The Pennsylvania State University
The Graduate School

NEW SCREENING PROCEDURE FOR ULTRAHIGH DIMENSIONAL VARYING-COEFFICIENT MODEL IN LONGITUDINAL DATA ANALYSIS

A Thesis in
Statistics
by
Wanghuan Chu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2014

The thesis of Wanghuan Chu was reviewed and approved* by the following:

Runze Li
Distinguished Professor of Statistics

Matthew Reimherr
Assistant Professor of Statistics

David Hunter
Department Head of Statistics

*Signatures are on file in the Graduate School.

Abstract

This thesis is concerned with feature screening methods for varying-coefficient models in ultrahigh dimensional longitudinal setting. Motivated by an empirical analysis of the Childhood Asthma Management Project, CAMP, we introduce a new screening procedure for time-varying coefficient models with ultrahigh dimensional longitudinal predictor variables. The performance of the proposed procedure is investigated via Monte Carlo simulation. Numerical comparisons indicate that it can outperform existing ones substantially, resulting in significant improvements in explained variability and prediction error. Applying these methods to CAMP, we are able to find a number of potentially important genetic mutations related to lung function, several of which exhibit interesting nonlinear patterns around puberty.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to sincerely thank my thesis advisors, Dr. Runze Li and Dr. Matthew Reihmerr, for their assistance and guidance in my research. They have offered me great help from selecting the research topic, developing the methodology, to analyzing the real data and interpreting the results. I have learned from them a tremendous amount about critical thinking and conducting scientific research.

This work was supported by the National Science Foundation under IGERT Award #DGE-1144860, Big Data Social Science, and Pennsylvania State University. This work was also supported by a National Institute on Drug Abuse (NIDA) grant P50-DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NIDA.

Chapter 1

Introduction

Various feature selection methods have been developed in high dimensional data analysis and widely used in diverse fields. The goal is to recover the underlying model structure when a large number of predictors are introduced at the initial stage, but only a small subset of them are truly associated with the response. The feature dimensionality p is usually much larger than the sample size n. These so called "large p, small n" problems require tools that are not only powerful, but also computationally efficient. While association methods such as the LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001) are powerful, they require significant computational resources for large-dimensional data sets. One of the main issues stems from having to handle all the predictors jointly, which is an enormous computational burden when dealing with hundreds of thousands or possibly millions of predictors simultaneously. An elegant and effective solution is to incorporate screening rules. A screening rule is a method which analyzes much smaller subsets of the predictors and attempts to filter out those that are clearly null. Such rules may attempt to pick the "best" subset of predictors, or just a substantially smaller subset which could in turn be analyzed by other methods. By using such screening rules, it is not unusual to see full day computation times reduced to minutes. The primary goal of this work is to develop an effective screening procedure for longitudinal genetic studies such as CAMP.

A number of feature screening procedures have been developed in various contexts. Fan and Lv (2008) developed a sure independence screening procedure (SIS) for ultrahigh dimensional linear models. Furthermore, they showed that SIS pos-

sesses the *sure screening property*, i.e. with probability tending to one, it produces a subset which contains the true underlying predictors. Fan et al. (2010) extended the SIS for ultrahigh dimensional generalized linear models by ranking the maximum marginal likelihood estimates. Fan et al. (2011) proposed an SIS for ultrahigh dimensional additive models by ranking the magnitude of each nonparametric component. In addition, model-free SIS procedures have been advocated in more recent literature. Zhu et al. (2011) proposed an SIS for the multi-index model setting. Li et al. (2012) developed a distance-correlation based SIS, which are directly applicable for a multivariate response and grouped predictors. He et al. (2013) proposed a quantile-adaptive model-free feature screening procedure for heterogeneous data. Screening procedures have also been developed for varying-coefficient models. Liu et al. (2014) developed an SIS for varying-coefficient models with ultrahigh dimensional predictor variables (ultrahigh dimensional varying-coefficient models for short) by using conditional Pearson correlation coefficient to rank the importance of predictors. Fan et al. (2013) proposed an SIS for ultrahigh dimensional varying-coefficient models by extending the B-spline techniques in Fan et al. (2011) for additive models. Song et al. (2014) further extended the proposal of Fan et al. (2013) for longitudinal data. Both Liu et al. (2014) and Fan et al. (2013) were developed based on independent and identically observed data, while the proposal of Song et al. (2014) did not incorporate within subject correlation and dynamic error variance at the screening stage; a key ingredient of our proposed methodology. Chapter 2 gives more detailed introductions of these screening methods.

1.1 Motivation

Over the last several decades we have seen the rapid development of high dimensional techniques fueled by the precipitous advancement of technology. As our computing power has increased, so has our ability to obtain and examine ever larger and more complicated data sets. One of the primary examples of such data come from genetic association studies. In traditional genome-wide association studies (GWAS) hundreds of thousands or even millions of single nucleotide polymorphisms (SNPs) are explored to find associations with some phenotypes of interest, e.g. blood pressure, height, asthma, etc. Companies are develop-

ing cheaper and cheaper sequencing technologies while also providing increasingly larger pictures of an individual's genome. Indeed, the next technological step consists of high throughput sequencing technologies which are capable of complete genome sequencing. Such studies result in millions of genetic mutations which include, not only SNPs, but also insertions or deletions of segments of DNA.

The present work was motivated by the Childhood Asthma Management Program (CAMP), a 4 year clinical trial which explored the impact of daily asthma medications on lung development in growing children. We consider 540 subjects each contributing 16 clinical visits. The goal is to determine which genetic markers, among hundreds of thousands, affect lung development. The specific outcome we focus on here is FEV1, a common proxy for lung development, which represents the volume of air one can expel out of their lungs in one second. Given that the subjects may change rather rapidly over the course of the trial, we also wish to understand how the effect of significant SNPs changes over time. Analyzing such longitudinal genetic data poses a substantial challenge for data scientists and necessitates the development of new data analytic tools to address scientific questions and test important hypotheses.

1.2 Contributions

While the vast majority of GWAS are cross-sectional, there are numerous longitudinal studies which also have genetic measurements. However, high dimensional methods for longitudinal outcomes have only been sparsely studied. In a longitudinal genetic study such as CAMP, it is typical that researchers collect many baseline variables, a huge number of genetic markers and longitudinal predictor variables/phenotypic traits. Some baseline variables and longitudinal predictor variables should be included in the analysis based on prior studies. None of the aforementioned works on feature screening for ultrahigh dimensional varying-coefficient models have studied this situation, and this work intends to fill this gap. This work also makes a substantial improvement to the B-spline methods in Fan et al. (2013) and Song et al. (2014) for ultrahigh dimensional varying-coefficient models, by effectively incorporating within subject correlation and dynamic error structure. This is now straightforward for standard multivariate regression models

because it is reasonable to assume that the working models are true or well approximate the truth. However, feature screening procedures focus on cycling through very small submodels, which are inherently misspecified. This poses a substantial challenge for constructing effective screening rules using longitudinal data. The main contributions of this thesis project is to present an effective screening rule based on B-spline regression and to demonstrate how within subject variability can be harnessed for increased screening accuracy by Monte Carlo simulation. Furthermore, we illustrate the proposed screening rule via an empirical analysis and comparison of the CAMP data. Our empirical analysis clearly shows that the proposed nonparametric approach is especially useful for such studies as children change quite extensively over a four-year period with highly nonlinear patterns. Chapter 3 gives detailed illustrations of our newly proposed procedure, the Monte Carlo simulations, and the real data application.

1.3 Organization

The rest of the thesis is organized as follows. In Chapter 2, we give a detailed review of the existing feature screening procedures with ultrahigh dimensional data for linear models, nonparametric additive models, and varying-coefficient models. We also introduce the statistical methods for varying-coefficient models with a focus on polynomial splines techniques. In Chapter 3, we propose a new screening procedure for time-varying-coefficient model under longitudinal data setting, and discuss how to incorporate within subject correlation and dynamic error variance into the screening procedure to increase screening accuracy. We conduct a Monte Carlo simulation to examine the finite sample performance of the proposed screening procedure, and to compare with existing ones. To show its application, we present an empirical analysis and comparison of CAMP data using the newly proposed procedure and existing procedures. Some conclusion remarks and future work are discussed in Chapter 4.

Chapter 2

Literature Review

2.1 Feature screening for ultrahigh dimensional data

High and ultrahigh dimensional data analysis plays an important role in modern scientific discovery and statistical research. By high dimension, it is assumed that the dimension p increases with sample size n at a polynomial rate: $p_n = O(n^\alpha)$ for some $\alpha > 0$. Compared to traditional statistical analysis, it brings new challenges such as noise accumulation and spurious correlations, and creates computational issues including heavy cost and algorithmic instability (Fan et al., 2014). Various variable selection and dimension reduction methods are developed to address the issue of noise accumulation in order to effectively and accurately make future prediction and gain scientific insight into the relationship between the features and response.

Nowadays, data with ultrahigh dimensionality are continuously produced with new technologies at much cheaper cost. Here, it means that p_n increases with sample size n at an exponential rate: $p_n = O(\exp(an))$ for some $a > 0$. In this scenario, variable selection and dimension reduction methods can be computationally infeasible. For instance, in genome-wide association studies (GWAS), the availability of inexpensive and high-throughput measurement of the whole genome and transcriptome enables the generation of hundreds of thousands of single-nucleotide polymorphisms (SNPs). The ultrahigh dimensionality of SNPs requires new data

analysis techniques to study their genetic associations with certain phenotypes. One potential solution is a two-stage approach, where a computationally efficient screening procedure is first employed to reduce the dimensionality to a moderate scale under sample size, and then more sophisticated variable selection techniques can be applied to identify important predictors and recover the sparse model. First proposed by Fan and Lv (2008), the idea of feature screening is that 1) marginal association of a predictor with the response is estimated and used as a criterion to measure the predictor's importance in the joint model, and 2) instead of selecting truly important variables, it aims at removing unimportant ones and generates a subset that contains all the active predictors with high probability.

2.1.1 Screening methods for linear models

Fan and Lv (2008) introduced the concept of sure screening and proposed Sure Independence Screening (SIS) method based on correlation learning in linear model setting. Consider the following linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is an $n \times 1$ response vector, $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})^T$ is an $n \times p$ design matrix where the rows are independent from each other. Here, it is assumed that $p >> n$, and only a small subset of $\boldsymbol{x} = (X_1, \ldots, X_p)^T$ are truly associated with the response. Therefore, $\boldsymbol{\beta}$ is sparse and has $d < n$ nonzero components. For simplicity, all predictors are standardized to have mean 0 and standard deviation 1. Then, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^T$ defined as

$$\boldsymbol{\omega} = \boldsymbol{X}^T\boldsymbol{y}, \tag{2.2}$$

is a vector of marginal correlations of predictors with the response. Thus, for any given $\gamma \in [0, 1]$, a submodel can be defined by

$$\mathcal{M}_\gamma = \{1 \leq j \leq p : |\omega_j| \text{ is among the first } [\gamma n] \text{ largest of all}\}, \tag{2.3}$$

where $[a]$ refers to the integer part of a. Thus, simply ranking the marginal cor-

relations and cutting off at certain point can effectively bring the full model with dimensionality p to a submodel with size $d = [\gamma n] < n$.

SIS is shown to have sure screening property, which means that the submodel after screening contains the true model with probability tending to one under certain conditions. To be more specific, let $\mathcal{M}_0 = \{1 \leq j \leq p : \beta_j \neq 0\}$ be the index set for nonzero components of $\boldsymbol{\beta}$, and define $\boldsymbol{z} = \Sigma^{-1/2}\boldsymbol{x}$ and $\boldsymbol{Z} = \boldsymbol{X}\Sigma^{-1/2}$ where $\Sigma = \text{cov}(\boldsymbol{x})$. The following regularity conditions are needed to establish sure screening property:

C1. $\boldsymbol{z}$ has a spherically symmetric distribution and $\boldsymbol{Z}$ has a concentration property, which states that there exists some $c, c_1 > 1$ and $C_1 > 0$ such that the following inequality

$$P(\lambda_{\max}(\widetilde{p}^{-1}\widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{Z}}^T) > c_1 \text{ and } \lambda_{\min}(\widetilde{p}^{-1}\widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{Z}}^T) < 1/c_1) \leq e^{-C_1 n} \tag{2.4}$$

holds for any $n \times \widetilde{p}$ submatrix $\widetilde{\boldsymbol{Z}}$ of $\boldsymbol{Z}$ with $cn < \widetilde{p} \leq p$, where $\lambda_{max}(A)$ and $\lambda_{min}(A)$ are the largest and smallest eigenvalue of matrix A. Also, the random noise $\varepsilon \sim N(0, \sigma^2)$ for some $\sigma > 0$.

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

$$\min_{i \in \mathcal{M}_0} |\beta_j| \geq \frac{c_2}{n^\kappa} \quad \text{and} \quad \min_{i \in \mathcal{M}_0} |\text{cov}(\beta_j^{-1}Y, X_i)| \geq c_3. \tag{2.5}$$

C3. There exist some $\tau \geq 0$ and $c_4 \geq 0$ such that $\lambda_{\max}(\Sigma) \leq c_4 n^\tau$. This condition rules out strong collinearity.

Suppose the above three conditions are satisfied, and if $2\kappa + \tau < 1$, then there exists some $\theta < 1 - 2\kappa - \tau$ such that when $\gamma \sim cn^{-\theta}$ with $c > 0$, it has for some $C > 0$,

$$P(\mathcal{M}_0 \subset \mathcal{M}_\gamma) = 1 - O(\exp(-Cn^{1-2\kappa}/\log n)) \tag{2.6}$$

The sure screening property guarantees that the submodel will contain the true model with an overwhelming probability. And since the size of the submodel $d = [\gamma n]$ is smaller than n, many standard variable selection methods such as SCAD (Fan and Li, 2001), adaptive lasso (Zou, 2006) and Dantzig selector (Candes and Tao, 2007) can be applied to further reduce the dimension and estimate the

coefficients. The simulation studies in Fan and Lv (2008) show that SIS followed by SCAD generates the most accurate model.

Although effective in reducing the dimensionality, SIS requires strong conditions for holding sure screening property, such as the concentration property of the design matrix, the identically normally distributed assumption for the predictors, etc. Under more relaxed conditions, Xue and Zou (2011) proposed a new method called aggressive betting for sparse noiseless signal recovery, which has the exact recovery property with overwhelming probability. Consider first an underdetermined linear equation system $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1}$, where $p >> n$ and $\boldsymbol{X}$ here is called a sensing matrix. The sparsest solution is obtained by the following optimization problem:

$$\min ||\boldsymbol{\beta}||_{L_0} \left(= \sum_{j=1}^{p} \mathbb{1}(\beta_j \neq 0) \right) \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}. \tag{2.7}$$

The aggressive betting method is based on SIS correlation learning, and it is computationally efficient to find the solution. The idea is to find a secure bet $\mathcal{M}$ such that $\mathcal{M}_0 \subset \mathcal{M}$, where $\mathcal{M}_0$ is the index set for true signals. Then, the linear system can be rewritten as $\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}}$, where $\boldsymbol{X}_{\mathcal{M}} = (\ldots, x_j, \ldots)^T$ and $\boldsymbol{\beta}_{\mathcal{M}} = (\ldots, \beta_j, \ldots)^T$ with $j \in \mathcal{M}_1$. And the nonzero components can be extracted from $\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{X}_{\mathcal{M}}^{-1}\boldsymbol{y}$ as the sparse signals. Xue and Zou (2011) proposed to use SIS correlation learning to find this secure bet. They use $\gamma = 1$ in (2.3), and generates an aggressive betting index set $\mathcal{M}_1$ that contains the first n largest $|\omega_j|$'s. It can be shown theoretically that $\mathcal{M}_1$ is a secure bet with overwhelming probability for a wide class of random sensing matrices.

Xue and Zou (2011) further considered sparse recovery in a contaminated linear system defined by $\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, where $\boldsymbol{\varepsilon}$ denotes the measurement error. The sparest solution can be found from the following minimization problem:

$$\min ||\boldsymbol{\beta}||_{L_1} \left(= \sum_{j=1}^{p} |\beta_j| \right) \quad \text{subject to} \quad ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_{L_2} \left(= \sqrt{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2} \right) \leq \nu, \tag{2.8}$$

where ν denotes the size of the error term $\boldsymbol{\varepsilon}$. They proposed to perform SIS first to reduce the dimension and then do robust compressed sensing (2.8), and

showed that the sure screening property holds with an overwhelming property. Compared to SIS in Fan and Lv (2008), the aggressive betting method relaxes the concentration property and identical distributions condition for $\boldsymbol{X}$ and only requires independence of its entries. Moreover, they provide theoretical insights on the impact of noise-to-signal ratio on the probability of sure screening.

Motivated by SIS, Wang (2009) studied the popular and classical variable screening method, forward regression (FR), under the ultrahigh dimensional setup. They proposed FR screening method, and established its screening consistency property. Consider again the linear model (2.1). Let $\mathcal{M}^{(k)}$ be the index set of the submodel at the kth step. Then, FR algorithm can be applied as follow.

Step 1. Set $\mathcal{M}^{(0)} = \emptyset$.

Step 2. At the kth step where $k \geq 1$, for every $j \in \{1, \ldots, p\} \backslash \mathcal{M}^{(k)}$, construct a candidate model based on the index set $\mathcal{M}_j = \mathcal{M}^{(k-1)} \bigcup \{j\}$. Then compute residual some of squares (RSS) from the candidate model $\boldsymbol{y} = \boldsymbol{X}_{\mathcal{M}_j}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ using least square estimation, i.e. $\text{RSS}_j = \boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{X}_{\mathcal{M}_j}(\boldsymbol{X}_{\mathcal{M}_j}^T\boldsymbol{X}_{\mathcal{M}_j})^{-1}\boldsymbol{X}_{\mathcal{M}_j}^T)\boldsymbol{y}$, where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix. Choose the model with the smallest RSS and update the index set to $\mathcal{M}^{(k)}$.

Step 3. Iterate **Step 2** for n times, which yields a solution path consisting of n nested models $\mathbb{M} = \{\mathcal{M}^{(k)} : k = 1, \ldots, n\}$.

The solution path $\mathbb{M}$ is defined to achieve screening consistency if

$$P(\mathcal{M}_0 \subset \mathcal{M}^{(k)} \in \mathbb{M} \quad \text{for some} \quad 1 \leq k \leq n) \to 1. \tag{2.9}$$

Suppose the following conditions are satisfied: 1) both $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ are normally distributed; 2) there exist $0 < \tau_{\min} < \tau_{\max} < \infty$ such that $2\tau_{\min} < \lambda_{\min}(\Sigma) < \tau_{\max}(\Sigma) < \tau_{\max}/2$; 3) $||\boldsymbol{\beta}||_{L_2} \leq C_{\boldsymbol{\beta}}$ for some $C_{\boldsymbol{\beta}} > 0$ and $\min_{j \in \mathcal{M}_0} |\beta_j| \geq \nu_{\boldsymbol{\beta}} n^{-\xi_{\min}}$ for some $\xi_{\min} > 0$; and 4) there exists constants ξ, ξ_0 and ν, such that $\log p \leq \nu n^{\xi}, |\mathcal{M}_0| \leq \nu n^{\xi_0}$ and $\xi + 6\xi_0 + 12\xi_{\min} < 1$. Then, Wang (2009) showed that as $n \to \infty$

$$P(\mathcal{M}_0 \subset \mathcal{M}^{(K\nu n^{2\xi_0 + 4\xi_{\min}})}) \to 1, \tag{2.10}$$

where $K = 2\tau_{\max}\nu C_{\boldsymbol{\beta}}^2 \nu_{\boldsymbol{\beta}}^{-4} \tau_{\min}^{-2}$. This means that with probability tending to one, FR can select all relevant predictors within $O(n^{2\xi_0 + 4\xi_{\min}})$ steps, which is a much

smaller number than n under the 4th condition. To further refine the model, the author suggested using the following BIC criterion (Chen and Chen, 2008) for a model defined by $\mathcal{M}$

$$\mathrm{BIC}(\mathcal{M}) = \log\left(\frac{1}{n}\mathrm{RSS}(\mathcal{M})\right) + \frac{1}{n}|\mathcal{M}|\left(\log n + 2\log p\right). \tag{2.11}$$

The best model denoted by $\widehat{\mathcal{M}}$ can be selected from the solution path $\mathbb{M}$ to be the one with the smallest BIC. It can be shown that the model $\widehat{\mathcal{M}}$ is screening consistent, i.e. $P(\mathcal{M}_0 \subset \widehat{\mathcal{M}}) \to 1$, as $n \to \infty$.

Remark Liang et al. (2012) extended this method to semiparametric partially linear models and proposed profiled forward regression (PRF) screening method. Specifically, they consider the partially linear model,

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + g(U) + \varepsilon, \tag{2.12}$$

where $g(\cdot)$ is an unknown smooth function, and U is an univariate explanatory variable in $[0, 1]$. To deal with the nonparametric part, the authors adopted the profile least squares approach used in Fan et al. (2005). They define the profiled response and the profiled predictor as $Y_i^* = Y_i - E(Y_i|U_i)$ and $\boldsymbol{X}_i^* = \boldsymbol{X}_i - E(\boldsymbol{X}_i|U_i)$ respectively, where each component of $\boldsymbol{X}_i^*$ is $X_{ij}^* = X_{ij} - E(X_{ij}|U_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Then the model reduces to classical linear regression model

$$Y_i^* = \boldsymbol{X}_i^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i. \tag{2.13}$$

The unknown functions $\mathrm{E}(Y_i|U_i)$ and $\mathrm{E}(\boldsymbol{X}_i|U_i)$ can be estimated nonparametrically by methods such as local linear regression (Fan and Gijbels, 1996). Then, the forward regression screening procedure can be applied using the profiled estimators.

Under the same setting of linear model (2.1), Wang (2012) considers in specific the correlation structure among predictors, and proposed factor profiled SIS (FP-SIS). It assumes that the correlation structure of the high-dimensional predictors can be well represented by a set of low-dimensional latent factors:

$$\boldsymbol{X}_i = \boldsymbol{B}\boldsymbol{Z}_i + \widetilde{\boldsymbol{X}}_i, \tag{2.14}$$

where $\boldsymbol{X}_i$ is the ith row of $\boldsymbol{X}$, $\boldsymbol{Z}_i$ is a d-dimensional latent factor and $\boldsymbol{B} = (b_{jk})$ is an $p \times d$ loading matrix. $\widetilde{\boldsymbol{X}}_i$ contains the information in $\boldsymbol{X}_i$ that is missed by $\boldsymbol{Z}_i$, and $\text{cov}(\widetilde{\boldsymbol{X}}_i)$ is a diagonal matrix. In addition, it is assumed that Y_i, X_{ij} and $\widetilde{X}_{ij}$ have mean 0 and variance 1, and $\text{cov}(\boldsymbol{Z}_i) = \boldsymbol{I}$. ε_i is allowed to be correlated with $\boldsymbol{X}_i$ through the latent factor $\boldsymbol{Z}_i$: $\varepsilon_i = \boldsymbol{Z}_i^T\boldsymbol{\alpha} + \widetilde{\varepsilon}_i$, and this might causes biased ordinary least squares estimates and affect the performance of SIS.

The idea of FP-SIS is to remove the common factor $\boldsymbol{Z}_i$ by profiling the response, predictor and noise. Specifically, define a profiled response as $\widetilde{Y}_i = Y_i - \boldsymbol{Z}_i^T\boldsymbol{\gamma}$ with $\boldsymbol{\gamma} = \boldsymbol{B}^T\boldsymbol{\beta} + \boldsymbol{\alpha}$. Then, model (2.1) can be written as

$$\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}, \tag{2.15}$$

where $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{\varepsilon}}$ are uncorrelated, and columns of $\widetilde{\boldsymbol{X}}$ are also mutually uncorrelated. In this case, $\boldsymbol{\beta}$ can be estimated consistently.

By the above results, the model can be rewritten into matrix forms of the observed variables:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\gamma} + \widetilde{\boldsymbol{Y}} = \boldsymbol{Z}\boldsymbol{\gamma} + \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}, \quad \text{and} \quad \boldsymbol{X} = \boldsymbol{Z}\boldsymbol{B}^T + \widetilde{\boldsymbol{X}}. \tag{2.16}$$

Let $\mathcal{S}(\boldsymbol{Z})$ denote the linear space spanned by the column vectors of $\boldsymbol{Z}$, then $\mathcal{H}(\boldsymbol{Z}) = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$ is an $n \times n$ projection matrix onto $\mathcal{S}(\boldsymbol{Z})$ and $\mathcal{Q}(\boldsymbol{Z}) = \boldsymbol{I} - \mathcal{H}(\boldsymbol{Z})$ is another projection matrix onto the space orthogonal to $\mathcal{S}(\boldsymbol{Z})$. Then model (2.15) can be well approximated by $\mathcal{Q}(\boldsymbol{Z})\boldsymbol{Y} = \mathcal{Q}(\boldsymbol{Z})\boldsymbol{X}\boldsymbol{\beta} + \mathcal{Q}(\boldsymbol{Z})\boldsymbol{\varepsilon}$, if a good estimate of $\mathcal{Q}(\boldsymbol{Z})$ is available. Wang (2012) proposed to first estimate the latent factor dimension d by a maximum eigenvalue ratio criterion, and use a least square type objective function to estimate the loading matrix $\boldsymbol{B}$ and then estimate $\boldsymbol{Z}$ by a profiled objective function. It has been shown that factor dimension d can be estimated consistently, and with a correctly specified d, $\mathcal{S}(\boldsymbol{Z})$ can be estimated accurately. Finally, the factor profiled response and predictor can be obtained by $\hat{\boldsymbol{Y}} = \mathcal{Q}(\hat{\boldsymbol{Z}})\boldsymbol{Y}$ and $\hat{\boldsymbol{X}} = \mathcal{Q}(\hat{\boldsymbol{Z}})\boldsymbol{X}$, and SIS can be applied. This procedure also has sure screening property. As suggested in Wang (2009), one can use BIC criterion defined in (2.11) to further reduce the model size.

All previous screening methods developed for linear model are based on Pearson correlation learning. Li et al. (2012) proposed a robust rank correlation screening

(RRCS) method based on Kendall τ correlation coefficient. Given pairs of data $\{Y_i, X_{ij}\}_{i=1}^n$, the marginal rank correlation coefficient between Y and X_j is defined as

$$\omega_j = \frac{1}{n(n-1)} \sum_{i \neq k}^{n} I(X_{ij} < X_{kj}) I(Y_i < Y_k) - \frac{1}{4}, \quad j = 1, \ldots, p. \tag{2.17}$$

And the importance of predictors can be ranked by the magnitudes of ω_j, so that a submodel can be defined the same way as (2.3).

There are several nice features about RRCS that Pearson correlation based SIS does not have. First, it is robust under outliers and influence points in the observations. Second, the Kendall τ is invariant under monotonic transformation, which allows RRCS to discover nonlinear relationship and deal with semiparametric models such as transformation regression models and single-index models under monotonic constraint to the link function without involving nonparametric estimation. Third, the use of ranking information greatly simplifies the theoretical derivation and allows RRCS to achieve sure screening property with only a moment condition.

2.1.2 Screening methods for additive model

Fan et al. (2011) extended SIS (Fan and Lv, 2008) to nonparametric model, and proposed nonparametric independence screening method for additive models. The idea is similar to SIS, i.e. to rank the importance of each predictor in the joint model based on a measure of the goodness of fit of their marginal model. The difference occurs at the marginal model of the response Y against each predictor X_j, which is nonparametric regression in this scenario. To be more specific, consider the following nonparametric additive model

$$Y = \sum_{j=1}^{p} m_j(X_j) + \varepsilon, \tag{2.18}$$

where $\{m_j(X_j)\}_{j=1}^p$ are unknown smooth functions and ε is the random error with conditional mean 0. For identifiability, it is assumed that $\{m_j(X_j)\}_{j=1}^p$ have mean

0. The index set for the true model is now defined by

$$\mathcal{M}_0 = \{1 \leq j \leq p : E[m_j(X_j)^2] > 0\}. \tag{2.19}$$

To identify important variables and recover sparsity, the following p marginal nonparametric regression problems are considered:

$$\min_{f_j \in L_2(P)} E(Y - f_j(X_j))^2, \tag{2.20}$$

where P denotes the joint distribution of $(\boldsymbol{X}, Y)$ with $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, and $L_2(P)$ is the class of square integrable functions under the measure P. The solution to the minimization problem (2.20) is $f_j(X_j) = E(Y|X_j)$, which is the projection of Y onto X_j. Then, the marginal utility of X_j can be measured by $Ef_j^2(X_j)$.

To estimate $Ef_j^2(X_j)$, Fan et al. (2011) used B-spline basis functions to approximate $\{f_j(\cdot)\}_{j=1}^p$. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geq 1$ and $\{\Psi_{jk}, k = 1, \ldots, d_n\}$ be a normalized B-spline basis with $||\Psi_{jk}||_\infty \leq 1$, where $||\cdot||_\infty$ is the sup norm. Then, for any $f_{nj} \in \mathcal{S}_n$ and $j = 1, \ldots, p$,

$$f_{nj}(x) = \sum_{k=1}^{d_n} \beta_{jk} \Psi_{jk}(x). \tag{2.21}$$

Under some smoothness conditions, $f_j(\cdot)$ can be well approximated by $f_{nj}(\cdot)$ for $j = 1, \ldots, p$, and the marginal nonparametric regression problem can be formulated into

$$\min_{f_{nj} \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^n (Y_i - f_{nj}(X_{ij}))^2 = \min_{\boldsymbol{\beta}_j \in \mathbb{R}^{d_n}} \frac{1}{n} \sum_{i=1}^n (Y_i - \boldsymbol{\Psi}_{ij}^T \boldsymbol{\beta}_j)^2, \tag{2.22}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_n})^T$ and $\boldsymbol{\Psi}_{ij} = (\Psi_1(X_{ij}), \ldots, \Psi_{d_n}(X_{ij}))^T$. Using ordinary least square method, the minimizer of (2.22) can be obtained as $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{\Psi}_j^T \boldsymbol{\Psi}_j)^{-1} \boldsymbol{\Psi}_j^T \boldsymbol{Y}$, where $\boldsymbol{\Psi}_j = (\boldsymbol{\Psi}_{1j}, \ldots, \boldsymbol{\Psi}_{nj})^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. Thus, $Ef_j^2(X_j)$ can be estimated by $||\hat{f}_{nj}||_n^2 = \frac{1}{n} \sum_{i=1}^n \hat{f}_{nj}^2(X_{ij})$ with $\hat{f}_{nj}(X_{ij}) = \boldsymbol{\Psi}_{ij}^T \hat{\boldsymbol{\beta}}_j$. Similar rule as (2.3) can be applied to select a submodel with $|\hat{\beta}_j^M|$ replaced by $||\hat{f}_{nj}||_n^2$. Note that it is also equivalent to rank the residual sum of squares of the marginal regression model, where $\text{RSS}_j = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}_{nj}(X_{ij}))^2$.

The theoretical basis of the sure screening is, as pointed out by Fan et al. (2011),

that the marginal signal of the true predictors $\{Ef_j^2, j \in \mathcal{M}_0\}$ does not vanish. The minimum signal of $\{||f_{nj}||\}_{j\in\mathcal{M}_0}$ is at the same level of the marginal projection $\{||f_j||\}_{j\in\mathcal{M}_0}$ when the approximation error is negligible, which can be controlled by the number of basis functions used. Under certain regularity conditions, the uniform convergence of $\{||\hat{f}_{nj}||_{j\in\mathcal{M}_0}\}$ to $\{||f_{nj}||_{j\in\mathcal{M}_0}\}$ and the sure screening property hold.

2.1.3 Screening methods for varying-coefficient models

Fan et al. (2013) extends NIS procedure (Fan et al., 2011) to screen variables in ultrahigh dimensional sparse varying-coefficient models with the following form:

$$Y = \beta_0(W) + \sum_{j=1}^{p} \beta_j(W)X_j + \varepsilon, \tag{2.23}$$

where $\{\beta_j(\cdot)\}_{j=0}^p$ are unknown smooth functions and W is some observable exposure variables. It is assumed that $\boldsymbol{\beta} = (\beta_1(\cdot), \ldots, \beta_p(\cdot))^T$ is sparse, and the index set for the true sparse model is defined as $\mathcal{M}_0 = \{1 \le j \le p : E[\beta_j^2(W)] > 0\}$.

The marginal strength of each predictor can be measured by the expected conditional correlation between Y tand X_j with respect to W. Consider the following marginal regression model for X_j:

$$\min_{\beta_{j0}^M(W),\beta_j(W)\in L_2(P)} E[(Y - \beta_{j0}^M(W) - \beta_j(W)X_j)^2|W]. \tag{2.24}$$

The minimizer of (2.24) is

$$\beta_{j0}^M(W) = \frac{\text{cov}(X_j, Y|W)}{\text{var}(X_j|W)}, \quad \beta_j^M(W) = E(Y|W) - \beta_{j0}^M(W)E(X_j|W). \tag{2.25}$$

Then the utility of X_j is defined by

$$u_j = E[\beta_{j0}^M(W) + \beta_j^M(W)X_j]^2 - E[\beta_0^M(W)]^2 = E\left\{\frac{[\text{cov}(X_j, Y|W)]^2}{\text{var}(X_j|W)}\right\}, \tag{2.26}$$

where $E[\beta_0^M(W)]^2 = E(Y|W)$. To estimate u_j, similar technique used by Fan et al. (2011) is applied, i.e. to approximate unknown coefficients $\{\beta_{j0}^M(W)\}_{j=1}^p$ and

$\{\beta_j^M(W)\}_{j=0}^p$ by B-splines basis functions:

$$\beta_{j0}^M(W) \approx \sum_{k=1}^{d_n} \eta_{jk}\Psi_{jk}(W) \quad \text{and} \quad \beta_j^M(W) \approx \sum_{k=1}^{d_n} \theta_{jk}\Psi_{jk}(W). \tag{2.27}$$

Then, $\boldsymbol{\eta}_j = (\eta_{j1}, \ldots, \eta_{jd_n})^T$ and $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jd_n})^T$ can be estimated by minimizing the ordinary least squares:

$$\min_{\boldsymbol{\eta}_j, \boldsymbol{\theta}_j} \frac{1}{n} \sum_{i=1}^{n} [Y_i - \boldsymbol{\Psi}_{ij}(W_i)\boldsymbol{\eta}_j - \boldsymbol{\Psi}_{ij}(W_i)\boldsymbol{\theta}_j]^2, \tag{2.28}$$

where $\boldsymbol{\Psi}_{ij}(W_i) = (\Psi_1(W_i), \ldots, \Psi_{d_n}(W_i))^T$. Then a sample estimate of the marginal utility u_j can be obtained by

$$\hat{u}_j = \frac{1}{n} \sum_{i=1}^{n} [\hat{\beta}_{j0}^M(W_i) + \hat{\beta}_j^M(W_i)X_{ij}]^2 - \frac{1}{n} \sum_{i=1}^{n} [\hat{\beta}_0(W_i)]^2 \tag{2.29}$$

where $\hat{\beta}_{j0}^M$, $\hat{\beta}_j^M$ and $\hat{\beta}_0^M$ are the estimates for β_{j0}^M, β_j^M and β_0^M using LSE from (2.28). An equivalent measure of marginal strength is the residual sum of squares of the marginal regression model, which can be calculated by

$$\hat{v}_j = \sum_{i=1}^{n} [Y_i - \hat{\beta}_{j0}^M(W_i) - \hat{\beta}_j^M(W_i)X_{ij}]^2. \tag{2.30}$$

By properly choosing thresholds τ_n or ν_n, a submodel can be defined by

$$\mathcal{M}_{\tau_n, \nu_n} = \{1 \le j \le p : \hat{u}_j \ge \tau_n\} = \{1 \le j \le p : \hat{v}_j \ge \nu_n\}. \tag{2.31}$$

Under certain regularity conditions, the sure screening property holds.

A very similar screening procedure is proposed in Liu et al. (2014), where the conditional correlations between Y and X_j's are estimated using kernel smoothing method. Consider again the varying-coefficient model (2.23). The importance of predictors can be measured by $E[\rho^2(X_j, Y|W)]$, where the conditional correlation

$\rho(X_j, Y|W)$ is defined as

$$\rho(X_j, Y|W) = \frac{\text{cov}(X_j, Y|W)}{\sqrt{\text{var}(X_j|W)\text{var}(Y|W)}}. \tag{2.32}$$

To estimate $\rho(X_j, Y|W)$, Liu et al. (2014) applied kernel smoothing method to estimate the five conditional means involved: $E(X_j|W)$, $E(Y|W)$, $E(X_j^2|W)$, $E(Y|W)$ and $E(X_jY|W)$, which are assumed nonparametric smoothing functions of W. Let $K(t)$ be a kernel function and $K_h(t) = K(t/h)/h$, where h is a bandwidth. Then the kernel regression estimate for $E(Y|W)$ is

$$\hat{E}(Y|W) = \sum_{i=1}^{n} \frac{K_h(W_i - W)Y_i}{\sum_{i=1}^{n} K_h(W_i - W)}. \tag{2.33}$$

Estimates for the other four conditional means can be similarly defined, and $\widehat{\text{cov}}(X_j, Y|W)$, $\widehat{\text{var}}(X_j|W)$ and $\widehat{\text{var}}(Y|W)$ can be obtained, and $\hat{\rho}(X_j, Y|W)$ can be calculated as well. Based on the observed i.i.d. data $\{y_i, w_i, x_{ij}\}_{i=1}^n$, a plug-in estimate for $E[\rho^2(X_j, Y|W)]$ is $u_j = \sum_{i=1}^n \hat{\rho}^2(x_{ij}, y_i|w_i)/n$ for $j = 1, \ldots, p$. Then, the screened submodel is defined by

$$\widehat{\mathcal{M}} = \{j : 1 \leq j \leq p : u_j \text{ ranks among the first } d\}, \tag{2.34}$$

where Liu et al. (2014) suggested using $d = [n^{4/5}/\log(n^{4/5})]$. Under certain regularity conditions, this method has both ranking consistency and sure screening properties, where the former states that the ranks of the true predictors are consistently higher than the ranks of the unimportant predictors.

Both methods proposed in Fan et al. (2013) and Liu et al. (2014) are developed for independent data. For longitudinal data with correlated response, Song et al. (2014) and Cheng et al. (2014) proposed to apply screening procedure in Fan et al. (2013) and assume working independence at the screening stage. The only difference is that they considered marginal weighted least square estimation, where equal weight for single observation or equal weight for single subject is used. The sure screening property is established in both works under certain conditions. After reducing the number of covariates to a moderate size by screening, Cheng et al. (2014) proposed to further identify varying coefficients using a group SCAD

estimator by accounting for within-subject correlation.

2.2 varying-coefficient model for longitudinal data

Parametric models such as linear model and generalized linear model are the most fundamental tools in statistical analysis. However, they are developed based on strict assumptions about the relationship between the response and the covariates. In practice, dynamic features exist in data from various scientific areas worth great attention and exploration, and its underlying mechanism cannot be fully understood via parametric modeling. One of the many attempts to increase flexibility and incorporate dynamic features is to fit the data with varying-coefficient model, which allows the regression coefficients to depend on certain covariates. First proposed in Hastie and Tibshirani (1993), the varying-coefficient model is defined as

$$y = \boldsymbol{X}^T \boldsymbol{\beta}(U) + \varepsilon, \tag{2.35}$$

where y is the response, $\boldsymbol{X} = (X_1, \ldots, X_p)$ is a $p-$dimensional predictor, U is the univariate index variable, and ε is the random error with $E(\varepsilon|\boldsymbol{X}, U) = 0$. Here, $\boldsymbol{\beta}(U) = (\beta_1(U), \ldots, \beta_p(U))^T$ consists of p unknown smooth functions of U, which need to be estimated by nonparametric approach. Model (2.35) can be easily extended to generalized linear model framework, by assuming

$$E(y|\boldsymbol{X}, U) = g^{-1}(\boldsymbol{X}^T \boldsymbol{\beta}) \tag{2.36}$$

where $g^{-1}(\cdot)$ is the inverse of the link function $g(\cdot)$.

One of the various applications of varying-coefficient model is to analyze longitudinal data. Longitudinal data occur frequently in biomedical research, where the subjects are measured repeatedly over a given period of time. By allowing the coefficients to vary with time, one can explore time-dependent effects and patterns without posing parametric constraints on the temporal changes of the relationship between response and covariates.

Specifically, consider a random sample from n subjects, and for the $i-$th

subject, we observe the response $y_i(t)$ along with its covariate vectors $\mathbf{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))^T$ at times t_{ij}, $j = 1, \cdots, J_i$, where J_i is the total number of observations from the ith subject. To explore potential time-varying effects, we consider the following time-varying coefficient model

$$y_i(t) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) x_{ik}(t) + \varepsilon_i(t), \tag{2.37}$$

where $\{\beta_k(t), k = 0, \ldots, q\}$ are nonparametric smooth coefficient functions, and $\varepsilon_i(t)$ is the error term with conditional mean $\mathrm{E}\{\varepsilon_i(t)|\boldsymbol{x}_i(t), \boldsymbol{z}_i(t)\} = 0$. The variance and covariance functions of $\varepsilon_i(t)$ is assumed to be time-varying and denoted by $\sigma^2_\varepsilon(t)$ and $C_\varepsilon(s,t)$ for $s \neq t$. Meanwhile, within-subject correlation is considered while observations from different subjects are independent. In general, it is assumed that $t \in \mathcal{T}$, where $\mathcal{T}$ is an interval in $\mathbb{R}$.

2.2.1 Coefficient functions estimation

There are mainly three different estimation methods for $\{\beta_k(t), k = 0, \ldots, q\}$ in (2.37). One is *kernel-local polynomial smoothing* proposed in Hoover et al. (1998), Fan and Zhang (1999) and Fan and Zhang (2008), etc. One is *smoothing spline estimation*, see Hastie and Tibshirani (1993), Hoover et al. (1998) and references therein. The last one is *polynomial spline* introduced in Huang et al. (2002, 2004) and Huang and Shen (2004), etc. In this section, we mainly describe the estimation method using polynomial splines proposed by Huang et al. (2002, 2004), and briefly introduce kernel-local polynomial smoothing in Section 2.2.1.6.

2.2.1.1 Polynomial splines

Polynomial splines are piecewise polynomials with the polynomial pieces jointing smoothly at a set of interior knots based on certain continuity and derivatives conditions. The knots are denoted by $\xi_0 < \xi_1 < \ldots < \xi_L < \xi_{L+1}$ where ξ_0 and ξ_{L+1} are two end points of the interval on $\mathcal{T}$. A spline of degree $d \geq 0$ consists of polynomials of degree d on each of the intervals $[\xi_l, \xi_{l+1})$, $0 \leq l \leq L-1$ and $[\xi_L, \xi_{L+1}]$, and globally has $d-1$ continuous derivatives for $d \geq 1$. Thus, the parameters need to be determined by the users include:

(a) The degree of the spline function, d;

(b) The number of knots, L;

(c) The positions of the interior knots, $\{\xi_l, l = 1, \ldots, L\}$;

(d) The number of free coefficients, i.e. degree of freedom of the spline function, $M = L + d + 1$.

One need to specify two out of (a), (b) and (d) to create the spline basis. Setting $d = 0, 1, 2, 3$ corresponds to, respectively, a piecewise constant function, linear, quadratic and cubic spline, among which cubic spline is the most commonly used.

2.2.1.2 B-spline approximation and least squares estimation

While there are different options of spline basis, we introduce Basis spline, or B-spline, which is a spline function that has the minimal support with respect to a given degree, smoothness and knots positions. See De Boor et al. (1978) and Schumaker (1981) for detailed construction and good properties of B-spline. Let $\{B_{km}, m = 1, \ldots, M_k\}$ be a B-spline basis for a linear space $\mathbb{G}_k$ of spline functions on $\mathcal{T}$, then $\beta_k(t)$ can be approximated by

$$\beta_k(t) \approx \sum_{m=1}^{M_k} \gamma_{km} B_{km}(t), \quad k = 0, 1, \ldots, p. \tag{2.38}$$

M_k, the number of basis functions used for $\beta_k(t)$, can be different for different k. Larger M_k leads to more accurate approximations of the varying coefficients but at the cost of higher variance (i.e. the tradeoff between bias and variance). Therefore, it is natural to allow M_k to increase with sample size. Then, model (2.37) becomes, approximately, a linear regression model:

$$y_i(t_{ij}) \approx \sum_{k=0}^{p} \sum_{m=1}^{M_k} \gamma_{km} B_{km}(t_{ij}) x_{ik}(t_{ij}) + \varepsilon_i(t_{ij}), \tag{2.39}$$

where $x_{i0}(t_{ij}) \equiv 1$ for all i and j. Using least square method, $\{\gamma_{km}, m = 1, \ldots, M_k; k = 1, \ldots, p\}$ can be esitmated by minimizing

$$\sum_{i=1}^{n} \omega_i \sum_{j=1}^{J_i} \left(y_i(t_{ij}) - \sum_{k=0}^{p} \sum_{m=1}^{M_k} \gamma_{km} B_{km}(t_{ij}) x_{ik}(t_{ij}) \right)^2, \tag{2.40}$$

where ω_i is the weight for the $i-$th subject. $\omega_i \equiv 1$ and $\omega_i = 1/J_i$ correspond to, respectively, equal weights for all observations and all subjects. Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ with $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kM_k})^T$,

$$\boldsymbol{B}(t) = \begin{pmatrix} B_{01}(t) & \cdots & B_{0M_0}(t) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & B_{p1}(t) & \cdots & B_{pM_p}(t) \end{pmatrix}, \tag{2.41}$$

$\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{iJ_i})^T$ with $\boldsymbol{Z}_{ij}^T = \boldsymbol{x}_i^T(t_{ij})\boldsymbol{B}(t_{ij})$, $\boldsymbol{\Omega}_i = \text{diag}(\omega_i, \ldots, \omega_i)$, and $\boldsymbol{y}_i = (y_i(t_{i1}), \ldots, y_i(t_{iJ_i}))^T$. Then, provided that $\sum_i \boldsymbol{Z}_i^T \boldsymbol{\Omega}_i \boldsymbol{Z}_i$ is invertible, the minimizer of (2.40) becomes

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{n} \boldsymbol{Z}_i^T \boldsymbol{\Omega}_i \boldsymbol{Z}_i \right)^{-1} \sum_{i=1}^{n} \boldsymbol{Z}_i^T \boldsymbol{\Omega}_i \boldsymbol{y}_i. \tag{2.42}$$

Hence, $\beta_k(t)$ can be estimated by $\hat{\beta}_k(t) = \sum_m \hat{\gamma}_{km} B_{km}(t)$.

2.2.1.3 Variance-covariance estimation for the spline estimators

Based on (2.42), the variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$ is

$$\text{var}(\hat{\boldsymbol{\gamma}}) = \left(\sum_{i=1}^{n} \boldsymbol{Z}_i^T \boldsymbol{\Omega}_i \boldsymbol{Z}_i \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{Z}_i^T \boldsymbol{\Omega}_i \boldsymbol{V}_i \boldsymbol{\Omega}_i \boldsymbol{Z}_i \right) \left(\sum_{i=1}^{n} \boldsymbol{Z}_i^T \boldsymbol{\Omega}_i \boldsymbol{Z}_i \right)^{-1}, \tag{2.43}$$

where $\boldsymbol{V}_i \equiv \text{var}(\boldsymbol{y}_i) = (C_\varepsilon(t_{ij}, t_{ij'}))$ and $C_\varepsilon(t, s)$ is the variance-covariance function of $\varepsilon(t)$. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}(t) = (\hat{\beta}_0(t), \ldots, \hat{\beta}_p(t))^T$ is $\text{var}(\hat{\boldsymbol{\beta}}(t)) = \boldsymbol{B}(t)\text{var}(\hat{\boldsymbol{\gamma}})\boldsymbol{B}^T(t)$.

2.2.1.4 Selection of smoothing parameters

Due to computational complexity, it is impractical to select in an optimal way for all three parameters: the degrees of splines, the number of basis functions (or numbers of knots), and the locations of knots. Huang et al. (2004) suggested using equally spaced knots and fixed degree, and only selecting the number of basis functions M by "leave-one-out" cross-validation (LooCV). This technique is also advocated in Rice and Silverman (1991), Hart and Wehrly (1993) and Hoover et al. (1998). Specifically, let $\hat{\boldsymbol{\beta}}^{(-i)}(t)$ be the spline estimator obtained from deleting all the data of the $i-$th subject. Define the cross-validation criterion to be

$$\mathrm{CV} = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \left\{ \omega_i \left(y_i(t_{ij}) - \boldsymbol{x}_i^T(t_{ij}) \hat{\boldsymbol{\beta}}^{(-i)}(t_{ij}) \right)^2 \right\}. \tag{2.44}$$

Then, $\{M_k, k = 0, \ldots, p\}$ is obtained by minimizing this cross-validation score. When the sample size is very large, one can also use the "K-fold" LooCV to reduce computation.

2.2.1.5 Asymptotic Theory

The asymptotic properties of spline estimator $\boldsymbol{\beta}$ are discussed in Huang et al. (2004). Let $M_n = \max_{0 \le k \le p} M_k$ and $\mathrm{dist}(\beta_k, \mathbb{G}_k) = \inf_{g \in \mathbb{G}_k} \sup_{t \in \mathcal{T}} |\beta_k(t) - g(t)|$ be the L_∞ distance between $\beta_k(\cdot)$ and $\mathbb{G}_k$. The following technical conditions are considered:

C1. The time points $\{t_{ij}, j = 1, \ldots, J_i, i = 1, \ldots, n\}$ are independently distributed on $\mathcal{T}$ as F_T with a Lebesgue density $f_T(t)$ which is bounded away from 0 and infinity uniformly over $t \in \mathcal{T}$. Moreover, they are also independent of the response and covariates $\{(y_i(t_{ij})), \boldsymbol{x}_i(t)), i = 1, \ldots, n\}$.

C2. The eigenvalues $\lambda_0(t) \le \ldots \le \lambda_p(t)$ of $\Sigma(t) = E[\boldsymbol{x}(t)\boldsymbol{x}^T(t))]$ are bounded away from 0 and infinity uniformly on $\mathcal{T}$, i.e. $K_1 \le \lambda_0(t) \le \ldots \le \lambda_p(t) \le K_2$ for some positive constants K_1 and K_2.

C3. There exists a positive constant K_3 such that $E[x_k(t)] \le K_3$ for $t \in \mathcal{T}$ and $k = 0, \ldots, p$.

C4. There exists a constant K_4 such that $\sigma_\varepsilon^2(t) \leq K_4 < \infty$ for $t \in \mathcal{T}$.

C5. $\limsup_n(\max_k M_k / \min_k M_k) < \infty$

C6. $\varepsilon(t)$ can be decomposed into two independent parts $\varepsilon^{(1)}(t)$ and $\varepsilon^{(2)}(t)$, where $\varepsilon^{(1)}(t)$ has mean zero with arbitrary covariance structure, and $\varepsilon^{(2)}(t)$ is measurement error with mean zero and variance σ^2 that are independent at different time points.

Theorem 1.(Consistency) Under conditions **C1-C5**, $\text{dist}(\beta_k, \mathbb{G}_k) = 0, k = 1, \ldots, p$, and $\lim_n M_n \log M_n / n = 0$, $\hat{\beta}_k, k = 1, \ldots, p$, are uniquely defined with probability tending to one. Moreover, $\hat{\beta}_k, k = 0, 1, \ldots, p$, are consistent, that is, $\lim_{n\to\infty} ||\hat{\beta}_k - \beta_k||_{L_2} = 0$

Let $\tilde{\beta}_k(t) = \text{E}[\hat{\beta}_k(t)]$ be the mean of $\hat{\beta}_k(t)$. The rates of convergence is established in the following theorem.

Theorem 2.(Rates of Convergence) Suppose conditions **C1-C5** hold, and $\lim_n M_n \log M_n / n = 0$. Then, $||\tilde{\beta}_k - \beta_k|| = O_p(\rho_n)$ and $||\hat{\beta}_k - \tilde{\beta}_k||_{L_2}^2 = O_p(1/n + K_n n^{-2} \sum_i J_i^{-1})$, where $\rho_n = \max_{0\leq k\leq p} \text{dist}(\beta_k, \mathcal{G}_k)$. Consequently, $||\hat{\beta}_k - \beta_k||_{L_2}^2 = O_p(1/n + K_n n^{-2} \sum_i J_i^{-1} + \rho_n^2)$.

Theorem 3.(Asymptotic Normality) Suppose conditions **C1-C6** hold, and $\lim_n M_n \log M_n / n = 0$ and $\lim_n M_n \max_i J_i / n = 0$. Then, $\{\text{var}[\hat{\boldsymbol{\beta}}(t)]\}^{-1/2}(\hat{\boldsymbol{\beta}}(t) - \tilde{\boldsymbol{\beta}}(t)) \to N(0, \boldsymbol{I})$ in distribution, where $\tilde{\boldsymbol{\beta}}(t) = (\tilde{\beta}_0, \ldots, \tilde{\beta}_p)^T$.

The asymptotic normality results can be used to construct asymptotic confidence intervals and confidence bands.

2.2.1.6 Other methods

The varying-coefficient model can be considered as a linear model at each given time point. Therefore, it is reasonable to estimate the coefficients using data from a local neighborhood, which is the idea of *kernel-local polynomial smoothing*. The coefficient functions $\{\beta_k(t), k = 1, \ldots, p\}$ are approximated locally by

$$\beta_k(t) \approx \beta_k(t_0) + \beta_k^{'}(t_0)(t - t_0) \equiv a_k + b_k(t - t_0) \tag{2.45}$$

for any t in a neighborhood of t_0. a_k and b_k can be estimated by minimizing the weighted least squares

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\left(y_i(t_{ij})-\sum_{k=1}^{p}(a_k+b_k(t_{ij}-t_0))\boldsymbol{x}_i(t_{ij})\right)^2 K_h(t_{ij}-t_0), \tag{2.46}$$

where $K_h(t) = K(t/h)/h$, $K(t)$ is a kernel function and h is a bandwidth. Frequently used kernels include Gaussian kernel ($K(t) = \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$), uniform kernel ($K(t) = I(|t| < 1/2)$), and Epanechikov kernel ($K(t) = 0.75(1-t^2)_+$, $t \in [-1, 1]$), etc. See Fan and Zhang (1999) for matrix form of the solution. Here, h is the only tuning parameter to control the extent of smoothing, so that the choice of h is essential. LooCV described in Section 2.2.1.4 can be applied, but it might cause heavy computation. Another popular method is to use the Mean Squared Error (MSE) criterion, which can be decomposed into the summation of variance and the squared bias. More details of estimation procedures for bias and variance can also be found in Fan and Zhang (1999).

2.2.2 Covariance structure estimation

An important issue with longitudinal data analysis is to study the within-subject correlation structure. Good estimation of the covariance structure improves the efficiency of estimated regression coefficients, makes better prediction of individual trajectory, and also shed valuable insights in practical problems (Sun et al., 2007). This issue has been thoroughly investigated in parametric modeling; see Diggle et al. (2002) and references therein. In nonparamtric setting, Wang (2003) proposed a marginal kernel method that incorporates the true correlation structure. This method has been further extended to marginal generalized semiparametric partially linear models by Wang et al. (2005), which achieves semiparametric information bound derived in Lin and Carroll (2001). Qu and Li (2006) proposed an estimation method for varying-coefficient model that uses penalized quadratic inference functions to incorporate within-subject correlation. In this section, we introduce two estimation procedures for the covariance structure. The first one (Huang et al., 2004) applies polynomial splines to approximate both variance and covariance function. The second one (Fan et al., 2007) uses a kernel estimator

for the nonparametric variance function, and assumes a parametric form for the correlation structure.

2.2.2.1 Polynomial splines method

Huang et al. (2004) proposed a spline based estimate of the covariance function. Let $\{B_m, m = 1, \ldots, M\}$ be a splines basis on $\mathcal{T}$ with a fixed knot sequence. $C_\varepsilon(t, s)$ can be approximated by a tensor product spline on $\mathcal{T} \times \mathcal{T}$:

$$C_\varepsilon(t, s) \approx \sum_m \sum_l u_{ml} B_m(t) B_l(s), \quad t, s \in \mathcal{T}, t \neq s. \tag{2.47}$$

Note that $\mathrm{E}[\varepsilon(t_{ij})\varepsilon(t_{ij'})] = C_\varepsilon(t_{ij}, t_{ij'})$ for $j \neq j'$ and $C(t, s) = C(s, t)$. Thus, $\{u_{ml} : u_{ml} = u_{lm}\}$ can be estimated by minimizing

$$\sum_{i=1}^{n} \sum_{j,j'=1,j<j'}^{J_i} \left(r_i(t_{ij}) r_i(t_{ij'}) - \sum_m \sum_l u_{ml} B_m(t_{ij}) B_l(t_{ij'}) \right)^2, \tag{2.48}$$

where $r_i(t_{ij}) \equiv y_i(t_{ij}) - \hat{\beta}_0(t_{ij}) - \sum_k \hat{\beta}_k(t_{ij})(t_{ij})$, and $\hat{\beta}_k(t_{ij})$ for $k = 0, 1, \ldots, p$ can be obtained using method described in Section 2.2.1.2. Let $\hat{u}_{ml}$ be the minimizers. Then, a spline estimator for $C_\varepsilon(t, s)$ is $\hat{C}_\varepsilon(t, s) = \sum_m \sum_l \hat{u}_{ml} B_m(t) B_l(s)$.

To estimate the variance function $\sigma^2_\varepsilon(t)$, it can be approximated by

$$\sigma^2_\varepsilon(t) \approx \sum_m v_m B_m(t), \tag{2.49}$$

where v_m can be estimated by minimizing

$$\sum_{i}^{n} \sum_{j}^{J_i} \left(r_i^2(t_{ij}) - \sum_m B_m(t_{ij}) \right)^2. \tag{2.50}$$

A spline estimator for $\sigma^2_\varepsilon(t)$ is denoted by $\hat{\sigma}^2_\varepsilon(t) = \sum_m \hat{v}_m B_m(t)$.

Good estimations of $C_\varepsilon(t, s)$ and $\sigma^2_\varepsilon(t)$ depend on appropriate choices of smoothing parameters. In practice, equally spaced knot sequences are usually preferred, and the number of knots can be chosen subjectively or through cross-validation procedures described in Section 2.2.1.4.

2.2.2.2 Semiparametric estimation method

Fan et al. (2007) considered a parametric correlation structure with a nonparametric variance function. To estimate $\sigma_\varepsilon^2(t)$ nonparametrically, they used a kernel estimator

$$\hat{\sigma}_\varepsilon^2(t) = \frac{\sum_i^n \sum_j^{J_i} r_i^2(t_{ij}) K_{h_\varepsilon}(t - t_{ij})}{\sum_i^n \sum_j^{J_i} K_{h_\varepsilon}(t - t_{ij})}. \tag{2.51}$$

For the correlation structure, let $C_i(\boldsymbol{\theta})$ be the correlation matrix for the $i-$th subject with (j, j')th element equaling $\rho(t_{ij}, t_{ij'}, \boldsymbol{\theta})$. The function form of $\rho(s, t, \boldsymbol{\theta})$ is known (e.g. working independence, ARMA(1, 1), etc.), but $\boldsymbol{\theta}$ is unknown and needs to be estimated. Although $\rho(s, t, \boldsymbol{\theta})$ might not be the true correlation function, one can always find a $\boldsymbol{\theta}$ to improve the efficiency of coefficient estimation. To estimate $\boldsymbol{\theta}$, Fan et al. (2007) proposed two minimization criterion: quasi-likelihood and minimum generalized variance of the regression coefficients. To further reduce modeling biases, Fan et al. (2007) suggested expanding the family of parametric functions by taking a linear combinations of different correlation structures.

Chapter 3

A new feature screening procedure

3.1 Methodology

In this section, we introduce a screening method for ultrahigh-dimensional time-varying coefficient model for longitudinal data. Suppose that we collect a random sample from n subjects, and for the i-th subject, we observe the response $y_i(t)$ along with its covariate vectors $\{\mathbf{z}_i(t), \mathbf{x}_i(t)\}$ at times t_{ij}, $j = 1, \cdots, J_i$, where J_i is the total number of observations from the ith subject. The covariate vector $\mathbf{z}_i(t)$ is a low-dimensional predictor consists of variables that are believed to impact the response based on empirical evidence or relevant theories. While we always include the argument t, a particular covariate need not change with time, such as gender. Thus, $\mathbf{z}_i(t)$ should be included into the model, and is not subject to be screened. The covariate vector $\mathbf{x}_i(t)$ is ultrahigh dimensional and contains a vast number of covariates such as hundreds of thousands of SNPs. It is believed that a relatively small number of x-variables have an impact on the response, and most of x-variables are likely to be irrelevant. To explore potential time-varying effects, we consider the following time-varying coefficient model

$$y_i(t) = \beta_0(t) + \sum_{l=1}^{q} \beta_l(t) z_{il}(t) + \sum_{k=1}^{p} \gamma_k(t) x_{ik}(t) + \varepsilon_i(t), \tag{3.1}$$

where $\{\beta_l(t), l = 0, \cdots, q\}$ and $\{\gamma_k(t), k = 1, \cdots, p\}$ are nonparametric smooth coefficient functions, and $\varepsilon_i(t)$ is the error term with conditional mean zero on the

covariates: $\mathrm{E}\{\varepsilon_i(t)|\boldsymbol{x}_i(t), \boldsymbol{z}_i(t)\} = 0$. It is assumed throughout this paper that the variance of $\varepsilon_i(t)$ varies across time, and the error $\varepsilon_i(t)$s are independent between subjects but correlated within subject. In model (3.1), t need not be calendar time. For example, we may set t to be the age of a subject in order to explore potential age-dependent genetic effects and examine whether genetic effect changes across developmental stages. In general, it is assumed that $t \in \mathcal{T}$, where $\mathcal{T}$ is an interval in $\mathbb{R}$.

The goal of a screening procedure is to effectively filter out as many unimportant x-variables as possible while retaining all important x-variables. To denote the significant variables, we define the index set

$$\mathcal{M}_0 = \{1 \leq k \leq p : \|\gamma_k(\cdot)\|_2 > 0\}. \tag{3.2}$$

The screening procedure proposed by Liu et al. (2014) based on conditional correlation cannot be used for feature screening for model (3.1) because of the inclusion of z-variables. The screening procedures developed in Fan et al. (2013) and Song et al. (2014) may be directly applicable for model (3.1) by assuming within-subject observations are independent. In this section, we introduce a more effective screening procedure, which improves the proposal of Fan et al. (2013) by incorporating within-subject correlation and taking into account the time-varying error variance. We next describe our procedure.

For each k, we define a marginal nonparametric regression model with the kth x-variable:

$$y_i(t_{ij}) = \beta^*_{0k}(t_{ij}) + \sum_{l=1}^{q} \beta^*_{lk}(t_{ij}) z_{il}(t_{lj}) + \gamma^*_k(t_{ij}) x_{ik}(t_{ij}) + \varepsilon^*_i(t_{ij}), \tag{3.3}$$

where $\{\beta^*_{lk}(t), l = 0, 1, \ldots, q\}$ and $\gamma^*_k(t)$ are smooth coefficient functions. Intuitively, the residual sum of squares of model (3.3) may be used to measure the importance of the k-th x-variable. A smaller residual sum of squares implies that the corresponding x-variable explains the more variation of the response variable, and therefore would be more important.

Model (3.3) is a nonparametric regression model. We employ a regression spline method to estimate its coefficient functions and obtain its residuals. Using cubic

B-splines bases, we approximate $\{\beta_{lk}^*(t), l = 0, 1, \ldots, q\}$ and $\gamma_k^*(t)$ as follows:

$$\beta_{lk}^*(t) \approx \sum_{m=1}^{M_{ln}} \eta_{lm} B_m(t) \quad \text{and} \quad \gamma_k^*(t) \approx \sum_{h=1}^{L_{kn}} \theta_{kh} B_{kh}(t), \tag{3.4}$$

where $\{B_{hm}(\cdot), m = 1, \ldots, M_{hn}\}$ is a set of B-splines which may differ across h, and M_{ln} and L_{kn} are the numbers of basis functions used for $\beta_{lk}^*(t)$ and $\gamma_k^*(t)$ respectively. Larger M_{kn} and L_{ln} lead to more accurate approximations of the varying coefficients but at the cost of higher variance (i.e., the tradeoff between bias and variance). Model (3.3) becomes, approximately, a linear regression model:

$$\begin{aligned} y_i(t_{ij}) \approx & \sum_{m=1}^{M_{0n}} \eta_{0m} B_{0m}(t_{ij}) + \sum_{l=1}^{q} \sum_{m=1}^{M_{ln}} \eta_{lm} B_{lm}(t) z_{il}(t_{ij}) \\ & + \sum_{h=1}^{L_{kn}} \theta_{kh} B_{kh}(t) x_{ik}(t_{ij}) + \varepsilon_i^*(t_{ij}). \end{aligned} \tag{3.5}$$

The error term $\varepsilon_i^*(t_{ij})$ is assumed to be independent between subjects and correlated within subject. Moreover, the variance of $\varepsilon_i^*(t_{ij})$ is assumed to be time-varying. Incorporating the error covariance structure into the model estimation is expected to increase screening accuracy. Intuitively, one may apply the techniques related to weighted least squares (WLS) methods or generalized estimating equation (GEE) method (Liang and Zeger, 1986) to construct an estimate for the coefficients. However, the situation here is much more challenging than the parametric GEE because (a) the working marginal model (3.5) is a misspecified model, and (b) the total computational cost for estimation of error variance and parameters in the error correlation matrix in each marginal model would be extremely expensive in the presence of ultrahigh dimensional x-covariates. Instead of estimating the covariance matrix of $\boldsymbol{\varepsilon}_i^* = (\varepsilon_i^*(t_{i1}), \ldots, \varepsilon_i^*(t_{iJ_i}))^T$, we propose an approach to construct the weighted matrix in the weighted least squares method.

We propose to construct $V(t_{ij})$, a working variance function for $\varepsilon_i^*(t_{ij})$, by applying the techniques in Huang et al. (2004). We apply the ordinary least

squares method and regression spline technique to the following working model

$$y_i(t_{ij}) = \beta_{0k}^w(t_{ij}) + \sum_{l=1}^{q} \beta_{lk}^w(t_{ij}) z_{il}(t_{lj}) + \varepsilon_i^w(t_{ij}), \tag{3.6}$$

and obtain the corresponding residuals $\{r_i(t_{ij})\}$. Assuming that $V(t)$ is a smooth function of t, we can approximate $V(t_{ij}) \approx \sum_{h=1}^{H_n} \alpha_h B_{hn}(t_{ij})$. Minimizing the following least squares function

$$\sum_{i=1}^{n} \sum_{j=1}^{J_i} \left(r_i^2(t_{ij}) - \sum_{h=1}^{H_n} \alpha_h B_{hn}(t_{ij}) \right)^2, \tag{3.7}$$

lead to a least squares estimate for the coefficients: $\{\hat{\alpha}_h, h = 1, \ldots, H_n\}$. Then, define $\hat{V}(t_{ij}) = \sum_{h=1}^{H_n} \hat{\alpha}_h B_{hn}(t_{ij})$.

In this paper, we consider a parametric model for the working correlation matrix. Denote by $\boldsymbol{R}_i(\boldsymbol{\lambda}) = (R_{jk})$, the $J_1 \times J_i$ working correlation matrix for the i-th subject, where $\boldsymbol{\lambda}$ is an $s \times 1$ vector that fully characterizes the correlation structure. Commonly used correlation structures include autoregressive (AR) correlation structure, stationary or nonstationary M-dependent correlation structures, as well as parametric families such as the Matérn. In practice, we propose to employ moment estimators for the parameters $\boldsymbol{\lambda}$ in the correlation structure based on the residuals $r_i(t_{ij})$s in feature screening procedures. Denote by $\hat{\boldsymbol{\lambda}}$ the resulting moment estimate of $\boldsymbol{\lambda}$.

We propose the following weighted matrix for the i-th subject

$$\boldsymbol{W}_i = \hat{\boldsymbol{V}}_i^{-\frac{1}{2}} \boldsymbol{R}_i^{-1}(\hat{\boldsymbol{\lambda}}) \hat{\boldsymbol{V}}_i^{-\frac{1}{2}}, \tag{3.8}$$

where $\hat{\boldsymbol{V}}_i$ is the $J_i \times J_i$ diagonal matrix consists of the time-varying variance

$$\hat{\boldsymbol{V}}_i = \begin{pmatrix} \hat{V}(t_{i1}) & 0 & \ldots & 0 \\ 0 & \hat{V}(t_{i2}) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \hat{V}(t_{iJ_i}) \end{pmatrix}. \tag{3.9}$$

Then, the generalized estimating equation becomes

$$\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\lambda})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = 0, \tag{3.10}$$

where $\boldsymbol{y}_i = (y_{i1}(t_{ij}), \ldots, y_{iJ_i}(t_{ij}))^T$. Note that $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ are different from what we define before. Here, $\boldsymbol{X}_i$ is a submatrix of the design matrix for model (3.5) consisting only data of the ith subject, and $\boldsymbol{\beta}$ contains all the coefficients to be estimated, i.e. $\{\eta_{lm}, m = 1, \ldots, M_{ln}, l = 0, 1, \ldots, q; \theta_{kh}, h = 1, \ldots, L_{kn}\}$. The unknown parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ can be estimated iteratively using Newton-Raphson algorithm. Note that one can choose different working structures based on the nature of data being analyzed. The simplest case would be working independence, $\boldsymbol{R}_i = \boldsymbol{I}_{J_i}$, which only accounts for varying variance but ignores within subject correlation that normally exists in longitudinal data. Another option is compound symmetry with $R_{jk} = \rho$ for any $j \neq k$, which assumes that observations from the same subject are equally correlated regardless of their time lags. This is equivalent to the correlation structure of a mixed effect model with a random intercept. More appropriate options for longitudinal data include first-order autoregressive (AR-1) with $R_{jk} = \rho^{|i-j|}$, and stationary stationary correlation structure. In our simulation studies and real data example, we use stationary M-dependent correlation structure. For instance, when the ith subject has $J_i = 5$ observations and we set $M = 3$, then its correlation matrix can be specified by

$$\boldsymbol{R}_i(\boldsymbol{\lambda}) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \alpha_3 & 0 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ 0 & \alpha_3 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}, \tag{3.11}$$

where $\boldsymbol{\lambda} = (\alpha_1, \alpha_2, \alpha_3)^T$. One would expect $\alpha_1 > \alpha_2 > \alpha_3$, since measurements taken closer in time would be more related. If there is not enough information about the data to impose any structure, one can choose fully unstructured correlation matrix with $J_i(J_i - 1)/2$ unknown parameters to be estimated.

Given $\hat{\boldsymbol{V}}_i$ and $\hat{\boldsymbol{R}}_i$, we can compute (3.8) by $\hat{\boldsymbol{\Sigma}}_i$, and WLSE of the coefficients in

(3.5) by $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n} \boldsymbol{X}_i \hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{X}_i)^{-1} (\sum_{i=1}^{n} \boldsymbol{X}_i \hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{y}_i)$. Plugging in these coefficients back into the B-spline approximations in (3.4), $\{\hat{\beta}_{lk}^*(t), l = 0, 1, \ldots, q\}$ and $\hat{\gamma}_k^*(t)$ can also be obtained.

We can then obtain the WLS estimate for regression coefficients in model (3.5), and calculate the fitted value $\hat{y}_i^{(k)}(t_{ij})$. This enables us to calculate the weighted mean squared errors denoted by u_k:

$$u_k = \frac{1}{n} \sum_{i=1}^{n} J_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i^{(k)})^T \mathbf{W}_i (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i^{(k)}), \tag{3.12}$$

where $\hat{\boldsymbol{y}}_i^{(k)} = (\hat{y}_i^{(k)}(t_{i1}), \ldots, \hat{y}_i^{(k)}(t_{iJ_i}))^T$. Note that smaller value of u_k indicates stronger marginal association between the $k-$th covariate and the response. Thus, we sort $\{u_k, k = 1, \ldots, p\}$ in an increasing order, and define the screened submodel as:

$$\hat{\mathcal{M}}_{\tau_n} = \{1 \leq k \leq p : u_k \text{ ranks among the first } \tau_n\}, \tag{3.13}$$

where τ_n is the sub model size chosen to be smaller than the sample size n. Following Fan and Lv (2008), we set $\tau_n = [n/\log(n)]$, where $[a]$ refers to the integer part of a.

3.2 Simulation studies

To make our simulation results more generalizable to real world applications, we generate data mimicking the CAMP data, and compare the finite sample performance of the new method with that of sure independence screening (SIS) (Fan and Lv, 2008) and nonparametric independence screening (NIS) with varying-coefficient models (Fan et al., 2013). The screening proposed in Song et al. (2014) is essentially equivalent to that in Fan et al. (2013) under our simulation setting since the number of observations for each subject is the same for all subjects (as is the case in CAMP). We do not include procedures proposed by Song et al. (2014) and Liu et al. (2014) in our numerical comparison as they cannot be applied to our setting.

We set the feature dimension, p, to 2000. We first randomly choose p SNPs from CAMP as the x-variables and set gender as the only z-variable. This is because only

gender among the baseline variable has a significant impact on the response based on our preliminary analysis of the CAMP data using an age-varying coefficient model. The distribution of the age variable are approximately normal over the range $[5, 17.2]$. To achieve better numerical stability, we make a transformation on the time points $\{\tilde{t}_{ij}, j = 1, \ldots, J_i; i = 1, \ldots, n\}$ so that they are approximately uniformly distributed on $[0, 1]$ by $t_{ij} = \Phi((\tilde{t}_{ij} - \bar{t})/s_t)$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$, and $\bar{t}$ and s_t are the sample mean and standard deviation of all time points $\tilde{t}_{ij}$ in the CAMP data. We generate the simulated data from

$$y_i(t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij})\text{Gender}_i + \sum_{k=1}^{p} \gamma_k(t_{ij})\text{SNP}_{ik} + \varepsilon_i(t_{ij}). \tag{3.14}$$

In each replication, we randomly select $n = 200$ subjects from the CAMP data, and directly take their Gender variable, time t_{ij} and the p selected SNPs to this replication.

The error term $\varepsilon_i(t_{ij})$ is generated from a zero mean Gaussian process with variance and correlation defined by

$$\text{Var}(\varepsilon_i(t_{ij})) \equiv V(t_{ij}) = 0.5 + 3t_{ij}^3, \quad \text{and} \quad \text{cor}(\varepsilon_i(t_{ij}), \varepsilon_i(t_{ik})) = 0.5\rho_1^{|j-k|} + 0.5\rho_2, \tag{3.15}$$

where we use a correlation structure as a combination of AR(1) and compound symmetry with equal weights. We set $(\rho_1, \rho_2) = (0.6, 0.4)$ and (0.8, 0.6) in our simulation.

We set x_1, x_2, x_3, x_4 to be significant, and all others are inactive. To make comparisons fair, we consider two examples for nonzero coefficients. In the first example, the nonzero coefficients for x-variables are time-varying, while they are time-invariant in the second example. The specific nonzero coefficient functions are given below.

- *Example I.* The nonzero coefficient functions are defined by

$$\gamma_1(t) = 0.5\cos(\pi t)\mathbf{1}_{\{t \le 0.5\}}, \quad \gamma_2(t) = -0.4\cos(2\pi t)\mathbf{1}_{\{t \le 0.5\}},$$

$$\gamma_3(t) = -0.3\sin(2\pi t), \quad \gamma_4(t) = 0.5(1.2 - t).$$

- *Example II.* The nonzero coefficient functions are defined by

$$\gamma_1(t) = 0.4, \quad \gamma_2(t) = 0.5, \quad \gamma_3(t) = -0.3, \quad \gamma_4(t) = -0.5.$$

We set $\beta_0(t)$ and $\beta_1(t)$ to be the coefficient functions estimated from $y_i(t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij})\text{Gender}_i + \varepsilon_i(t_{ij})$ using the CAMP data. Their plots are shown in Figure 3.1. The baseline predictor Gender is also considered in SIS and NIS method in our numerical comparison.

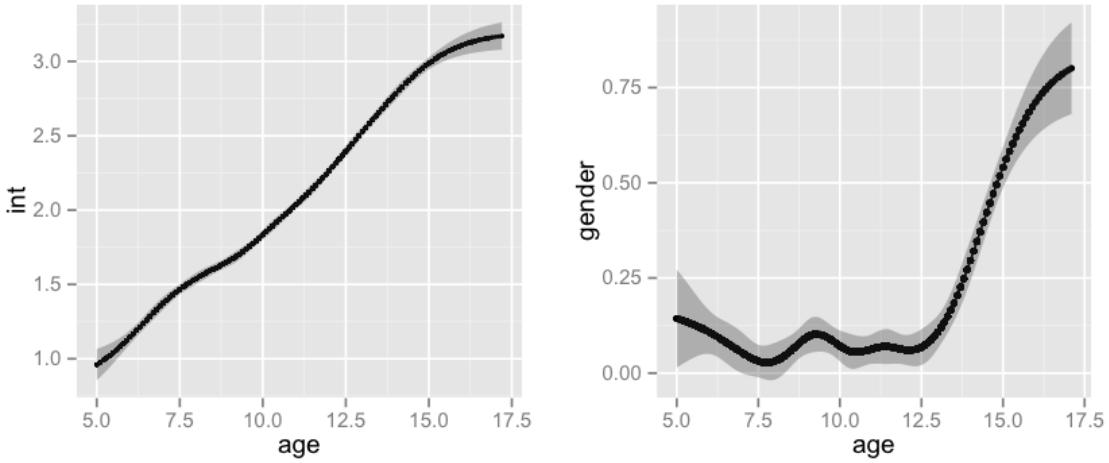

Figure 3.1. Coefficient Functions for Intercept and Gender

Following Liu et al. (2014), the following four criteria are used to evaluate the performance of different screening methods.

- R_k: The average of ranks of x_k (or SNP_k in our case) in terms of the screening criterion based on 1000 replications.

- M: The minimum size of the submodel so that all true predictors can be selected. The 5%, 25%, 50%, 75% and 95% quantiles of M are reported from 1000 replications.

- p_a: The proportion of 1000 replications where all true predictors are being selected into $\hat{\mathcal{M}}_{\tau_n}$.

- p_k: The proportion of x_k being selected into the submodel $\hat{\mathcal{M}}_{\tau_n}$ over 1000 replications.

Table 3.1: R_j of the Active SNPs

	Example 1: $\gamma(t)$'s are time-varying				*Example 2*: $\gamma(t)$'s are time-invariant			
Method	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
	$\rho_1 = 0.6$, $\rho_2 = 0.4$							
SIS	140.605	1054.662	1013.490	1.440	4.341	3.576	1.124	5.588
NIS	17.610	139.569	92.412	1.555	4.436	3.859	1.127	6.453
new method	4.392	4.530	13.165	1.539	7.035	4.457	1.059	14.905
	$\rho_1 = 0.8$, $\rho_2 = 0.6$							
SIS	261.827	1020.349	1022.777	7.236	4.896	3.939	1.226	6.874
NIS	53.680	231.055	176.300	5.263	5.988	4.611	1.221	10.543
new method	7.109	2.967	12.778	2.787	11.446	6.853	1.102	25.455

To calculate p_a and p_k, we set the selected submodel size $\tau_n = \nu[n/\log n]$, $\nu = 1, 2, 3$ (Fan and Lv, 2008). All the simulation results are summarized over 1000 replications.

Results of R_j's are reported in Table 3.1, quantiles of M in Table 3.2, and p_j's and p_a in Table 3.3. Outputs of the first example shows that SIS is able to identify SNP_4, with an average rank (R_4) of 1.555 and 5.263, and selection proportion (p_4) 0.998 and 0.972 under $\tau_n = 38$, for the two correlation cases respectively. But it fails to select other three SNPs. This is because $\gamma_4(t)$ provides the strongest and most stable signal among all with small coefficient (-0.5) and relatively large intercept (0.6), which is the closest to what SIS is designed for.

NIS can also identify SNP_4 very well. In addition, NIS selects SNP_1 into the submodel with relatively large probability, especially under $(0.6, 0.4)$ correlation scenario and using more conservative submodel size ($\tau_n = 76$ or 114). However, it gives bad ranking to SNP_2 and SNP_3 (R_2 and R_3 of NIS from Table 3.1) and low selection rates over 1000 replications (p_2 and p_3 of NIS from Table 3.3). This is because NIS can correctly specify the time-varying effects of $\gamma_1(t)$, but the signal magnitudes of $\gamma_2(t)$ and $\gamma_3(t)$ are not large enough for NIS to detect under the non-negligible within subject correlation and time-varying variance.

Table 3.2: The quantiles of M

	Example 1: $\gamma(t)$'s are time-varying					*Example 2*: $\gamma(t)$'s are time-invariant				
Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
	$\rho_1 = 0.6$, $\rho_2 = 0.4$									
SIS	563.8	1074.25	1453	1759.25	1955.10	4	5	7	8	10
NIS	18	57	118	255.25	632.10	4	5	8	9	13
new method	4	5	6	10	57.05	4	6	9	14	53.1
	$\rho_1 = 0.8$, $\rho_2 = 0.6$									
SIS	534.65	1077.5	1466.5	1758.25	1954.0	4	5	8	9	17.05
NIS	43.00	135.0	246.0	447.25	874.4	4	5	8	11	40.05
new method	4	5	7	13	68.2	4	7	13	29	126.25

Our procedure has excellent performance for all four SNPs, generating consistently high ranking and large selection rates under all scenarios. This indicates that under longitudinal settings, better screening results can be gained from accounting for varying variance and within subject correlation.

As for the second example, all methods have good performance with SIS performing the best. Thus, our screening method is also valid for linear models. However, if the underlying model is known to be linear, SIS would be the best option due to its small computational cost. By comparing the results of the two correlation scenarios, we can observe that all methods perform slightly worse when the error correlations get larger.

Given the performance of the three methods, we can definitively recommend our procedure in practice for longitudinal data. Especially in the setting where further analyses are to be performed, our method truly shines. While our rankings for constant effects are slightly worse, they are still very high and thus very likely to make it past any reasonable screening threshold. Our performance for truly time varying effects and dynamic errors is substantially better, and it is clear that SIS and NIS run the risk of missing such signals.

Table 3.3: Selection proportion p_j's and p_a for true SNPs

		Example 1: $\gamma(t)$'s are time-varying					*Example 2*: $\gamma(t)$'s are time-invariant				
τ_n	Method	p_1	p_2	p_3	p_4	p_a	p_1	p_2	p_3	p_4	p_a
		$\rho_1 = 0.6$, $\rho_2 = 0.4$									
	SIS	0.569	0.010	0.011	0.998	0.000	1.000	1.000	1.000	0.997	0.997
38	NIS	0.885	0.321	0.534	0.998	0.156	1.000	1.000	1.000	0.993	0.993
	new method	0.997	0.987	0.943	1.000	0.927	0.984	0.995	1.000	0.936	0.918
	SIS	0.669	0.030	0.022	1.000	0.001	1.000	1.000	1.000	0.999	0.999
76	NIS	0.955	0.515	0.677	0.999	0.336	1.000	1.000	1.000	0.997	0.997
	new method	0.997	0.997	0.972	1.000	0.966	0.991	0.999	1.000	0.976	0.966
	SIS	0.727	0.048	0.034	1.000	0.002	1.000	1.000	1.000	1.000	1.000
114	NIS	0.972	0.627	0.775	0.999	0.491	1.000	1.000	1.000	0.999	0.999
	new method	0.997	0.998	0.983	1.000	0.978	0.997	0.999	1.000	0.984	0.980
		$\rho_1 = 0.8$, $\rho_2 = 0.6$									
	SIS	0.380	0.018	0.014	0.972	0.000	0.997	0.999	1.000	0.992	0.988
38	NIS	0.653	0.171	0.318	0.971	0.040	0.989	0.997	1.000	0.959	0.945
	new method	0.969	0.996	0.931	0.996	0.896	0.949	0.982	1.000	0.867	0.808
	SIS	0.499	0.037	0.028	0.980	0.000	0.999	1.000	1.000	0.996	0.995
76	NIS	0.807	0.309	0.459	0.993	0.127	0.997	1.000	1.000	0.983	0.980
	new method	0.993	0.996	0.965	0.999	0.954	0.983	0.993	1.000	0.936	0.915
	SIS	0.567	0.054	0.040	0.987	0.003	0.999	1.000	1.000	0.997	0.996
114	NIS	0.864	0.419	0.556	0.996	0.207	0.997	1.000	1.000	0.992	0.989
	new method	0.997	0.998	0.982	1.000	0.977	0.988	0.995	1.000	0.957	0.941

3.3 Application

The Childhood Asthma Management Program (CAMP) was a longitudinal study designed to explore the long-term impact of several daily treatments for mild to moderate asthma in children (The Childhood Asthma Management Program Research Group, 1999, 2000). Here, we consider $n = 540$ Caucasian subjects, each of whom contributed 16 clinical visits over 4 years. The primary outcome variable examined here is lung growth, as assessed by the change in forced expiratory volume in one second (FEV1, expressed as a percentage of the predicted value), which is used as the response variable in our analysis. Genomewide SNP data and phenotype information are downloaded from dbGaP (http://www.ncbi.nlm.nih.gov/gap) study accession phs000166.v2.p1. There are in total eight hundred and seventy thousand SNPs to be screened. We set the age of the i-th subject at the j-th measurements to be the time variable t_{ij}, and consider the following model

$$\begin{aligned} \mathrm{FEV}_i(\mathrm{age}_{ij}) =& \beta_0(\mathrm{age}_{ij}) + \beta_1(\mathrm{age}_{ij})\mathrm{Gender}_i \\ &+ \sum_{k=1}^{p} \gamma_k(\mathrm{age}_{ij})\mathrm{SNP}_{ik} + \varepsilon_i(\mathrm{age}_{ij}), \end{aligned} \tag{3.16}$$

where gender is the only baseline predictor, and $\{\mathrm{SNP}_{ik}\}$ are the SNP variables. Throughout this empirical analysis, it is assumed that $\varepsilon_i(\mathrm{age}_{ij})$ is a Gaussian process with mean zero and variance $\mathrm{Var}(\varepsilon_i(\mathrm{age}_{ij})) = V(\mathrm{age}_{ij})$, a smoothing function of age.

We apply the feature screening procedure introduced in Section 3.1 and NIS method to this data set. Both methods select $\tau_n = [540/\log(540)] = 85$ SNPs. The two submodels obtained have 15 overlapping SNPs. Since the purpose of screening procedures is to remove as many irrelevant SNPs as possible, and to retain all important SNPs, the screening procedures are typically conservative. Thus, we apply further confirmatory analyses to remove more irrelevant SNPs.

We next employ stepwise regression techniques to further remove unimportant SNPs. In the forward step, we choose the SNP which results in the greatest decrease in the weighted residual sum of squares (WRSS), and then use F-test to determine if this SNP should be added to the model. The F statistic can be

calculated by

$$F = \frac{(\text{WRSS}_1 - \text{WRSS}_2)/(\text{df}_2 - \text{df}_1)}{\text{WRSS}_2/\text{df}_2}, \tag{3.17}$$

where WRSS_1 and WRSS_2 are the weighted residual sum of squares of the model without and with the candidate SNP respectively, and defined as

$$\text{WRSS} = \sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \hat{\boldsymbol{\Sigma}}_i^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i), \tag{3.18}$$

where $\hat{\boldsymbol{\Sigma}}_i^{-1}$ is the estimated covariance matrix for subject i from its corresponding model. In the backwards step, to determine if an existing SNP should be excluded, we check if its contribution is smaller than the newly added SNP. Specifically, let $\{\text{SNP}_{(k)},\ k = 1, \ldots, K\}$ be the existing SNPs and $\text{SNP}_{(K+1)}$ be the new one. Then, delete $\text{SNP}_{(j)}$ if WRSS of model based on $\{\text{SNP}_{(k)},\ k = 1, \ldots, K\}$ is greater than WRSS of model based on $\{\text{SNP}_{(k)},\ k \in \{1, \ldots, K+1\} \setminus \{j\}\}$. This procedure automatically stops when no SNP can make a significant contribution to the model. By applying this procedure to the two submodels obtained from screening, a final model with 23 SNPs is selected for the new method and a model with 6 SNPs for NIS.

To further compare these two models, we conduct leave one subject out cross validation (LooCV) and assess their predication performance. At each evaluation, we leave the data of one subject out, and predicated his/her FEV. Let $y_i(t_{ij})$ and $\hat{y}_i^{(i)}(t_{ij})$, $j = 1, \ldots, J_i$ be the observed and predicted value for subject i, then we calculate the prediction sum of squares (PRESS):

$$\text{PRESS} = \sum_{i=1}^{n} \sum_{j=1}^{J_i} (y_i(t_{ij}) - \hat{y}_i^{(i)}(t_{ij}))^2. \tag{3.19}$$

Table 3.4 shows the results of the two models selected and the new method outperforms NIS by more than 10%.

We show in Figure 3.2 the estimated coefficient functions of the best model selected for the new method. The first two panels are the coefficients for the intercept and gender (with female as the baseline); the others are for the 23 SNPs. Detailed information about these 23 SNPs is also shown in Table 3.5. The shape

Table 3.4: LooCV Results

	Number of SNPs	PRESS
new method	23	873.37
NIS	6	992.01

of the intercept function is as expected; as subjects age their lungs develop and FEV1 increases. We see that there is a slight tapering around 16-17 years old as teenagers get closer to their adult heights. The shape of the gender function is especially interesting. We see that at younger ages, boys have slightly higher (recall female is the baseline) lung function. However, we see a dip and the two groups begin to converge starting around age 10 which is right around the time girls begin entering puberty. Boys, on average, enter puberty about a year after girls which we can also see as the plot rebounds around age 12 as the boys begin growing larger than the girls. Finally, around age 16 when both groups are closer to their adult heights, we see the plots settle on a more pronounced difference between the genders.

The shapes we see in the SNP functions take a variety of forms. Most are primarily protective (1, 6, 10, 14, 16, 17, 18, 20, 23) or deleterious (2, 3, 4, 5, 7, 9, 11, 13, 15, 19, 21 22) though SNPs 8 and 12 don't clearly fall into one category. We also see that the impact of many of the SNPs seems to fundamentally change before and after puberty. The plot we see for SNP14 might be what one would expect for a protective SNP; a steady increase which accelerates during puberty and then tapers off. Shapes that are more surprising are ones like SNP1. This SNP starts off as protective, but when children hit puberty, it seems to decrease in effect. SNP3 only seems to be active during puberty, but otherwise doesn't seem to have an effect. In many of the plots we see more chaotic or rapid behavior around puberty. This makes sense as a rapid growth in the children should rapidly change how SNPs are affecting lung function. What isn't so obvious is that puberty also seems to fundamentally change the nature of certain SNPs. Some seem to change the direction of the effect while some seem most active during puberty. It is these types of patterns which make nonparametric longitudinal methods so powerful.

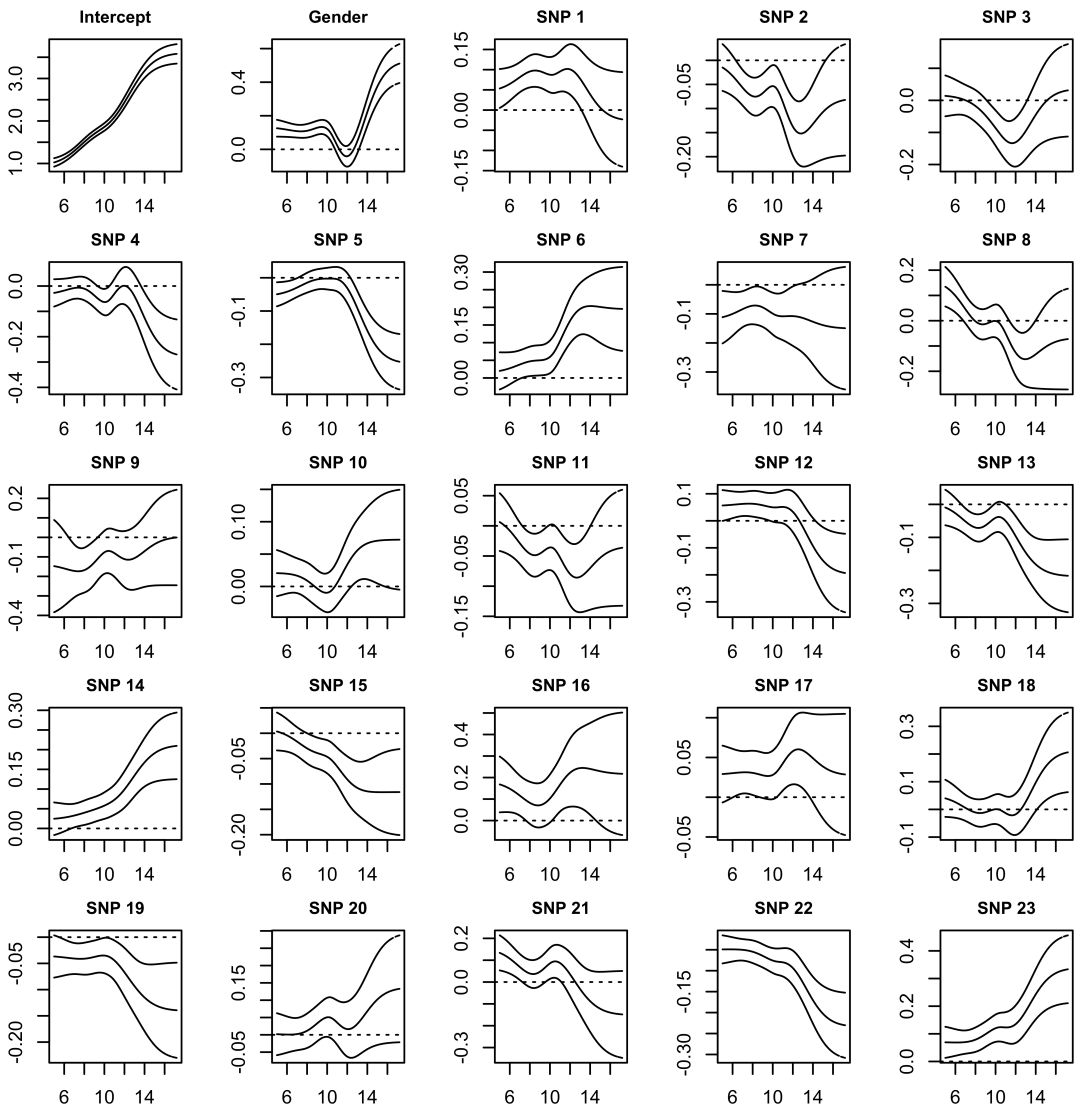

Figure 3.2. Estimated Coefficient Functions for Best Model Selected by Our Procedure

By allowing very general structures for the coefficient functions, we can better find nonlinear patterns.

We conclude this section by examining the heritability discovered by the models, as well as the heritability explained by individual SNPs. Heritability is a concept that summarizes the proportion of variation in a trait due to genetic factors. Examining heritability is an important step in understanding the genetic architecture of complex diseases. Since we are selecting a relatively small subset of SNPs, the heritability we examine here is not the over all heritability of the

Table 3.5: Information of the 23 SNPs selected by the new method

No.	Chromosome	SNP Name	Chr. Position	No.	Chromosome	SNP Name	Chr. Position
SNP_1	22	rs5992809	16601985	SNP_{13}	2	rs2894456	222765340
SNP_2	16	rs17766975	74968977	SNP_{14}	1	rs1499663	55830578
SNP_3	5	rs4704894	157136938	SNP_{15}	1	rs7530486	64955414
SNP_4	8	rs16924622	60197787	SNP_{16}	5	rs16902245	85806442
SNP_5	10	rs293286	52889045	SNP_{17}	19	rs11673302	9462362
SNP_6	4	rs17444879	41429386	SNP_{18}	11	rs10501066	26724528
SNP_7	15	rs12050625	30751109	SNP_{19}	13	rs12716713	67431310
SNP_8	5	rs17167077	98947056	SNP_{20}	14	rs4904757	41274666
SNP_9	2	rs12469442	195233905	SNP_{21}	4	rs10433674	71980590
SNP_{10}	18	rs1459497	52150550	SNP_{22}	1	rs12734254	77180853
SNP_{11}	2	rs1481387	157598327	SNP_{23}	6	rs7751381	117037951
SNP_{12}	5	rs1013193	169131901				

disease but only the heritability due to our sub model. The heritability of FEV1 was explored in Reimherr et al. (2014), where they found that the heritability of FEV1 in asthmatic children was around 46%. However, they also discovered that heritability can vary substantially with age. In their methods, "time" was study time (i.e. number of weeks of the trial), where as here we let time be the age of the child. This is especially important as we can get a more direct handle on how heritability changes with age. For model (3.16), we consider the total heritability of all selected SNPs and heritability of a single SNP. The heritability of all SNPs is calculated by

$$\begin{aligned} \mathrm{H(FEV)} =& \frac{\mathrm{RSS(FEV|Gender)}}{\mathrm{RSS(FEV|Gender)}} \\ & - \frac{\mathrm{RSS(FEV|Gender, SNP_1, \ldots, SNP_p)}}{\mathrm{RSS(FEV|Gender)}}. \end{aligned} \tag{3.20}$$

Here, RSS is the unweighted residual sum squares defined by

$$\text{RSS} = \sum_{i=1}^{n} \sum_{j=1}^{J_i} (y_i(t_{ij}) - \hat{y}_i(t_{ij}))^2,$$

where $\hat{y}_i(t_{ij})$ is the fitted value from the model using weighted least square estimation, i.e. accounting for time-varying variance and within subject correlation. The total heritability for our model and best model of NIS are, respectively, 34.673% and 17.977%. We also estimate the time-varying heritability for all SNPs using a B-splines approximation and the results are shown in Figure 3.3. There we see a similar result of Reimherr et al. (2014) that the heritability seems to change quite substantially with age. In particular, we see rapid increases in the heritability as children enter puberty. It seems to level off at around ages 16-17. While we know that the heritability of the NIS sub model is lower than ours, we see another remarkable difference in their time varying heritability patterns. The NIS model plot looks similar to ours except at the later ages as it decreases as puberty ends. This suggests that the NIS model has missed SNPs which play a larger role at the later ages.

Finally, we calculate the heritability of single SNPs. This is determined by the order in which each SNP is selected into the model in the stepwise selection procedure. Let $\text{SNP}_{(k)}$ be the kth SNP to be selected into the model, then its heritability is calculated by

$$\begin{aligned} \text{H}(\text{SNP}_{(k)}) = & \frac{\text{RSS}(\text{FEV}|\text{Gender}, \text{SNP}_{(1)}, \ldots, \text{SNP}_{(k-1)})}{\text{RSS}(\text{FEV}|\text{Gender})} \\ & - \frac{\text{RSS}(\text{FEV}|\text{Gender}, \text{SNP}_{(1)}, \ldots, \text{SNP}_{(k-1)}, \text{SNP}_{(k)})}{\text{RSS}(\text{FEV}|\text{Gender})}. \end{aligned}$$

Table 3.6 and Table 3.7 show the heritability of single SNP in the two best models. We see that the heritability of the SNPs ranges fairly evenly between zero and four percent. Interestingly, SNP22 or rs12734254 on gene ST6GALNAC5 was also discovered in Reimherr et al. (2014) using a very different and stringent statistical approach, which reaffirms that this gene is influencing lung function.

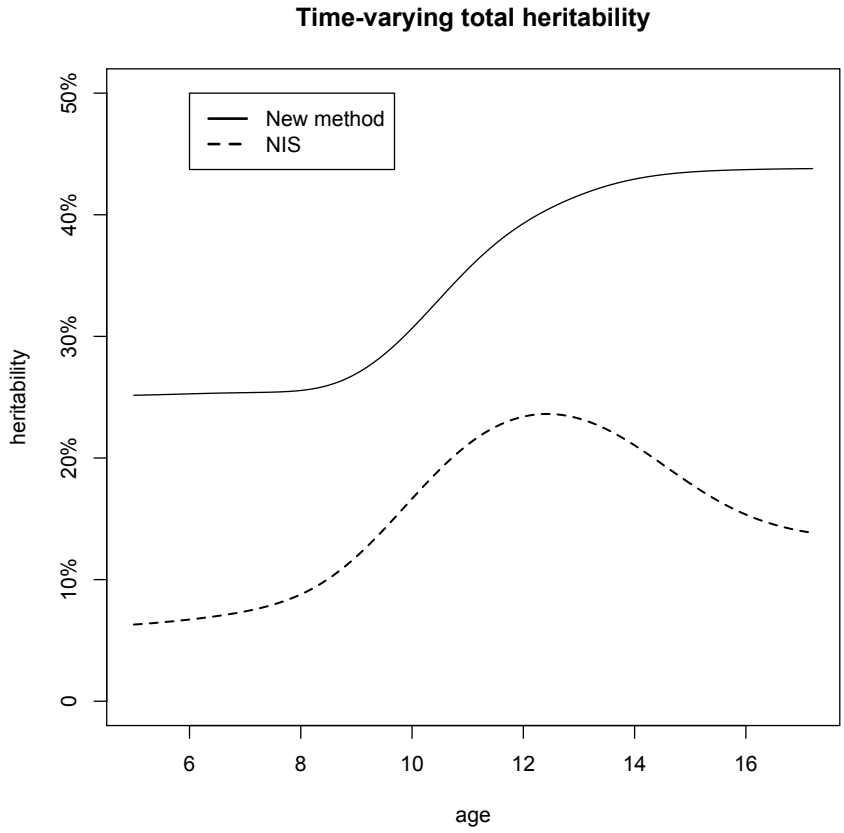

Figure 3.3. Time-varying Total Heritability

Table 3.6: Heritability of Single SNPs by New Method

Selecting Order	SNP Name	H($SNP_{(k)}$)	Selecting Order	SNP Name	H($SNP_{(k)}$)
1	rs5992809	1.321%	13	rs2894456	3.196%
2	rs17766975	3.494%	14	rs1499663	1.53%
3	rs4704894	0.68%	15	rs7530486	1.267%
4	rs16924622	1.515%	16	rs16902245	0.825%
5	rs293286	2.412%	17	rs11673302	0.53%
6	rs17444879	4.231%	18	rs10501066	0.198%
7	rs12050625	1.408%	19	rs12716713	1.689%
8	rs17167077	1.327%	20	rs4904757	0.204%
9	rs12469442	0.286%	21	rs10433674	0.388%
10	rs1459497	1.245%	22	rs12734254*	3.454%
11	rs1481387	0.9%	23	rs7751381	2.051%
12	rs1013193	0.524%			

Table 3.7: Heritability of Single SNPs by NIS

Selecting Order	SNP Name	H($SNP_{(k)}$)	Selecting Order	SNP Name	H($SNP_{(k)}$)
1	rs1522621	4.201%	4	rs2894456	3.423%
2	rs17766975	3.137%	5	rs4323745	2.698%
3	rs17444879	4.183%	6	rs12734069	0.336%

Chapter 4

Conclusions and Future Work

We developed a screening procedure for ultrahigh dimensional varying-coefficient models motivated by longitudinal genetic studies. From our numerical comparison, the proposed procedure can outperform the SIS proposed in Fan and Lv (2008) and NIS proposed in Fan et al. (2013) for longitudinal data. This implies that incorporating within-subject variability and within-subject correlation may increase the accuracy of a screening rule. We further applied the proposed procedure for an empirical analysis of CAMP data. The newly proposed screening procedure is able to select a model with much higher heritability and lower prediction error.

There are a number of extensions of the proposed methodology can be expanded. One that we briefly explored, is allowing the correlation structure to also take a smooth nonparametric form. However, our initial attempts showed that the resulting estimates were too noisy to be of much use, and resulted in inconsistent screening results. Thus, finding a nonparametric estimation method for the correlation structure which results in efficient and stable screening would be useful. Another useful generalization would be to allow for more smoothing procedures such as local polynomial smoothing, smoothing splines, etc. Regression splines allow for nice statistical tests which we exploit in the application section. To achieve a similar effect, other smoothing methods would need to be incorporated with care.

Bibliography

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–2351.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Cheng, M.-Y., T. Honda, J. Li, and H. Peng (2014). Nonparametric independence screening and structural identification for ultra-high dimensional longitudinal data. *Annals of Statistics*, to appear.

De Boor, C. et al. (1978). *A practical guide to splines*. Springer Verlag.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger (2002). *Analysis of longitudinal data*. Oxford University Press.

Fan, J., Y. Feng, and R. Song (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association 106*(494), 544–557.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Volume 66. CRC Press.

Fan, J., F. Han, and H. Liu (2014). Challenges of big data analysis. *National Science Review*, nwt032.

Fan, J., T. Huang, et al. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli 11*(6), 1031–1057.

Fan, J., T. Huang, and R. Li (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association 102*(478), 632–641.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Fan, J., Y. Ma, and W. Dai (2013). Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association* (just-accepted).

Fan, J., R. Song, et al. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics 38*(6), 3567–3604.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *Annals of Statistics 27*(5), 1491–1518.

Fan, J. and W. Zhang (2008). Statistical methods with varying coefficient models. *Statistics and its Interface 1*(1), 179.

Hart, J. D. and T. E. Wehrly (1993). Consistency of cross-validation when the data are curves. *Stochastic processes and their applications 45*(2), 351–361.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological) 55*(4), 757–796.

He, X., L. Wang, H. G. Hong, et al. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics 41*(1), 342–369.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

Huang, J. Z. and H. Shen (2004). Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scandinavian journal of statistics 31*(4), 515–534.

Huang, J. Z., C. O. Wu, and L. Zhou (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica 14*(3), 763–788.

Li, G., H. Peng, J. Zhang, L. Zhu, et al. (2012). Robust rank correlation based screening. *The Annals of Statistics 40*(3), 1846–1877.

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Liang, H., H. Wang, and C.-L. Tsai (2012). Profiled forward regression for ultrahigh dimensional variable screening in semiparametric partially linear models. *Statistica Sinica 22*(2), 531.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Lin, X. and R. J. Carroll (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association 96*(455), 1045–1056.

Liu, J., R. Li, and R. Wu (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association 109*(505), 266–274.

Qu, A. and R. Li (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics 62*(2), 379–391.

Reimherr, M., D. Nicolae, et al. (2014). A functional data analysis approach for genetic association studies. *The Annals of Applied Statistics 8*(1), 406–429.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological) 53*(1), 233–243.

Schumaker, L. L. (1981). *Spline functions: basic theory*. Cambridge University Press.

Song, R., F. Yi, and H. Zou (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sincia*, to appear.

Sun, Y., W. Zhang, and H. Tong (2007). Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics 35*(6), 2795–2814.

The Childhood Asthma Management Program Research Group (1999). The childhood asthma management program (CAMP): design, rationale, and methods. *Controlled Clinical Trials 20*, 91–120.

The Childhood Asthma Management Program Research Group (2000). Long–term effects of budesonide or nedocromil in children with asthma. *New England Journal of Medicine 343*, 1054–1063.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association 104*(488), 1512–1524.

Wang, H. (2012). Factor profiled sure independence screening. *Biometrika 99*(1), 15–28.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika 90*(1), 43–52.

Wang, N., R. J. Carroll, and X. Lin (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association 100*(469), 147–157.

Xue, L. and H. Zou (2011). Sure independence screening and compressed random sensing. *Biometrika 98*(2), 371–380.

Zhu, L.-P., L. Li, R. Li, and L.-X. Zhu (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association 106*(496), 1464–1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.