right) X\right]^T$.

Theory 4 shows that our variable selection procedures of minimizing penalized least squares objection function enjoy the oracle property.

2.5 Practical Implementations

2.5.1 Selection for the Bandwidth and Tuning Parameters

To do the nonparametric estimation and variable selection simultaneously, we should choose suitable regularization parameters, bandwidth h for nonparametric estimator and λ's for penalty terms. For simplicity, we just consider globally bandwidth selection rather point-wise in this Chapter. There are several popular methods to choose these two parameters, for example, plug-in bandwidth selector (Liang, and Li, 2009) for bandwidth selection, and K-fold cross validation (Breiman, 1995; Fan and Li, 2001), generalized cross validation (Tibshirani, 1996; Fu, 1998; Fan and Li, 2001), BIC (Liang and Li, 2009; Ma and Li, 2010; Liang et al., 2010) for tuning parameters, and so on.

However, Wang, Li and Tsai (2007) showed that BIC can select and estimate the true model consistently, where generalized cross validation cannot and it comes with an over fitting effect in the resulting model. Further, Zhang, Li and Tsai (2010) presented that the BIC-type selector identifies the true model consistently, and the resulting estimator possesses the oracle property. In contrast, the AIC-type selector tends to be less efficient and over fitting in the final model.

This motivates us to select the bandwidth h and tuning parameters λ's simultaneously with BIC-type criterion. We define our BIC criterion as

$$\mathrm{BIC}(h,\lambda) = \log\,\mathrm{SSE}(h,\lambda) + \mathrm{df}(h,\lambda)\log(n)/n,$$

where $\mathrm{SSE}(h,\lambda)$ is the sum of squared errors obtained from the penalized least squares object function with parameters (h,λ), and $\mathrm{df}(h,\lambda)$ is the number of nonzero coefficients of $\hat{\beta}$ conditional on parameters h and λ. However, it is still computationally expensive to choose d-dimensional tuning parameters λ. Fan and Li (2004) suggested to let tuning parameters λ_k be proportional to the corresponding standard deviation of un-penalized estimator $\hat{\beta}_k^{(0)}$ e.g. $\lambda_k = \lambda_0\hat{\sigma}(\hat{\beta}_k^{(0)})$, where $\hat{\sigma}(\hat{\beta}_k^{(0)})$ is the standard deviation of un-penalized estimator $\hat{\beta}_k^{(0)}$. Another difficulty is that we should standardize all variables before doing variable selection. It will be complicated in application if we want to find the coefficient of original variables. Adopted the idea of Fan and Li (2004), we let $\lambda_k = \lambda_0 \hat{\sigma}(\hat{\beta}_k^{(0)})$ reduce the dimension of λ. In fact, this method can overcome two difficulties above since $\hat{\sigma}(\hat{\beta}_k^{(0)})$ is the proportion of $\hat{\sigma}^{-1}(z_k)$ if variables z_k are orthogonal. The theoretical property of BIC(h, λ) and dimension reduction with $\lambda_k = \lambda_0 \hat{\sigma}(\hat{\beta}_k^{(0)})$ need further research.

2.5.2 Computational Algorithms

Algorithms for Step 1

To select significant covariates and estimate the functional coefficients simultaneously, we present algorithms as follows.

(1) Find an initial value $\hat{\beta}^{(0)}$ such that $\| \hat{\beta}^{(0)} - \beta_0 \| = O_p(1/\sqrt{n})$ and initial values $g_0^{(0)}$ by the un-penalized object function algorithms in Fan, Yao and Cai (2003).

(2) Local quadratic approximation: For given $\hat{\beta}^{(0)}$ and initial values $g_0^{(0)}$, the object function $Q(g, \hat{\beta}, h)$ can be locally approximated by (Wang and Xia, 2009)

$$\begin{aligned} Q(g, \hat{\beta}, h) \approx \quad & \sum_{j=1}^{n} \sum_{i=1}^{n} \{y_i - g^T\left(\hat{\beta}^T Z_j\right) X_i\}^2 w(\hat{\beta}^T Z_i) K_h\left(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j\right) \\ & + n \sum_{k=1}^{p} P'_{\lambda_k}\left(\| g_{0\cdot k}^{(0)} \|\right) \frac{\| g_{\cdot k} \|^2}{\| g_{0\cdot k}^{(0)} \|} \\ = \quad & \sum_{j=1}^{n} \Bigg\{ \sum_{i=1}^{n} [y_i - g^T\left(\hat{\beta}^T Z_j\right) X_i]^2 w(\hat{\beta}^T Z_i) K_h\left(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j\right) \\ & + n \sum_{k=1}^{p} P'_{\lambda_k}\left(\| g_{0\cdot k}^{(0)} \|\right) \frac{g_k^2(\hat{\beta}^T Z_j)}{\| g_{0\cdot k}^{(0)} \|} \Bigg\}. \end{aligned}$$

Then, the minimizer of $Q(g, \hat{\beta}, h)$ is $g^{(1)}$ with j-th row given by

$$\begin{aligned} g^{(1)}(\hat{\beta}^T Z_j) = & \quad \left\{ \sum_{i=1}^{n} X_i X_i^T w(\hat{\beta}^T Z_i) K_h \left(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j \right) + n\Sigma(\| \ g^{(0)} \ \|) \right\}^{-1} \\ & \times \sum_{i=1}^{n} X_i y_i w(\hat{\beta}^T Z_i) K_h \left(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j \right), \end{aligned}$$

where $\Sigma(\| \ g^{(0)} \ \|) = diag \left(P'_{\lambda_1}(\| \ g_{\cdot 1}^{(0)} \ \|)/ \ \| \ g_{\cdot 1}^{(0)} \ \|, \cdots, P'_{\lambda_p}(\| \ g_{\cdot p}^{(0)} \ \|)/ \ \| \ g_{\cdot p}^{(0)} \ \| \right)$, and $w(\hat{\beta}^T Z_j)$ is the weight function of $\hat{\beta}^T Z_j$, which avoids the boundary effect. We let $w(\hat{\beta}^T Z_j) = 1$ if $\hat{\beta}^T Z_j$ is between 5% sample quantile and 95% sample quantile of $\hat{\beta}^T Z$, otherwise $w(\hat{\beta}^T Z_j) = 0$.

(3) Repeat the above steps until the convergence is achieved.

Algorithms for Step 2

First, for computing the local constant estimator of functional coefficients with given β, we need to find the minimizer of the penalized least squares object function,

$$\begin{aligned} \hat{g} &= arg \min_{g \in G} Q(\beta, g) \\ &= arg \min \Big\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{g}^T(\beta^T Z_i) X_i)^2 w(\beta^T Z_i) + n \sum_{k=1}^{d} P_{\lambda_n}(|\beta_k|) \Big\}, \end{aligned}$$

where $w(\beta^T Z_j)$ is the weight to avoid the boundary problem, $\hat{g} = (\hat{g}^T(\beta^T Z_1), \hat{g}^T(\beta^T Z_2), \cdots, \hat{g}^T(\beta^T Z_n))^T$ is an $n \times p$ matrix, and $\hat{g}^T(\beta^T Z_j) = (\hat{g}_1(\beta^T Z_j), \cdots, \hat{g}_p(\beta^T Z_j))^T$ is a p-dimensional functional coefficients vector. The penalized term is constant for a given β. This leads to the local constant estimators of functional coefficients.

$$\hat{g}^T(\beta^T Z_j) = arg \min \Big\{ \sum_{i=1}^{n} (y_i - \hat{g}^T(\beta^T Z_j) X_i)^2 K_h(\beta^T Z_i - \beta^T Z_j) w(\beta^T Z_j) \Big\},$$

which implies that

$$\hat{g}^T(\beta^T Z_j) = \{X^T \mathcal{K}(Z_j) X\}^{-1} X^T \mathcal{K}(Z_j) Y, \tag{2.15}$$

where $Y = (y_1, \cdots, y_n)^T$, X is an $n \times p$ matrix, and $\mathcal{K}(Z_j)$ is an $n \times n$ diagonal matrix with $K_h(\beta^T Z_i - \beta^T Z_j)w(\beta^T Z_j)$ as its i-th diagonal element. Note that in our simulations, we let $w(\beta^T Z_j) = 1$ if $\beta^T Z_j$ is between 5% sample quantile and 95% sample quantile of $\beta^T Z$, otherwise $w(\beta^T Z_j) = 0$.

Next is about the Newton-Raphson estimator of β with given $\hat{g}(\cdot)$. Following the algorithm of local quadratic approximation in Fan and Li (2001), the penalty term can be locally approximated by a quadratic function as

$$P_\lambda(|\beta_k|) \approx P_\lambda(|\beta_{k0}|) + \frac{1}{2}\{P'_\lambda(|\beta_{k0}|)/|\beta_{k0}|\}(\beta_k^2 - \beta_{k0}^2).$$

Then, we only need to minimize

$$Q_1(\beta, \hat{g}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{g}^T(\beta^T Z_i)X_i)^2 w(\beta^T Z_i) + n\sum_{k=1}^{p}\frac{1}{2}\{P'_\lambda(|\beta_{k0}|)/|\beta_{k0}|\}\beta_k^2.$$

Given any initial value β_0 that is close to the minimizer of $Q_1(\beta, g)$, the objective function can be locally approximated by

$$Q_1(\beta, \hat{g}) \approx \ell(\beta_0, \hat{g}) + \nabla\ell(\beta_0, \hat{g})^T(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)^T\nabla^2\ell(\beta_0, \hat{g})(\beta - \beta_0) + \frac{1}{2}n\beta^T\Sigma_\lambda(\beta_0)\beta,$$

where

$$\ell(\beta_0, \hat{g}) = \frac{1}{2}\sum_{i=1}^{n}\{y_i - \hat{g}^T(\beta_0^T Z_i)X_i\}^2 w(\beta_0^T Z_i),$$

$$\nabla\ell(\beta_0, \hat{g}) = -\sum_{i=1}^{n}\{y_i - \hat{g}^T(\beta_0^T Z_i)X_i\}\{\hat{\dot{g}}^T(\beta_0^T Z_i)X_i\}Z_i w(\beta_0^T Z_i),$$

$$\begin{aligned}\nabla^2\ell(\beta_0, \hat{g}) = \quad & \sum_{i=1}^{n}\{\hat{\dot{g}}^T(\beta_0^T Z_i)X_i\}^2 Z_i Z_i^T w(\beta_0^T Z_i) \\ & -\sum_{i=1}^{n}\{y_i - \hat{g}^T(\beta_0^T Z_i)X_i\}\{\hat{\ddot{g}}^T(\beta_0^T Z_i)X_i\}Z_i Z_i^T w(\beta_0^T Z_i),\end{aligned}$$

and

$$\Sigma_\lambda(\beta_0) = \mathrm{diag}\{P'_\lambda(|\beta_{10}|)/|\beta_{10}|, \cdots, P'_\lambda(|\beta_{p0}|)/|\beta_{p0}|\}.$$

Here, $\hat{\dot{g}}(\cdot)$ and $\hat{\ddot{g}}(\cdot)$ are the estimators of first and second derivative of $g(\cdot)$ respectively. In the derivation, as suggested by Fan, Yao and Cai (2003), the derivative of the weight of function $w(\cdot)$ is assumed to be 0. This leads to

$$\beta = \beta_0 - \{\nabla^2 \ell(\beta_0, \hat{g}) + n\Sigma_\lambda(\beta_0)\}^{-1}\{\nabla \ell(\beta_0, \hat{g}) + n\Sigma_\lambda(\beta_0)\beta_0\}. \qquad (2.16)$$

The detailed algorithm is summarized as the following steps:

S1: Given an initial value β_0, estimate $\hat{g}^T(\beta_0^T Z_j)$ for $j = 1 \cdots n$ by (2.15).

S2: Estimate $\hat{\dot{g}}(\cdot)$ and $\hat{\ddot{g}}(\cdot)$ by local cubic estimator.

S3: Estimate new β by (2.16) and replace β_0 with its standardized form of the new β, which has unite norm and positive first component.

S4: Repeat Step 1 - Step 3 until β converges.

Remark 4: $\hat{\dot{g}}(\cdot)$ and $\hat{\ddot{g}}(\cdot)$ can be estimated by local cubic estimator. An alternative standardization method is to let the first coefficient of Z be one. However, we do not adopt this method since we are not sure the first coefficient of Z is zero or not.

2.6 Monte Carlo Simulation Studies

2.6.1 Simulation of the Variable Selection for Covariates

We study the performance of the variable selection for the covariates with functional coefficients. The program is implemented with R software. In our simulations, the optimal bandwidth and the tuning parameter λ_n are chosen by BIC criterion described in Section 2.5.1. The Epanechnikov kernel $K(x) = 0.75(1 - x^2)(|x| \leq 1)$ is used, and we let the value of a in SCAD be 3.7, as suggested in Fan and Li (2001).

In this following example, the data generating process is

$$y_i = (Z_{1i} + Z_{2i}) + (Z_{1i} + Z_{2i})^2 X_{1i} + \sigma\varepsilon_i, \quad 1 \leq i \leq n,$$

and our working model is

$$y_i = g_0(\beta^T Z_i)+g_1(\beta^T Z_i)X_{1i}+g_2(\beta^T Z_i)X_{2i}+g_3(\beta^T Z_i)X_{3i}+g_4(\beta^T Z_i)X_{4i}+e_i, 1 \leq i \leq n,$$

where ε_i is generated from standard normal distribution, $Z = (Z_1, Z_2)^T$, $Z_1 = \Phi(Z_1^*)$, $Z_2 = \Phi(Z_2^*)$, and $\Phi(\cdot)$ is the cumulated standard normal distribution function. The six dimensional vector $(Z_1^*, Z_2^*, X_1, X_2, X_3, X_4)^T$ is generated from vector auto-regression process

$$\begin{pmatrix} Z_i^* \\ X_i \end{pmatrix} = \mathbb{A} \begin{pmatrix} Z_{i-1}^* \\ X_{i-1} \end{pmatrix} + \xi_i,$$

where $Z^* = (Z_1^*, Z_2^*)^T$, $X = (X_1, X_2, X_3, X_4)^T$, $\mathbb{A}$ is a 6×6 matrix with the diagonal elements being 0.15 and all others being 0.05. The initial values $(Z_1^*, X_1)^T$ and each component of the random vector term ξ_i are generated from identically independently standard normal distribution. We consider three sample sizes as $n = 100$, $n = 200$ and $n = 400$ and we repeat the simulation 300 for each sample. Also, we consider both cases of $\sigma = 3$ and $\sigma = 6$. For each replication, we first find the un-penalized estimator $\hat{\beta}$ and $\hat{g}(\cdot)$ as our initial estimators. Similar to Fan and Li (2001), the average number of correct and incorrect shrinkage to zero are reported in Table 2.1, in which "Correct" represents the average number of three zero coefficients correctly shrinkage to 0, and "Incorrect" stands for the average number of two no-zero functional coefficients erroneously set to 0.

Table 2.1 reports the simulation results of SCAD variable selection for the covariates. It can be also seen that "Correct" and "Incorrect" numbers perform better with large sample size and smaller noise and it performs as good as oracle estimator if the sample size $n = 400$ and $\sigma = 3$. In a sum, this simulation shows that the proposed variable selection procedures perform fairly well in the finite sample cases.

Table 2.1: Simulation results for the covariates with functional coefficients

	Average number of 0 coefficients			
	$\sigma = 3$		$\sigma = 6$	
	Correct[a]	Incorrect	Correct	Incorrect
N=100				
SCAD	2.7	0.1	1.82	0.3
Oracle	3	0	3	0
N=200				
SCAD	2.91	0	2.27	0.2
Oracle	3	0	3	0
N=400				
SCAD	3	0	2.69	0.03
Oracle	3	0	3	0

[a] The “Correct” represents the average number of three zero coefficients correctly shrinkage to 0; the “Incorrect” represents the average number of two no-zero coefficients erroneously set to 0.

2.6.2 Simulation of the Variable Selection for the Local Variables

To examine the performance of the variable selection for the local variables with parametric coefficients, similar to Tibshirani (1996) and Fan and Li (2001), our data generating process is given below

$$y_i = u_i + u_i^2 X_i + \sigma \varepsilon_i,$$

where $u_i = Z_i^T \beta$, $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, ε_i is generated from standard normal distribution. Furthermore, the nine dimensional vector $(Z_i^T, X_i)^T$ is generated from vector auto-regression process

$$\begin{pmatrix} Z_i \\ X_i \end{pmatrix} = \mathbb{A}^* \begin{pmatrix} Z_{i-1} \\ X_{i-1} \end{pmatrix} + e_i,$$

where $\mathbb{A}^*$ is a 9×9 matrix with the diagonal elements being 0.15 and all others being 0.05. The initial values $(Z_1^T, X_1)^T$ and each elements of the random vector e_i are generated from identically independently distributed (i.i.d) normal with mean 0 and standard deviation 1. We consider three sample sizes as $n = 100$, $n = 200$ and

$n = 400$ and we repeat the simulation 300 for each sample. We also consider both cases of $\sigma = 7.5$ and $\sigma = 15$.

In the simulation, the optimal bandwidth and the tuning parameter λ_n are chosen by BIC criterion described in Section 2.5.1. The Epanechnikov kernel $K(x) = 0.75(1 - x^2)(|x| \leq 1)$ is used. The average number of correct and incorrect shrinkage to zero are reported in Table 2.2.

Table 2.2: Simulation results for the local variables with parametric coefficients

	Average number of 0 coefficients			
	$\sigma = 7.5$		$\sigma = 15$	
	Correct[a]	Incorrect	Correct	Incorrect
N=100				
SCAD	4.84	0.05	3.51	0.25
Oracle	5	0	5	0
N=200				
SCAD	5	0	4.21	0.10
Oracle	5	0	5	0
N=400				
SCAD	5	0	4.87	0.01
Oracle	5	0	5	0

[a] The“Correct” represents the average number of five zero coefficients correctly shrinkage to 0; the “Incorrect” represents the average number of two no-zero coefficients incorrectly shrinkage to 0.

Table 2.2 shows that the simulation results of SCAD variable selection for the local variables with parametric coefficients. It can also be seen that “Correct” and “Incorrect” numbers perform better with large sample size and smaller noise. Specifically, it performs as good as oracle estimator if the sample size $n \geq 200$ and $\sigma = 7.5$. In a sum, this simulation shows that the proposed variable selection procedures perform fairly well in the finite sample cases.

Remark 5: The difficulty is that, if the sample size is small and the initial value is far away from the true value, the iteration procedure may not convergence to the true value, even divergence. This phenomena is similar to Fan, Yao and Cai (2003).

They suggested that we may detect whether an estimated $\hat{\beta}$ is likely to be the global minimum by using multiple initial values. To our best knowledge if the sample size is larger, it will be easy to converge.

2.7 Empirical Studies

2.7.1 Functional Index Coefficient Autoregressive Models

Linear time series models such as linear autoregressive moving average models (Box and Jenkins, 1970) were well developed in last century. However, it may not capture some nonlinear features. Many nonlinear time series models have been proposed. The early work includes the bilinear models (Granger and Andersen, 1978; Subba Rao and Gabr, 1984; Liu and Brockwell, 1988), the threshold autoregressive (TAR) models (Tong, 1978), the smooth transition AR (STAR) models (Chan and Tong 1986; Ter$\ddot{a}$svirta, 1994), Markov switching models (Hamilton, 1989), and so on. One of the extensions is functional coefficient autoregressive (FAR) model, which is proposed by Chen and Tsay (1993). The coefficients in FAR models are unknown vector functional form and depend on lagged terms and the FAR models satisfy

$$r_t = g_1(\mathbf{r}^*_{t-1})r_{t-1} + \cdots + g_p(\mathbf{r}^*_{t-1})r_{t-p} + \varepsilon_t,$$

where $\mathbf{r}^*_{t-1} = (r_{t-i_1}, r_{t-i_2}, \cdots, r_{t-i_d})^T$ for $j = 1, \cdots, d$, $g_i(\cdot)$'s are unknown functions in R^d for $1 \leq i \leq p$, $\{\varepsilon_t\}$ is a noise term with mean 0 and variance σ^2, and $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ with $\mathcal{F}_{t-1}$ being an $\sigma-$algebra generated by the past information set I_{t-1}.

In fact, the above FAR model covers several traditional varying coefficient models, such as the threshold autoregressive (TAR) models in Tong (1983, 1990), the smooth transition AR (STAR) models in Chan and Tong (1986) and Ter$\ddot{a}$svirta (1994), and the exponential autoregressive (EXPAR) models of Haggan and Ozaki (1981).

Due to the curse of dimensionality, Chen and Tsay (1993) just considered one single threshold variable case $\mathbf{r}^*_{t-1} = r_{t-k}$, and they proposed an arranged local regression to estimate the functional coefficient β's with an iterative algorithm. In fact, their method is similar to the local constant estimator as pointed out by Cai, Fan and Yao (2000). For the efficient estimation of FAR model, we refer to the papers by Cai, Fan and Yao (2000), Cai, Fan and Li (2000), Fan and Zhang (1999) and Huang and Shen (2004).

To overcome the curse of dimensionality and incorporate more variables in the functional coefficients β's, we assume that $\mathbf{r}^*_{t-1}$ is a linear combination of r_{t-i_j}'s, e.g. $\mathbf{r}^*_{t-1} = \beta^T\mathbf{r}$, where $\mathbf{r} = (r_{t-1}, \cdots, r_{t-d})^T$. The FAR models can be reduced to a special case of functional index coefficient models of Fan, Yao and Cai (2003). We name it as functional index coefficient autoregressive models (FIAR) as

$$r_t = g_1(\beta^T\mathbf{r})r_{t-1} + \cdots + g_p(\beta^T\mathbf{r})r_{t-p} + \varepsilon_t.$$

For the above model, Fan, Yao and Cai (2003) provided algorithms to estimate local parameters β and functional coefficients $g(\cdot)$, and they proposed a combination of the t-statistic and the Akaike information criterion (AIC) to select significant variables of $\mathbf{r}$. They deleted the least significant variables in a given model according to t-value, and selected the best model according to the AIC. However, as mentioned in Fan and Li (2001), this stepwise deletion procedure may suffer stochastic errors inherited in the multiple stages. Meanwhile, there is no theory on this variable selection procedure and the authors did not mention how to select the regressors r_{t-j}.

In this section, we use two step variable selection procedures to select significant variables and to estimate unknown coefficients simultaneously. Firstly, we do variable selection on the regressors based on penalized local maximum likelihood and then we do variable selection on the local variables based on penalized global maximum likelihood.

Our data consists of daily, weekly and monthly returns on the Dow Jones

Industrial Average, NASDAQ Composite and $S\&P$ 500 INDEX. The Dow Jones Industrial Average is from October 1, 1928 to November 30, 2011, the NASDAQ Composite is from February 5, 1971 to November 30, 2011 and the $S\&P$ 500 INDEX is from January 3, 1980 to November 30, 2011. All the data are downloaded from the web site http://www.finance.yahoo.com.

Table 2.3 shows the description of returns for one day horizon, one week horizon and one month horizon. All horizons show the negative skewness. For one day and one week horizons, they appear to have high kurtosis which is higher than 3. The Box-Pierce test shows that the autocorrelations of one month horizon of NASDAQ and $S\&P$ 500 are zero. However, others are none zero. We also present the autocorrelations of one period lagged terms for these indexes with three horizons.

To explore the performance of functional index coefficient autoregressive models, we simply assume our working model as listed below

$$r_t = g_1(\mathbf{z}_t)r_{t-1} + g_2(\mathbf{z}_t)r_{t-2} + g_3(\mathbf{z}_t)r_{t-3} + g_4(\mathbf{z}_t)r_{t-4} + g_5(\mathbf{z}_t)r_{t-5} + g_6(\mathbf{z}_t)r_{t-6} + \varepsilon_t,$$

where $\mathbf{z}_t = \beta_1 r_{t-1} + \beta_2 r_{t-2} + \beta_3 r_{t-3}$ and we assume $\beta_1^2 + \beta_2^2 + \beta_3^2 = 1$ in order to satisfy the identification assumption. When two step estimations and variable selections are employed in our above model, the estimated coefficients of local variables and the norms of co-variates are reported in Table 2.4.

In the local variables part, It is interesting that one day lagged return does not have any effects for one day return of three index. And two day lagged return and three day lagged return perform the same with similar parameter coefficients. However, only one week lagged return contributes for one week return of DOW and $S\&P$ 500, and three week lagged return does not have any contribution. Specially, one week lagged return and two week lagged return have the same coefficients for one week return of NASDAQ. And only two month lagged return has significant effect for one month return of DOW, and only month lagged return has significant effect

Table 2.3: Description of returns for different horizons

	Mean	Median	StdDev	Skewness	Kurtosis	Min	Max	ρ_1	Box-Pierce test
					One day horizon				
DOW	0.0004	0.0004	0.0117	-0.5835	24.2770	-0.2563	0.1427	0.0141	0.0000
NASDAQ	0.0003	0.0011	0.0127	-0.2896	9.7335	-0.1204	0.1325	0.0552	0.0000
$S\&P$ 500	0.0003	0.0004	0.0098	-1.0417	27.8785	-0.2290	0.1096	0.0299	0.0000
					One week horizon				
DOW	0.0009	0.0025	0.0249	-0.5405	6.0324	-0.2003	0.1673	0.0122	0.0000
NASDAQ	0.0015	0.0031	0.0284	-1.0690	9.6448	-0.2918	0.1738	0.0664	0.0009
$S\&P$ 500	0.0013	0.0028	0.0210	-0.5677	5.7557	-0.2008	0.1320	-0.0189	0.0000
					One month horizon				
DOW	0.0039	0.0084	0.0541	-0.8000	6.8783	-0.3667	0.3057	0.0790	0.0009
NASDAQ	0.0067	0.0132	0.0639	-0.8262	2.6339	-0.3179	0.1987	0.1351	0.5985
$S\&P$ 500	0.0058	0.0090	0.0424	-0.6454	2.3579	-0.2454	0.1510	0.0537	0.6689

Table 2.4: Coefficients for local variables and covariates

	local variables			covariates[a]					
	X1	X2	X3	X1	X2	X3	X4	X5	X6
				One day horizon					
DOW	0	0.51	-0.86	2.11	1.29	2.31	3.04	1.16	1.56
NASDAQ	0	0.57	-0.82	0	0	0	2.22	0	0
S&*P* 500	0	0.63	-0.77	1.83	2.81	1.96	3.44	2.32	1.71
				One week horizon					
DOW	1	0	0	5.96	0	0	0	0	0
NASDAQ	0.71	0.70	0	1.29	1.41	1.95	2.40	2.48	2.16
S&*P* 500	1	0	0	6.06	0	0	0	0	0
				One month horizon					
DOW	0	1	0	3.29	2.66	2.20	2.05	2.75	2.05
NASDAQ	1	0	0	5.01	0	0	0	0	0
S&*P* 500	0.6	0	0.79	1.56	1.10	2.02	1.19	2.53	1.35

[a] we calculate the norm of the functional coefficient for covariates.

for one month return of NASDAQ. In the covariates part, only lagged one covariate (X1) has significant effect on one week horizon of DOW and *S*&*P* 500, meanwhile, on the one month horizon of NASDAQ. And only lagged four covariate (X4) is an factor for one day horizon of NASDAQ. All lagged covariates (X1 $\sim$ X6) contribute in other cases. We plot none zero coefficients for DOW, NASDAQ and *S*&*P* 500 with one day, one week and one month horizons respectively, see Figure 2.2$\sim$ Figure 2.10.

2.7.2 Functional Index Coefficient Models for the Stock Return Predictability

In the last section, we consider a regression model for the stock return with its lagged terms, which is in the framework of functional index coefficient autoregressive models. In this section, we focus on a regression model of the stock return with its lagged terms and other lagged financial variables. This leads to another interesting topic - the so called predictability for the stock return in the finance literature. It is very important in empirical finance since it is the center issue to the asset allocation for practitioners in finance markets.

In the 1980's and 1990's, people usually employed a classical linear models to study the predictability for the stock return. They used different financial variables, such as the the default spread, the term spread, the one-month bill rate, the dividend-price ratio, the earnings-price ratio, the book-to-market ratio and so on (Rozeff, 1984; Chen, Ross and Ross, 1986; Campbell, 1987; Campbell and Shiller, 1988a, 1988b; Fama and French, 1988, 1989; Cochrane, 1991; Hodrick, 1992; Goetzmann and Jorion, 1993; Kothari and Shanken, 1997; Pontiff and Schallm 1998; among others). Most of them found some evidence that the stock return can be predictable.

However, there arise some interesting questions and inconsistent results. The first one is that the predictability for stock returns varies with different return horizons. The regression in Fama and French (1988) indicated that the dividend yields typically explain less than 5% of the variance of stock return with short-horizon, either monthly or quarterly returns, whereas the dividend yields can explain more than 25% of the variance of stock return with long-horizon, such as two to four year returns. Another one is the inconsistent results of in-sample predictability and out-sample predictability. Some papers (Bossaerts and Hillion, 1999; Campbell, 2007; Goyal and Welch, 2003, 2008; Butler, Grullon, and Weston, 2005) showed the evidence of in sample predictability, however they found that there is little out of sample forecasting power in the stock return models.

Hence, the question of whether the stock is predictable is still controversial right now in the finance literature. During the resent years, many data-analytic techniques have been developed, and the researchers have tied to explain the above new phenomena and to answer this question in two directions.

The first one is to incorporate the effect of persistency, non-stationary or unit root. When the predictor variable is persistent or a unit root process, conventional t-test for the predictability of stock returns may be failed and spurious regression may arise (Ferson, Sarkissian and Simin, 2003; Boudoukh, Richardson and Whitelaw, 2008). It will produce "significant" results even when there is no relationship between stock return and predictable variables. It is because that persistence or unit root will lead to biased estimator and the t-test does not converge to the true t-distribution even the sample size is large in the prediction models. During the recent years, several papers (Ferson, Sarkissian, and Simin, 2003; Valkanov, 2003; Lewellen, 2004; Torous, Valkanov and Yan, 2004; Hjalmarsson, 2004; Campbell and Yogo, 2006; Jansson and Moreira, 2006; Polk, Thompson, and Vuolteenaho, 2006; Ang and Bekaert, 2007; Cochrane, 2007; Boudoukh, Richardson and Whitelaw, 2008, Cai and Wang, 2011a) focused on estimation and inference with persistent variables in the stock return prediction models.

Another direction is to consider non-constant coefficients. Some empirical studies provided the strong evidence that there may exist time varying parameters. For example, Bossaerts and Hillion (1999) mentioned that, the poor external validity of the prediction models indicates the parameters of the best prediction model change over time. Some papers proposed different approaches to identify structural breaks or parameter instability (Viceira, 1997; Pesaran and Timmermann, 2002; Campbell and Yogo, 2006; Paye and Timmermann, 2006; Ang and Bekaert, 2007; Lettau and Van, 2008; Pettenuzzo and Timmermann, 2011; Ang and Timmermann, 2011, Cai and Wang, 2011b). For example, Paye and Timmermann (2006) examined evidence of instability in models of ex post predictable components in stock returns. They considered structural breaks in the coefficients and different state variables in their models such as the lagged dividend yield, short interest rate, term spread and default premium. Ang and Bekaert (2007) did a test for time variation in coefficients by splitting their entire sample into different sub-periods. Pettenuzzo and Timmermann (2011) studied the effect of rare and large structural break. Comparing to the instantly changed break, Dangl and Halling (2009) allowed gradual changes of coefficients and considered the gradually varying coefficients of coefficients with random walk process. They found a strong relationship between out-of-sample predictability and the business cycle.

However, as mentioned in Granger (2005), "*It is likely that there will be structural breaks in the present framework, but such breaks are difficult to forecast, which is the basic element of their nature*".

To select significant variables for predicting stock returns, Bossaerts and Hillion (1999) implemented several selection criteria, which include R^2, AIC, BIC, FiC, PIC, PLS and PLS-MDC, to verify the predictability of stock return. They confirmed the presence of in sample predictability in the international stock market and no out of sample forecasting power. Dangl and Halling (2009) estimated the $2^k - 1$ dynamic linear models, which result from all possible combinations of predictive variables, and they used a Bayesian model selection criterion to select these variables. One of difficulties is that the selection procedures would be complicated if the number of predictors k is large. Another one is that there is no theory on this work. As mentioned in Fan and Li (2001), this stepwise deletion procedure may suffer stochastic errors inherited in the multiple stages.

In fact, some empirical studies in literatures have revealed that the coefficients of predictors may depend on some financial variables. For example, Fama and French (1989) showed that the slopes for the default spread and the dividend yield increase from high-grade to low-grade bonds and from bonds to stocks. This finding indicates that the coefficients may depend on some variables. In this section, we consider the predictability for the stock return with the functional index coefficient models (Fan, Yao and Cai, 2003). To avoid the curse of dimensionality, it specifies an index form in the functional coefficient part, which is a linear combination of multiple financial variables,

$$r_t = g^T(\beta^T Z_{t-1})X_{t-1} + \varepsilon_t,$$

where r is the stock return, $X_t = (X_{1t}, X_{2t}, \ldots X_{pt})^T$ is a $p \times 1$ dimensional vector of financial variables and Z is a $d \times 1$ dimensional vector of financial variables. $\beta \in R^d$ are

$d-$dimensional unknown parameters and $g(\cdot) = (g_1(\cdot) \ldots g_p(\cdot))^T$ are $p-$dimensional unknown functional coefficients. We assume that $\parallel \beta \parallel = 1$ and the first element of β is positive for identification.

If p=2 and the only one regressor is market return, the above model reduces to the case studied by Cai and Ren (2011). They considered a nonparametric estimate of time-varying beta and alpha in the conditional capital asset pricing model (CAPM), and they developed a procedure that can estimate and select the local variables simultaneously with smoothly clipped absolute deviation penalty. However, they did not provide any theory for their variable selection procedures.

The attractive point of varying-coefficient model is that, the coefficients of regressors are functional form of other variables rather than constant in ordinary linear models. It can capture many financial features in the predictability for the stock return models. First, this model can capture parameter instability with the coefficients that are allowed to change with other economic variables. And it is easy to do forecasting compare to other structure break models. Second, it can incorporate some nonlinear relationship between stock return and financial predictors. For example, let $g^T(\beta^T Z_{t-1}) = \beta^T Z_{t-1}$, then the above model reduces to bilinear model (Granger and Andersen, 1978) and the model in Ferson and Harvey (1999).

In this empirical study, we consider the predictability for stock index returns. The dependent variables include monthly returns on the Dow Jones Industrial Average, NASDAQ Composite and $S\&P$ 500 INDEX, and the covariates and local variables include “BamAa”, the spread between Moody’s Baa corporate bond yield and Moody’s Aaa corporate bond yield, “Bam3m”, the spread between Moody’s Baa corporate bond yield and a three-month Treasury bill, “term1year”, the term spread between the one year and three-month Treasury yields, and “term10year”, the term spread between the ten year and three-month Treasury yields. To match the predictors, we let the lagged data as our local variables. Dow Jones Industrial Average and

$S\&P$ 500 INDEX are from July 1, 1953 to November 1, 2011, NASDAQ Composite is from February 5, 1971, and local variables are from June 1, 1953 to October 1, 2011. The sample size of NASDAQ Composite is 489 and all others are 702.

From the descriptive statistics in Table 2.5, we can find that all the index returns DOW, NASDAQ and $S\&P$ 500 are skewed and their Kurtosis are less then 3. This phenomenon coincides with statement that the returns are usually not normally distributed. We also check the stationarity of these seven variables. We reject the hypothesis that these variables are unit root respectively by using the augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) unit root test. The assumption of stationarity is automatically satisfied.

To study the performance of functional index coefficient for the stock return predicability, we simply assume two working models are set up as below

$$\text{Model 1:}\quad r_t = g_1(\mathbf{z}_t)z_{1,t-1}+g_2(\mathbf{z}_t)z_{2,t-1}+g_3(\mathbf{z}_t)z_{3,t-1}+g_4(\mathbf{z}_t)z_{4,t-1}+\varepsilon_t,$$

and

$$\text{Model 2:}\quad r_t = g_1(\mathbf{z})r_{t-1}+g_2(\mathbf{z}_t)r_{t-2}+g_3(\mathbf{z}_t)r_{t-3}+g_4(\mathbf{z}_t)r_{t-4}+g_5(\mathbf{z}_t)r_{t-5}+g_6(\mathbf{z}_t)r_{t-6}+\varepsilon_t,$$

where $\mathbf{z}_t = \beta_1 z_{1,t-1} + \beta_2 z_{2,t-1} + \beta_3 z_{3,t-1} + \beta_4 z_{4,t-1}$ and we assume $\beta_1^2+\beta_2^2+\beta_3^2+\beta_4^2 = 1$ in order to satisfy the identification assumption.

In Model 1, we simply assume that the local variables and covariates are the same, which include “BamAa”, “Bam3m”, “term1year” and “term10year”. However, we let the covariates be six lagged terms of stock index returns in the model 2. We do estimation and variable selection simultaneously on the above two models by two step estimation procedures. The estimated coefficients of local variables and the norms of covariates are reported in Tables 2.6 and 2.7. An interesting finding is that, all the coefficients of covariates in Model 1 are zero. This implies that the stock index DOW, NASDAQ and $S\&P$ 500 are not predictable with these four variables

Table 2.5: Description of monthly returns for different predictors and local variables

	Mean	Median	StdDev	Skewness	Kurtosis	Min	Max	ADF test	pp test
DOW	0.0063	0.0085	0.0424	-0.4303	1.9202	-0.2322	0.1441	<0.01	<0.01
NASDAQ	0.0087	0.0132	0.0450	-0.4680	1.7130	-0.2723	0.2198	<0.01	<0.01
$S\&P$ 500	0.0065	0.0091	0.0427	-0.4094	1.6479	-0.2176	0.1630	<0.01	<0.01
BamAa	0.9792	0.8500	0.4544	1.7901	4.3264	0.3200	3.3800	0.0165	<0.01
Bam3m	2.2860	2.0850	1.5523	0.0701	-0.5637	-2.2800	5.9300	<0.01	<0.01
term1year	0.5398	0.4600	0.4083	1.1675	3.4631	-0.9400	2.9300	<0.01	<0.01
term10year	1.4470	1.3600	1.2123	-0.1004	-0.3431	-2.6500	4.4200	<0.01	<0.01

"BamAa", "Bam3m", "term1year" and "term10year" in the framework of functional index coefficients.

In Model 2, we find all the functional index coefficients are non-zero if the covariates are lagged returns. As demonstrated in Table 2.7, "Bam3m" and "term10year" are significant local variables with the monthly data of DOW and NASDAQ, and "Bam3m", "term1year" and "term10year" are important predictors for the data of $S\&P$ 500. It is interesting that the local variable "BamAa" is not significant in prediction of three stock indexes index DOW, NASDAQ and $S\&P$ 500 with one month horizon. Meanwhile, the plots for the functional coefficients show non-linearity. Detail can be found in the Figures 2.11, 2.12 and 2.13.

2.8 Conclusion

Variable selection technology and its algorithms are well developed in the last decade. Variable selection for semiparametric models have become more and more popular in the recent years. In this Chapter, we consider variable selection in functional index coefficient models under strong mixing context. Our variable selection procedures include two steps. First, we select significant covariates with functional coefficients, and then do variable selection for local significant variables with parametric coefficients. Simulations show that our two steps procedures perform good. The predictability in stock returns are always interesting and hot topics. In the empirical studies, we consider two financial examples, which include functional index coefficient autoregressive models and functional index coefficient models for the stock return predictability.

The persistence and instabilities of predictive variables, and the nonlinearities of the time series models are hot and hard topics for the stock return predictability. For the further research, we may do variable selection for the time series prediction models with persistent variables. Meanwhile, a "good" prediction model may work for some data and some period. We can investigate some time varying non-linear models, such as selecting predictive periods with variable selection procedures.

Table 2.6: Coefficients for local variables and covariates in model 1(One month horizon)

	local variables				covariates[a]			
	BamAa	Bam3m	term1year	term10year	BamAa	Bam3m	term1year	term10year
DOW	-	-	-	-	0	0	0	0
NASDAQ	-	-	-	-	0	0	0	0
$S\&P$ 500	-	-	-	-	0	0	0	0

[a] we calculate the norm of the functional coefficient for covariates

Table 2.7: Coefficients for local variables and covariates in model 2(One month horizon)

	local variables				covariates[a]					
	BamAa	Bam3m	term1year	term10year	r_{t-1}	r_{t-2}	r_{t-3}	r_{t-4}	r_{t-5}	r_{t-6}
DOW	0	0.72	0	0.69	3.18	2.50	2.45	3.62	2.90	2.12
NASDAQ	0	0.47	0	0.88	3.59	1.58	1.93	1.32	2.99	2.20
$S\&P$ 500	0	0.84	0.33	0.43	1.69	1.47	1.86	2.53	2.63	1.17

[a] we calculate the norm of the functional coefficient for covariates

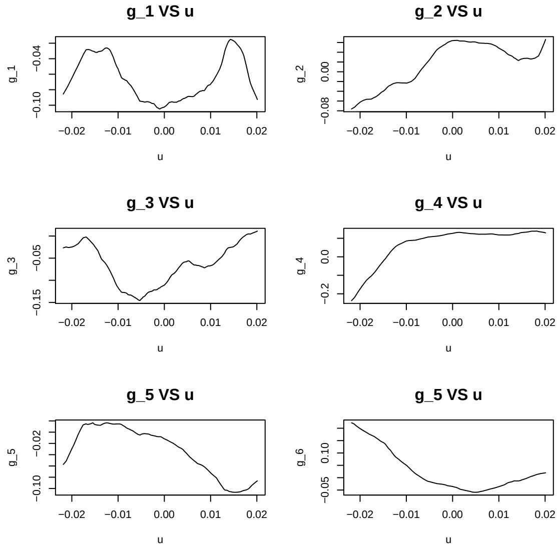

Figure 2.2: Non-zero functional coefficients for FIAR with daily DOW data

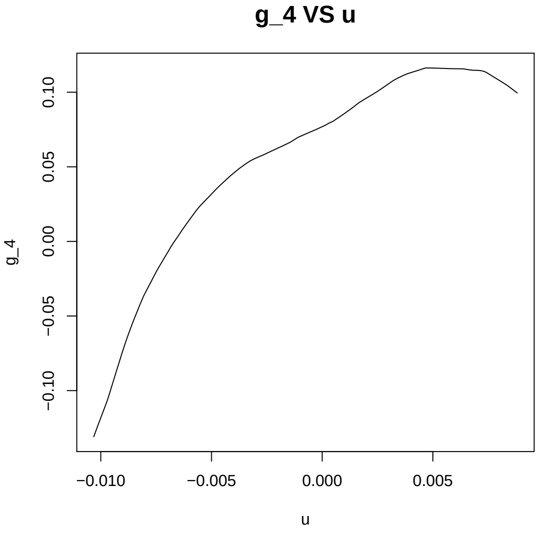

Figure 2.3: Non-zero functional coefficients for FIAR with daily NASDAQ data

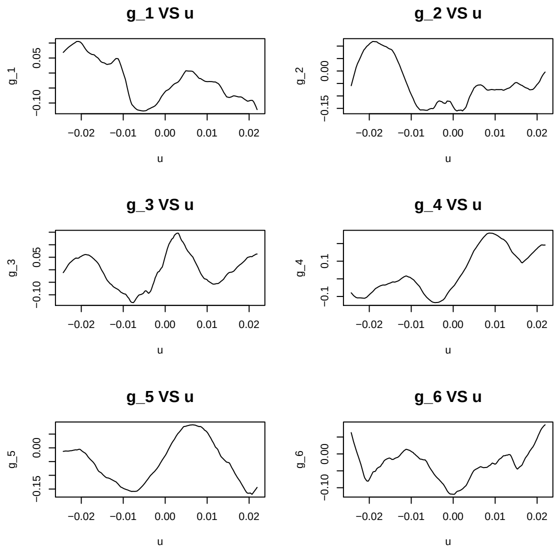

Figure 2.4: Non-zero functional coefficients for FIAR with daily SP data

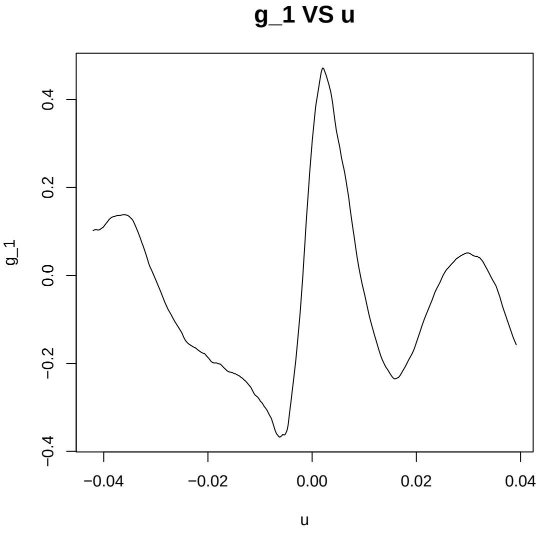

Figure 2.5: Non-zero functional coefficients for FIAR with weekly DOW data

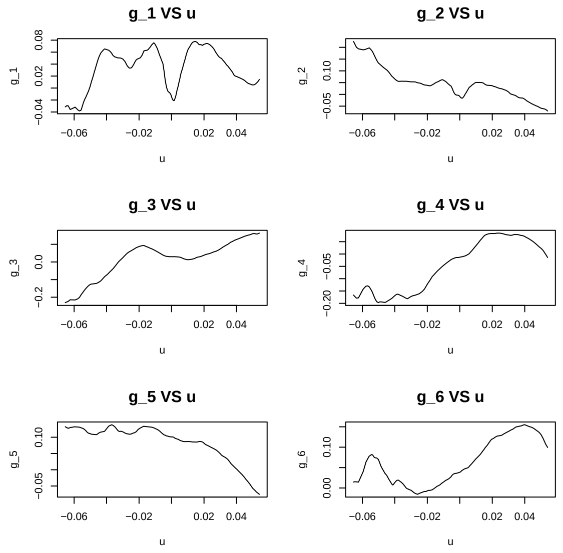

Figure 2.6: Non-zero functional coefficients for FIAR with weekly NASDAQ data

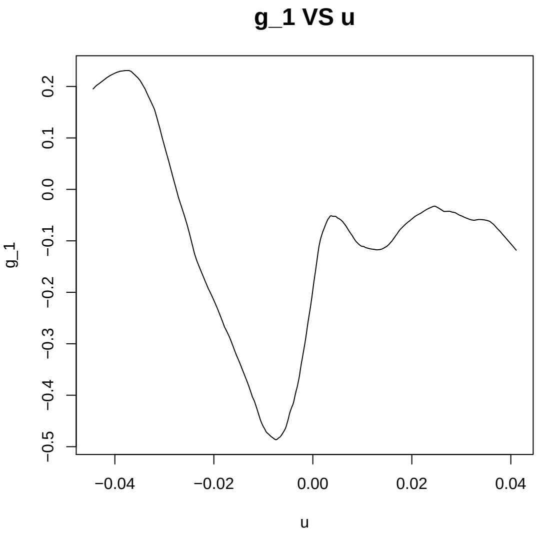

Figure 2.7: Non-zero functional coefficients for FIAR with weekly SP data

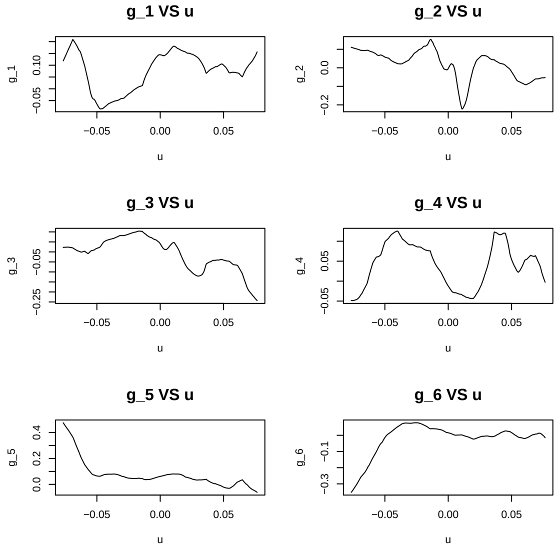

Figure 2.8: Non-zero functional coefficients for FIAR with monthly DOW data

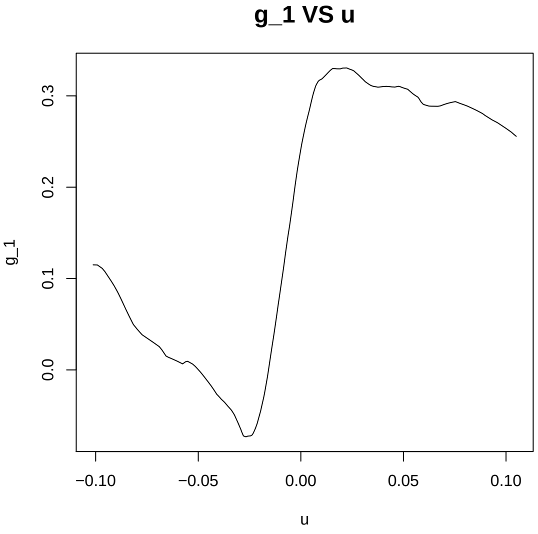

Figure 2.9: Non-zero functional coefficients for FIAR with monthly NASDAQ data

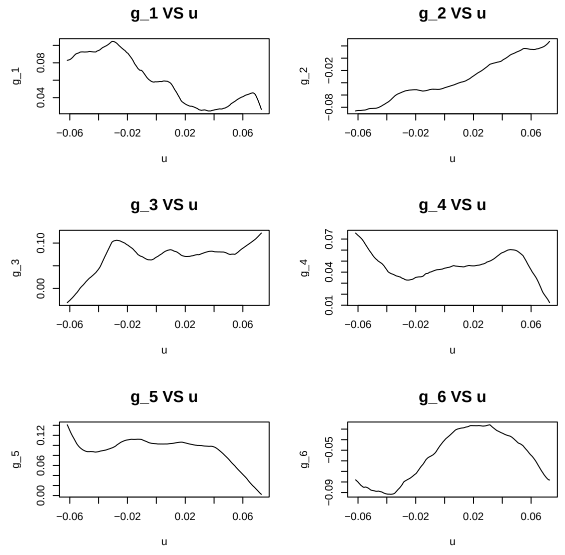

Figure 2.10: Non-zero functional coefficients for FIAR with monthly SP data

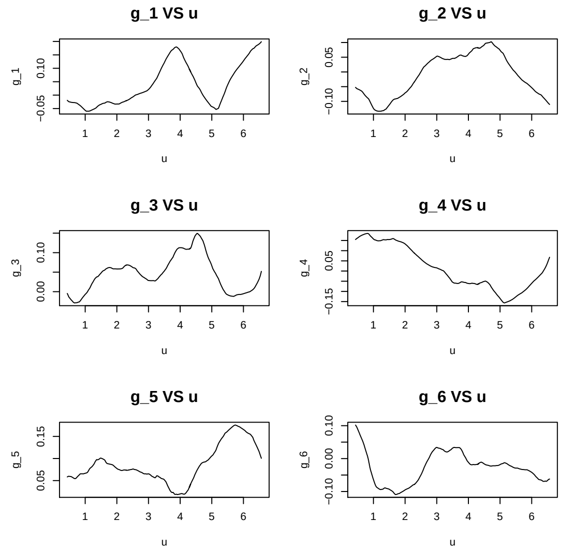

Figure 2.11: Non-zero functional coefficient for prediction with monthly DOW data

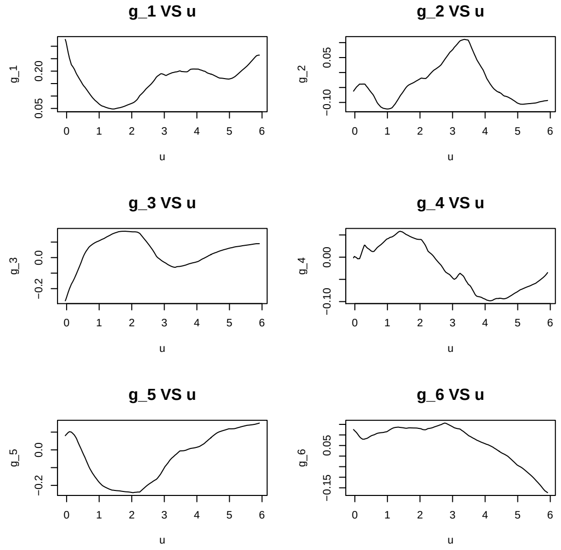

Figure 2.12: Non-zero functional coefficient for prediction with monthly NASDAQ data

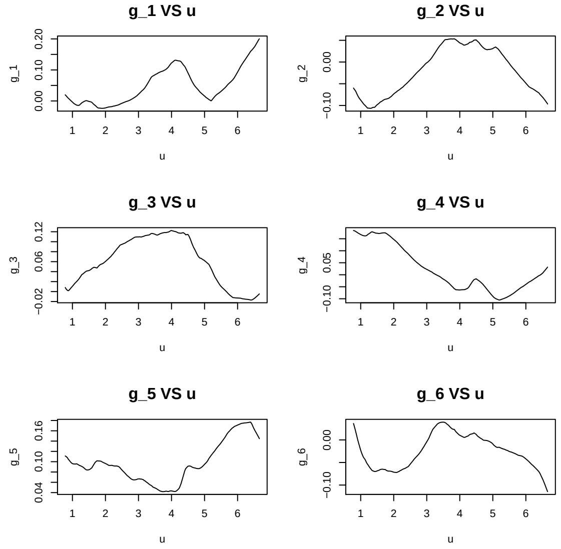

Figure 2.13: Non-zero functional coefficient for prediction with monthly SP data

CHAPTER 3: EFFICIENT LOCAL AADT ESTIMATION VIA SCAD VARIABLE SELECTION BASED ON REGRESSION MODELS

3.1 Introduction

It is well known that the Annual Average Daily Traffic (AADT) information is very important to the VMT (Vehicle Miles of Travel) calculation, thus it is very important for the decision making, planning, air quality analysis, etc., including accident analysis, design and operation analysis of highway facilities, energy consumption, vehicle emissions estimate, air quality analysis, traffic impact assessing, budget estimate, and revenue allocation. AADT and VMT data are required by the Federal Highway Administration (FHWA of USA). The traffic volumes at most of the interstate highways, US and NC routes are collected on an annual basis, and most of the secondary road traffic volumes are collected on a biennial cycle with approximately half being counted each year. However, for most of the local loads, they are lack of detail information and even without any measures due to the cost. Even though the large amount AADT and VMT are on the high functional class roads, the majority percentage of roads is local area roads and rural minor collectors. For example, the document of “North Carolina Highway and Road Mileage Reports 2007” reports that 72% of the statewide road mileage is local area roads (NCDOT, 2010). Thus, the local AADT also makes certain percentage contribution to the total VMT. Therefore, how to estimate the local AADT is a tough and urgent issue for accurate AADT and VMT in state-wide and nation-wide due to lack of observation counters to provide measurements.

Regression analysis may be one of the most popular methods to estimate

AADT. There are many papers on choosing different variables that contribute to AADT. Mohammad et al. (1998) incorporated relevant demographic variables for county roads into a traffic prediction model. Xia et al. (1999) found roadway characteristics such as the number of lanes, functional classification and area type which are contributing predictors to the AADT estimation of non-state roads in urbanized areas in Florida. Zhao and Chung (2001) well developed and compared four multiple linear regression models using geographic information system technology. Four groups of independent variables are considered, including roadway characteristics, socioeconomic characteristics, expressway accessibility, and accessibility to regional employment centers. Zhao and Park (2004) allowed model parameters to be estimated locally by geographically weighted regression (GWR) methods. They argued that the GWR models were comparable with ordinary least square models. Kingan and Westhuis (2006) suggested robust regression methods for AADT forecasting.

Some other information may be incorporated to estimate AADT. Jiang, McCord and Goel (2006) proposed to utilize weighted information of both imaged-based and ground-based traffic data. Eom, Park, Heo and Hunstiger (2006) took into account both spatial trend (mean) and spatial correlation using the spatial regression model. In recent years, many new methods and algorithms are proposed, such as K-nearest neighbor algorithm (Li and Fricker, 2008), co-clustering based collaborative filtering (Wu and Zhang, 2009), support vector machine (Manoel, Jeong, Jeong and Han, 2009), and neural networks (Sharma, Lingras, Xu and Liu, 1999; William and Xu, 2000; Sharma, Lingras, Xu and Kilburn, 2001). To see the performance of some popular methods and algorithms, we refer to the comparison papers (Fricker, Xu and Li, 2008; Sharma, Lingras, Liu and Xu, 2000).

After we collect the large amount of explanatory variables, one critical step is to keep significant variables and exclude the non-significant variables in the final model. As mentioned in Zhao and Chung (2001), although most variables were statistically significant, few added enough explanatory power to be practical and useful. So it is very important to find a criterion to maximize the explanatory power. Thus it is fundamental important to do the variable selection in AADT estimation effectively for the efficiency of estimation and accuracy of prediction. However, there is little literature about the detail of variable selection in the AADT estimation. They may select significant variables by t-test, and F-test, or select the best model according the Akaike information criterion (Akaike, 1973) and Bayesian information criterion (Schwarz, 1978). As mentioned in Fan and Li (2001), this stepwise deletion procedures may suffer stochastic errors inherited in the multiple stages, and there is no theory on the validity of these multiple selecting steps. Modern various shrinkage methods are more and more popular recently years, which include LASSO (Tibshirani, 1996; Knight and Fu, 2000), the bridge regression (Fu, 1998), SCAD (Fan and Li, 2001), the one-step sparse estimator (Zou and Li, 2007). Here we refer to the SCAD (Fan and Li, 2001), since it has oracle property, namely, the resulting procedures perform as well as if the subset of significant variables were known in advance.

In this paper, we focus on the local AADT estimation with effective SCAD variable selection based on regression models. Initially, four groups of 19 variables are collected, which include satellite information and the topological structure of roads. Then, we develop a variable selection procedure by the smoothly clipped absolute deviation penalty (SCAD) procedure, which can simultaneously select significant variables and estimate unknown regression coefficients. Thus, it avoids multiple selecting steps, and guarantees efficiency with theoretical support (Fan and Li, 2001). Further, the algorithm and tuning parameters are explicitly studied. The proposed method shows the validity of our selection procedure. The method further improves the local AADT estimation by incorporating satellite information. It outperforms some other regression method if it is applied to local AADT estimation.

The remainder of this paper is organized as follows. In section 3.2, four groups of variables are collected. Model estimation and variable selection procedures are set up in section 3.3. Section 3.4 compares the suggested method with one previous method. Finally, the conclusion is in section 3.5.

3.2 Groups of Variables

Here we present four groups of variables that we have collected as listed as A through D in the followings. They include driving behavior of individuals, characteristics of the roads, information from satellite and socioeconomic variables.

A. General driving behavior

We assume that (i) the supply of each household from its location is evenly distributed in its community area; (ii) individuals who drive out usually find a short and quick way in a local road to a high class road, US/NC or interstate route, and (iii) individuals who drive in a local community usually take a short and quick way from a high class road, US/NC or interstate route to a destination. Thus we can get the contribution of households to each section of roads by a weighted shortest path algorithm. We take the length of road as distance, and the inverse of number of lanes as a weight. We define this contribution as the loading factor.

B. Characteristics of the roads

We take the following variables for characteristics of the roads.

Number of lanes: if a road has different numbers of lanes, we can define the average number of lanes as the number of lanes of this road.

Road length: if a road is longer, most probably, it would be the “main road” in the local roads, and more cars and trucks would pass by.

X axis and Y axis: we define east-west to be the x axis and north-south to be the y axis.

Local road connectivity: number of other local roads which connect to the studied road.

High road connectivity: number of high level roads (such as secondary road, NC road, US road, and so on) which connect to the studied road.

Collector: we define the nearest collector as a variable for the studied road. It is based on a fact that if the AADT of a collector is high, then the AADT of the nearby road of this collector should be large as usually.

C. Information from satellite

We notice that the traffic information from satellite may be available, such as sampled in Google maps. Thus, we introduce the following variables further.

Cars on the road: number of cars on the studied road. It can be obtained from satellite photos based on digital image processing.

Cars intensity: cars on the road/length of the road, i.e., intensity of cars on the road, calculated as cars on the road divided by the length of the road.

D. Socioeconomic variables

For different zip code areas, we consider the variables of population, population density, housing units, land area (square mile), water area (square mile), median of income, percentage of unemployed, and percentage of people who are below the poverty line.

We may consider other variables such as dummy variables of rural or urban area as predictors if we consider the state-wide AADT estimation because it includes rural areas and urban areas.

3.3 Linear Regression Models and Variable Selection

We define $Y = (y_1, y_2, \cdots, y_n)$ as a dependent variable-vector, $X = (X_1, X_2, \cdots, X_n)$ as an $n \times d$ design matrix, where X_k is d dimensional covariates with the first component as 1. The loss function of linear regression models can be expressed as

$$L(\beta) = \sum_{i=1}^{n} (y_i - X_i^T \beta)^2 \tag{3.1}$$

To shrink the coefficients of non-significant variables to 0, we add smoothly clipped absolute deviation penalty (SCAD) to the above loss function $L(\beta)$.

$$Min \quad L(\beta) + n\sum_{k=1}^{d} P_{\lambda_k,a}(|\beta_k|) \tag{3.2}$$

where the first-order derivative $P'_{\lambda_k,a}(|\beta_k|)$ of the continuous differentiable function $P_{\lambda_k,a}(|\beta_k|)$ is defined as

$$P'_{\lambda_k,a}(|\beta_k|) = \lambda_k I(\beta_k \leq \lambda_k) + \frac{(a\lambda_k - \beta_k)_+}{(a-1)} I(\beta_k > \lambda_k) \text{ for some } a > 2 \text{ and } |\beta_k| > 0 \tag{3.3}$$

$(a\lambda - \beta_k)_+$ takes its positive value if it is positive, otherwise 0, and $\lambda_k, k = 1, \cdots, d$, are tuning parameters. We select $a = 3.7$ as suggested in Fan and Li (2001). Notice that this choice gives pretty good practical performance for various variable selection problems. For the algorithm, we can calculate $\hat{\beta}$ iteratively by

$$\beta^{(1)} = [X^T X + n\sum_{\lambda}(\beta^{(0)})]^{-1} X^T Y \tag{3.4}$$

and

$$\sum_{\lambda}(\beta^{(0)}) = \text{diag}\{\frac{P'_{\lambda_1,a}(|\beta_1^{(0)}|)}{|\beta_1^{(0)}|}, \cdots, \frac{P'_{\lambda_d,a}(|\beta_d^{(0)}|)}{|\beta_d^{(0)}|}\} \tag{3.5}$$

There are three popular methods to estimate tuning parameters λ_k: (i) leaving one out cross-validation, (ii) generalized cross-validation, and (iii) fivefold cross-validation. For computation simplicity, we follow fivefold cross-validation (Fan and Li, 2001), and choose λ_k as follows. Denote the full dataset of X and Y by T, and the training set and the test set by $T - T^q$ and T^q for $q = 1, 2, \cdots, 5$, respectively. For each q, we find the estimator $\hat{\beta}^{(-q)}(\lambda_1, \lambda_2, \cdots, \lambda_d)$ from the training set $T - T^q$. Then we find tuning parameters λ_k to minimize

$$CV(\lambda_1, \lambda_2, \cdots, \lambda_d) = \sum_{q=1}^{5} \sum_{(y_i, X_i) \in T^q} [y_i - X_i^T \hat{\beta}^{(-q)}(\lambda_1, \lambda_2, \cdots, \lambda_d)]^2 \qquad (3.6)$$

But if the dimension d is large, the minimization of equation (3.6) is still difficult. Follow the idea from Zou (2006) who used adaptive weights for penalizing different coefficient by assigning $\lambda_k = \lambda / |\hat{\beta}_k(OLS)|^\gamma$, where $\gamma > 0$, $\hat{\beta}(OLS)$ is the ordinary least square estimator of the equation (3.1). Thus, we let $\lambda_k = \lambda / |\hat{\beta}_k(OLS)|$ to decrease the dimension d of λ_k to dimension 1 of λ.

The data we studied is 243 cases in Mecklenburg County of North Carolina in 2007. We collected 19 explanatory variables as explained in section 3.2. All cases are in the kind of local functional class roads. The statistical software we use is R. Firstly, we do regression of the AADT on the above variables respectively, and then find that some of the variables are significant. The contribution (R^2) of significant variables to the AADT is listed in table 3.1. From table 3.1, we can see that variables of the cars, cars intensity and lanes have large contributions to the AADT, from 0.41 to 0.52. Y axis has moderate contributions to the AADT, as 0.12. The road length, local/high road connectivity, collector and median income have some contributions to AADT. It is noticed that the y-axis is significant. It may be due to the fact that people who live in the south of Charlotte (Mecklenburg County) are richer than in the north, which makes a higher AADT in the south community.

Secondly, we standardize dependent variable Y and covariates X_k by $z_i = (x_i - \bar{x})/\hat{\sigma}$, where $\bar{x}$ is the mean of x and $\hat{\sigma}$ is the standard deviation of x. Then we estimate $\hat{\beta}$ iteratively by equation (3.4). The tuning parameters are chosen by

Table 3.1: The contribution of significant variables to AADT

Variables	R^2
Loading factor	0.07
Road length	0.09
Local road connectivity	0.08
High road connectivity	0.09
Cars	0.52
Cars intensity	0.46
Lanes	0.41
Collector	0.03
Y axis	0.12
Median income	0.08

equation (3.6) and $\lambda_k = \lambda/|\hat{\beta}_k(OLS)|$. Then we can get the following linear model.

$$\begin{aligned} s\_AADT = \quad & 0.2959 * s\_Cars + 0.3217 * s\_Lanes + 0.1200 * s\_Housingunits \\ & +0.2765 * s\_Income + 0.2648 * s\_Belowpovline \\ & +0.2729 * s\_Carintensity + \varepsilon \end{aligned}$$

Table 3.2 below presents the non-zero coefficients from above model. The standard errors are computed by bootstrap with replacement. We can see that only six variables including cars, lanes, housingunits, income, belowpovline and car intensity have significant impact on the local AADT, due to the Multi-Colinearity. The adjusted R-square of the model is 0.65. That is to say, the covariates can explain about 65% of the total variation. If we don't consider satellite information, i.e., to exclude cars intensity and cars in model (3.2), the covariates only explain about 48% of the total variation.

3.4 Comparison

In Zhao and Chung's paper (2001) four regression models are presented. In their paper, model 1 includes the road functional class as a variable, and can explain 82% of the total variation in the high functional class roads. If this model is just

Table 3.2: Regression result of the linear model

Coefficients	Estimate	Std. Error[a]
s_Cars	0.2959	0.0936
s_Lanes	0.3217	0.0552
s_housingunits	0.1200	0.0485
s_income	0.2765	0.0698
s_belowpovline	0.2648	0.0776
s_Carsintensity	0.2729	0.1011

[a] Std. Errors are computed by bootstrap with replacement

applied to local functional class roads, it is equivalent to do regression on variables of LANE, REACCESS, DIRECTAC and BUFFEMP. But it is hard to implement directly, because it is not easy to measure the above last three variables in local areas. In their paper, the ratio of partial R^2 of the last three variables to the one of the first variable LANE is about 0.25. If we assume this ratio remains the same in the local functional class roads, we could get the estimated $R^2 = 0.5$ for these four variables if it is applied to local functional class roads, since variable LANE can explain 41% of the total variation.

In our presented method, we randomly choose 200 cases from 243 cases as the predictor data, and the remaining 43 cases as the prediction data. The prediction error is defined as

$$\text{Prediction Error} = \frac{|\text{predicted value} - \text{true value}|}{\text{true value}} \tag{3.7}$$

In the above formula, we let the predicted value be 0 if it is negative. By doing this procedure until the preset accuracy is reached by the L1 norm of the output difference between the current step and the last step, or until the total iteration number reaches a preset limit, say 500 or 1000 times, we can get the median prediction error for different percentiles. Table 3.3 shows a comparison for two different methods with the percentile of prediction error in 30%, 50%, 80% and 90% respectively.

Table 3.3: Percentile of prediction error for two different methods

Percentile	30%	50%	80%	90%
Prediction Error (this paper)	0.20	0.37	0.98	2.5
Prediction Error (Zhao and Chung, 2001)[a]	0.32	0.45	1.41	2.74

[a] the prediction procedure does not include three covariates REACCESS, DIRECTAC and BUFFEMP.

From the above Table 3.3, with the method we proposed in the local AADT estimation, the half of the prediction error is below 0.37, and about 80% of the prediction error is below 0.98. From the above, we observer the better results of our method in the prediction error comparing to the model 1 if three covariates REACCESS, DIRECTAC and BUFFEMP are not included.

3.5 Conclusion

For the local AADT estimation, we present the smoothly clipped absolute deviation penalty (SCAD) procedure to the regression method. The SCAD can simultaneously select significant variables and estimate unknown regression coefficients at one step. The advantage of the presented method is to avoid multiple selecting steps, and guarantee efficiency with theoretical support.

We consider four groups of 19 variables including statistical general driving behavior, characteristics of the roads, information from satellite and socioeconomic variables. The incorporated satellite information has a great improvement in our model, and makes R-square to go up from 0.48 to 0.65. According to the R-square and the prediction error, if to estimate AADT in the local functional class roads, our method presents the better results. In addition, prediction error is also reasonable although the AADT in local function roads may vary sharply. The further work can be to extend our variable selection procedure to high functional roads and to develop other new algorithms.

CHAPTER 4: NONPARAMETRIC APPROACH TO CALCULATE SEASONAL FACTORS FOR AADT ESTIMATION

4.1 Introduction

Seasonal factors are very important to the estimation of AADT which is useful to decision making, planning, air quality analysis, etc. They are used to transfer the measured traffic volume data of one or two days at the portable traffic monitoring sites to the AADT. The factors are usually calculated based on the traffic information of permanent traffic monitoring sites.

The Federal Highway Administration (FHWA) recommends three factor grouping methods, i.e., the cluster analysis, the geographic/functional assignment of roads to groups, and same road factor application (FHWA, 2001). The early work (Sharma and Werner, 1981; Sharma, 1983; Weinblatt, 1996; Wright et al., 1997) shows that this seasonal adjustment is needed to reduce the significant temporal bias introduced by short duration traffic counts. But as mentioned in the traffic monitoring guide (FHWA, 2001) and a research report in Florida (Zhao, Li and Chow, 2004), there are some problems with this method, such as the difficulties to define groups of roads, to assign groups, and to select a representative sample of roads from which to collect data for calculating the mean values used as factors. In fact, it strongly depends on the judgment of engineers in practice.

By assuming that all roads within a group behave similarly, FHWA suggested that the mean value of randomly selected sample is used as the "best" measure of how all roads in the group behave (FHWA, 2001). In the literature (Davis, 1997; Mohammad et al, 1998; Xia et al, 1999; Zhao and Chung, 2001; Zhao and Park, 2004; Kingan and Westhuis, 2006), regression techniques and their various extensions may be one of the popular tools to estimate AADT. The detail can be found in the review of Zhao and Park (2004). Some authors also adopted regression methods to analyze the relationship between the seasonal factors and some covariates. Faghri and Hua (1995) concluded that variables representing the physical and functional characteristics or their combinations are statistically significant, and they can provide better results than cluster analysis, whereas Zhao, Li and Chow (2004) incorporated roadway functional classification, land use, and other relevant factors into data collection and processing, and argued that it was possible to reduce the data collection effort while improving the accuracy of seasonal factor estimations. In the meanwhile, some other methods are developed to seasonal classification and seasonal factor assignment. For example, Faghri and Hua (1995) applied neural networks to roadway seasonal classification, and Li, Zhao and Chow (2006) proposed a data-driven procedure for assigning a seasonal factor category to a given portable count site. By using a fuzzy decision tree, they considered the similarities between the characteristics of permanent count sites in the seasonal factor group and portable count site. Finally, Bassan (2009) presented a practical statistic methodology of state-wide traffic pattern grouping, in which he combined roadways with similar traffic characteristics such as volume, seasonal variation and land use in Delaware.

Although the progress has been made, there are still many difficulties in the process, e.g., seasonal groups are only constructed based on monthly seasonal factors. As mentioned in Zhao, Li and Chow (2004), it is the traffic monitoring spatial sample location that is the key for the season factor estimation in urban areas. Furthermore, they advocated that there are needs to develop new modelling techniques such as nonlinear regression models since the current regression models for estimating monthly seasonal factors on rural roads have relatively low R squares.

In this paper, to calculate the seasonal factors, we propose a nonlinear regression model based on the nonparametric method by introducing the distance kernel and by using the local weights. The factors utilize the similarity of seasonal variability and traffic characteristics at the count sites in a nearby area. They are decomposed into monthly factors and weekly factors. Then, we introduce a nonlinear distance weighting kernel to estimate the weekly factors. It puts more weight on the observation points which are much closer to the interested point, and puts less weight on the far away observation points. Thus, it makes the seasonal factor estimation more reasonable and accurate to follow the fact. Moreover, the proposed approach can be extended to grouping cases if prior information of grouping is available.

The remainder of this paper is organized as follows. Section 4.2 presents some assumptions and the tests for the assumptions. In section 4.3, we derive the estimation of the seasonal factors including monthly factors and weekly factors by using the nonparametric method. Section 4.4 depicts a theoretical analysis and test for our proposed location effect. Section 4.5 provides a real example to demonstrate our method using the distance kernel with the traffic data observed in the Mecklenburg County of North Carolina. In addition, the results are further compared to the method without the distance kernel. Finally, conclusions are given in Section 4.6.

4.2 Assumptions and Their Validation

4.2.1 Assumptions

To establish and derive the seasonal factor, we present the following assumptions:

A1: There definitely exist monthly effects.

A2: Monthly factors and weekly factors are not linearly interactive.

A3: The weekly factors in a local neighbourhood are similar.

Assumption A1 is commonly used in other methods, which does also reflect a fact. Assumption A2 is weaker than the independence between the monthly factors and weekly factors. This no linearly interaction is observed and tested below. We do the tests for both Assumptions A1 and A2. Assumption A3 means that the weekly factors are similar in the local area. In fact this assumption is reasonable because the traffic has its pattern in the local area.

4.2.2 Tests

In the literature, many papers have addressed how to calculate seasonal factors, and how to apply it when estimating AADT. They usually impose Assumptions A1 and A2 without test or verification. Here we test these presented assumptions by Two-Way ANOVA. The null hypotheses that there exist no monthly effects on the average and no linear interactive effect between monthly factors and weekly factors will be tested. The computational aspect involves computing F-statistic for the hypothesis.

In Section 4.4, we analyze the variance of the seasonal factors as listed in Table 4.1. We can observe from the results in Table 4.1 that the monthly effects do exist, and indeed, there is no linear interaction term between the monthly factors and the weekly factors. This shows that the above assumptions are valid.

4.3 Seasonal Factors Estimation

Under Assumption A1 and A2, we can calculate seasonal factors as follows, which is similar to the method in FHWA (2001)

$$F_{mw} = F_m \cdot F_w \tag{4.1}$$

where F_{mw} is the seasonal factor for the m-th month and the w-th week; F_m is the monthly factor for the m-th month; F_w is the weekly factor for the w-th day in a week. Therefore, if we can get the estimation of the monthly factors and the weekly factors, then it will be easy to obtain the seasonal factors for the m-th month and w-th week by (4.1).

Our goal is to estimate the seasonal factors for our interesting point (x_0, y_0), where we do not have a permanent counter, by the available measurement data at the points $(x_i, y_i), i = 1, \cdots, n$, near the interesting point.

4.3.1 Weekly factors estimation

Here we propose a new approach that is a consistent estimation of weekly factor under some mild assumptions, e.g., the conditional mean is smooth enough, and the residuals are independently identically distributed. Under Assumption A3, our nonparametric regression model can be written as:

$$F_w(x_i, y_i) = g_w(x_i, y_i) + u_w(x_i, y_i), i = 1, \cdots, n, \tag{4.2}$$

where $F_w(x_i, y_i)$ is the weekly factor for the w-th day in a week, and can be calculated by

$$F_w(x_i, y_i) = VOL_{\bar{w}}(x_i, y_i)/VOL_w(x_i, y_i), \tag{4.3}$$

in which $VOL_{\bar{w}}(x_i, y_i)$ is the average traffic volume of observation (x_i, y_i) for a week, $VOL_w(x_i, y_i)$ is the traffic volume of observation (x_i, y_i) for the w-th day in a week, $g_w(x_i, y_i)$ is the estimator of $F_w(x_i, y_i)$, $u_w(x_i, y_i)$ is the residual term, and (x_i, y_i) is the observed location coordinates.

Note that if prior information is available, our proposed nonparametric regression model also can be extended to grouping cases as follows:

$$F_w(x_i, y_i) = g_w(x_i, y_i) + u_w(x_i, y_i), i \in \{\text{group}j\}, \tag{4.4}$$

i.e., we just do nonparametric regression within group j. Further if more variables are available, we can generalize the nonparametric model (4.2) to semi-parametric model (Hardle, Liang and Gao, 2000) as

$$F_w(x_i, y_i) = g_w(x_i, y_i) + z_i'\beta + u_w(x_i, y_i), i = 1, \cdots, n \tag{4.5}$$

where z_i is the vector of other variables and β is the coefficient vector of z_i.

For simplicity, we just consider model (4.2) in this paper. There are many nonlinear or nonparametric methods that can be used to estimate $g_w(x_i, y_i)$, such as spline, series method, two-dimensional kernel method, and so on. For large sample size, the results of different methods are similar. Here we propose to use distance d_i as covariate since x and y have the same measure,

$$d_i = \sqrt{((x_i - x_0)^2 + (y_i - y_0)^2)} \tag{4.6}$$

where (x_0, y_0) is the interesting point, and (x_i, y_i) is the available point with the measurement data for estimation of the seasonal factors as mentioned above.

There are some merits for this suggested method. Firstly, it eases the so-called "curse of dimensionality" (Bellman 1961) if the sample size is not large enough. Secondly, it can overcome the difficulty of choosing two bandwidths simultaneously if a two-dimensional kernel is used. The third is that, if we use two-dimensional data directly, then we always assume that our interested point is in a range of both x and y. If a one-dimensional d_i is used, we just need the condition as our interest point is in the range of either x or y. Denote $K_h(d_i)$ as the kernel, where

$$K_h(d) = K(d/h)/h \tag{4.7}$$

and h is the bandwidth. By adopting from the method of Nadaraya (1965) and Watson (1964), we obtain the estimator of $F_w(x_0, y_0)$, given by

$$\hat{g}_w(x_0, y_0) = \sum_{i=1}^{n} w_i F_w(x_i, y_i) \tag{4.8}$$

$$w_i = K(d_i/h) / \sum_{j=1}^{n} K(d_j/h) \tag{4.9}$$

where w_i is the weight attached to $F_w(x_i, y_i)$, and n is the sample size. It is clear that the weights are nonnegative, and the sum is to one. For the bandwidth selection, we can use plug-in method or cross-validation method (Jones and Sheather, 1991). We also can use local linear method (Cleveland, 1979). For simplicity, we use local constant method (Nadaraya, 1965; Watson, 1964) here. For kernel selection, we choose the commonly used Epanechnikov kernel. In fact, the grouping method is similar to the uniform kernel method.

The formula of Epanechnikov kernel is listed below, and the plots of Epanechnikov and uniform kernels are shown in Figure 4.1.

$$K(u) = (3/4)(1 - u^2)I_{|u| \leq 1} \tag{4.10}$$

When we implement this method in the statistical software, we find that some of interested points are on the boundary of covariate d. To overcome the boundary problem, we use the reflection method proposed in Schuster (1985) and Hall and Wehrly (1991). The reflection method is to construct the synthetic data $\{F_w(x_i, y_i), d_i\}$, where the original data are $\{F_w(x_i, y_i), d_i\}$, and the "reflected" data are $\{F_w(x_i, y_i), -d_i\}$. Also, some of the interested points are far away from the boundary. To solve this problem, we can use fixed proportion (for example, 30%) of data which are the most close to the interest point to estimate the weekly factors.

4.3.2 Monthly factors estimation

In many cases, monthly factor is more sensitive than weekly factor, since the weather conditions contribute a lot to the volume changes. If we have sufficient large amount of traffic monthly data, the monthly factors with location parameter (x_i, y_i) can be estimated by regression model:

$$F_m(x_i, y_i) = g_m(x_i, y_i) + u_m(x_i, y_i), \quad i = 1, \cdots, n \tag{4.11}$$

where $F_m(x_i, y_i)$ is the monthly factor for the m-th month in a year, and can be calculated by $VOL_{\bar{m}}(x_i, y_i)/VOL_m(x_i, y_i)$, in which $VOL_{\bar{m}}(x_i, y_i)$ is the average traffic volume of observation (x_i, y_i) for a year, $VOL_m(x_i, y_i)$ is the traffic volume of observation (x_i, y_i) for the m-th month in a year, and $g_m(x_i, y_i)$ is the estimator of $F_m(x_i, y_i)$; $u_m(x_i, y_i)$ is the measurement error term.

Then we obtain the estimator of $g_m(x_0, y_0)$ by

$$\hat{g}_m(x_0, y_0) = \sum_{i=1}^{n} w_i F_m(x_i, y_i) \tag{4.12}$$

where the weight w_i is defined by equation (4.9).

The above monthly factors estimation procedures are based on a large sample size. However, the sample size for monthly factors is usually small. Furthermore, it is usual to see that some of the data may be missed or even not fully available. Thus, we do the followings:

(i) first to compute the factor of an average day of a week for each month;

(ii) then to compute an annual average value from those monthly averages; and

(iii) finally to compute a single annual average daily value.

This process effectively removes most biases that result from missing days of data, especially when those missing days are unequally distributed across the days of the week or months. Therefore, we have

$$AADT = \frac{1}{12} \sum_{m=1}^{12} MADT_m \tag{4.13}$$

and

$$MADT_m = \frac{1}{7} \sum_{w=1}^{7} (\frac{1}{D_{mw}} \sum_{d=1}^{D_{mw}} VOL_{mwd}) \tag{4.14}$$

where VOL_{mwd} is the d-th daily traffic volume of the w-th day in a week of the m-th month, D_{mw} is the number of the w-th day in the m-th month. For example, $D_{mw} = D_{35} = 4$, it means that there are four Fridays in March in view of the 5th day of a week as Friday and the 3rd month as March; and $MADT_m$ is the monthly average daily traffic. It is noticed that the formulas (4.13) and (4.14) without the distance kernel are similar to those in Wang and Teng (2004).

Then, an average monthly factor F_m is defined for each month as

$$F_m = AADT/MADT_m, \quad m = 1, \cdots, 12. \tag{4.15}$$

and the seasonal factor estimator can be calculated by

$$\hat{F}_{mw} = \hat{F}_m \cdots \hat{F}_w = \hat{g}_m \cdots \hat{g}_w \tag{4.16}$$

If two consecutive daily counting numbers are collected, we can estimate AADT by

$$AA\hat{D}T(x_0, y_0) = 1/2[\hat{F}_{m,w}(x_0, y_0)VOL_{m,w}(x_0, y_0) + \hat{F}_{m,w+1}(x_0, y_0)VOL_{m,w+1}(x_0, y_0)] \tag{4.17}$$

where $VOL_{m,w}(x_0, y_0)$ is the traffic volume in the w-th day of the m-th month at point (x_0, y_0).

4.4 Test for Location Effect

An important problem arises here, i.e., does nonlinear location effect exist, or not? Or is it sufficient to estimate seasonal effect without considering location effect? To answer these questions, we can construct the hypothesis testing formulated as

$$H_0 : g_w(x_i, y_i) = \mu \quad \text{versus} \quad H_1 : g_w(x_i, y_i) \neq \mu \tag{4.18}$$

where μ is the true mean of the weekly factors, which is unknown.

We notice that the so-called generalized likelihood ratio test (GLR test), proposed by Cai, Fan and Yao (2000) and studied by Fan, Zhang and Zhang (2001), can be applied here for testing the hypothesis given in (4.18). To test the hypothesis of

(4.18), we denote $\hat{\mu}$ as the simple average estimator of μ in (4.18), i.e.,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} F_w(x_i, y_i) \tag{4.19}$$

and $\hat{g}_w(x_i, y_i)$ as the nonparametric estimator of $g_w(x_i, y_i)$ in equation (4.8) by introducing the distance kernel. Further, we define

$$RSS_0 = \sum_{i=1}^{n}(F_w(x_i, y_i) - \hat{\mu})^2 \tag{4.20}$$

which is the sum of squared errors (SSE) under the null hypothesis and

$$RSS_1 = \sum_{i=1}^{n}(F_w(x_i, y_i) - \hat{g}_w(x_i, y_i))^2 \tag{4.21}$$

which is the SSE under alternative. Then, the GLR test statistics is defined as

$$\lambda_n = (n/2)log(RSS_0/RSS_1). \tag{4.22}$$

The null distribution of the GLR statistic can be estimated by using the following wild bootstrap method as proposed in Cai, Fan and Yao (2000):

(1) Estimate $\hat{\mu}$ by (19) and $\hat{g}_w(x_i, y_i)$ by (4.8). Compute the GLR statistic λ_n in (4.22) and residuals e_i of F_w from the estimate $\hat{g}_w$ at (x_i, y_i) from our model.

(2) Resample e_i^* from the above e_i set with replacement, and calculate $F_w^*(x_i, y_i) = \hat{\mu} + e_i^*$.

(3) Construct new sample $F_w^*(x_i, y_i)$ and obtain the GLR statistic λ_n^*.

(4) Repeat Steps 2 and 3 B times (for example, B=1000), and get B values of the statistic λ_n^*.

(5) Find the p-value which is the percentage of B values greater than λ_n.

If the p-value in step (5) is less than α (for example, $\alpha = 0.05$), we reject the null hypothesis H_0, and conclude that location effect exists in weekly factors. The test result is presented in the next example section.

Similarly, we can test location effect in monthly factors F_m if data are available.

4.5 Example

The data we use in the first test is 133 cases of the permanent count sites and 359 cases for 7 days of Mecklenburg County, North Carolina from 2002 to 2007. The total effective sample size is 3444. We do a linear regression of the weekly volume on the year, month, week and their linear interaction term. The ANOVA analysis of the results is listed in Table 4.1.

Table 4.1: Analysis of the Variance

instrument variables

	DF	MSE	F value	Pr(>F)
year	5	3.72e+09	15.80	2.20e-15***
months	11	1.46e+10	62.22	<2.2e-16***
week	6	5.25e+10	22.33	<2.2e-16***
month: week	66	1.11e+08	0.47	1
year: week	30	2.58e+07	0.11	1
year: month	53	3.66e+09	15.55	<2.2e-16***
year: month:week	318	2.69e+07	0.11	1
Residuals	2954	6.95e+08		

*** 0.001. DF: Degree of freedom, and MSE: Mean Squares Error.

From analysis of the variance given in Table 4.1, we can see that there definitely exist yearly, monthly and weekly effects in the weekly volumes. Assumption A2 holds since the interaction term between monthly and weekly factors is not significant. And it is interesting to observe that the interaction term between yearly and monthly factors is significant. It may be from the fact that the weather is different among years. Based on Assumption A2 and the above equations (4.13) - (4.17) in Section 4.3, we can calculate the average monthly factors as listed below.

Table 4.2: Average monthly seasonal factors for North Carolina

Month	1	2	3	4	5	6
factor	1.0882	1.0246	0.9906	0.9808	0.9713	0.9770
Month	7	8	9	10	11	12
factor	1.0071	0.9808	0.9843	0.9758	1.0045	1.0316

From the average monthly seasonal factors listed in Table 4.2, we can see that the factors in Month 1, 2, 7, 11 and 12 are larger than 1, and the others are less than 1. It implies that people may go out more in these months because more holidays are in January, February, November and December, and people may take more vocations in the summer time such as July.

The Figure 4.2 shows three dimensional weekly predicted factors of Sunday and Monday. For example, in 2006, the traffic volume in the 1-st day of the 7-th month (Sunday of July) at a location point (1477167, 556127), which locates on Old Concord Road in Charlotte of NC, is 7,805. The traffic volume in the 2-nd day of the 7-th month (Monday of July) at the same location is 10,024. From Table 4.2, we can see that the estimated monthly factor $\hat{g}_7(x, y)$ is 1.007088. With the weekly factor estimation procedure we present above, the estimated weekly factor $\hat{g}_1(x, y)$ and $\hat{g}_2(x, y)$ are 1.273243 and 0.9650776, respectively. Then we can get the estimated AADT 9,875 at the point (1477167, 556127) by (4.17) as follows:

$$(78051.273243 + 100240.9650776)1.0071/2 = 9875$$

Table 4.3 presents the weekly factors estimation at the location $(1477167, 556127)$ by the method with distance kernel and without distance kernel. This table does show the difference between our presented method and the previous methods in the literature. To check whether there exist location effects or not in the weekly factor of North Carolina, we do test as described in Section 4.4. By using bootstrap 1000 times, the critical value is 12.41 under $\alpha = 0.05$. It leads to $\lambda_n = 242.72$ which greater than 12.41. Also we can calculate the mean square error $MSE_0 = RSS_0/(n-1)$ in null model and $MSE_1 = RSS_1/(n-1)$ in our model. It can be found that $MSE_1 = 0.0191$ which is much smaller than $MSE_0 = 0.0537$. We can conclude that there definitely exists location effect. It shows that our new method to introduce the distance kernel is reasonable and necessary.

Table 4.3: Weekly factors estimation at the location (1477167, 556127) with grouping and nonparametric method.

Week	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Method without Kernel	1.179	1.017	1.025	1.013	0.968	0.854	1.071
Method with Kernel	1.273	0.965	0.945	0.921	0.942	0.866	1.142

4.6 Conclusions

In this paper, we present a new method to calculate the seasonal factors for estimating the AADT and the VMT by introducing the distance kernel as (4.6) - (4.9). The proposed method decomposes the seasonal factors into monthly factors and weekly factor by assuming these two factors are not linear interactive. The method we proposed is a data-driving approach method to calculate the weekly factors based on the similarity of seasonal variability and traffic characteristics at the short-term count sites and permanent count sites. Two assumptions are verified by the tests in the example.

The detail estimation formulas of the seasonal factors are clearly presented. Comparing to the seasonal factors derived by other methods, our method would be convenient to be implemented in computer. Furthermore, if we have more data on different categories of roads on the continuous dates, we may extend the above nonparametric model to a semi-parametric model by adding the covariates of the characteristics of roads.

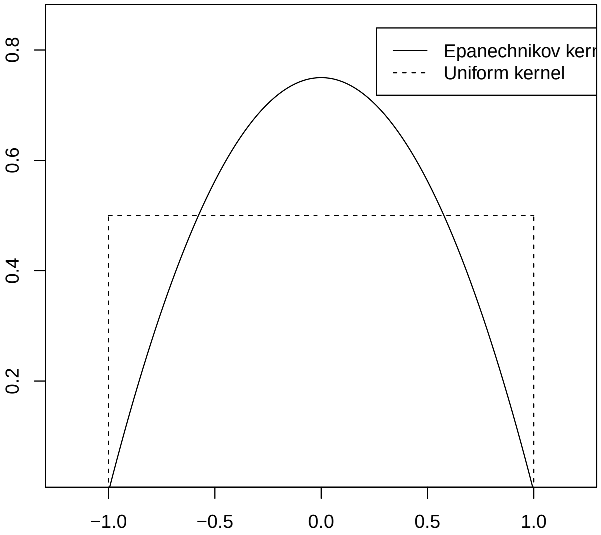

Figure 4.1: Epanechnikov kernel (solid line) and Uniform kernel (dotted line)

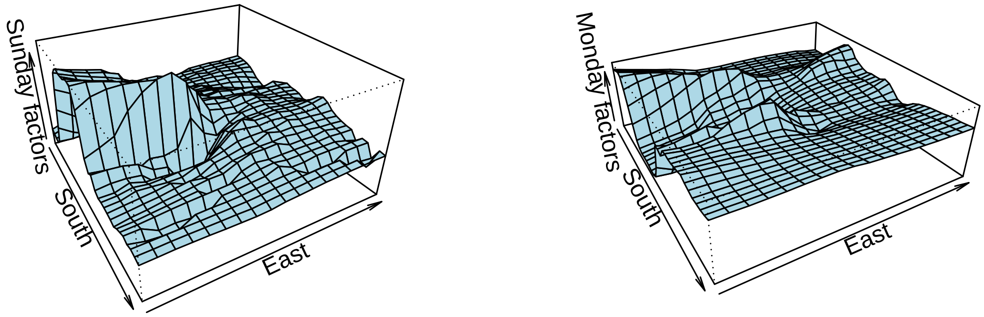

Figure 4.2: Three dimension weekly predicted factors (a) Sunday, (b) Monday.

REFERENCES

Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. Biometrika, 60, 255-265.

Ang, A. and Bekaert, G. (2007). Stock return predictability: is it there? The Review of Financial Studies, 20, 651-707.

Ang, A. and Timmermann, A. (2011). Regime changes and financial markets. Working Paper, Columbia University.

Bassan, S. A. (2009). Statistical practical methodology of statewide traffic pattern grouping and precision analysis. Canadian Journal of Civil Engineering, 336, 427-437.

Bellman, R. (1961). Adaptive control processes: a guided tour. Princeton University Press, New Jersey.

Bossaerts, P. and Hillion, P. (1999). Implementing statistical criteria to select return forecasting models: What do we learn? The Review of Financial Studies, 12, 405-428.

Boudoukh, J., Richardson, M., and Whitelaw, R. F. (2008). The myth of long-horizon predictability. The Review of Financial Studies, 21, 1577-1605.

Box, G. E. P. and Jenkins, G. M. (1970). Time series analysis: forecasting and control. San Francisco: Holden-Day.

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

Brent, A. J., Lin, D. Y. and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. Journal of the American Statistical Association, 103, 672-680.

Butler, A. W., Grullon, G. and Weston, J. P. (2005). Can managers forecast aggregate market returns? Journal of Finance, 60, 963-986.

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

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. Journal of the American Statistical Association, 95, 941-956.

Cai, Z. (2002). Regression quantiles for time series. Econometric Theory, 18, 169-192.

Cai, Z. and Wang, Y. (2011a). Testing predictive regression models with nonstationary regressors. Working Paper, University of North Carolina at Charlotte.

Cai, Z. and Wang, Y. (2011b). Instability of predictability of asset returns. Working Paper, University of North Carolina at Charlotte.

Cai, Z. and Ren, Y. (2011). A new estimation on time-varying betas in conditional capital asset pricing model. Working Paper, the Wang Yanan Institute for Studies in Economics, Xiamen University.

Campbell, J. Y. (1987). Stock returns and the term structure. Journal of Financial Economics, 18, 373-399.

Campbell, J. Y. (2007). Estimating the equity premium. Working Paper, Harvard University.

Campbell, J. Y. and Shiller, R. J. (1988a). Stock prices, earnings, and expected dividends. Journal of Finance, 43, 661-676.

Campbell, J. Y. and Shiller, R. J. (1988b). The dividend-price ratio and expectations of future dividends and discount factors. The Review of Financial Studies, 1, 195-227.

Campbell, J. Y. and Yogo, M. (2006). Efficient tests of stock return predictability. Journal of Financial Economics, 81, 27-60.

Chan, K. S. and Tong, H. (1986). On estimating thresholds in autoregressive models. Journal of Time Series Analysis, 7, 179-190.

Chen, N. F., Ross, R. and Ross, S. (1986). Economic forces and the stock market. Journal of Business, 59, 383-404.

Chen, R. and Tsay, R. S. (1993). Functional coefficient autoregressive model. Journal of the American Statistical Association, 88, 298-308.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatter plots. Journal of the American Statistical Association, 74, 829-836.

Cochrane, J. H. (1991). Explaining the variance of price-dividend ratios. The Review of Financial Studies, 5, 243-280.

Cochrane, J. H. (2007). The dog that did not bark: a defense of return predictability. The Review of Financial Studies, 21, 1533-1575.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of soothing by the method of generalized cross-validation. Numerical Mathematics, 31, 377-403.

Dangl, T. and Halling, M. (2009). Predictive regressions with time-varying coefficients. Working Paper, School of Business, University of Utah.

Davis, G. A. (1997). Estimation theory approach to monitoring and updating average daily traffic. Final report, Minnesota Department of Transportation, St. Paul, Minnesota.

Eom, J. K., Park, M. S., Heo, T. Y. and Hunstiger, L. F. (2006). Improving the prediction of annual average daily traffic for nonfreeway facilities by applying spatial statistic method, Transportation Research Record, 20-29.

Faghri, A. and Hua J. (1995). Roadway seasonal classification using neural networks. Journal of Computing in Civil Engineer, 79, 209-215.

Fama, E. F. and French, K. R. (1988). Dividend yields and expected stock returns. Journal of Financial Economics, 22, 3-27.

Fama, E. F. and French, K. R. (1989). Business conditions and expected returns on stocks and bonds. Journal of Financial Economics, 25, 23-49.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli, 11, 1031-1057.

Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96, 1348-1360.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data Analysis. Journal of the American Statistical Association, 99, 710-723.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica, 20, 101-148.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. Journal of Royal Statistical Society, Series B, 65, 57-80.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. The Annals of Statistics, 29, 153-193.

Fan, J. and Zhang, W. Y. (1999). Statistical estimation in varying coefficient models. The Annals of Statistics, 27, 1491-1518.

Federal Highway Administration. (1994). Travel estimation procedures for the local functional system. [Online]. http://isddc.dot.gov/OLPFiles/FHWA/013434.pdf.

Federal Highway Administration. (2001). Traffic monitoring guide. Washington, D.C. [Online]. http://www.fhwa.dot.gov/ohim/tmguide/.

Ferson, W. E., Sarkissian, S. and Simin, T. (2003). Spurious regressions in financial economics? Journal of Finance, 58, 1393-1413.

Ferson, W. E., and Harvey, C. R. (1999). Conditioning variables and the cross-section of stock returns. Journal of Finance, 54, 1325-1360.

Fricker, J. D., Xu, C. and Jin, L. (2008). Comparison of annual average daily traffic estimates: traditional factor, statistical, artificial neural network, and fuzzy basis neural network approach. Transportation Research Board Annual Meeting.

Fu, W. J. (1998). Penalized regressions: the Bridge versus the LASSO. Journal of Computational and Graphical Statistics, 7, 397-416.

Goetzmann, W. N. and Jorion, P. (1993). Testing the predictive power of dividend yields. Journal of Finance, 48, 663-679.

Goyal, A. and Welch, I. (2003). Predicting the equity premium with dividend ratios. Management Science, 49, 639-654.

Goyal, A. and Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. The Review of Financial Studies, 21, 1455-1508.

Granger, C. W. J. (2005). The past and future of empirical finance: some personal comments. Journal of Econometrics, 129, 35-40.

Granger, C. W. J. and Andersen, A. P. (1978). An introduction to bilinear time series models. Vanderhoek and Ruprecht, Gottingen.

Haggan, V. and Ozaki, T. (1981). Modeling nonlinear vibrations using an amplitude dependent autoregressive time series model. Biometrika, 68, 189-196.

Hall, P. and Wehrly, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. Journal of the American Statistical Association, 86, 665-672.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57, 357-384.

Hardle, W., Liang, H. and Gao, J. (2000). Partially linear models. New York: Springer Series in Contributions to Statistics Physica-Verlag.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models (with discussion). Journal of the Royal Statistical Society, Series B, 55, 757-796.

Hjalmarsson, E. (2004). On the predictability of global stock returns. Working Paper.

Hodrick, R. (1992). Dividend yields and expected stock returns: alternative procedures for inference and measurement. The Review of Financial Studies, 5, 357-386.

Horowitz, J. L. (2009). Semiparametric and nonparametric methods in econometrics. New York: Springer-Verlag.

Huang, J., Joel, L. H. and Wei, F. R. (2010). Variable selection in nonparametric additive models. The Annals of Statistics, 38, 2282-2313.

Huang, J. Z. and Shen, H. (2004). Functional coefficient regression models for non-linear time series: a polynomial spline approach. Board of the Foundation of the Scandinavian Journal of Statistics. Published by Blackwell Publishing Ltd, 515-534.

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.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. The Annals of Statistics, 33, 1617-1642.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single index model, Journal of Econometrics, 58, 71-120.

Janson, S. (1987). Maximal spacing in several dimensions. Annals of Probability, 15, 274-280.

Jansson, M. and Moreira, M. J. (2006). Optimal inference in regression models with nearly integrated regressors. Econometrica, 74, 681-715.

Jiang, Z., McCord, M. R. and Goel, P. K. (2006). Improved AADT estimation by combining information in mage-and ground-based traffic data. Journal of Transportation Engineering, 132, 523-530.

Jones, M. C. and Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. Statistics and Probability Letters, 11, 511-514.

Kingan, R. J. and Westhuis, T. B. (2006). Robust regression methods for traffic growth forecasting. National, State, and Freight Data Issues and Asset Management, Publisher: Transportation Research Board, 51-55.

Knight, K. and Fu, W. J. (2000). Asymptotic for LASSO-type estimators. The Annals of Statistics, 28, 1356-1378.

Kong, E. and Xia, Y. (2007). Variable selection for the single index model. Biometrika, 94, 217-229.

Kothari, S. P. and Shanken, J. (1997). Book-to-market time series analysis. Journal of Financial Economics, 44, 169-203.

Lettau, M. and Van, N. S. (2008). Reconciling the return predictability evidence. The Review of Financial Studies, 21, 1607-1652.

Lewellen, J. W. (2004). Predicting returns with financial ratios. Journal of Financial Economics, 74, 209-235.

Li, M. T., Zhao, F. and Chow, L. F. (2006). Assignment of seasonal factor categories to urban coverage count stations using a fuzzy decision tree. Journal of Transportation Engineering, 8, 654-662.

Li, J. and Fricker, J. D. (2008). Applying K-nearest neighbor algorithm for statewide annual average daily traffic estimates. Transportation Research Board Annual Meeting.

Li, Q. and Jeffrey, S. R. (2007). Nonparametric econometrics: theory and practice. Princeton University Press.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. The Annals of Statistics, 36, 261-286.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. Journal of American Statistical Association, 104, 234-248.

Liang, H., Liu, X., Li, R. and Tsai, C. L. (2010). Estimation and testing for partially linear single-index models. Working Paper, Department of Statistics, Pennsylvania State University.

Lin, Y. and Zhang, H. (2006). Component selection and smoothing in multivariate nonparametric regression. The Annals of Statistics, 34, 2272-2297.

Liu, J. and Brockwell, P. J. (1998). On the general bilinear time-series model. Journal of Applied Probability, 25, 553-564.

Ma, Y. and Li, R. (2010). Variable selection in measurement error models. Bernoulli, 16, 274-300.

Manoel, C. N., Jeong, Y., Jeong, M. K. and Han, L. D. (2009). AADT prediction using support vector regression with data-dependent parameters. Expert Systems with Applications, 36, 2979-2986.

Mohammad, D., Sinha, K. C., Kucek, T. and Scholer, C. F. (1998). Annual average daily traffic prediction model for county roads. Transportation Research Record, 69-77.

Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves. Theory of Applied Probability, 10, 186-190.

NCDOT. North carolina highway and road mileage reports 2007. [Oneline]. http://www.ncdot.org/travel/statemapping/download/highwayroadmileage_2007.pdf, Retrieved July 5th, 2010.

Paye, B. S. and Timmermann, A. (2006). Instability of return prediction models. Journal of Empirical Finance, 13, 274-315.

Pesaran, M. H. and Timmermann, A. (2002). Market timing and return prediction under model instability. Journal of Empirical Finance, 9, 495-510.

Pettenuzzo, D. and Timmermann, A. (2011). Predictability of stock returns and asset allocation under structural breaks. Journal of Econometrics, 164, 60-78.

Polk, C., Thompson, S. and Vuolteenaho, T. (2006). Cross-sectional forecasts of the equity premium. Journal of Financial Economics, 81, 101-141.

Pontiff, J. and Schall, L. D. (1998). Book-to-market ratios as predictors of market returns. Journal of Financial Economics, 49, 141-160.

Rozeff, M. S. (1984). Dividend yields are equity risk premia. Journal of Portfolio Management, 49, 141-160.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. Communications in Statistics: Theory and Methods, 14, 1123-1136.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461-464.

Sharma, S. C. (1983). Improved classification of Canadian primary highways according to type of road use. Canadian Journal of Civil Engineering, 10, 497-509.

Sharma, S. C., Lingras, P., Liu, G. X. and Xu, F. (2000). Estimation of annual average daily traffic on low-volume roads: factor approach versus neural networks. Transportation Research Record, 103-111.

Sharma, S. C., Lingras, P., Xu, F. and Kilburn, P. (2001). Application of neural networks to estimate AADT on low-volume roads. Journal of Transportation Engineering, 127, 426-432.

Sharma, S. C., Lingras, P., Xu, F. and Liu, G. X. (1999). Neural networks as an alternative to the traditional factor approach of AADT estimation from traffic counts. Transportation Research Record, 24-31.

Sharma, S. C. and Werner A. (1981). Improved method of grouping province wide permanent traffic counters. Transportation Research Record, Transportation Research Board, National Research Council, Washington, D.C., 12-18.

Subba Rao, T. and Gabr, M. M. (1984). An introduction to bispectral analysis and bilinear time series model. Lecture Notes in Statistics, 24, Spinger-Verla. New York.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. Journal of the American Statistical Association, 89, 208-218.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B, 58, 267-288.

Tong, H. (1978). On a threshold model. Pattern Recognition and Signal Processing. Sijhoff: Noordhoff, Amsterdam.

Tong, H. (1983). Threshold models in nonlinear time series analysis. Lecture Notes in Statistics, Springer-Verlag, New York.

Tong, H. (1990). Non-linear time series: a dynamical system approach. Oxford University press, Oxford, UK.

Torous, W., Valkanov, R. and Yan, S. (2004). On predicting stock returns with nearly integrated explanatory variables. Journal of Business, 77, 937-966.

Valkanov, R. (2003). Long-Horizon regressions: theoretical results and applications. Journal of Financial economics, 68, 201-232.

Viceira, L. M. (1997). Testing for structural change in the predictability of asset returns. Working Paper, Harvard University.

Wang, H., Li, G. and Tsai, C. L. (2007). Regression coefficient and autoregressive order shrinkage and selection via LASSO. Journal of the Royal Statistical Society, Series B, 69, 63-68.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, 94, 553-568.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. Journal of the American Statistical Association, 104, 747-757.

Wang, L. F., Li, H. Z and Huang, J. H. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. Journal of the American Statistical Association, 103, 1556-1569.

Wang, N. and Teng, H. (2004). VMT estimation associated with ITS data and maintenance of loop detectors. Research Report UVACTS-15-0-88.

Watson, G. S. (1964). Smooth regression analysis. The Indian Journal of Statistics, Series A. 359-372.

Weinblatt, H. (1996). Using seasonal and day-of-week factoring to improve estimates of truck vehicle miles traveled. Transportation Research Record, Transportation Research Board, National Research Council, Washington D.C., 1-8.

William, H. K. and Xu, J. (2000). Estimation of AADT from short period counts in Hong Kong: a comparison between neural network method and regression analysis, Journal of Advanced Transportation, 34, 249-268.

Wright, T., Hu, P. S., Young, J. and Lu, A. (1997). Variability in traffic monitoring data. Final Summary Report, prepared for U.S. Department of Energy, prepared by Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Wu, H. and Zhang, Z. (2009). Framework for estimating AADT using coclustering-based collaborative filtering. Transportation Research Board Annual Meeting.

Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. Journal of the American Statistical Association, 94, 1275-1285.

Xia, Q., Zhao, F., Chen, Z., Shen, L. D. and Ospina, D. (1999). Estimation of annual average daily traffic for nonstate road in a Florida county. Transportation Research Record, 32-40.

Yang, B. and Cai, Z. (2012). Index copulas and nonparametric time varying mixture copula models. Working Paper.

Yang, B., Wang, S. G. and Bao, Y. (2011). Efficient local AADT estimation via SCAD variable selection based on regression models. Chinese Control and Decision Conference, Mianyang (China), 1898-1902.

Yang, B., Wang, S. G. and Cai, Z. (2011). Nonparametric approach to calculate seasonal factors for AADT estimation. International Federation for Automatic Control, the 18th IFAC World Congress, Milano (Italy), 10727-10732.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B, 68, 49-57.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. Journal of the Royal Statistical Society, Series B, 69, 143-161.

Zhang, H. H. and Lin, Y. (2006). Component selection and smoothing for nonparametric regression in exponential families. Statistica Sinica, 16, 1021-1041.

Zhang, Y., Li, R. and Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion. Journal of the American Statistical Association, 105, 312-323.

Zhao, F. and Chung, S. (2001). Estimation of annual average daily traffic in a Florida county using GIS and regression. Transportation Research Board Annual Meeting, 32-40.

Zhao, F., Li, M. T. and Chow, L. F. (2004). Alternatives for estimating seasonal factors on rural and urban roads Florida. [Online]. http://www.dot.state.fl.us/researchcenter/ Completed_Proj/Summary_PL/FDOT_BD015_03_rpt.pdf.

Zhao, F. and Park, N. (2004). Using geographically weighted regression models to estimate annual average daily traffic. Transportation Research Record, 99-107.

Zhao, P. X. and Xue, L. (2010). Variable selection for semi-parametric varying coefficient partially linear errors-in-variables models. Journal of Multivariate Analysis, 101, 1872-1883.

Zou, H. (2006). The adaptive LASSO and its oracle properties. Journal of the American Statistical Association, 101, 1481-1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in non-concave penalized likelihood models. The Annals of Statistics, 36, 1509-1533.

APPENDIX A: PROOF OF LEMMA AND THEOREM

Proof of Lemma 2: By triangle inequality $n^{-1}\sum_{i=1}^{n} \parallel \hat{g}\left(\hat{\beta}^T Z_i\right) - g_0\left(\beta_0^T Z_i\right) \parallel^2 \leq$ $n^{-1}\sum_{i=1}^{n} \parallel \hat{g}\left(\hat{\beta}^T Z_i\right) - g_0\left(\hat{\beta}^T Z_i\right) \parallel^2 + n^{-1}\sum_{i=1}^{n} \parallel g_0\left(\hat{\beta}^T Z_i\right) - g_0\left(\beta_0^T Z_i\right) \parallel^2$. By continuous mapping theorem, the order of the second term on the right hand side is $O_p(n^{-1})$. Only we should show $n^{-1}\sum_{i=1}^{n} \parallel \hat{g}\left(\hat{\beta}^T Z_i\right) - g_0\left(\hat{\beta}^T Z_i\right) \parallel^2 = O_p(n^{-4/5})$. Follow the proof in Wang and Xia (2009), let $u = u_{jk} \in R^{n\times p}$ be an arbitrary $n \times p$ matrix with rows $u_{i\cdot}$ and columns $u_{\cdot k}$

$$u = (u_{1\cdot}, u_{2\cdot}, \cdots, u_{n\cdot})^T = (u_{\cdot 1}, u_{\cdot 2}, \cdots, u_{\cdot p})$$

For any small $\varepsilon > 0$, if we can show that there is a large constant C such that

$$p\{\inf_{n^{-1}\|u\|^2=C} Q(G_0 + (nh)^{-1/2}u, \hat{\beta}) > Q(G_0, \hat{\beta})\} > 1 - \varepsilon$$

then the proof is finished.

$$\begin{aligned}
D \equiv \quad & n^{-1}h\{Q(G_0 + (nh)^{-1/2}u, \hat{\beta}) - Q(G_0, \hat{\beta})\} \\
= \quad & n^{-1}h\{\sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - g_0^T(\hat{\beta}^T Z_j)X_i - (nh)^{-1/2}u_{j\cdot}^T X_i\right]^2 k_h\left(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j\right) \\
& - \sum_{j=1}^{n}\sum_{i=1}^{n}\left[y_i - g_0^T(\hat{\beta}^T Z_j)X_i\right]^2 k_h\left(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j\right)\} \\
& + h\sum_{k=1}^{p}\left[P_{\lambda_k}\left(\parallel g_{0\cdot k} + (nh)^{-1/2}u_{\cdot k} \parallel\right) - P_{\lambda_k}\left(\parallel g_{0\cdot k} \parallel\right)\right] \\
\geq \quad & n^{-1}\sum_{j=1}^{n}\left[u_{j\cdot}^T\hat{\Sigma}(\hat{\beta}^T Z_j)u_{j\cdot} - 2u_{j\cdot}^T\hat{e}_j\right] \\
& + h\sum_{k=1}^{p_0}\left[P_{\lambda_k}\left(\parallel g_{0\cdot k} + (nh)^{-1/2}u_{\cdot k} \parallel\right) - P_{\lambda_k}\left(\parallel g_{0\cdot k} \parallel\right)\right],
\end{aligned}$$

where $\hat{\Sigma}(\hat{\beta}^T Z_j) = n^{-1}\sum_{i=1}^{n} X_i X_i^T k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j)$ and $\hat{e}_j = n^{-1/2}h^{1/2}\sum_{i=1}^{n}[X_i X_i^T$ $(g_0(\beta_0^T Z_i) - g_0(\hat{\beta}^T Z_i)) + X_i X_i^T(g_0(\hat{\beta}^T Z_i) - g_0(\hat{\beta}^T Z_j)) + X_i\varepsilon_i]k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j)$. Let $\hat{\lambda}_j^{\min}$ be the smallest eigenvalue of $\hat{\Sigma}(\hat{\beta}^T Z_j)$,$\hat{\lambda}_{\min} = \min\{\hat{\lambda}_j^{\min}, j = 1, \cdots, n\}$ and $\hat{e} = (\hat{e}_1, \cdots, \hat{e}_n)^T = R^{n\times p}$. Then, $D \geq n^{-1}\sum_{j=1}^{n}(\parallel u_{j\cdot} \parallel^2 \hat{\lambda}_j^{\min} - 2 \parallel u_{j\cdot} \parallel\parallel \hat{e}_j \parallel) -$

$n^{-1/2}h^{1/2}\sum_{k=1}^{p_0} P'_{\lambda_k}(\parallel g_{0\cdot k} \parallel) \parallel u_{\cdot k} \parallel$ where the first term on the right hand side is followed by Cauchy-Schwarz inequality and the second term is followed by Taylor expansion and Cauchy-Schwarz inequality.

$$\begin{aligned} D \geq & \quad \hat{\lambda}_{\min} n^{-1} \sum_{j=1}^{n} \parallel u_{j\cdot} \parallel^2 - 2(n^{-1} \parallel u \parallel^2)^{1/2}(n^{-1} \parallel \hat{e} \parallel^2)^{1/2} - n^{-1/2}h^{1/2}\alpha_n \sum_{k=1}^{d_0} \parallel u_{\cdot k} \parallel \\ \geq & \quad \hat{\lambda}_{\min} n^{-1} \parallel u \parallel^2 - 2(n^{-1} \parallel u \parallel^2)^{1/2}(n^{-1} \parallel \hat{e} \parallel^2)^{1/2} - h^{1/2}\alpha_n (n^{-1} \sum_{k=1}^{d_0} \parallel u_{\cdot k} \parallel)^{1/2} \\ = & \quad \hat{\lambda}_{\min} C - 2\sqrt{C}(n^{-1} \parallel \hat{e} \parallel^2)^{1/2} - h^{1/2}\alpha_n \sqrt{C}. \end{aligned}$$

As we will show below

$$n^{-1} \parallel \hat{e} \parallel^2 = O_p(1) \quad \text{and} \quad \hat{\lambda}_{\min} \rightarrow^P \lambda_0^{\min} \quad \text{as} \quad n \rightarrow \infty$$

where $\lambda_0^{\min} = \inf_{z\in[0,1]} \lambda_{\min}(f(\hat{\beta}Z)\Omega(\hat{\beta}Z))$, $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalues of an arbitrary positive definite matrix. By Assumption A2, A4 and (2.13), $\lambda_0^{\min} > 0$ and $h^{1/2}\alpha_n \rightarrow 0$. For a sufficient large C, $D > 0$. Q.E.D

To show $n^{-1} \parallel \hat{e} \parallel^2 = O(1)$

$$n^{-1} \parallel \hat{e} \parallel^2 \rightarrow^P E \parallel \hat{e}_j \parallel^2$$

And

$$\begin{aligned} E_j \parallel \hat{e}_j \parallel^2 \leq & \quad n^{-1}hE_j \parallel \sum_{i=1}^{n} [X_i X_i^T (g_0(\hat{\beta}^T Z_i) - g_0(\hat{\beta}^T Z_j))] k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j) \parallel^2 \\ & + n^{-1}hE_j \parallel \sum_{i=1}^{n} [X_i \varepsilon_i k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j) \parallel^2 \\ & + n^{-1}hE_j \parallel \sum_{i=1}^{n} [X_i X_i^T (g_0(\beta_0^T Z_i) - g_0(\hat{\beta}^T Z_i))] k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j) \parallel^2 \\ \equiv & \quad A + B + C \end{aligned}$$

where the expectation is with respect to z_j. By (2.13) in Section 2.3.2 and continuous mapping theorem, $C = O_p(h)$, one has

$$\begin{aligned} A = \quad & n^{-1}hE_j\{\sum_{i\neq s\neq j}[(g_0(\hat\beta^T Z_i) - g_0(\hat\beta^T Z_j))^T X_i X_i^T X_s X_s^T (g_0(\hat\beta^T Z_s) - g_0(\hat\beta^T Z_j)) \\ & k_h(\hat\beta^T Z_i - \hat\beta^T Z_j)k_h(\hat\beta^T Z_s - \hat\beta^T Z_j)]\} + n^{-1}hE_j \sum_{(i=s)\neq j}\{\cdots\} \\ \equiv \quad & A_1 + A_2 \end{aligned}$$

$$\begin{aligned} A_1 = \quad & nhE\{(g_0(\hat\beta^T Z_i) - g_0(\hat\beta^T Z_j))^T X_i X_i^T X_s X_s^T (g_0(\hat\beta^T Z_s) - g_0(\hat\beta^T Z_j)) \\ & k_h(\hat\beta^T Z_i - \hat\beta^T Z_j)k_h(\hat\beta^T Z_s - \hat\beta^T Z_j)\} + R_m \\ = \quad & nhE\{(g_0(\hat\beta^T Z_i) - g_0(\hat\beta^T Z_j))^T \Omega(\beta^T Z_i, \beta^T Z_s, \beta^T Z_j)(g_0(\hat\beta^T Z_s) - g_0(\hat\beta^T Z_j)) \\ & k_h(\hat\beta^T Z_i - \hat\beta^T Z_j)k_h(\hat\beta^T Z_s - \hat\beta^T Z_j)\} + R_m \\ = \quad & nhE\{(g_0(z_i) - g_0(z_j))^T \Omega(z_i, z_s, z_j)(g_0(z_s) - g_0(z_j))k_h(z_i - z_j)k_h(z_s - z_j)\} \\ & +R_m \\ = \quad & nh\int E\{(g_0(z_i) - g_0(z_j))^T \Omega(z_i, z_s, z_j)(g_0(z_s) - g_0(z_j))k_h(z_i - z_j)k_h(z_s - z_j) \\ & |z_j\}f(z_j)dz_j + R_m \\ = \quad & A_{11} + R_m \end{aligned}$$

Let $z_i = z_j + uh$ and $z_s = z_j + vh$

$$\begin{aligned} A_{11} = \quad & nh\int\{\int\int(\dot g_0(z_j)uh + \frac{1}{2}C_1u^2h^2)^T\Omega(z_j + uh, z_j + vh, z_j)(\dot g_0(z_j)vh \\ & +\frac{1}{2}C_2v^2h^2)k(u)k(v)f((z_j + uh, z_j + vh)|z_j)dudv\}f(z_j)dz_j \\ = \quad & nh\int A_{12}(z_j)f(z_j)dz_j \end{aligned}$$

$$\begin{aligned} A_{12}(z_j) = \quad & \int\int(\dot g_0(z_j)uh + \frac{1}{2}C_1u^2h^2)^T[\Omega(z_j, z_j, z_j) + \Omega_1(z_j, z_j, z_j)uh + \Omega_2(z_j, z_j, \\ & z_j)vh + o_p(u^2h^2) + o_p(v^2h^2)](\dot g_0(z_j)vh + \frac{1}{2}C_2v^2h^2)[f((z_j, z_j)|z_j) + f_1((\\ & z_j, z_j)|z_j)uh + f_2((z_j, z_j)|z_j)vh + o_p(u^2h^2) + o_p(v^2h^2)]k(u)k(v)dudv \\ = \quad & C(Z_j)h^4\int u^2v^2k(u)k(v)dudv + o_p(h^4). \end{aligned}$$

Then, $A_1 = O_p(nh^5) = O_p(1)$. Also,

$$\begin{aligned}
A_2 = \quad & n^{-1}hE_j\{\sum_{i\neq j}[(g_0(\hat{\beta}^T Z_i) - g_0(\hat{\beta}^T Z_j))^T X_i X_i^T X_i X_i^T (g_0(\hat{\beta}^T Z_i) - g_0(\hat{\beta}^T Z_j)) \\
& k_h^2(\hat{\beta}_0^T Z_i - \hat{\beta}_0^T Z_j)]\} \\
= \quad & hE\{(g_0(\hat{\beta}^T Z_i) - g_0(\hat{\beta}^T Z_j))^T X_i X_i^T X_i X_i^T (g_0(\hat{\beta}^T Z_i) - g_0(\hat{\beta}^T Z_j)) \\
& k_h^2(\hat{\beta}_0^T Z_i - \hat{\beta}_0^T Z_j)\} + R_m \\
= \quad & hE\{(g_0(z_i) - g_0(z_j))^T \Omega(z_i, z_j)(g_0(z_i) - g_0(z_j))k_h^2(z_i - z_j)\} + R_m \\
= \quad & h\int E\{(g_0(z_i) - g_0(z_j))^T \Omega(z_i, z_j)(g_0(z_i) - g_0(z_j))k_h^2(z_i - z_j)|z_j\}f(z_j)dz_j \\
& +R_m \\
= \quad & h\int A_{21} f(z_j)dz_j + R_m,
\end{aligned}$$

where

$$A_{21}(z_j) = \int (g_0(z_i) - g_0(z_j))^T \Omega(z_i, z_j)(g_0(z_i) - g_0(z_j))k_h^2(z_i - z_j)f(z_i|z_j)dz_i$$

Let $z_i = z_j + uh$

$$\begin{aligned}
A_{21}(z_j) = \quad & \frac{1}{h}\int (\dot{g}_0(z_j)uh + Cu^2h^2)^T \Omega(z_j + uh, z_j)(\dot{g}_0(z_j)uh + Cu^2h^2)k^2(u) \\
& f(z_j + uh|z_j)du \\
= \quad & C(z_i)h\int u^2k^2(u)du + R_m.
\end{aligned}$$

Then, $A_2 = O_p(h^2) = o_p(1)$. Hence, $A = O_p(1)$ and

$$\begin{aligned} B = \quad & n^{-1}hE_j\{(\sum_{i=1}^{n} X_i\varepsilon_i k_h(z_i - z_j))^T(\sum_{s=1}^{n} X_s\varepsilon_s k_h(z_s - z_j))\} \\ = \quad & n^{-1}hE_j\{(\sum_{(i\neq s)\neq j} X_i^T X_s\varepsilon_i\varepsilon_s k_h(z_i - z_j)k_h(z_s - z_j)\} \\ & +2n^{-1}hE_j\{(\sum_{(i=j)\neq s} X_i^T X_s\varepsilon_i\varepsilon_s k_h(z_i - z_j)k_h(z_s - z_j)\} \\ & +n^{-1}hE_j\{\sum_{(i=s)\neq j} X_i^T X_s\varepsilon_i\varepsilon_s k_h(z_i - z_j)k_h(z_s - z_j)\} \\ & +n^{-1}hE_j\{\sum_{i=s=j} X_i^T X_s\varepsilon_i\varepsilon_s k_h(z_i - z_j)k_h(z_s - z_j)\} \\ = \quad & B_1 + B_3 + B_2 + B_4, \end{aligned}$$

where

$$\begin{aligned} B_1 = \quad & nh\{E[X_i^T X_s\varepsilon_i\varepsilon_s k_h(z_i - z_j)k_h(z_s - z_j)] + O_p(h^4)\} + R_m \\ = \quad & nh\{E[X_i^T X_s k_h(z_i - z_j)k_h(z_s - z_j)E(\varepsilon_i\varepsilon_s|X_i, X_s, z_i, z_s)]\} + R_m \\ = \quad & O_p(1) \end{aligned}$$

$$\begin{aligned} B_2 = \quad & hE[X_i^T X_i\varepsilon_i^2 k_h^2(z_i - z_j))] + R_m \\ = \quad & hE[X_i^T X_i k_h^2(z_i - z_j)E(\varepsilon_i^2|X_i, z_i, z_j)] + R_m \\ = \quad & h\sigma^2 E[X_i^T X_i k_h^2(z_i - z_j)] + R_m \\ = \quad & h\sigma^2 E[\Omega(z_i, z_j)k_h^2(z_i - z_j)] + R_m \\ = \quad & h\sigma^2 E\{E[\Omega(z_i, z_j)k_h^2(z_i - z_j)|z_j]\} + R_m. \end{aligned}$$

Let $z_i = z_j + uh$. Then, we have

$$\begin{aligned} E[\Omega(z_i, z_j)k_h^2(z_i - z_j)|z_j] = \quad & \frac{1}{h^2}\int \Omega(z_i, z_j)k^2(\frac{z_i - z_j}{h})f_{z_i|z_j}(z_i|z_j)dz_i \\ = \quad & \frac{1}{h}\int \Omega(z_j + uh, z_j)k^2(u)f_{z_i|z_j}(z_j + uh|z_j)du \\ = \quad & C(z_j)O_p(1/h)\int k^2(u)du, \end{aligned}$$

then, $B_2 = O_p(1)$ and

$$\begin{aligned} B_3 = & \quad 2n^{-1}hE_j\{\sum_{s\neq j} X_j^T X_s \varepsilon_j \varepsilon_s k_h(0)k_h(z_s - z_j)\} \\ = & \quad 2h\{E_j[X_j^T X_s \varepsilon_j \varepsilon_s k_h(0)k_h(z_s - z_j)]\} + R_m \\ = & \quad 2h\{E_j[X_j^T X_s E(\varepsilon_j \varepsilon_s | z_s - z_j)k_h(0)k_h(z_s - z_j)]\} + R_m \\ = & \quad O_p(h^3) \\ = & \quad o_p(1) \end{aligned}$$

$$\begin{aligned} B_4 = n^{-1}hE[X_j^T X_j \varepsilon_j^2 k_h^2(0)] & \\ = & \quad n^{-1}hE[X_j^T X_j E(\varepsilon_j^2|X_j)k_h^2(0)] \\ = & \quad n^{-1}h\sigma^2 k_h^2(0)E[X_j^T X_j)] \\ = & \quad O_p(n^{-1}) \\ = & \quad o_p(1). \end{aligned}$$

Hence $B = O_p(1)$. Q.E.D

Proof of Lemma 3: Assume $\| \hat{g}_{.k} \| \neq 0$, then

$$\frac{\partial Q(G, \hat{\beta}, h)}{\partial g_{.k}} = J_1 + J_2 = 0$$

$$\| J_1 \| = \| J_2 \|$$

where

$J_2 = nP'_{\lambda_k}(\| g_{.k} \|)\frac{g_{.k}}{\|g_{.k}\|}$,

$J_1 = (J_{11}, J_{12}, \cdots, J_{1n})^T$

and $J_{1j} = -2\sum_{i=1}^n X_{ik}\left(y_i - \hat{g}^T(\hat{\beta}^T Z_j)X_i\right)k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j)$

Similar to the proof of (A.7) in Wang and Xia (2009), we can derive that $\| J_1 \| = O_p(nh^{-1/2})$ and we know $\| J_2 \| = nP'_{\lambda_k}(\| g_{.k} \|) = \frac{P'_{\lambda_k}(\|g_{.k}\|)}{\lambda_n} \cdot \sqrt{h}\lambda_n \cdot nh^{-1/2}$. Since $\frac{P'_{\lambda_k}(\|g_{.k}\|)}{\lambda_n} > 0$ and $\sqrt{h}\lambda_n \to 0$, then $P(\| J_2 \| < \| J_1 \|) \to 1$ as $n \to \infty$. It contradicts with the assumption, hence, $\| \hat{g}_{.k} \|= 0$ as $n \to \infty$. Q.E.D.

Proof of Theorem 2: (a) From lemma above $\| \hat{g}_{.k} \|= 0$, then $\hat{g}_k(\hat{\beta}^T Z_j) = 0$ for $j = 1, \cdots , n \quad k = p_0 + 1 \cdots , p$. Then $\| \hat{g}_b(\hat{\beta}^T Z_j) \|= 0 \qquad j = 1, \cdots , n$.

(b)From part(a) we know that $\| \hat{g}_b(\hat{\beta}^T Z_j) \|= 0$. We can find there exists a $\hat{G}_a$ that is the minimizer of $Q((G_a, 0), \hat{\beta}, h)$. Take the first derivative of $Q((G_a, 0), \hat{\beta}, h)$ with respective to the $\hat{g}_a(\hat{\beta}^T Z_j)$ we can get the normal equation.

$$\sum_{i=1}^{n} X_{ia} \left(y_i - \hat{g}_a^T(\hat{\beta}^T Z_j) X_{ia}\right) k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j) + n\Pi = 0$$

where Π is a a-dimensional vector with its k-th component given by

$$P_k = P'_{\lambda_k}(\| \hat{g}_{.k} \|) \frac{\hat{g}_k(\hat{\beta}^T Z_j)}{\| \hat{g}_{.k} \|}$$

As we know that$P'_{\lambda_k}(\| \hat{g}_{.k} \|) = 0$ when $\| \hat{g}_{.k} \|\neq 0$ and n is large, then $\Pi = 0$ follows when n is large.

$$\sum_{i=1}^{n} X_{ia} \left(y_i - \hat{g}_a^T(\hat{\beta}^T Z_j) X_{ia}\right) k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j) = 0$$

$$\hat{g}_a(\hat{\beta}^T Z_j) = \quad [\sum_{i=1}^{n} X_{ia} X_{ia}^T k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j)]^{-1} \sum_{i=1}^{n} X_{ia} y_i k_h(\hat{\beta}^T Z_i - \hat{\beta}^T Z_j)$$

$$\hat{g}_a(\hat{\beta}^T Z_j) - g_{0a}(\beta_0^T Z_j) = \quad \{\hat{g}_a(\hat{\beta}^T Z_j) - g_{0a}(\hat{\beta}^T Z_j)\} + \{g_{0a}(\hat{\beta}^T Z_j) - g_{0a}(\beta_0^T Z_j)\}$$

The first term in the right hand side is the order of $O_p(n^{-2/5})$ and the second term in the right hand side is the order of $O_p(n^{-1/2})$. Thus the asymptotic property of the $\hat{g}_a(\hat{\beta}^T Z_j) - g_{0a}(\beta_0^T Z_j)$ is the same as that of $\hat{g}_a(\hat{\beta}^T Z_j) - g_{0a}(\hat{\beta}^T Z_j)$. Li and Jeffrey (2007) presented the asymptotic distribution of the first term under i.i.d case. Similar to the local linear estimator of varing coefficient model under strong mixing case in Cai, Fan and Yao (2000), they showed that the asymptotic distribution of the local linear estimators are identical to the independent data case. It is not hard to show the same result by local constant method under strong mixing condition. By the assumption $Z \in \mathcal{A}'_z$ and similar argument in Wand and Xia (2009), it suffices to approximate the entire coefficient curve $g_0(\hat{\beta}^T Z)$by $\{\hat{g}(\hat{\beta}^T Z_j)|\hat{\beta}^T Z_j \in [a, b]\}$. Q.E.D.

Proof of Theorem 3: Let $\delta_n = n^{-1/2} + a_n, t = (t_1, \cdots t_d)^T$. For any small $\varepsilon > 0$, if we can show there exists a large constant C,such that

$$P\{\inf_{\|t\|=C} Q(\beta_0 + \delta_n t, \hat{g}) > Q(\beta_0, \hat{g})\} > 1 - \varepsilon,$$

then

$$\| \hat{\beta} - \beta_0 \| = O_p(\delta_n)$$

. Define:

$$\begin{aligned} D_n = \quad & Q(\beta_0 + \delta_n t, \hat{g}) - Q(\beta_0, \hat{g}) \\ \geq \quad & \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{g}^T(\beta_0^T Z_i + \delta_n t^T Z_i)X_i)^2 - \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{g}^T(\beta_0^T Z_i)X_i)^2 \\ & + n\sum_{k=1}^{d_0} P_{\lambda_n}(|\beta_{10k} + \delta_n t_k|) - n\sum_{k=1}^{d_0} P_{\lambda_n}(|\beta_{10k}|) \qquad (\text{ by } \beta_{20} = 0) \end{aligned}$$

$$\begin{aligned} & n\sum_{k=1}^{d_0} P_{\lambda_n}(|\beta_{10k} + \delta_n t_k|) - n\sum_{k=1}^{d_0} P_{\lambda_n}(|\beta_{10k}|) \\ = \quad & n\sum_{k=1}^{d_0}\left[\delta_n P'_{\lambda_n}(|\beta_{10k}|)sgn(\beta_{10k})t_k + \frac{1}{2}\delta_n^2 P''_{\lambda_n}(|\beta_{10k}|)t_k^2\right] + o_p(n\delta_n^2) \\ \leq \quad & \sqrt{d_0} n\delta_n a_n \| t \| + \frac{1}{2} n\delta_n^2 max_{1\leq k\leq d_0}\{P''_{\lambda_n}(|\beta_{10k}|)\} \| t \|^2 + o_p(n\delta_n^2) \\ \leq \quad & n\delta_n^2\sqrt{d_0}C + O_p(n\delta_n^2) \quad \text{as } n \to \infty \text{ and } max_{1\leq k\leq d_0}\{P''_{\lambda_n}(|\beta_{10k}|)\} \to 0 \end{aligned}$$

$$\begin{aligned}
& \frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{g}^T(\beta_0^T Z_i+\delta_n t^T Z_i)X_i)^2-\frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{g}^T(\beta_0^T Z_i)X_i)^2 \\
= & \frac{1}{2}n[\tilde{V}_0^{1/2}\delta_n t-n^{-1/2}\sigma\varepsilon]^T[\tilde{V}_0^{1/2}\delta_n t-n^{-1/2}\sigma\varepsilon]-\frac{1}{2}n[n^{-1/2}\sigma\varepsilon]^T[n^{-1/2}\sigma\varepsilon] \\
& +R_1(\beta_0+\delta_n t,h)-R_1(\beta_0,h)+o_p(1) \qquad \text{(by Theorem 2)} \\
= & \frac{1}{2}n\delta_n^2 t^T\tilde{V}_0 t-n^{1/2}\delta_n t^T\tilde{V}_0^{1/2}\sigma\varepsilon+R_1(\beta_0+\delta_n t,h)-R_1(\beta_0,h)+o_p(1) \\
= & \frac{1}{2}n\delta_n^2 t^T\tilde{V}_0 t-\delta_n t^T V_n+R_1(\beta_0+\delta_n t,h)-R_1(\beta_0,h)+o_p(1)
\end{aligned}$$

Since R_1 are negligible terms as $n \to \infty$ and $\frac{1}{\sqrt{n}}V_n = O_p(1)$. then $-\delta_n t^T V_n = C\cdot O_p(\delta_n\sqrt{n}) = C\cdot O_p(\delta_n^2 n)$. By choosing a sufficient large C, then the term $\frac{1}{2}n\delta_n^2 t^T\tilde{V}_0 t$ will dominate others. Hence $D_n \geq 0$ holds. Q.E.D

Proof of Theorem 4: Let $\widehat{\beta}_1-\beta_{10}=O_p\left(n^{-1/2}\right)$, we want to show

$$\left(\widehat{\beta}_1,0\right)^T = argmin_{\left(\beta_1^T,\beta_2^T\right)^T\in\mathbb{B}} Q\left(\left(\beta_1^T,\beta_2^T\right)^T,\hat{g}\right).$$

Only we should show for some constant C and $k=q_0+1,\cdots q$:

$$\frac{\partial Q\left(\left(\beta_1^T,\beta_2^T\right)^T,\hat{g}\right)}{\partial\beta_k} \quad \begin{array}{ll} >0 & \text{for } 0<\beta_k<Cn^{-1/2} \\ <0 & \text{for } -Cn^{-1/2}<\beta_k<0 \end{array}$$

$$\text{Note} \qquad \begin{aligned}
\frac{\partial\hat{S}\left(\beta,h\right)}{\partial\beta_k} & = \frac{\partial\tilde{S}\left(\beta\right)}{\partial\beta_k}+o_p\left(1\right) \\
& = e_k^T\frac{\partial\tilde{S}\left(\beta\right)}{\partial\beta}+o_p\left(1\right) \\
& = 2ne_k^T\tilde{V}_0\left(\beta-\beta_0\right)-2n^{1/2}\sigma e_k^T\tilde{V}_0^{1/2}\varepsilon+o_p\left(1\right) \\
& = 2ne_k^T\tilde{V}_0\left(\beta-\beta_0\right)-2e_k^T V_n+o_p\left(1\right)
\end{aligned}$$

where e_k is a d dimensional vector with k-th element is one and all others are zero. And We know $\beta-\beta_0=O_p\left(1/\sqrt{n}\right)$ and $V_n=O_p\left(\sqrt{n}\right)$, then

$$\frac{\partial\hat{S}\left(\beta,h\right)}{\partial\beta_k}=O_p\left(\sqrt{n}\right)$$

$$\begin{aligned}\frac{\partial Q\left(\left(\beta_1^T,\beta_2^T\right)^T,\hat{g}\right)}{\partial\beta_k} &= \frac{1}{2}\frac{\partial\hat{S}\left(\beta,h\right)}{\partial\beta_k}+nP'_{\lambda_n}\left(|\beta_k|\right)sgn\left(\beta_k\right)\\ &= n\lambda_n\left[O_p\left(\frac{1}{\sqrt{n}\lambda_n}\right)+\frac{P'_{\lambda_n}\left(|\beta_k|\right)}{\lambda_n}sgn\left(\beta_k\right)\right]\end{aligned}$$

Since $\sqrt{n}\lambda_n\to\infty$ and $\lim\inf_{n\to\infty,\beta_k\to0^+}\frac{P'_{\lambda_n}(|\beta_k|)}{\lambda_n}>0$, the sign of $\frac{\partial Q}{\partial\beta_k}$ is determined by the sign of β_k

(2) From part(1) we know that

$$\frac{\partial Q\left(\left(\beta_1^T,\beta_2^T\right)^T,\widehat{g}\right)}{\partial\beta}|_{\beta=\binom{\widehat{\beta}_1}{0}}=0$$

$$\frac{1}{2}\frac{\partial\widehat{S}\left(\left(\widehat{\beta}_1,0\right),h\right)}{\partial\beta_1}+n\sum_{k=1}^{d_0}P'_{\lambda_n}\left(|\beta_k|\right)sgn\left(\beta_k\right)=0$$

Note as $n\to\infty$ and $\lambda_n\to0$,$P'_{\lambda_n}\left(|\beta_k|\right)=0$ for $k=1,\cdots,d_0$.

$$\begin{aligned}&\frac{1}{2}\frac{\partial\widehat{S}\left(\left(\widehat{\beta}_1,0\right),h\right)}{\partial\beta_1}=0\\ &n\widetilde{V}_{10}\left(\widehat{\beta}_1-\beta_{10}\right)-n^{1/2}\sigma\widetilde{V}_{10}^{1/2}\varepsilon_1+o_p\left(1\right)=0\\ &\sqrt{n}\widetilde{V}_{10}\left(\widehat{\beta}_1-\beta_{10}\right)-\sigma\widetilde{V}_{10}^{1/2}\varepsilon_1+o_p\left(\frac{1}{\sqrt{n}}\right)=0\\ &\sqrt{n}\left(\widehat{\beta}_1-\beta_{10}\right)\to^D N\left(0,\sigma^2\widetilde{V}_{10}^{-1}\right)\end{aligned}$$

Q.E.D.