High/Ultra-High Dimensional Single-Index Models

by

Guannan Wang

(Under the direction of Lily Wang and Jaxk Reeves)

Abstract

Single-index models are useful and fundamental tools for handling "curse of dimensionality" problems in nonparametric regression. In addition to that, variable selection also plays an important role in such model building processes when the index vectors are high-dimensional. Several procedures have been developed for estimation and variable selection for single-index models when the number of index parameters is fixed.

In many high-dimensional model selection problems, the number of parameters is increasing along with the sample size. In the first part of this work, we consider weakly dependent data and propose a class of variable selection procedures for single-index prediction models. We apply polynomial spline basis function expansion and smoothly clipped absolute deviation penalty to perform estimation and variable selection in the framework of a diverging number of index parameters. Under stationary and strong mixing conditions, the proposed variable selection method is shown to have the "oracle"

property when the number of index parameters tends to infinity as the sample size increases. A fast and efficient iterative algorithm is developed to simultaneously estimate parameters and select significant variables. The finite sample behavior of the proposed method is evaluated with simulation studies and illustrated by some river flow data from Iceland.

Most recently, among numerous modern problems in multiple scientific fields, a noteworthy characteristic feature is that the dimension of the explanatory variable, p, is large, and potentially much larger than the sample size, n. For those problems of large scale or dimensionality, variable selection again plays an important role in the modeling process. Under the sparsity assumption, a variable screening procedure was proposed by [16] to reduce the ultra-high dimensionality to a moderate level. However, for practical data analysis, without any prior knowledge, both the true model and the marginal regression can be highly non-linear. To address the above issues in the second part of this work, we investigate ultra-high dimensional penalized single-index models. We further extend the sure independence screening method into a nonparametric independence screening procedure. In addition, a data-driven thresholding determination procedure is proposed to enhance the finite sample performance. New theoretical results are also derived for oracle parameters. Both the numerical results and the real data application demonstrate that the proposed procedure works very well, even for moderate sample size and large dimensionality.

Index words: B-spline, diverging parameters, NP dimensionality, SCAD, semiparametric regression, weakly dependent data

High/Ultra-High Dimensional Single-Index Models

by

Guannan Wang

B.S., Nankai University, China 2008

M.S., University of Georgia 2010

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Department of Statistics

Athens, Georgia

2015

©2015

Guannan Wang

All Rights Reserved

High/Ultra-High Dimensional Single-Index Models

by

Guannan Wang

Approved:

Major Professors: Lily Wang
Jaxk Reeves

Committee: Cheolwoo Park
David Lowenthal
Gauri Datta
William McCormick

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015

Acknowledgments

First and foremost, I would like to thank my PhD advisors, Dr. Lily Wang and Dr. Jaxk Reeves, for their long-term strong support and tremendous patience in mentoring my PhD research, revising my manuscripts, and providing me with numerous opportunities to engage with the statistical research community, such as attending workshops and conferences, peer-reviewing scientific articles, etc. Under their guidance and influence, I am not only growing to be a great and enthusiastic statistician, but also striving to be a person with great personality and confidence to embrace everyday life with broader vision, bigger heart and grateful attitude.

I would like to thank my committee members Dr. Cheolwoo Park, Dr. David Lowenthal, Dr. Gauri Datta, and Dr. William McCormick, for contributing their time, efforts and scientific insights in supporting my PhD studies at UGA. I would like to extend my special thanks to Dr. Gauri Datta, Dr. Abhyuday Mandal and Dr. Cheolwoo Park who wrote very supportive recommendation letters for me when I applied to the Statistics Department in 2009. In addition, I would like to thank Dr. Jaxk Reeves for granting me an excellent opportunity to be a member of the Statistics Department in 2010. I also want to thank Dr. Jaxk Reeves and Dr. Kimberly Love-Myers for recruiting me as a graduate consultant in our Statistical Consulting Center; this is an invaluable experience for me. There is a long list of professors I am grateful for due to their excellent teaching: Dr. Lynne Billard, Dr. Gauri Datta, Dr. Daniel Hall, Dr. Abhyuday Mandal, Dr. William McCormick, Dr. Cheolwoo Park, Dr. Jaxk Reeves, Dr. Lynne Seymour, Dr. T. N. Sriram, Dr. Lily Wang, Dr. Xiangrong Yin, and many others. They helped me to lay a solid foundation in methodology and theory in statistics. I also want to express my gratitude to my two advisors for giving me constructive career advice and support. I am thankful for the tremendous assistance provided by administrative staff in the Statistics Department, with special thanks to Tim Cheek, Daphney Smith and Megan Weatherford.

My time in the US was made enjoyable in large part due to many talented colleagues and close friends I met here, including Cheng Qiu, Fei Liu, Jin Tang, Jing Wan, Lina Liao, Nan Zhang, Xinyi Li, Ping Yin, Qiangjia Zhong, Wenbo Wu, Wenhui Sheng, Xijue Tan, Yijie Xue, Zhen Yan, and many others.

Last but not least, I would like to thank my beloved parents, Jinhong Wang and Qing Chu, my husband, Haiming Wang, and my lovely daughter, Anna, for their perpetual and absolute love and trust.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Single-Index Models

In Statistics, a linear regression is an approach for modeling the relationship between a response variable, Y, and one or more explanatory variables denoted as X. The model can be expressed as

$$E(Y|X) = X^T\beta.$$

Unfortunately, in practice, the relationship between response variable and explanatory variables is not limit to linearity. The generalized linear model is a flexible generalization of ordinary linear regression by allowing the linear model to be related to the response variable via a link function, i.e.,

$$E(Y|X) = m(X^T\beta). \tag{1.1.1}$$

These models typically assume that $m(\cdot)$ is known up to a finite number of parameters, such as binary logit or probit. When m is unknown, model (1.1.1) can provide a specification which is more flexible than a purely parametric model, and it can become a single-index model (SIM) under certain conditions. Single-index models have applications to a variety of fields. For example, econometric studies use these models as a compromise between too restrictive parametric models and flexible but hardly estimable purely non-parametric models.

Another excellent feature of the single-index model is that it is an attractive dimension reduction method. Single-index models are similar to the first step of projection pursuit regression, a dimension reduction method, (see [24], [33] and [6]). The basic appeal of the single-index model is its simplicity: the p-variate function $m(x) = m(x_1, \ldots, x_p)$ is expressed as a univariate function of $x^T\theta_0 = \sum_{j=1}^{p} x_j\theta_{0,j}$.

Over the last two decades, much effort has been focused on research into estimation of the single-index coefficients, as well as into the non-parametric link function, with concentration on proofs of root-n consistency and demonstrations of efficiency. Examples can be found in [43], [29], [4], [54] and [30]. Among these methods of estimation, the most popular are the average derivative estimation method proposed by [29] and [28]. More recently, [55] proposed the minimum average variance estimation (MAVE) for several index vectors. [52] proposed the polynomial spline estimator for the single-index prediction model (SIP), which is more robust against deviations from SIMs. [5] studied the SIMs with heteroscedastic errors and recommended an estimating equation method in terms of transferring restricted least squares to un-restricted least squares.

[59] derived inference for the index parameters by the local linear method. [7] suggested an estimating function method to study the SIMs.

1.2 Variable Selection Techniques

When the dimension of X is high, one unavoidable issue is the "curse of dimensionality", which refers to phenomena that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. To circumvent this difficulty, variable selection techniques play pivotal roles. The statistical literature contains numerous procedures on variable selection for linear models and other parametric models. Akaike's information criterion (AIC), Mallows' C_p, and the Bayesian information criterion (BIC) are several examples of traditional variable selection procedures. They all use a fixed penalty on the size of a model. To replace fixed penalties, [1] and [45] suggest the use of a data adaptive penalty in the variable selection procedures. However, as pointed out in [15] and [19], all these procedures follow stepwise or subset selection procedures, which are extremely computationally intensive, hard to derive sampling properties for, and unstable. On the other hand, most convex penalties, such as quadratic penalties, often produce shrinkage estimators of parameters that make trade-offs between bias and variance. To avoid the unnecessary biases and the inefficiency of traditional variable selection procedures, [15] proposed a unified approach via non-concave penalized least squares. Such methods can select variables and estimate the coefficients of variables automatically and simultaneously. There are four fundamental good features of this method: (1) it keeps the appealing features of subset selection and ridge regression; (2) it produces sparse solutions; (3) it ensures continuity of the selected models, and (4) it has unbiased estimates for large coefficients. All of the above-mentioned appealing features can be achieved by choosing suitable penalized functions, such as the smoothly clipped absolute deviation (SCAD) penalty that was proposed by [12], the Lasso [46], the Dantzig selector [3], the Elastic net (Enet) penalty [65], the MCP [58] and related methods proposed in [64] and [66].

In Chapter 2, we consider a class of single-index prediction models with diverging number of index parameters. We propose to use polynomial spline basis function expansion and SCAD penalty to perform estimation and variable selection.

1.3 Independence Screening Techniques

With rapid improvement of computing power, high-throughput data of unprecedented size and complexity are frequently collected in many scientific fields. As discussed in the previous section, to handle these large-scale data, variable selection plays an important role in high dimensional statistical modeling. However, when the number of variables p grows much faster than the sample size n, the aforementioned variable selection techniques face the following three tremendous challenges: (1) computational expediency, (2) statistical accuracy, and (3) algorithmic stability. To tackle these problems, [16] introduced a sure independence screening (SIS) method to select important variables in the framework of ultra-high dimensional linear regression via marginal correlation learning. Later, [25] extended the SIS method to the generalized correlation ranking. [21] extended the SIS idea to ultra-high dimensional generalized linear models. In addition, a useful technical tool for establishing the sure screening results and bounding false selection rates is derived in the same paper. In 2011, [13] further extended the SIS method to nonparametric independence screening (NIS) which can be implemented onto ultra-high dimensional additive models.

In the meantime, several other methods have been developed to handle such ultra-high dimensional problems, such as the data-tiling method proposed by [26], the marginal partial likelihood method [57], robust screening methods using rank correlation by [37], and distance correlation [38]. Inspired by all these previous works, in Chapter 3, we will focus on variable nonparametric independence screening in single-index models with non-polynomial (NP) dimensionality.

Chapter 2

High Dimensional Single-Index Models [1]

2.1 Introduction

For the past two decades, high dimensional problems are becoming increasingly common in many scientific areas, including biostatistics, medicine, economics and financial econometric. When the dimension of covariates increases, one unavoidable issue is the "curse of dimensionality", which refers to the poor convergence rate. Much effort has been devoted to tackling of this difficulty. As an attractive dimension reduction method, single-index models (SIMs) play a useful and fundamental role for handling "curse of

[1] Wang, G. and Wang, L. (2015). Spline estimation and variable selection for single-index prediction models with diverging number of index parameters. *Journal of Statistical Planning and Inference* **162**, 1–19. Reprinted here with permission of publisher.

dimensionality" problems. Various intelligent estimators of the single-index coefficients have been derived by many researchers. Examples can be found in [43], [29], [4], [54] and [30]. [55] introduced the minimum average variance estimation (MAVE) for several index vectors. [52] proposed the polynomial spline estimator for the SIP, which is more robust against deviations from SIMs. [5] studied the SIMs with heteroscedastic errors and recommended an estimating equation method in terms of transferring restricted least squares to un-restricted least squares. [59] derived inference for the index parameters by the local linear method. [7] suggested an estimating function method to study the SIMs.

Along with the SIMs, when the index vectors are high-dimensional, variable selection for significant predictors is very practical in such model building processes. For example, in time series modeling, we often need to select significant explanatory lagged variables. Most traditional variable selection procedures, such as Akaike's information criterion (AIC), Mallow's C_p, and the Bayesian information criterion (BIC), use a fixed penalty on the size of a model. To overcome the inefficiency of traditional variable selection procedures, [15] proposed a unified approach via non-concave penalized likelihood and demonstrated that penalized likelihood estimators are asymptotically as efficient as the ideal "oracle" estimator for certain penalty functions, such as the smoothly clipped absolute deviation (SCAD) penalty. [19] further extended the method to the situation with a diverging number of parameters, which substantially enlarges the scope of applicability of the shrinkage methods. We refer to [19], [31] and [53] for more works in the high-dimensional framework where the number of covariates increases with the sample size.

Several procedures have been developed for estimation and variable selection for SIMs when the number of index parameters is fixed. Examples include the dissected cross-validation (DCV) method in [36], the profile least squares (PrLS) estimation procedure in [39], the adaptive lasso with kernel smoothing in [63], the penalized least squares method in [41], and the lasso with local linear smoothing method in [60]. Unfortunately, in practice, many variables are sometimes introduced in an effort to reduce possible modeling biases. In many high-dimensional model selection problems, the number of introduced variables depends on the sample size, which reflects the instability of the parametric problem. For example, when running regressions on time-series data, it is often important to include many lagged values of the dependent variable as predictor variables. Sometimes, to capture the persistence of a time series, the lag length can be very long, or even close to the entire length of the time series.

When a diverging number of predictors are involved in SIM, [62] proposed a method based on slice inverse regression (SIR) to select variables. However, the SIR based method imposes a strong assumption on the predictors: the distribution of the covariates must be elliptically symmetric distributions. In time series analysis, the covariates are typically the lagged values of a time series. As discussed in [55], the elliptical symmetry of the covariates implies the time series itself is time reversible [47], which is an exceptional feature in most time series. Therefore, their method will not work for most time series data; see the discussions in [55] and [41].

In this work, we consider weakly dependent data, and focus on variable selection and estimation for single-index prediction models, as in [52]. We apply the SCAD penalty and polynomial spline basis function expansion to simultaneously perform variable selection and estimation in the framework of a diverging number of index parameters. Under a mixing condition and some other regularity conditions, the proposed variable selection method is shown to have the "oracle" property when the number of parameters diverges as the sample size increases. A fast and efficient algorithm is developed to simultaneously estimate parameters and select significant variables. Our method is applicable to selecting significant variables when modeling time series data, which may include endogenous variables (lagged variables) as well as exogenous variables.

The rest of the article is organized as follows. Section 2.2 first provides the background of the single-index prediction model, then introduces the polynomial spline smoothing and the penalized SCAD estimators. Section 2.3 shows the main theoretical results in the framework of a diverging number of index parameters. Section 2.4 presents an algorithm to implement the proposed method. Section 2.5 reports our findings in three simulation studies. The proposed method is applied in Section 2.6 to the Iceland river flow data. All technical proofs are given in Section 2.7.

2.2 Methodologies

2.2.1 Single-Index Prediction Model

Let $\{X_i, Y_i\}_{i=1}^n$ be a length n realization of a $(d+1)$-dimensional (strictly) stationary process with $X_i = \{X_{i,1}, \cdots, X_{i,d}\}$ being $\mathbb{R}^d$ valued $(d \geq 1)$ and Y_i being real valued. In particular, X_i may consist of lagged values of Y_i, and X_i may also include some exogenous variables. We assume $\{X_i, Y_i\}_{i=1}^n$ follow the single-index model

$$Y_i = m\left(X_i^T \theta_0\right) + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{2.2.1}$$

in which $E\left(\varepsilon_i | X_i\right) = 0$, $E\left(\varepsilon_i^2 | X_i\right) = \sigma_0^2$. Without loss of generality, we assume the predictors considered in this article are standardized to have mean zero and variance one. In what follows, let $\left(\mathcal{X}^T, \mathcal{Y}, \varepsilon\right)$ represent the stationary distribution of $\left(X_i^T, Y_i, \varepsilon_i\right)$. In (2.2.1), the unknown parameter θ_0 is the single-index coefficient used for simple interpretation once estimated, and m is a smooth but unknown function used for further data summary. For model identifiability, we assume the Euclidean norm for θ_0, $\|\theta_0\| = 1$.

2.2.2 Estimation and Variable Selection for Single-Index Model

The dimension d of predictors can be large, and here we consider the case that d increases as the sample size n, so we write it as d_n. The goal of this article is to select a proper subset of significant variables $\{X_{i,j}, j \in s\}$, $s \subset \{1, ..., d_n\}$ while estimating $\theta_0 \in \Theta = \{(\theta_1, \cdots, \theta_{d_n}) | \sum_{j=1}^{d_n} \theta_j^2 = 1, \theta_1 > 0\}$ and m simultaneously.

For simplicity, given a fixed θ, denote $\mathcal{X}_\theta = \mathcal{X}^T\theta$, $X_{\theta,i} = X_i^T\theta$, $1 \leq i \leq n$. Let

$$m_\theta(u) = E(\mathcal{Y}|\mathcal{X}_\theta = u) = E\{m(\mathcal{X}_{\theta_0})|\mathcal{X}_\theta = u\}, \tag{2.2.2}$$

then θ_0 is the minimizer of the following population least squares criterion function

$$R(\theta) = \frac{1}{2}E\left[\{\mathcal{Y} - m_\theta(\mathcal{X}_\theta)\}^2\right]. \tag{2.2.3}$$

To select significant variables, we need some nonparametric techniques to estimate the unknown function m in (2.2.1). We consider the use of polynomial spline smoothing introduced in [52]. The appeal of polynomial splines is that they often provide good approximations of smoothing functions with a simple linear combination of spline basis; see more discussions in [56]. Let $N \in \mathbb{N}$ be the number of interior knots, and let

$$a = t_0 < t_1 < \cdots < t_N < t_{N+1} = b$$

be a knot sequence. Denote by $\Gamma^{(r-2)} = \Gamma^{(r-2)}[a, b]$ the polynomial spline space of order r on $[a, b]$, i.e., the space of all $\Gamma^{(r-2)}[a, b]$ functions that are polynomials of degree $r - 1$ on each interval $[t_k, t_{k+1})$, $k = 0, \cdots, N$. For any given θ, the polynomial spline estimator of order r for m_θ can be obtained from solving the least squares problem over

$\Gamma^{(r-2)}[a,b]$:

$$\hat{m}_{\theta}(u) = \arg \min_{m(u)\in\Gamma^{(r-2)}} \sum_{i=1}^{n} \{Y_i - m(X_{\theta,i})\}^2 . \tag{2.2.4}$$

Note that Θ is not a compact set, so we consider the minimization problem of (2.2.3) over all $\theta \in \Theta_c$, where

$$\Theta_c = \left\{ (\theta_1, ..., \theta_{d_n}) \mid \sum_{j=1}^{d_n} \theta_j^2 = 1, \theta_1 \geq c \right\},\ c \in (0,1).$$

We define the empirical least squares criterion function of θ as

$$\hat{R}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \{Y_i - \hat{m}_{\theta}(X_{\theta,i})\}^2 .$$

In practice, many variables can be introduced to reduce possible modeling biases. To perform simultaneous selection and estimation for the single-index model, we propose minimizing the following penalized sum of squares

$$\hat{Q}(\theta) = \hat{R}(\theta) + \sum_{j=1}^{d_n} p_{\lambda_n}(|\theta_j|) I\{|\theta_j| \neq \max_{1\leq k\leq d_n}(|\theta_k|)\}, \tag{2.2.5}$$

which shrinks small components of estimated functions to zero. Note that the above minimization in (2.2.5) is for all $\theta \in \Theta_c$, so we don't penalize the largest element of θ.

[12] proposed a continuous differentiable penalty function called SCAD penalty, which is defined in terms of its first derivative by

$$p'_{\lambda}(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_{+}}{(a-1)\lambda} I(\theta > \lambda) \right\}$$

for some $a > 2$ and $\theta > 0$. In this article, we consider the SCAD penalty, and use $a = 3.7$ as suggested in [15].

The penalized estimator of the single-index coefficient θ_0 is then defined as follows:

$$\hat{\theta} = \arg \min_{\boldsymbol{\theta} \in \Theta_c} \hat{Q}\left(\theta\right),$$

and the polynomial spline estimator of order r for m is $\hat{m}_{\boldsymbol{\theta}}$ with θ replaced by $\hat{\theta}$, i.e.

$$\hat{m}_{\boldsymbol{\theta}}\left(\cdot\right) = \arg \min_{m(\cdot) \in \Gamma^{(r-2)}[0,1]} \sum_{i=1}^{n} \left\{ Y_i - m(X_{\hat{\theta},i}) \right\}^2.$$

2.3 Main Results

In this section, we establish the asymptotic properties of the estimators for the penalized single-index model in the following theorems. We state only the main results here. The regularity conditions and proofs are given in Section 2.7.

Note that one can always arrange the predictors, $X_{i,1}, \cdots, X_{i,d_n}$, in a non-increasing order of $|\theta_{0,1}|, \cdots, |\theta_{0,d_n}|$. Without loss of generality, we assume θ_0 belongs to a compact set

$$\widetilde{\Theta}_c = \left\{ (\theta_1, ..., \theta_{d_n}) \,|\, \sum_{j=1}^{d_n} \theta_j^2 = 1, |\theta_1| \geq |\theta_2| \geq \cdots \geq |\theta_{d_n}|, \theta_1 \geq c \right\}, \; c \in (0,1).$$

For $\theta_0 \in \widetilde{\Theta}_c$, let s_n be the number of non-zero components of θ_0. Write $\theta_0 = (\theta_{0,1}, \cdots, \theta_{0,d_n})^T$ $= (\theta_{01}^T, \theta_{02}^T)^T$, where θ_{01} consists of all s_n non-zero components of θ_0, and $\theta_{02} \equiv 0$. Further we denote $\theta_{01}^* = (\theta_{0,2}, \cdots, \theta_{0,s_n})^T$. Similarly, we define θ^*, $\hat{\theta}^*$ and θ_0^* as the regular θ vectors, but without the first element.

Note that for fixed $\theta \in \Theta_c$, the least squares criterion function $R(\theta)$ depends only on θ^*, so in the following, with a slight abuse of notation, we use $R(\theta^*)$ and $\hat{R}(\theta^*)$ instead of $R(\theta)$ and $\hat{R}(\theta)$. Similarly, we write $Q(\theta^*)$ and $\hat{Q}(\theta^*)$ rather than $Q(\theta)$ and $\hat{Q}(\theta)$ respectively.

The first theorem provides the existence and consistency of the penalized estimator when d_n diverges.

Theorem 2.1. *(Existence of penalized local minimizer). Suppose Conditions (A1)-(A7) and (P2)-(P4) in Section 2.7 are satisfied. If the number of predictors $d_n = n^\delta$ for some $0 < \delta < 1/4(1 - 3/(2r+1))$ $(r > 1)$, then there is a local minimizer $\hat{\theta}^*$ of $\hat{Q}(\theta^*)$ such that $\|\hat{\theta}^* - \theta_0^*\| = O_P\{d_n^{1/2}(n^{-1/2}N^{3/2}\log(n) + a_n)\}$, where $a_n = \max_{1 \leq j \leq s_n - 1}\{p'_{\lambda_n}(|\theta_{0,j}^*|), \theta_{0,j}^* \neq 0\}$.*

Remark 1. *Note that [19] assume that $d_n = n^\delta$ $(\delta < 1/4)$ for linear regression models with independent data, and our condition of d_n for single-index models with weakly dependent data is in parallel with their requirement. Here d_n depends on both the sample* *size and the smoothness assumption of the true link function m. If we assume that m is infinitely differentiable or smooth, i.e., m has infinitely many derivatives, then only $\delta < 1/4$ is required.*

Let the Score function, $S()$, and Hessian, $\mathbf{H}()$, be defined as:

$$S\left(\theta^{*}\right) = \frac{\partial}{\partial \theta^{*}} R\left(\theta^{*}\right), \mathbf{H}\left(\theta^{*}\right) = \frac{\partial^{2}}{\partial \theta^{*} \partial \theta^{*T}} R\left(\theta^{*}\right),$$

and denote S, $\mathbf{H}$ and $\dot{m}_j$ as the values of $S\left(\theta^{*}\right)$, $\mathbf{H}\left(\theta^{*}\right)$ and $\frac{\partial}{\partial \theta_j} m_\theta$ evaluated at $\theta^{*} = \theta_0^{*}$. We further define

$$\mathbf{\Sigma}_{\lambda_n} = \text{diag}\{p''_{\lambda_n}(|\theta^{*}_{0,1}|), \cdots, p''_{\lambda_n}(|\theta^{*}_{0,s_n-1}|)\}$$

and

$$b_n = \{p'_{\lambda_n}(|\theta^{*}_{0,1}|)\text{sgn}(\theta^{*}_{0,1}), \cdots, p'_{\lambda_n}(|\theta^{*}_{0,s_n-1}|)\text{sgn}(\theta^{*}_{0,s_n-1})\}.$$

Theorem 2.2 below shows that the "oracle" property holds for the penalized estimator when d_n diverges.

Theorem 2.2. *Assume Assumptions (A1)-(A8) and (P1)-(P4) in Section 2.7 are satisfied. If $d_n = n^{\delta}$ for some $0 < \delta < 1/5(1 - 3/(r-1))$ $(r > 4)$, $\lambda_n \to 0$ and $d_n^{-1/2} n^{1/2} N^{-3/2} \lambda_n \to \infty$, then, with probability tending to 1, the consistent local minimizer*

$\hat{\theta} = \left\{\left(1 - \|\hat{\theta}^{*}_1\|^2 - \|\hat{\theta}_2\|^2\right)^{1/2}, \hat{\theta}_1^{*T}, \hat{\theta}_2^{T}\right\}^{T}$ *in Theorem 2.1 must satisfy:*

1. *(Sparsity)* $\hat{\theta}_2 = 0$.

2. (Asymptotic normality) Let $\mathbf{A}_n$ be a $q \times (s_n - 1)$ matrix such that $\mathbf{A}_n\mathbf{A}_n^T$ converges to a nonnegative symmetric $q \times q$ matrix $\mathbf{\Sigma}_A$. Then

$$\sqrt{n}\mathbf{A}_n\mathbf{\Omega}^{-1/2}\left\{(\hat{\theta}_1^* - \theta_{01}^*) + (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1} b_n\right\} \to N\left(0, \mathbf{\Sigma}_A\right)$$

in distribution, where $\mathbf{\Omega} = (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\mathbf{\Psi}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}$, $\mathbf{\Psi} = \left\{\psi_{jk}\right\}_{j,k=1}^{s_n-1}$ with

$$\psi_{jk} = \sum_{i=-\infty}^{\infty} E\left[\left\{\left(\dot{m}_j - \theta_{0,j}^*\theta_{0,1}^{-1}\dot{m}_1\right)\left(X_{\theta_0,1}\right)\left(\dot{m}_k - \theta_{0,k}^*\theta_{0,1}^{-1}\dot{m}_1\right)\left(X_{\theta_0,1+i}\right)\right\}\varepsilon_1\varepsilon_{1+i}\right]$$

for any $j, k = 1, \cdots, s_n - 1$.

Remark 2. *When $\{X_i, Y_i\}_{i=1}^n$ are i.i.d.,*

$$\psi_{jk} = E\left[\left\{\left(\dot{m}_j - \theta_{0,j}^*\theta_{0,1}^{-1}\dot{m}_1\right)\left(\dot{m}_k - \theta_{0,k}^*\theta_{0,1}^{-1}\dot{m}_1\right)\right\}\left(X_{\theta_0}\right)\varepsilon_i^2\right]$$

for any $j, k = 1, \cdots, s_n - 1$.

Remark 3. *Our condition on the number of index variables $d_n = n^\delta$ for some $0 < \delta < 1/5(1 - 3/(r-1))$ is an analog to the assumption in [19], in which they require $\delta < 1/5$ for linear regression models when the observations are independent. We require $0 < \delta < 1/5(1 - 3/(r-1))$ terms because we need to consider the smoothness of the true link function and the approximation power of polynomial splines.*

The results in Theorems 2.1, 2.2 and Lemma 3.1 in Section 2.7 lead to the following Corollary.

Corollary 2.1. *Assume Assumptions (A1)-(A7) and (P2)-(P4) in the Section 2.7 are satisfied. If* $d_n = n^{\delta}$ *for some* $0 < \delta < 1/5(1 - 3/(r-1))$*, then*

$$\|\hat{m} - m\|_{\infty} = O\{n^{-1/2} N^{1/2} \log(n) + N^{-r}\}.$$

2.4 Algorithm

In this section, we develop the estimation algorithm for the single-index coefficient as well as the unknown link function in (2.2.1). As discussed in Section 2.2, polynomial spline approximations are used to estimate the unknown functions. Let $B_{k,r}(u)$, $k = 1-r, ..., N$, be the spline basis functions of order r, $u \in [a, b]$. For any given θ, the polynomial spline estimator $\hat{m}_{\theta}$ in (2.2.4) can be obtained via

$$\hat{m}_{\theta}(u) = B_r(u) \left(\mathbf{B}_{\boldsymbol{\theta}}^T \mathbf{B}_{\boldsymbol{\theta}}\right)^{-1} \mathbf{B}_{\boldsymbol{\theta}}^T Y. \tag{2.4.1}$$

where $Y = (Y_1, \ldots, Y_n)^T$, $B_r(u) = \{B_{k,r}(u)\}_{k=1-r}^{N}$ and $\mathbf{B}_{\boldsymbol{\theta}} = \{B_{k,r}(X_{\boldsymbol{\theta},i})\}_{i=1,k=-(r-1)}^{n,\ N}$ for any fixed θ.

For any $\nu \le r-2$, $k = 1-r, ..., N$, let $B_{k,r}^{(\nu)}(u)$ be the ν-th order derivative of $B_{k,r}(u)$ with respect to u, and let $B_r^{(\nu)}(u) = \{B_{k,r}^{(\nu)}(u)\}_{k=1-r}^{N}$. According to B-spline property in

[8], $B_r^{(\nu)}(u) = \mathbf{D}_{(\nu)}^T B_{r-\nu}(u)$, where $\mathbf{D}_{(\nu)} = \mathbf{D}_1 \cdots \mathbf{D}_{\nu-1}\mathbf{D}_\nu$, with matrix

$$\mathbf{D}_l = (r-l)\begin{pmatrix} \frac{-1}{t_1 - t_{1-r+l}} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{t_1 - t_{1-r+l}} & \frac{-1}{t_2 - t_{2-r+l}} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{t_2 - t_{2-r+l}} & \frac{-1}{t_3 - t_{3-r+l}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{t_{N+r-l} - t_N} \end{pmatrix}, \ 1 \le l \le \nu.$$

Next, we denote two $n \times (N+r)$ matrices $\dot{\mathbf{B}}_j = \left\{B_r^{(1)}(X_{\boldsymbol{\theta},i}) X_{i,j}\right\}_{i=1}^n$ and $\ddot{\mathbf{B}}_{jj'} = \left\{B_r^{(2)}(X_{\boldsymbol{\theta},i}) X_{i,j} X_{i,j'}\right\}_{i=1}^n$. For any fixed θ, let $\mathbf{P}_{\boldsymbol{\theta}} = \mathbf{B}_{\boldsymbol{\theta}}\left(\mathbf{B}_{\boldsymbol{\theta}}^T \mathbf{B}_{\boldsymbol{\theta}}\right)^{-1} \mathbf{B}_{\boldsymbol{\theta}}^T$ be the projection matrix onto the polynomial spline space $\Gamma_n^{(r-2)}$. For any $j, j' = 1, ..., d_n$, let $\dot{\mathbf{P}}_j$ and $\ddot{\mathbf{P}}_{jj'}$ be the first and second order partial derivatives of $\mathbf{P}_{\boldsymbol{\theta}}$ with respect to θ_j and $\theta_{j'}$. Simple algebra shows that

$$\begin{aligned}
\dot{\mathbf{P}}_j &= (\mathbf{I} - \mathbf{P}_\theta)\dot{\mathbf{B}}_j \left(\mathbf{B}_\theta^T \mathbf{B}_\theta\right)^{-1} \mathbf{B}_\theta^T, \\
\ddot{\mathbf{P}}_{j,j'} &= (\mathbf{I} - \mathbf{P}_\theta)\{\ddot{\mathbf{B}}_{j,j'} - \dot{\mathbf{B}}_j \left(\mathbf{B}_\theta^T \mathbf{B}_\theta\right)^{-1} \mathbf{B}_\theta^T \dot{\mathbf{B}}_{j'}\} \left(\mathbf{B}_\theta^T \mathbf{B}_\theta\right)^{-1} \mathbf{B}_\theta^T \\
&\qquad +\{(\mathbf{I} - \mathbf{P}_{\boldsymbol{\theta}})\dot{\mathbf{B}}_j - \mathbf{B}_\theta(\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \dot{\mathbf{B}}_j \mathbf{B}_\theta\} \left(\mathbf{B}_\theta^T \mathbf{B}_\theta\right)^{-1} \dot{\mathbf{B}}_{j'}^T (\mathbf{I} - \mathbf{P}_{\boldsymbol{\theta}}) \\
&\qquad -\mathbf{B}_\theta \left(\mathbf{B}_{\boldsymbol{\theta}}^T \mathbf{B}_\theta\right)^{-1} \dot{\mathbf{B}}_j^T (\mathbf{I} - \mathbf{P}_{\boldsymbol{\theta}})\dot{\mathbf{B}}_{j'} \left(\mathbf{B}_\theta^T \mathbf{B}_\theta\right)^{-1} \mathbf{B}_\theta^T.
\end{aligned}$$

Then, the score vector can be written as

$$\hat{S}(\theta^*) = \frac{\partial}{\partial \theta^*}\hat{R}(\theta^*) = -\frac{1}{n}\sum_{i=1}^n \hat{S}_i(\theta^*) = -\frac{1}{n}\left\{Y^T \dot{\mathbf{P}}_j Y - \theta_j \theta_1^{-1} Y^T \dot{\mathbf{P}}_1 Y\right\}_{j=2}^{d_n},$$

and the Hessian matrix is

$$\hat{\mathbf{H}}(\theta^*) = \frac{\partial^2}{\partial\theta^*\partial\theta^{*T}}\hat{R}(\theta^*) \; = \; -\frac{1}{n}\left\{Y^T\ddot{\mathbf{P}}_{j,j'}Y - \theta_1^{-1}\left(\theta_{j'}Y^T\ddot{\mathbf{P}}_{j,1}Y + \theta_j Y^T\ddot{\mathbf{P}}_{1,j'}Y\right)\right\}_{j,j'=2}^{d_n}$$
$$+\frac{1}{n}\{Y^T\dot{\mathbf{P}}_1 Y(\theta_1^{-1}\mathbf{I} + \theta_1^{-3}\theta^*\theta^{*T}) - Y^T\ddot{\mathbf{P}}_{1,1}Y(\theta_1^{-2}\theta^*\theta^{*T})\}.$$

In addition, given a tuning penalty parameter λ, we denote

$$\mathbf{\Sigma}_\lambda(\theta^*) = \text{diag}\left\{\frac{p'_\lambda(|\theta_1^*|)}{\varepsilon + |\theta_1^*|}, \cdots, \frac{p'_\lambda\left(|\theta^*_{d_n-1}|\right)}{\varepsilon + |\theta^*_{d_n-1}|}\right\}, \; \varepsilon \text{ is a small number,}$$

which is an approximation of $\mathbf{\Sigma}_{\lambda_n}$ and

$$b_\lambda(\theta^*) = \left\{p'_\lambda(|\theta_1^*|)\,\text{sgn}(\theta_1^*), \cdots, p'_\lambda\left(|\theta^*_{d_n-1}|\right)\text{sgn}\left(\theta^*_{d_n-1}\right)\right\}^T.$$

We outline our algorithm based on the local quadratic approximation [15] to solve the penalized least squares problem in (2.2.5). Note that the unpenalized estimator of [52] is still consistent if spline basis functions are appropriately chosen, thus we use it as the initial value in our estimating algorithm. To satisfy the assumption $\theta \in \Theta_c$, for small $c = 10^{-6}$, we first arrange $\tilde{\theta}_j$ and $X_{i,j}$, $j = 1, ..., d_n$, according to the non-increasing order of the absolute values of $\tilde{\theta}_j$. Then we set $\hat{\theta}^{(0)} = \text{sgn}(\tilde{\theta}_1) \times \tilde{\theta}/\|\tilde{\theta}\|$, where $\text{sgn}(\tilde{\theta}_1)$ is the sign of the first parameter in the rearranged $\tilde{\theta}$. Using this initial estimator $\hat{\theta}^{(0)}$, the algorithm iterates through the following steps.

1. $k \leftarrow k + 1$.

2. By the local quadratic approximations for penalty functions, a better approximation is given by

$$\hat{\theta}^{*(k)} = \hat{\theta}^{*(k-1)} - \left\{ \hat{\mathbf{H}}(\hat{\theta}^{*(k-1)}) + \boldsymbol{\Sigma}_\lambda(\hat{\theta}^{*(k-1)}) \right\}^{-1} \left\{ \hat{S}(\hat{\theta}^{*(k-1)}) + b_\lambda(\hat{\theta}^{*(k-1)}) \right\}.$$

3. If $\|\hat{\theta}^{*(k)}\| > \sqrt{1-c^2}$, then $\hat{\theta}^{*(k)} = \hat{\theta}^{*(k)}/\|\hat{\theta}^{*(k)}\| \times \sqrt{1-c^2}$.

4. Set the first index parameter $\hat{\theta}_1^{(k)} = \sqrt{1 - \|\hat{\theta}^{*(k)}\|^2}$.

5. If $\hat{\theta}_j^{(k)}$ is close to 0, say $|\hat{\theta}_j^{(k)}| < \delta_1$, for a small number δ_1 (for example, $\delta_1 = 10^{-3}$), then we set $\hat{\theta}_j^{(k)} = 0$. Rescale $\hat{\theta}^{(k)} = (\hat{\theta}_1^{(k)}, \hat{\theta}^{*(k)T})^T$ by $\hat{\theta}^{(k)} = \hat{\theta}^{(k)}/\|\hat{\theta}^{(k)}\|$.

6. Obtain the difference between $\hat{\theta}^{(k)}$ and $\hat{\theta}^{(k-1)}$: $\text{diff}_\theta = \|\hat{\theta}^{(k)} - \hat{\theta}^{(k-1)}\|$.

7. Arrange $\hat{\theta}^{(k)}$ and the predictors in a non-increasing order of $|\hat{\theta}^{(k)}|$ and set $\hat{\theta}^{(k)} = \text{sgn}(\hat{\theta}_1^{(k)}) \times \hat{\theta}^{(k)}$.

8. Repeat Steps 1 and 7 until we have $\text{diff}_\theta < \delta_2$, for a small number δ_2 (for example, $\delta_2 = 10^{-6}$).

The tuning parameter, λ, plays an important role in the performance of model selection. It is well known that for a fixed predictor dimension that SCAD estimator can identify the true model consistently when one chooses the tuning parameter using a BIC-type criterion. For example, [39] show that BIC can identify the true model consistently for penalized partially linear single-index models. However, as shown in [48], the traditional BIC does not work very well for diverging number of parameters because the number of candidate models increases rapidly and can easily exceed the sample size. To overcome this challenge, in this article, we adopt the modified BIC approach proposed by [48] to select the tuning parameter. Such modified BIC has been proved to be consistent in model selection, even with a diverging number of parameters.

Let $\hat{\theta}_\lambda$ and d_λ be the estimator and the effective number of parameters respectively in the last iteration of the our algorithm above. Then the modified BIC can be defined as

$$\mathrm{BIC}\,(\lambda) = \log\left\{\hat{R}(\hat{\theta}_\lambda)\right\} + d_\lambda n^{-1}\log(n)C_n.$$

In our simulations and application, we choose C_n to be $\log\{\log(d_n)\}$, as suggested in [48].

The spline approximation for the regression function requires an appropriate selection of the knot sequences. For the ease of computation, we consider equally spaced knots after conducting the transformation introduced the above. Note that Assumption (A.5) requires $n^{1/\{2(r-1)\}} \ll N \ll \min\{n^{1/6}\log^{-2/3}(n)d_n^{-5/6}, n^{1/8}\log^{-1/2}(n)d_n^{-3/8}\}$ for some integer $r > 5$. Therefore, we choose $r = 6$ for the simulations and real data application in the article. In our numerical studies, we find that the variable selection result is less sensitive to the choice N compared with the function estimation result, so we suggest the following simple formula to compute the number of interior knots:

$$N = [\tau n^{1/\{2(r-1)\}}\log n],$$

for some positive tuning parameter τ. For example, $\tau \in [0.5, 1]$ usually works very well, and in our simulations and application below, we choose $\tau = 0.8$.

The standard errors for the estimated parameters can be obtained as follows. Given a PSIM estimator $\hat{\theta}^*$, a good estimator of Ψ is given by

$$\hat{\Psi} = \sum_{1 \le i, i' \le n} (n - |i - i'|)^{-1} \hat{S}_i(\hat{\theta}^*) \hat{S}_{i'}^T(\hat{\theta}^*). \tag{2.4.2}$$

When $\{X_i, Y_i\}_{i=1}^n$ are i.i.d., the above estimator can be reduced to

$$\hat{\Psi} = \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i(\hat{\theta}^*) \hat{S}_i^T(\hat{\theta}^*).$$

Our limited simulation results indicate that this variance estimator performs very well.

2.5 Simulations

In this section, three simulation studies are carried out to illustrate the finite-sample behavior of our estimation and variable selection method for the single-index models. All the codes for these simulations are written in R and the computing environment is x64 PC with Intel Dual Core i5.

2.5.1 Example 1

In this example, we compare our method (PSIM) with the penalized least squares (PLS) method [41] and the penalized slice inverse regression (PSIR) method [62]. We consider the following single-index model

$$Y_i = \sin\left(\frac{\pi}{4} X_i^T \theta_0\right) + \sigma_0 \varepsilon_i. \tag{2.5.1}$$

Here, $X_i = (X_{i,1}, \cdots, X_{i,d})^T$, ε_i's are independently and identically distributed as $N(0,1)$, for all $i = 1, \cdots, n$ and $\sigma_0 = 0.2$. In this simulation, the true parameter is $\theta_0^T = (1,1,1,1,1,0,\cdots,0)/\sqrt{5}$, i.e., the first five elements of θ_0 are $1/\sqrt{5}$ and the remaining $d-5$ elements are zero. We consider the selection and estimation for model (2.5.1) with $d = 25$, 50, or 100 which is smaller than or close to the sample size used in this example. We draw samples of size $n = 100$ and $n = 200$ and implement 500 Monte Carlo experiments.

We consider SCAD penalty for all three methods. We use the five-fold generalized cross validation to choose the tuning parameter for PLS and PSIR, as suggested in both [41] and [62]. The results are summarized in Table 2.1. The column labeled "TPN" presents the average number of zero coefficients estimated from among the $d-5$ true zero coefficients, "FPN" shows the average number of 5 non-zero coefficients erroneously set to zero, and "C" demonstrates the percentage of runs for which the correct model has been chosen. The "oracle" (ORACLE) method always identifies the five non-zero coefficients and $d-5$ zero coefficients correctly. The medians of model errors (MMEs),

$(\hat{\theta}-\theta_0)^T E(X^T X)(\hat{\theta}-\theta_0)$, of the "oracle" estimators and our penalized estimators (PSIM) are used to measure the effectiveness of the methods. In addition, Table 2.1 also provides the average computing time ("TIME") in seconds and the average number of iterations ("ITER") of our PSIM method. Table 2.2 presents the bias (BIAS), standard error (SD) and the mean squared error (MSE) of the estimates of θ_0. In terms of the computing time, PSIR method is the fastest, followed by our PSIM method, and the slowest is the PLS method. However, in terms of the accuracy of selection and estimation, the behavior of our PSIM method is the closest to that of the "oracle". Table 2.2 lists the bias and the mse of various estimators for the five non-zero index parameters. From the tables, one can see that regardless of sample size and the dimension of the parameter, the PSIM estimator is superior to the PLS and PSIR estimators.

[Tables 2.1 and 2.2 about here]

We now test the accuracy of the standard error formula in (2.4.2) for the PSIM estimators. Table 2.3 presents the results for the first five coefficients. Similar to [19], the standard deviations of the estimated index parameters are computed over 500 simulations. These can be regarded as the true standard errors (column labeled "SD") and compared with the median of the 500 estimated standard errors calculated using (2.4.2) (column labeled "SD$_m$"). The column labeled "SD_{mad}" is interquartile range of the 500 estimated standard errors divided by 1.349, which is a robust estimate of the standard deviation. When n is small and d is large, the variances are a little under-estimated, but the estimation becomes better as we increase sample size. For example, when $d = 25$, the estimated standard error based on sample size $n = 200$ is very accurate.

[Table 2.3 about here]

2.5.2 Example 2

We use another simulation study to augment our theoretical results on time series. To make a fair comparison, we use a similar model to [62] but in a time series setting. Specifically, we consider the following nonlinear autogressive (NAR) model:

$$X_i = 2\sin(\theta_1 X_{i-1} + \theta_2 X_{i-2} + \cdots + \theta_{d_n} X_{i-d_n}) + \varepsilon_i,\ i = 1, 2, \cdots, n, \tag{2.5.2}$$

where $\theta_1 = 11/4,\ \theta_2 = -23/6,\ \theta_3 = 37/12,\ \theta_4 = -13/9$ and $\theta_5 = 4/3$, so the standardized $\theta_0 = (0.461, -0.642, 0.517, -0.242, 0.223, 0, \cdots, 0)^T$. The ε_i's are white noise with $\sigma_0 = 0.5$. In time series modeling, we often must explore many models with various sets of lagged values to reduce possible modeling biases, so the number of predictors usually depends on n. In our simulation, the dimension is calculated by $d_n = [4n^{1/4}] - 5$ which is also used in both [19] and [62].

We generate 500 Monte Carlo time series of length 100, 200, 400 and 800 from model (2.5.2). In each replication, the first 1000 observations are discarded to make the time series $\{X_i\}_{i=1}^n$ behave like a stationary time series. Figure 2.1 is one typical plot of a simulated time series of length 800 ($d_n = 16$), which shows an evident stationary feature.

Tables 2.4 and 2.5 present the selection and estimation results of various methods: PSIM, PLS and PSIR. From the table, one sees that the comparison is even more favorable to our PSIM method. The PSIM method performs significantly better than the PLS and PSIR regardless of the dimension and sample size. The models selected by the PSIM is very close to the true model, and the differences between the MMEs of the PSIM and "oracle" are small. Note that the PSIR proposed by [62] is not very suitable for time series data, so it is not surprising that the PSIR does not perform well in this example.

[Figure 2.1 and Tables 2.4 to 2.5 about here]

We now investigate the performance of the variance estimators of the PSIM estimators. Similar to Example 2, we give the SD, the SD$_m$, and the SD$_{mad}$ of the PSIM estimators; see Table 2.6. These numerical results suggest that the proposed estimator in (2.4.2) yields very reasonable standard error estimates.

[Table 2.6 about here]

2.5.3 Example 3

In Examples 1 and 2, the underlying models that the data are generated from are genuine single-index models. In this example, we want to examine the behavior of our proposed method when the model is misspecified. We implement the proposed methods under two different scenarios: one with a genuine single-index function and one without. We consider a similar example to Example 1 in [52], and let

$$Y_i = m(X_i) + \sigma_0 \varepsilon_i,\ i = 1, ..., n,$$

$$m(x) = \sum_{j=1}^{5} x_j + \exp\left\{-\left(\sum_{j=1}^{5} x_j\right)^2\right\} + \delta\left(\sum_{j=1}^{5} x_j^2\right)^{1/2},$$

where X_i's are generated from a d-variate standard normal distribution, ε_i's are generated from $N(0,1)$, and $\sigma_0 = 0.5$. When $\delta = 0$, the underlying true function m can be written as

$$m(x) = \sqrt{5}x^T\theta_0 + \exp\{-5(x^T\theta_0)^2\},$$

where $\theta_0^T = (1,1,1,1,1,0,\cdots,0)/\sqrt{5}$. It is obvious that m is a genuine single-index in this case. In contrast, if $\delta \neq 0$, m is not a single-index function.

For both $\delta = 0$ and $\delta = 1$, we draw 500 random samples of size $n = 100,\ 200$ with number of predictors $d = 25,\ 50,\ 100$. The variable selection and estimation results are summarized in Tables 2.7 and 2.8, respectively. The columns labeled as “TPN”, “FPN”, “C”, “MME”, “TIME” and “ITER” are similarly defined in Table 2.1. From Tables 2.7 and 2.8, one sees that the proposed method still works very well when the underlying regression function is not a single-index function.

[Tables 2.7 and 2.8 about here]

2.6 Application

In this section, we adopt the proposed PSIM method to the river flow data of Jökulsá Eystri of Iceland [47]. The dataset contains the daily river flow, temperature and precipitation observations collected from January 1, 1972 to December 31, 1974. The response variable in this analysis is the daily river flow $\{Y_t\}_{t=1}^{1096}$, measured in meter cubed per second of Jökulsá Eystri River. There are two exogenous variables: temperature $\{X_t\}_{t=1}^{1096}$ in degrees Celsius and daily precipitation $\{Z_t\}_{t=1}^{1096}$ in millimeters collected at the meteorological station at Hveravellir. See the time series plots in [52].

[52] used the SIP model to forecast the river flow series and discussed the advantages of SIP over the linear regression model (LM). In our analysis, we are more interested in finding significant predictors that help to forecast the river flow $\{Y_t\}$. We pre-select all the lagged values in the past seven days (one week), i.e., the predictors are these 23 variables: $Y_{t-1}, \cdots, Y_{t-7}, X_t, X_{t-1}, \cdots, X_{t-7}, Z_t, Z_{t-1}, \cdots, Z_{t-7}$. Following [52], we remove the trend by a simple quadratic spline regression and work on the residual series. All three residual series pass the unit-root test, so we treat them as stationary time series. We then apply the PSIM method with the SCAD penalty to select significant predictors and estimate the the index parameters. We compare our PSIM method with the BIC method (BIC-SIP) proposed in [52].

Table 2.9 lists the variable selection and estimation results for both methods. The PSIM method selects the following seven explanatory variables: Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, X_t, Z_t

and Z_{t-1}. The BIC-SIP selects nine variables, and seven variables are common to both methods.

[Table 2.9 about here]

In order to evaluate the prediction performance of different methods, we use the observations of the first two years to fit the model and compute the out sample forecast error over the last year:

$$\text{MSPE} = \left\{ \frac{1}{365} \sum_{t=732}^{1096} (Y_t - \hat{Y}_t)^2 \right\}^{1/2}.$$

We show in Table 2.8, the MSPEs for PSIM, BIC-SIP, BIC based linear regression model ("BIC-LM") and the full SIP model (FULL-SIP) with all the lagged values in the last seven days. In terms of the MSPEs from Table 2.10, our PSIM produces the best forecast among all these methods. In addition, Figure 2.2 shows the estimated nonparametric function for river flow based on the single-index model. The estimated function and plotted points are for the two years (1972-1973) used in the training set.

[Table 2.10 and Figure 2.2 about here]

2.7 Proof of Theorems

2.7.1 Assumptions

We state our assumptions below.

(A1) *The least squares criterion function R is locally convex at θ_0^*, i.e., for any $\varepsilon > 0$, there exists $\delta > 0$ such that $R(\theta^*) - R(\theta_0^*) < \delta$ implies $\|\theta^* - \theta_0^*\| < \varepsilon$. The Hessian matrix $\mathbf{H}(\theta_0^*)$ defined in Section 2.3 is positive definite and its eigenvalues are bounded below and above from ∞.*

(A2) For any $\theta_1, \theta_2 \in \Theta_c$, the joint density function $f_{\theta_1,\theta_2}(x_{\theta_1}, x_{\theta_2})$ of X_{θ_1} and X_{θ_2} has r-th order ($r > 5$) continuous partial derivatives and is bounded below and above on $[a, b]^2$. The marginal density function of X_θ, $f_\theta(x_\theta) \in \Gamma^{(1)}[a, b]$ and is bounded below, for any $\theta \in \Theta_c$.

(A3) The true link function $m \in \Gamma^{(r)}[a, b]$ for some $r > 5$.

(A4) There exist positive constants K_0 and λ_0 such that $\alpha(n) \leq K_0 e^{-\lambda_0 n}$ holds for all n, with the α-mixing coefficient for $\left\{Z_i = \left(X_i^T, \varepsilon_i\right)\right\}_{i=1}^n$ defined as

$$\alpha(k) = \sup_{B \in \sigma\{Z_s, s \leq t\}, C \in \sigma\{Z_s, s \geq t+k\}} |P(B \cap C) - P(B) P(C)|, \ k \geq 1.$$

(A5) The number of interior knots N satisfies:

$$n^{1/\{2(r-1)\}} \ll N \ll \min\{n^{1/6} \log^{-2/3}(n) d_n^{-5/6}, n^{1/8} \log^{-1/2}(n) d_n^{-3/8}\}.$$

(A6) There is a large enough open subset ω_n of $\widetilde{\Theta}_c$ which contains the true parameter point θ_0, such that for all $\theta \in \omega_n$ *and* $j, k, l = 2, \cdots, d_n$, the third order derivative satisfies

$$\left| E\left\{ \frac{\partial^3 R(\theta)}{\partial\theta_j \partial\theta_k \partial\theta_l} \right\} \right| < C_3 < \infty. \tag{2.7.1}$$

(A7) Let the values of $\theta_{0,1}$, $\theta_{0,2}$, ..., θ_{0,s_n} be nonzero, θ_{0,s_n+1}, θ_{0,s_n+2}, ..., θ_{0,d_n} be zero, and $\theta_{0,1}$, $\theta_{0,2}$, ..., θ_{0,s_n} satisfy $\min_{1 \le j \le s_n} \theta_{0,s_n}/\lambda_n \to \infty$ as $n \to \infty$.

Remark 4. *Assumption (A1)-(A3) are also assumed in [52]. For Assumptions (A2) and (A3), [52] requires only $r = 4$. In our article, we consider diverging number of parameters, which requires the investigation of the third order derivative of $R(\theta)$ in order to derive the "oracle" properties. Therefore, we need to require higher order smoothness of the underlying regression function. Assumption (A4) is suitable to model time series data. [42] shows that a geometrically ergodic time series is a strongly mixing sequence. Assumption (A5) gives the requirement for the number of interior knots, which depends not only on the smoothness of the underlying regression function but also on the growing rate of the dimension of covariates. If d_n is finite, then we have $n^{1/\{2(r-1)\}} \ll N \ll n^{1/8} \log^{-1/2}(n)$. This is slightly different from the assumption in [52] because we consider higher order spline approximation ($r > 5$) rather than cubic spline approximation ($r = 4$). Assumptions (A6) and (A7) are similar to Conditions (G) and (H) in [19].*

(P1) $\liminf_{n \to +\infty} \liminf_{\theta \to 0_+} p'_{\lambda_n}(\theta)/\lambda_n > 0$.

(P2) $a_n = \max_{2 \le j \le d_n} \{p'_{\lambda_n}(|\theta_{0j}|), \theta_{0j} \neq 0\} = O\{d_n^{1/2} n^{-1/2} N^{3/2} \log(n)\}$.

(P3) $u_n = \max_{2 \le j \le d_n} \{p''_{\lambda_n}(|\theta_{0j}|), \theta_{0j} \neq 0\} \to 0$ *as* $n \to +\infty$.

(P4) *There exists constants C_1 and C_2 such that, when $\theta_1, \theta_2 > C_1\lambda_n$, $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \leq C_2|\theta_1 - \theta_2|$.*

Remark 5. *Conditions (P1), (P3) and (P4) are also assumed in [19]. Assumption (P2) ensures the unbiasedness property for large parameters and the existence of a consistent penalized estimator.*

2.7.2 Preliminary Results

Before we prove all the theorems, we first state several lemmas.

Lemma 2.1 (See page 149 of [8])**.** *There is a positive constant C_r such that for every $m \in \Gamma^{(r)}[a,b]$, there exists a function $g \in \Gamma^{(r-2)}[a,b]$ that satisfies $\|g - m\|_\infty \leq C_r \left\|m^{(r)}\right\|_\infty N^{-r}$.*

According to Theorem 7.7.4 in [9], the following lemma holds.

Lemma 2.2. *There exists a constant $C > 0$, such that for $0 \leq k \leq 2$ and $m \in \Gamma^{(r)}[a,b]$*

$$\left\|(m - Q_{T,r}(m))^{(k)}\right\|_\infty \leq C \left\|m^{(r)}\right\|_\infty N^{-(r-k)},$$

where $Q_{T,r}(m)$ is the r-th order quasi-interpolant of m corresponding to a sequence of knots T; see the definition of $Q_{T,r}$ on Page 146 of [9].

The following lemma gives the uniform convergence rate of the r-th order polynomial spline estimator $\hat{m}_\theta$ in (2.2.4) to m_θ in (2.2.2) as well as its derivative approximation rate.

Lemma 2.3. *Under Assumptions (A2)-(A4), we have that*

$$\sup_{\theta \in \Theta_c} \left\| \hat{m}_{\theta}^{(k)} - m_{\boldsymbol{\theta}}^{(k)} \right\|_{\infty} = O_P \left\{ n^{-1/2} N^{1/2+k} \log(n) + N^{-(r-k)} \right\}, \tag{2.7.2}$$

for any $k = 0, \ldots, r-2$.

Proof of Lemma 2.3 is the same as the proof of Proposition A.1 in [52] where we replace the approximation rate of cubic spline smoothing by the more general polynomial spline approximation results given in Lemmas 2.1 and 2.2, and is thus omitted.

Lemma 2.4. *Under Assumptions (A1)-(A4), we have*

$$\begin{aligned}
\sup_{\theta \in \Theta_c} \max_{1 \le j \le d_n - 1} \left| \frac{\partial}{\partial \theta_j^*} \{ \hat{R}(\theta^*) - R(\theta^*) \} \right| &= O_P \left\{ n^{-1/2} N^{3/2} \log(n) + N^{-(r-1)} \right\}, \\
\sup_{\theta \in \Theta_c} \max_{1 \le j,k \le d_n - 1} \left| \frac{\partial^2}{\partial \theta_j^* \partial \theta_k^*} \{ \hat{R}(\theta^*) - R(\theta^*) \} \right| &= O_P \left\{ n^{-1/2} N^{5/2} \log(n) + N^{-(r-2)} \right\}, \\
\sup_{\theta \in \Theta_c} \max_{1 \le j,k,l \le d_n - 1} \left| \frac{\partial^3}{\partial \theta_j^* \partial \theta_k^* \partial \theta_l^*} \{ \hat{R}(\theta^*) - R(\theta^*) \} \right| &= O_P \left\{ n^{-1/2} N^{7/2} \log(n) + N^{-(r-3)} \right\}
\end{aligned}$$

Proof of Lemma 2.4 is the same as the proof of Lemma A.15 in [52] replacing the approximation rate of cubic spline smoothing by the more general polynomial spline approximation results, and is thus omitted.

2.7.3 Proof of Theorem 2.1

Proof of Theorem 2.1. Let $\alpha_n = d_n^{1/2} n^{-1/2} N^{3/2} \log(n)$ and set $\|u\| = C$, where C is a large enough constant. To show the existence of such penalized local minimizer, it is equivalent to prove that for any given ε there is a large constant C such that, for large n we have

$$P\left\{\inf_{\|u\|=C} \hat{Q}(\theta_0^* + \alpha_n u) > \hat{Q}(\theta_0^*)\right\} \geq 1 - \varepsilon.$$

This implies that with probability tending to 1 there is a local minimizer $\hat{\theta}^*$ in the ball $\{\theta_0^* + \alpha_n u : \|u\| \leq C\}$ such that $\|\hat{\theta}^* - \theta_0^*\| = O_P(\alpha_n)$.

Using $p_{\lambda_n}(0) = 0$, we have

$$\begin{aligned} D(u) &= \hat{Q}(\theta_0^* + \alpha_n u) - \hat{Q}(\theta_0^*) \\ &\geq \left\{\hat{R}(\theta_0^* + \alpha_n u) - \hat{R}(\theta_0^*)\right\} + \sum_{j=1}^{s_n - 1} \left\{p_{\lambda_n}\left(\left|\theta_{0,j}^* + \alpha_n u_j\right|\right) - p_{\lambda_n}\left(\left|\theta_{0,j}^*\right|\right)\right\} \\ &= D_1(u) + D_2(u), \end{aligned}$$

where s_n is the number of parameters for which the true values are not 0. Then, by Taylor's expansion, we obtain

$$\begin{aligned} D_1(u) &= \hat{R}(\theta_0^* + \alpha_n u) - \hat{R}(\theta_0^*) \\ &= \alpha_n \left\{\frac{\partial}{\partial \theta^*} \hat{R}(\theta_0^*)\right\} u + \frac{1}{2} \alpha_n^2 u^T \left\{\frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \hat{R}(\theta_0^*)\right\} u \\ &\quad + \frac{1}{6} \alpha_n^3 \frac{\partial}{\partial \theta^*} \left[u^T \left\{\frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \hat{R}(\bar{\theta})\right\} u\right] u \\ &= \alpha_n \hat{S}(\theta_0^*) u + \frac{1}{2} \alpha_n^2 u^T \hat{\mathbf{H}}(\theta_0^*) u + \frac{1}{6} \alpha_n^3 \frac{\partial}{\partial \theta^*} \left[u^T \left\{\frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \hat{R}(\bar{\theta})\right\} u\right] u \\ &= D_{11}(u) + D_{12}(u) + D_{13}(u), \end{aligned}$$

where the vector $\overline{\theta}$ lies between θ_0^* and $\theta_0^* + \alpha_n u$, and

$$\begin{aligned} D_2\left(u\right) &= \sum_{j=1}^{s_n-1} \left\{ p_{\lambda_n}(\theta_{0,j}^* + \alpha_n u_j|) - p_{\lambda_n}(|\theta_{0,j}^*|) \right\} \\ &= \sum_{j=1}^{s_n-1} \left[\alpha_n p'_{\lambda_n}(\theta_{0,j}^*)\mathrm{sgn}(\theta_{0,j}^*)u_j + \alpha_n^2 p''_{\lambda_n}(\theta_{0,j}^*)u_j^2\{1+o(1)\} \right] \\ &= D_{21}\left(u\right) + D_{22}\left(u\right). \end{aligned}$$

Note that $\frac{\partial}{\partial \theta^*} R\left(\theta^*\right) = 0$, by Assumptions (A2)-(A5) and Lemma 2.4 we have

$$\begin{aligned} |D_{11}| &\leq \alpha_n \left\| \frac{\partial}{\partial \theta^*} \left\{ \hat{R}\left(\theta^*\right) - R\left(\theta^*\right) \right\} \right\| \|u\| \\ &= \alpha_n \|u\| \times O_P \left\{ d_n^{1/2} n^{-1/2} N^{3/2} \log(n) + d_n^{1/2} N^{-r+1} \right\} \\ &= O_P\left(\alpha_n^2\right) \|u\|. \end{aligned} \tag{2.7.3}$$

Next, we consider D_{12},

$$\begin{aligned} D_{12} &= \frac{1}{2} u^T \left\{ \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \hat{R}(\theta_0^*) - \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} R(\theta_0^*) \right\} u \alpha_n^2 + \frac{1}{2} u^T \left\{ \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} R(\theta_0^*) \right\} u \alpha_n^2 \\ &= \frac{1}{2} u^T \left\{ \hat{\mathbf{H}}(\theta_0^*) - \mathbf{H}(\theta_0^*) \right\} u \alpha_n^2 + \frac{1}{2} u^T \mathbf{H}(\theta_0^*) u \alpha_n^2. \end{aligned}$$

According to Lemma 2.4 and Assumption (A5), we have

$$\begin{aligned} |D_{12}| &\leq \frac{1}{2} u^T \mathbf{H}(\theta_0^*) u \alpha_n^2 + O\left\{ \left(n^{-1/2} N^{5/2} \log(n) + N^{-r+2} \right) d_n \right\} \alpha_n^2 \|u\|^2 \\ &= \frac{1}{2} u^T \mathbf{H}(\theta_0^*) u \alpha_n^2 + o_P\left(1\right) \times \alpha_n^2 \|u\|^2. \end{aligned} \tag{2.7.4}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} D_{13} &= \frac{1}{6}\alpha_n^3 \frac{\partial}{\partial\theta^*}\left[u^T\left\{\frac{\partial^2}{\partial\theta^*\partial\theta^{*T}}\hat{R}(\theta^*)\right\}u\right]u \\ &\leq \frac{1}{6}\alpha_n^3 \frac{\partial}{\partial\theta^*}\left[u^T\frac{\partial^2}{\partial\theta^*\partial\theta^{*T}}\left\{\hat{R}(\theta^*)-R(\theta^*)\right\}u\right]u \\ &\quad +\frac{1}{6}\alpha_n^3 \frac{\partial}{\partial\theta^*}\left[u^T\left\{\frac{\partial^2}{\partial\theta^*\partial\theta^{*T}}R(\theta^*)\right\}u\right]u. \end{aligned}$$

Using the result in Lemma 2.4 again, together with Assumption (A5), implies that

$$\begin{aligned} |D_{13}| &\leq O_P\left(d_n^{3/2}\alpha_n\right)\alpha_n^2\|u\|^3 + O_P\left\{\left(n^{-1/2}N^{7/2}\log(n)+N^{-r+3}\right)d_n^{3/2}\alpha_n\right\}\alpha_n^2\|u\|^3 \\ &= o_P(1)\times\alpha_n^2\|u\|^2. \end{aligned} \tag{2.7.5}$$

Furthermore, by Assumptions (P2)-(P4), the terms D_{21} and D_{22} satisfy the following

$$|D_{21}| = \sum_{j=1}^{s_n-1}\left|\alpha_n p'_{\lambda_n}\left(|\theta^*_{0,j}|\right)\operatorname{sgn}\left(\theta^*_{0,j}\right)u_j\right| \leq \sqrt{s_n}\alpha_n a_n\|u\| \leq \alpha_n^2\|u\|, \tag{2.7.6}$$

and

$$|D_{22}| = \sum_{j=1}^{s_n-1}\alpha_n^2 p''_{\lambda_n}\left(\left|\theta^*_{0,j}\right|\right)u_j^2\{1+o(1)\} \leq 2\max_{1\leq j\leq s_n-1} p''_{\lambda_n}\left(\left|\theta^*_{0,j}\right|\right)\alpha_n^2\|u\|^2. \tag{2.7.7}$$

By equations (2.7.3)-(2.7.7), when $\|u\|$ is large enough, all terms D_{11}, D_{13}, D_{21} and D_{22} are dominated by a positive term D_{12}. Hence, Theorem 2.1 holds. ▮

2.7.4 Proof of Sparsity

To prove Theorem 2.2, we first show the sparsity property using Lemma 2.5.

Lemma 2.5. *Suppose Assumptions (A1)-(A7) and (P1) are satisfied. If* $d_n = n^{\delta}$ *for some* $0 < \delta < 1/5(1 - 3/(r-1))$, $\lambda_n \to 0$ *and* $\lambda_n d_n^{-1/2} n^{1/2} N^{-3/2} \log^{-1}(n) \to \infty$ *as* $n \to \infty$, *then with probability tending to 1, for any given* θ_1^* *satisfying* $\|\theta_1^* - \theta_{01}^*\| = O_P\{d_n^{1/2} n^{-1/2} N^{3/2} \log(n)\}$ *and any constant* C, *we have*

$$\hat{Q}\left\{\left(\theta_1^{*T}, 0^T\right)^T\right\} = \min_{\|\theta_1^* - \theta_{01}^*\| \leq C d_n^{1/2} n^{-1/2} N^{3/2} \log(n)} \hat{Q}\left\{\left(\theta_1^{*T}, \theta_2^T\right)^T\right\}.$$

Proof. Let $\varepsilon_n = C d_n^{1/2} n^{-1/2} N^{3/2} \log(n)$, then to prove Lemma 2.5, it is sufficient to show that with probability tending to 1, as $n \to \infty$, for any θ_1^* satisfying $\|\theta_1^* - \theta_{01}^*\| = O_P\{d_n^{1/2} n^{-1/2} N^{3/2} \log(n)\}$, we have, for any $j = s_n, \cdots, d_n - 1$

$$\frac{\partial \hat{Q}(\theta^*)}{\partial \theta_j^*} < 0, \text{ for } -\varepsilon_n < \theta_j^* < 0; \qquad (2.7.8)$$

$$\frac{\partial \hat{Q}(\theta^*)}{\partial \theta_j^*} > 0, \text{ for } 0 < \theta_j^* < \varepsilon_n. \qquad (2.7.9)$$

Using Taylor expansion, we have for any $j = s_n, \ldots, d_n$

$$\begin{aligned} K &= \frac{\partial \hat{Q}(\theta^*)}{\partial \theta_j^*} = \frac{\partial \hat{R}(\theta^*)}{\partial \theta_j^*} + p'_{\lambda_n}(|\theta_j^*|)\mathrm{sgn}(\theta_j^*) \\ &= \frac{\partial \hat{R}(\theta_0^*)}{\partial \theta_j^*} + \sum_{k=1}^{d_n-1} \frac{\partial^2 \hat{R}(\theta_0^*)}{\partial \theta_j^* \partial \theta_k^*} \left(\theta_k^* - \theta_{0,k}^*\right) + \sum_{k,l=1}^{d_n-1} \frac{\partial^3 \hat{R}(\bar{\theta})}{\partial \theta_j^* \partial \theta_k^* \partial \theta_l^*} \left(\theta_k^* - \theta_{0,k}^*\right)\left(\theta_l^* - \theta_{0,l}^*\right) \\ &\quad + p'_{\lambda_n}(|\theta_j^*|)\mathrm{sgn}(\theta_j^*) \\ &= K_1 + K_2 + K_3 + K_4, \end{aligned}$$

where $\overline{\theta}^*$ lies between θ^* and θ_0^*. Next, we consider the terms K_1, K_2 and K_3. Based on the proof of Theorem 2.1, we have

$$\begin{aligned} |K_1| &= \left| \frac{\partial}{\partial \theta_j^*} \left\{ (\hat{R} - R)(\theta_0^*) \right\} \right| + \left| \frac{\partial R(\theta_0^*)}{\partial \theta_j^*} \right| = O_P\left\{ n^{-1/2} N^{3/2} \log(n) + N^{-r+1} \right\} \\ &= o_P\left\{ d_n^{1/2} \left(n^{-1/2} N^{3/2} \log(n) + N^{-r+1} \right) \right\}. \qquad (2.7.10) \end{aligned}$$

The term K_2 can be written as

$$\begin{aligned} K_2 &= \sum_{k=1}^{d_n-1} \frac{\partial^2 \hat{R}(\theta_0^*)}{\partial \theta_j^* \partial \theta_k^*} \left(\theta_k^* - \theta_{0,k}^*\right) \\ &= \sum_{k=1}^{d_n-1} \frac{\partial^2 \left\{ \hat{R}(\theta_0^*) - R(\theta_0^*) \right\}}{\partial \theta_j^* \partial \theta_k^*} \left(\theta_k^* - \theta_{0,k}^*\right) + \sum_{k=1}^{d_n-1} \frac{\partial^2 R(\theta_0^*)}{\partial \theta_j^* \partial \theta_k^*} \left(\theta_k^* - \theta_{0,k}^*\right) \\ &= K_{21} + K_{22}. \end{aligned}$$

Based on the proof of Theorem 2.1, using the Cauchy-Schwarz inequality and $\|\theta^* - \theta_0^*\| = O_P\{d_n^{1/2} n^{-1/2} N^{3/2} \log(n)\}$, we have

$$\begin{aligned} |K_{21}| &\leq \left\|\theta_k^* - \theta_{0,k}^*\right\| \left| \sum_{k=1}^{d_n-1} \frac{\partial^2 \left\{\hat{R}(\theta_0^*) - R(\theta_0^*)\right\}}{\partial\theta_j^* \partial\theta_k^*} \right| \\ &= O_P\left\{d_n\left(n^{-1/2}N^{5/2}\log(n) + N^{-r+2}\right)\right\} \times O_P\left\{n^{-1/2}N^{3/2}\log(n) + N^{-r+1}\right\} \\ &= O_P\left\{d_n n^{-1} N^4 \log^2(n)\right\} = o_P\left\{d_n^{1/2} n^{-1/2} N^{3/2} \log(n) + d_n^{1/2} N^{-r+1}\right\}. \quad (2.7.11) \end{aligned}$$

On the other hand, we have

$$\begin{aligned} |K_{22}| &= \left| \sum_{k=1}^{d_n-1} \frac{\partial^2 R(\theta_0^*)}{\partial\theta_j^* \partial\theta_k^*} (\theta_k^* - \theta_{0,k}^*) \right| \\ &\leq O_P\left\{d_n^{1/2} n^{-1/2} N^{3/2} \log(n) + d_n^{1/2} N^{-r+1}\right\} \times \left| \sum_{k=1}^{d_n-1} H_{j,k}^*(\theta_0^*) \right| \\ &= O_P\left\{d_n^{1/2} n^{-1/2} N^{3/2} \log(n) + d_n^{1/2} N^{-r+1}\right\}. \quad (2.7.12) \end{aligned}$$

Next, we consider K_3, and we can write it as follows:

$$\begin{aligned} K_3 &= \sum_{k,l=1}^{d_n-1} \left\{ \frac{\partial^3 \hat{R}(\theta^*)}{\partial\theta_j^* \partial\theta_k^* \partial\theta_l^*} - \frac{\partial^3 \hat{R}(\theta^*)}{\partial\theta_j^* \partial\theta_k^* \partial\theta_l^*} \right\} \left(\theta_k^* - \theta_{0,k}^*\right) \left(\theta_l^* - \theta_{0,l}^*\right) \\ &+ \sum_{k=1}^{d_n-1} \frac{\partial^3 \hat{R}(\theta_0^*)}{\partial\theta_j^* \partial\theta_l^* \partial\theta_k^*} \left(\theta_k^* - \theta_{0,k}^*\right) \left(\theta_l^* - \theta_{0,l}^*\right) \\ &= K_{31} + K_{32}. \end{aligned}$$

However, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
|K_{31}| &\leq \left| \sum_{k,l=1}^{d_n-1} \left\{ \frac{\partial^3 \hat{R}(\theta^*)}{\partial\theta_j^* \partial\theta_k^* \partial\theta_l^*} - \frac{\partial^3 R(\theta^*)}{\partial\theta_j^* \partial\theta_k^* \partial\theta_l^*} \right\} \right| \|\theta^* - \theta_0^*\|^2 \\
&= O_P\left\{ d_n \left(n^{-1/2} N^{7/2} \log(n) + N^{-r+3} \right) \right\} \times O_P\left\{ \left(d_n^{1/2} n^{-1/2} N^{3/2} \log(n) \right)^2 \right\} \\
&= O_P\left\{ d_n^2 n^{-3/2} N^{13/2} \log^3(n) \right\} \\
&= o_P\left\{ d_n^{1/2} n^{-1/2} N^{3/2} \log(n) + d_n^{1/2} N^{-r+1} \right\}. \qquad (2.7.13)
\end{aligned}$$

By Assumption (A5),

$$\begin{aligned}
|K_{32}| &\leq O_P(d_n) \|\theta_n^* - \theta_0^*\|^2 = O_P(d_n) \times O_P\left\{ d_n n^{-1} N^3 \log^2(n) \right\} \\
&= o_P\left\{ d_n^{1/2} n^{-1/2} N^{3/2} \log(n) \right\}. \qquad (2.7.14)
\end{aligned}$$

From equations (2.7.10) to (2.7.14), we have

$$K_1 + K_2 + K_3 = O_P\left\{ d_n^{1/2} n^{-1/2} N^{3/2} \log(n) \right\}.$$

According to Assumptions (A7), (P1) and $\{d_n^{1/2} n^{-1/2} N^{3/2} \log(n)\} \lambda_n^{-1} \to 0$, we have

$$\frac{\partial \hat{R}(\theta_n^*)}{\partial \theta_j^*} = \lambda_n \left[\frac{p'_{\lambda_n}(|\theta_j^*|)}{\lambda_n} \operatorname{sgn}(\theta_j^*) + O_P\left\{ \left(d_n^{1/2} n^{-1/2} N^{3/2} \log(n) \right) \lambda_n^{-1} \right\} \right].$$

Hence, it is easy to see that the sign of θ_j^* completely determines the sign of $\frac{\partial \hat{R}(\theta^*)}{\partial \theta_j^*}$ and Lemma 2.5 holds. ∎

2.7.5 Proof of Theorem 2.2

As shown in Theorem 2.1, there is a α_n-consistent local minimizer $\hat{\theta}^*$ of $\hat{Q}\left(\theta^*\right)$. By Lemma 2.5, part (i) of Theorem 2.2 holds, thus, $\hat{\theta}^*$ has the form $\left\{(1-\|\hat{\theta}_1^*\|^2)^{1/2}, \hat{\theta}_1^{*T}, 0^T\right\}^T$. To prove part (ii) in Theorem 2.2, it is equivalent to show that

$$(\mathbf{H}+\mathbf{\Sigma}_{\lambda_n})(\hat{\theta}_1^* - \theta_{01}^*) + b_n = \hat{S}(\theta_{01}^*) + o_P\left(n^{-1/2}\right).$$

With a slight abuse of notation, let $\hat{Q}\left(\theta_1^*\right) = \hat{Q}\left\{((1-\|\theta_1^*\|^2)^{1/2})^T, \theta_1^{*T}, 0^T\right\}$. As $\hat{\theta}_1^*$ must satisfy the penalized equation $\frac{\partial}{\partial\theta_1^*}\hat{Q}(\hat{\theta}_1^*) = 0$, using the Taylor expansion on $\frac{\partial}{\partial\theta_1^*}\hat{Q}(\hat{\theta}_1^*)$ at point θ_{01}^* component-wisely, we have

$$\begin{aligned}
&\left[\left\{\frac{\partial^2}{\partial\theta_1^*\partial\theta_1^{*T}}\hat{R}\left(\theta_{01}^*\right) + p_{\lambda_n}''(\bar{\theta}_1)\right\}(\hat{\theta}_1^* - \theta_{01}^*) + p_{\lambda_n}'\left(\theta_{01}^*\right)\right] \\
= \quad & -\frac{\partial}{\partial\theta_1^*}\hat{R}(\theta_{01}^*) - \frac{1}{2}\left[(\hat{\theta}_1^* - \theta_{01}^*)^T\frac{\partial^2}{\partial\theta_1^*\partial\theta_1^{*T}}\left\{\frac{\partial}{\partial\theta_j^*}\hat{R}(\bar{\bar{\theta}}_1^*)\right\}(\hat{\theta}_1^* - \theta_{01}^*)\right]_{j=1}^{s_n-1},
\end{aligned}$$

where $\bar{\theta}_1$ and $\bar{\bar{\theta}}_1$ lie between $\hat{\theta}_1^*$ and θ_{01}^*. Now, we define

$$\begin{aligned}
U &= \frac{\partial^2}{\partial\theta_1^*\partial\theta_1^{*T}}\left\{\hat{R}\left(\theta_{01}^*\right) - R\left(\theta_{01}^*\right)\right\}(\hat{\theta}_1^* - \theta_{01}^*), \\
T &= \frac{1}{2}\left[(\hat{\theta}_1^* - \theta_{01}^*)^T\frac{\partial^2}{\partial\theta_1^*\partial\theta_1^{*T}}\left\{\frac{\partial}{\partial\theta_j}\hat{R}(\bar{\bar{\theta}}_1^*)\right\}(\hat{\theta}_1^* - \theta_{01}^*)\right]_{j=1}^{s_n-1}.
\end{aligned}$$

Similar to the proof of Theorem 2.1 and by the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\|T\| &\leq O_P\left\{\left(d_n^{1/2}n^{-1/2}N^{3/2}\log(n) + d_n^{1/2}N^{-r+1}\right)^2\right\} \\
&\quad \times O_P\left\{d_n^{3/2}n^{-1/2}N^{7/2}\log(n) + d_n^{3/2}N^{-r+3}\right\} \\
&\quad + O_P\left\{\left(d_n^{1/2}n^{-1/2}N^{3/2}\log(n) + d_n^{1/2}N^{-r+1}\right)^2\right\} \times O_P\left(d_n^{3/2}\right) \\
&= O_P\left\{d_n^{5/2}n^{-3/2}N^{13/2}\log^3(n)\right\} + O_P\left\{d_n^{5/2}n^{-1}N^3\log^2(n)\right\} \\
&= o_P(n^{-1/2}). \qquad (2.7.15)
\end{aligned}
$$

We also have that

$$|U| = O_P\left\{d_n^{3/2}n^{-1}N^4\log^2(n)\right\} = o_P(n^{-1/2}). \qquad (2.7.16)$$

Finally, from (2.7.15) and (2.7.16), we have

$$(\hat{\mathbf{H}} + \mathbf{\Sigma}_{\lambda_n})(\hat{\theta}_1^* - \theta_{01}^*) + b_n = \hat{S}(\theta_{01}^*) + o_P\left(n^{-1/2}\right).$$

Let $\mathbf{\Psi} = \left\{\psi_{jk}\right\}_{j,k=2}^{s_n}$ be the asymptotic covariance matrix of $\sqrt{n}\hat{S}(\theta_{01}^*)$. Following [52], we have

$$\psi_{jk} = \sum_{i=-\infty}^{\infty} E\{(\dot{m}_j - \theta_{0,j}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,1}})(\dot{m}_k - \theta_{0,k}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,i+1}})\varepsilon_1\varepsilon_{i+1}\},$$

in which $\dot{m}_j$ is the value of $\frac{\partial}{\partial\theta_j}m_\theta$ taking at $\theta^* = \theta_0^*$, for any $j, k = 2, \cdots, s_n$.

Let $\mathbf{\Omega} = (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\mathbf{\Psi}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}$ and $\mathbf{A}_n$ be a $q \times (s_n - 1)$ matrix such that $\mathbf{A}_n\mathbf{A}_n^T$ converges to a nonnegative symmetric $q \times q$ matrix $\mathbf{\Sigma}_A$. We now prove the asymptotic normality of $\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H}+\mathbf{\Sigma}_{\lambda_n})^{-1}\sqrt{n}\hat{S}(\theta_{01}^*)$. To achieve such aim, we have to show that for any vector $a = (a_1, a_2, \cdots, a_q)^T \in R^q$,

$$a^T\{\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\sqrt{n}\hat{S}(\theta_{01}^*)\} \to N(0, a^T\mathbf{\Sigma}_A a). \tag{2.7.17}$$

in distribution.

By the first order derivative approximation result in Lemma 2.4 and Assumption (A5), we have for any j,

$$\begin{aligned}\hat{S}_j(\theta_{01}^*) &= \frac{1}{n}\sum_{i=1}^{n}(\dot{m}_j - \theta_{0,j}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,i}})\varepsilon_i \\ &+o_p\{N^{-(r-1)} + n^{-1}N^2\log^2(n) + (nN)^{-1/2}\log(n)\}.\end{aligned}$$

According to Assumptions (A2) and (A3),

$$\hat{S}_j(\theta_{01}^*) = \frac{1}{n}\sum_{i=1}^{n}(\dot{m}_j - \theta_{0,j}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,i}})\varepsilon_i + o_p(n^{-1/2}).$$

For simplicity, we let $W_i = \{(\dot{m}_j - \theta_{0,j}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,i}})\}_{j=2}^{s_n}$ and write

$$\begin{aligned}a^T\{\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\sqrt{n}\hat{S}(\theta_{01}^*)\} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}a^T\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}W_i\varepsilon_i + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\varepsilon_i + o_p(1),\end{aligned}$$

where $Z_i = a^T \mathbf{A}_n \mathbf{\Omega}^{-1/2} (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1} W_i$. Note that $E(Z_i \varepsilon_i) = 0$, and

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \varepsilon_i \right) \\
&= a^T \mathbf{A}_n \mathbf{\Omega}^{-1/2} (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1} \mathrm{Var}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \varepsilon_i \right) (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1} \mathbf{\Omega}^{-1/2} \mathbf{A}_n^T a \\
&= a^T \mathbf{A}_n \mathbf{\Omega}^{-1/2} (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1} \mathbf{\Psi} (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1} \mathbf{\Omega}^{-1/2} \mathbf{A}_n^T a \\
&\to a^T \mathbf{\Sigma}_A a.
\end{aligned}$$

Applying Theorem 2.21 in [22], we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \varepsilon_i \to N(0, a^T \mathbf{\Sigma}_A a)$$

in distribution. Slutsky's theorem entails that

$$\sqrt{n} \mathbf{A}_n \mathbf{\Omega}^{-1/2} \left\{ (\hat{\theta}_1^* - \theta_{01}^*) + (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1} b_n \right\} \to N\left(0, \mathbf{\Sigma}_A\right).$$

This completes the proof.

Table 2.1: Selection results for Example 1

n	d	METHOD	TPN	FPN	C(%)	MME ($\times 10^{-2}$)	TIME (s)	ITER
100	25	ORACLE	20.00	0.00	100.0	0.94	0.01	–
		PSIM	19.77	0.00	81.4	1.17	1.29	3.56
		PLS	17.90	0.00	55.4	2.25	4.33	–
		PSIR	15.90	0.00	53.8	2.77	1.05	–
100	50	ORACLE	45.00	0.00	100.0	1.43	0.01	–
		PSIM	44.56	0.01	71.8	3.26	3.30	5.78
		PLS	33.83	0.00	33.2	10.49	11.44	–
		PSIR	35.90	0.00	35.0	8.38	3.20	–
200	25	ORACLE	20.00	0.00	100.0	0.43	0.02	–
		PSIM	19.92	0.00	92.0	0.50	2.35	2.50
		PLS	18.18	0.00	68.2	0.93	5.48	–
		PSIR	17.24	0.00	64.6	1.29	1.16	–
200	50	ORACLE	45.00	0.00	100.0	0.69	0.02	–
		PSIM	44.80	0.00	84.4	0.86	6.28	3.78
		PLS	41.40	0.00	63.4	3.84	13.33	–
		PSIR	37.50	0.01	62.2	5.27	3.13	–
200	100	ORACLE	95.00	0.00	100.0	1.24	0.03	–
		PSIM	94.64	0.00	75.8	2.19	17.12	4.80
		PLS	80.83	0.00	39.6	13.80	39.30	–
		PSIR	83.51	0.00	41.6	11.77	14.23	–

Table 2.2: Bias and MSE of coefficients of Example 1

n	d	EST	BIAS			MSE		
			PSIM	PLS	PSIR	PSIM	PLS	PSIR
100	25	θ_1	-0.002	-0.010	-0.010	0.003	0.006	0.008
		θ_2	-0.004	-0.007	-0.010	0.002	0.005	0.008
		θ_3	-0.005	-0.010	-0.009	0.002	0.006	0.008
		θ_4	-0.003	0.010	-0.017	0.002	0.004	0.010
		θ_5	-0.005	-0.010	-0.011	0.002	0.007	0.008
100	50	θ_1	-0.019	-0.040	-0.039	0.015	0.028	0.030
		θ_2	-0.012	-0.036	-0.037	0.016	0.036	0.029
		θ_3	-0.019	-0.043	-0.036	0.016	0.028	0.030
		θ_4	-0.032	0.042	-0.044	0.019	0.032	0.030
		θ_5	-0.019	-0.038	-0.035	0.017	0.033	0.029
200	25	θ_1	0.001	-0.005	-0.008	0.001	0.004	0.005
		θ_2	-0.002	-0.006	-0.008	0.001	0.004	0.006
		θ_3	-0.003	-0.004	-0.006	0.001	0.003	0.004
		θ_4	-0.002	0.006	-0.009	0.001	0.005	0.007
		θ_5	0.000	-0.005	-0.007	0.001	0.004	0.005
200	50	θ_1	-0.001	-0.016	-0.012	0.001	0.008	0.008
		θ_2	-0.003	-0.016	-0.015	0.002	0.008	0.006
		θ_3	-0.004	-0.013	-0.019	0.002	0.007	0.008
		θ_4	-0.002	0.012	-0.010	0.002	0.006	0.004
		θ_5	-0.007	-0.015	-0.014	0.002	0.007	0.006
200	100	θ_1	-0.019	-0.060	-0.048	0.013	0.044	0.032
		θ_2	-0.019	-0.026	-0.046	0.013	0.030	0.035
		θ_3	-0.019	-0.123	-0.044	0.013	0.062	0.035
		θ_4	-0.015	0.042	-0.044	0.012	0.033	0.035
		θ_5	-0.018	-0.045	-0.042	0.011	0.035	0.035

Table 2.3: Standard deviations of the estimators for Example 1

	$\hat{\theta}_1$		$\hat{\theta}_2$		$\hat{\theta}_3$		$\hat{\theta}_4$		$\hat{\theta}_5$	
$n(d)$	SD	SD$_m$	SD	SD$_m$	SD	SD$_m$	SD	SD$_m$	SD	SD$_m$
	(SD$_{mad}$)		(SD$_{mad}$)		(SD$_{mad}$)		(SD$_{mad}$)		(SD$_{mad}$)	
100	0.050	0.032	0.048	0.031	0.048	0.031	0.050	0.031	0.049	0.032
(25)	(0.009)		(0.009)		(0.009)		(0.008)		(0.008)	
100	0.119	0.036	0.126	0.036	0.126	0.036	0.134	0.036	0.127	0.036
(50)	(0.018)		(0.018)		(0.018)		(0.016)		(0.016)	
200	0.032	0.023	0.031	0.024	0.033	0.023	0.033	0.024	0.033	0.024
(25)	(0.005)		(0.005)		(0.004)		(0.004)		(0.004)	
200	0.039	0.027	0.043	0.027	0.042	0.027	0.044	0.026	0.042	0.026
(50)	(0.006)		(0.007)		(0.006)		(0.006)		(0.006)	
200	0.113	0.033	0.113	0.034	0.114	0.033	0.110	0.033	0.101	0.032
(100)	(0.019)		(0.021)		(0.021)		(0.020)		(0.018)	

Table 2.4: Selection results for Example 2

n	d	METHOD	TPN	FPN	C(%)	MME ($\times 10^{-2}$)	TIME (s)	ITER
100	7	ORACLE	2.00	0.00	100.0	0.65	0.35	–
		PSIM	1.84	0.10	75.4	0.82	0.79	2.54
		PLS	1.28	0.85	31.6	9.08	2.95	–
		PSIR	1.19	1.21	15.0	11.72	0.65	–
200	10	ORACLE	5.00	0.00	100.0	0.25	0.38	–
		PSIM	4.89	0.05	91.2	0.27	0.97	2.07
		PLS	3.67	0.04	26.8	15.92	10.13	–
		PSIR	3.71	0.88	19.4	24.70	0.84	–
400	12	ORACLE	7.00	0.00	100.0	0.11	0.41	–
		PSIM	6.95	0.00	94.6	0.12	3.61	1.88
		PLS	6.49	0.46	46.8	5.46	13.46	–
		PSIR	5.60	0.41	31.0	6.41	1.36	–
800	16	ORACLE	11.00	0.00	100.0	0.08	0.46	–
		PSIM	10.98	0.00	98.2	0.09	15.06	1.57
		PLS	10.81	0.12	73.6	1.12	43.11	–
		PSIR	9.68	0.09	54.8	1.38	3.32	–

Table 2.5: Bias and MSE for the coefficients in Example 2

n	d	EST	BIAS			MSE		
			PSIM	PLS	PSIR	PSIM	PLS	PSIR
100	7	θ_1	0.005	0.004	0.008	0.0035	0.009	0.010
		θ_2	-0.010	0.038	-0.078	0.0036	0.012	0.018
		θ_3	-0.027	-0.086	-0.159	0.0039	0.036	0.065
		θ_4	0.019	0.113	0.125	0.0097	0.044	0.045
		θ_5	-0.034	-0.058	-0.115	0.0086	0.030	0.044
200	10	θ_1	0.003	0.068	0.020	0.0010	0.012	0.006
		θ_2	-0.003	-0.056	-0.051	0.0006	0.012	0.011
		θ_3	-0.008	-0.114	-0.121	0.0009	0.036	0.042
		θ_4	0.006	0.182	0.102	0.0027	0.041	0.038
		θ_5	-0.012	-0.115	-0.191	0.0025	0.038	0.044
400	12	θ_1	0.001	0.005	0.019	0.0004	0.004	0.005
		θ_2	-0.001	-0.023	-0.032	0.0002	0.005	0.006
		θ_3	-0.002	-0.048	-0.067	0.0002	0.019	0.024
		θ_4	-0.001	0.067	0.066	0.0005	0.021	0.023
		θ_5	-0.004	-0.069	-0.067	0.0005	0.022	0.024
800	16	θ_1	0.003	0.004	0.010	0.0002	0.002	0.005
		θ_2	0.004	-0.007	-0.010	0.0002	0.003	0.005
		θ_3	0.000	-0.014	-0.018	0.0002	0.003	0.006
		θ_4	-0.002	0.036	0.035	0.0003	0.011	0.012
		θ_5	0.000	-0.032	-0.033	0.0003	0.011	0.012

Table 2.6: Standard deviations of estimators for Example 2

	$\hat{\theta}_1$		$\hat{\theta}_2$		$\hat{\theta}_3$		$\hat{\theta}_4$		$\hat{\theta}_5$	
$n(d)$	SD	SD$_m$	SD	SD$_m$	SD	SD$_m$	SD	SD$_m$	SD	SD$_m$
	(SD$_{mad}$)		(SD$_{mad}$)		(SD$_{mad}$)		(SD$_{mad}$)		(SD$_{mad}$)	
100	0.059	0.037	0.044	0.032	0.057	0.035	0.84	0.046	0.081	0.043
(7)	(0.012)		(0.013)		(0.016)		(0.024)		(0.014)	
200	0.031	0.025	0.024	0.020	0.030	0.021	0.052	0.035	0.048	0.025
(10)	(0.005)		(0.005)		(0.006)		(0.009)		(0.005)	
400	0.019	0.019	0.014	0.014	0.014	0.015	0.026	0.022	0.022	0.018
(13)	(0.003)		(0.002)		(0.003)		(0.005)		(0.002)	
800	0.015	0.013	0.012	0.011	0.012	0.011	0.017	0.017	0.018	0.015
(16)	(0.001)		(0.001)		(0.001)		(0.002)		(0.001)	

Table 2.7: Selection results for Example 3

n	d	METHOD	δ	TPN	FPN	C(%)	MME ($\times 10^{-2}$)	TIME (s)	ITER
100	25	ORACLE	0	20.00	0.00	100.0	0.06	0.03	–
			1	20.00	0.00	100.0	0.11	0.03	–
		PSIM	0	19.50	0.00	93.8	0.06	1.32	5.80
			1	19.40	0.00	93.4	0.12	1.31	5.80
100	50	ORACLE	0	45.00	0.00	100.0	0.05	0.03	–
			1	45.00	0.00	100.0	0.11	0.03	–
		PSIM	0	42.04	0.01	74.2	0.07	3.88	7.91
			1	41.92	0.00	73.8	0.13	3.79	7.49
200	25	ORACLE	0	20.00	0.00	100.0	0.02	0.05	–
			1	20.00	0.00	100.0	0.05	0.05	–
		PSIM	0	19.97	0.00	99.2	0.02	3.07	4.15
			1	19.97	0.00	99.2	0.05	2.94	3.93
200	50	ORACLE	0	45.00	0.00	100.0	0.02	0.05	–
			1	45.00	0.00	100.0	0.05	0.05	–
		PSIM	0	44.82	0.00	96.8	0.02	9.35	5.28
			1	44.82	0.00	96.8	0.05	8.76	5.02
200	100	ORACLE	0	95.00	0.00	100.0	0.02	0.05	–
			1	95.00	0.00	100.0	0.05	0.05	–
		PSIM	0	93.95	0.00	78.8	0.04	16.92	6.42
			1	93.81	0.00	77.4	0.07	17.55	6.58

Table 2.8: Bias and MSE of coefficients of Example 3

n	d	EST	BIAS		SD		MSE	
			$\delta = 0$	$\delta = 1$	$\delta = 0$	$\delta = 1$	$\delta = 0$	$\delta = 1$
100	25	θ_1	-0.006	-0.006	0.061	0.066	0.004	0.004
		θ_2	-0.011	-0.002	0.080	0.060	0.007	0.004
		θ_3	-0.006	-0.006	0.073	0.061	0.005	0.004
		θ_4	-0.003	0.009	0.071	0.074	0.005	0.006
		θ_5	-0.002	-0.007	0.058	0.075	0.003	0.006
100	50	θ_1	-0.019	-0.023	0.124	0.132	0.016	0.018
		θ_2	-0.020	-0.019	0.125	0.131	0.016	0.018
		θ_3	-0.039	-0.029	0.140	0.125	0.021	0.016
		θ_4	-0.021	0.023	0.117	0.124	0.014	0.016
		θ_5	-0.016	-0.020	0.124	0.120	0.016	0.015
200	25	θ_1	-0.001	0.000	0.023	0.027	0.001	0.001
		θ_2	-0.001	-0.001	0.024	0.016	0.001	0.000
		θ_3	-0.002	-0.003	0.030	0.037	0.001	0.001
		θ_4	0.000	0.000	0.013	0.016	0.000	0.000
		θ_5	0.000	0.000	0.026	0.026	0.001	0.001
200	50	θ_1	-0.001	-0.002	0.039	0.038	0.002	0.001
		θ_2	-0.002	0.000	0.043	0.040	0.002	0.002
		θ_3	-0.002	-0.004	0.037	0.044	0.001	0.002
		θ_4	-0.001	0.002	0.029	0.028	0.001	0.001
		θ_5	-0.004	-0.001	0.043	0.039	0.003	0.001
200	100	θ_1	-0.022	-0.018	0.144	0.117	0.021	0.014
		θ_2	-0.021	-0.021	0.136	0.128	0.019	0.017
		θ_3	-0.011	-0.019	0.134	0.134	0.018	0.018
		θ_4	-0.021	0.022	0.139	0.129	0.020	0.017
		θ_5	-0.021	-0.015	0.138	0.127	0.020	0.016

Table 2.9: Variable selection and estimation for the river flow dataset

	Y_{t-1}	Y_{t-2}	Y_{t-3}	Y_{t-4}	Y_{t-5}	Y_{t-6}	Y_{t-7}
PSIM	0.885	-0.408	0.179	-0.085			
BIC-SIP	0.877	-0.382	0.208	-0.125			
	X_t	X_{t-1}	X_{t-2}	X_{t-3}	X_{t-4}	X_{t-5}	X_{t-6}
PSIM	0.043						
BIC-SIP	0.046	0.034	-0.004				
	Z_t	Z_{t-1}	Z_{t-2}	Z_{t-3}	Z_{t-4}	Z_{t-5}	Z_{t-6}
PSIM	0.096	-0.012					
BIC-SIP	0.126	-0.079					

Table 2.10: Mean squared prediction errors (MSPEs) for river flow dataset

METHOD	PSIM	BIC-SIP	FULL-SIP	BIC-LM
MSPE	49.09	60.52	62.11	81.99

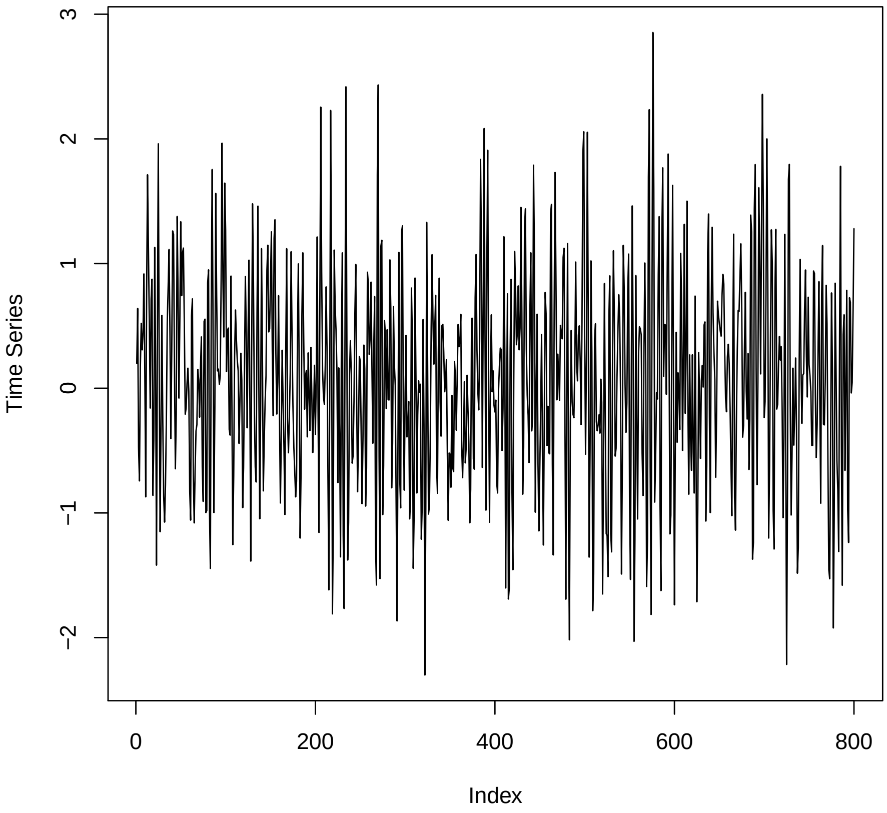

Figure 2.1: A simulated time series from NAR model ($n = 800$, $d = 16$).

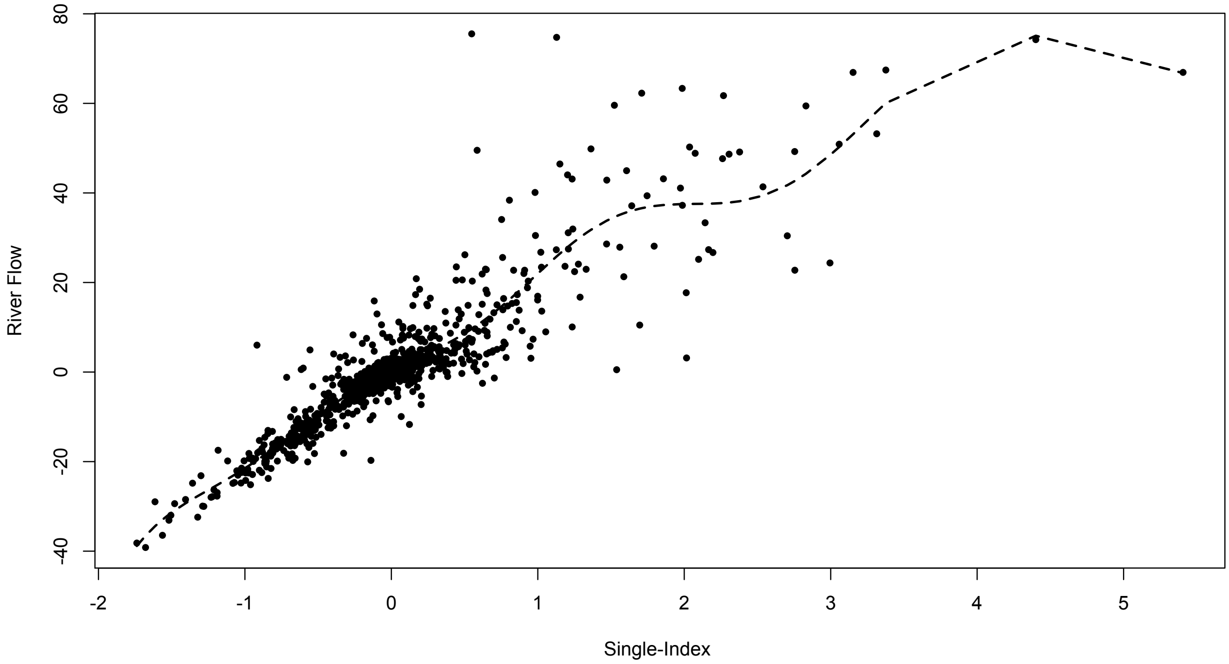

Figure 2.2: Estimated single-index function for river flow (based on 1972-1973).

Chapter 3

Ultra-High Dimensional Single-Index Models

3.1 Introduction

Advancements in information technology have enabled scientists to collect data of unprecedented size as well as complexity. Nowadays, high-dimensional data commonly arise in such diverse fields as biology, engineering, health sciences, economics and information technology. Here, the word "high-dimensional" refers to the case where the number of potential explanatory variables, p, is one or several orders of magnitude larger than the sample size, n, of the data. The analysis of high-dimensional data gives rise to many new challenges and opportunities for statistical methodology. Take a simple linear model for example. When it has many more unknown parameters than the number of observations, the least-squares fitting is ill-posed.

To make high dimensional statistical inference possible, it is often useful and reasonable to assume that the p-dimensional regression parameters are sparse, with many components being zero, and this assumption is well known as the "sparsity" assumption. With sparsity, a widely used class of approaches for analyzing high-dimensional data is regularized or penalized regression. Methods that have been proposed include Bridge regression [23], L_1 penalized regression or LASSO [46], smoothly clipped absolute deviation (SCAD) [15], Dantzig selector [3], and minimax concave penalty (MCP, [58]). Related to these methods, there has been a great deal of theoretical study as well as algorithmic development recently, see [10], [11] as well as [49].

Several recent papers have considered variable selection in non-convex penalized high-dimensional regression. For example, [15] first suggested the use of SCAD penalty for model selection in the non-convex penalized likelihood with fixed and finite number of coefficients and covariates p. [19] extended these results by allowing p to diverge slowly with the sample size n at the rate $p = o(n^{1/5})$ or $p = o(n^{1/3})$ in a general likelihood framework. While all of the above methods focus on the case that $n > p$, recently more and more researches target on the scenario of $p \gg n$. In the context of the linear model, [35] proved that the SCAD estimator still has the oracle property on ultrahigh dimensional problems with $p \gg n$, specifically, they have shown that the oracle estimator itself is a local minimum of SCAD penalized least squares regression, for high dimensional non-convex penalized regression with $p \gg n$.

Meanwhile, a series of methods based on independent learning has become increasingly popular. [16] proposed a method of independent learning via a two-stage procedure for linear model. In the first stage, independence screening is adopted as a fast but crude method of reducing the dimensionality from ultra-high to a more moderate size (usually below the sample size); a suitable feature selection technique can be applied in the second stage. Using a similar idea, later, [21] extended the previous procedure into a more general framework – generalized linear models. More recently [17] studied penalized likelihood methods for ultrahigh dimensional variable selection, and in the context of generalized linear models, they demonstrated the proposed method possesses model selection consistency with oracle properties for $p \gg n$. In [13] and [18], they proposed a class of nonparametric independence screening for ultra-high-dimensional additive models and varying coefficient models, respectively.

In practice, without prior knowledge about the relationship between response variable y and independent variables X, the regression function $g(x) = E(y|X = x)$ often needs to be modeled in a flexible nonparametric fashion. Therefore, one of our goals is to approximate $g(x)$ by a function having simplifying structure which makes both estimation and interpretation possible. Due to such concerns, single-index models (SIMs) seem to be very appealing. Another advantage of SIMs is that they are very useful and fundamental tools for handling "curse of dimensionality". Intensive research on estimation in single-index models has given rise to a large amount of literature; [43], [29], [4], [54] and [30] are among those frontiers. Later, [55] developed the minimum average variance estimation method to prevent the under-smoothing of the nonparametric link function. To make the estimation more robust against deviations from SIMs, [52] proposed a polynomial spline estimator for the single-index prediction model. Furthermore, [5] and [7] introduced estimating equation based methods to study SIMs. In the former, [5] transferred restricted least squares to unrestricted least squares to reduce the limiting variance. In the latter, [7] further relaxed the constraint that all the index parameters lie in an open unit ball. Most recently, [40] proposed a robust and efficient estimation procedure based on local model regression in terms of the sensitivity outliers and heavy tailed error.

As the number of index parameters in SIMs increases, it becomes more and more difficult to identify the significant explanatory variables efficiently. To conquer such challenge, several procedures have been developed. [36] proposed the dissected cross-validation method which outperformed the traditional leave-m-out cross-validation. Later, [63] implemented the adaptive lasso with kernel smoothing to estimate and select important predictors without assuming the error term as additive. Furthermore, [41] introduced the penalized least squares method and [60] proposed the LASSO with local linear smoothing method. Both methods are able to simultaneously estimate parameters and select variables.

All the above methods aim at the case of fixed number of index parameters. In practice, the number of introduced variables typically varies with the sample size. Therefore, [62] developed a sliced inverse regression (SIR) based method which can handle the case of a diverging number of index parameters. When the dimension of index variables is high and even ultra-high, how to find the relationship between the response and the index variables efficiently becomes a serious scientific endeavor. This motivates us to consider feature selection for single-index models in high dimensional, even ultra-high-dimensional settings, with the goal being to identify the oracle estimator with high probability. To achieve this, a nonparametric independent screening is developed and a new theoretical result for the oracle property to hold is derived within this article. The rest of this chapter is organized as follows. In Section 3.2, we briefly review the single-index models and the methodology for polynomial spline estimation. In Section 3.3, we introduce a new nonparametric independence screening algorithm. We establish the properties of the proposed estimators in Section 3.4. In Section 3.5, we report numerical results from Monte Carlo simulations and we present a real data example in Section 3.6. The proofs are given in Section 3.7.

3.2 Single-Index Model and Its Estimation

Suppose we have an i.i.d. random sample, $\{(X_i, Y_i)\}_{i=1}^n$, from the following single-index model

$$Y_i = g\left(X_i^T \theta_0\right) + \varepsilon_i,\ i = 1, \ldots, n, \tag{3.2.1}$$

in which Y_i are the response variables and $X_i = (X_{i1}, \cdots, X_{ip})^T$ are p-dimensional ($p \geq 1$) vectors of covariates. Without loss of generality, we assume the covariates are standardized to have mean zero and variance one. The function g is some smooth but unknown univariate link function, and θ_0 is a vector of some unknown parameters, often referred to as the single-index coefficient. Errors ε_i are i.i.d. random noise with

$E\left(\varepsilon_i\right)=0$ and $E\left(\varepsilon_i^2\right)=\sigma^2<\infty$. For model identifiability, we assume that the single-index coefficients $\theta_0 \in \Theta=\{(\theta_{01},\cdots,\theta_{0p})|\sum_{j=1}^{p}\theta_{0j}^2=1,\theta_{01}>0\}$.

In the following, we assume that the observed data $\{(X_i,Y_i)\}_{i=1}^n$ are i.i.d. copies of $(\mathcal{X},\mathcal{Y})$. Before describing the estimation method of the single-index parameters, we define the following notations for simplicity. Given a fixed θ, denote $\mathcal{X}_\theta=\mathcal{X}^T\theta$ and $X_{\theta,i}=X_i^T\theta$ for any $i=1,\ldots,n$. Let

$$g_\theta(z)=E(\mathcal{Y}|\mathcal{X}_\theta=z)=E\{g(\mathcal{X}^T\theta_0)|\mathcal{X}_\theta=z\}$$

then θ_0 is the minimizer of the following population least squares criterion function

$$R(\theta)=\frac{1}{2}E[\{Y-g_\theta(\mathcal{X}_\theta)\}^2]=\frac{1}{2}E[\{g(\mathcal{X}^T\theta_0)-g_\theta(\mathcal{X}_\theta)\}^2]+\frac{1}{2}\sigma^2. \tag{3.2.2}$$

Since Θ is not a compact set, we consider the minimization problem in (3.2.2) over all $\theta\in\Theta_c$, where $\Theta_c=\{(\theta_1,\cdots,\theta_p)|\sum_{j=1}^{p}\theta_j^2=1,\theta_1\geq c\}$ for some $c\in(0,1)$.

In the following we discuss the estimation method for the single-index coefficient $\theta_0\in\Theta_c$ and the unknown function g in (3.2.1). To estimate the unknown functional parameters, we use spline basis approximations. In principle, any basis functions can be used, but in this article we consider the polynomial splines to estimate the unknown function g_θ for any given θ. The appeal of polynomial splines is that they often provide good approximations of smoothing functions with a simple linear combination of spline bases; see more discussions in [56].

For any given θ, suppose $g_\theta(z)$ can be approximated by $\sum_{k=1-r}^{N} B_{k,r}(z)$, where N is the number of interior knots, and $B_{k,r}(z)$, $k = 1-r, ..., N$, are the B-spline basis functions of order r; see [8]. Denote next the $(N+r)$-dimensional space $\mathcal{G}^{(r-2)}$ of spline basis functions as the linear space spanned by $\{B_{k,r}(z), k = 1-r, \ldots, N\}$. Then, for any given θ, the polynomial spline estimator of order r for g_θ is defined as

$$\widehat{g}_\theta(\cdot) = \underset{g(\cdot)\in\mathcal{G}^{(r-2)}}{\arg\min} \sum_{i=1}^{n} \{Y_i - g(X_{\theta,i})\}^2.$$

Let $Y = (Y_1, \ldots, Y_n)^T$ be the response vector. For any fixed θ, denote $B_r(z) = \{B_{k,r}(z)\}_{k=1-r}^{N}$, and $\mathbf{B}_{\boldsymbol{\theta}} = \{B_{k,r}(X_{\boldsymbol{\theta},i})\}_{i=1,k=-(r-1)}^{n,\ N}$. Then one can obtain the spline estimator of $g_\theta(z)$ by

$$\widehat{g}_\theta(z) = B_r(x_\theta)(\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}})^{-1}\mathbf{B}_{\boldsymbol{\theta}}^T Y.$$

Then, the single-index parameters θ_0 can be estimated by minimizing

$$\widehat{R}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\{Y_i - \widehat{g}_\theta(z)\}^2. \tag{3.2.3}$$

Now suppose some of the variables are not relevant in the single-index, i.e., the corresponding single-index coefficients are zero. In the following we introduce a regularization penalty to (3.2.3). Since the potential number of explanatory variables increases at an exponential rate of the sample size n, we denote it as p_n. To perform simultaneous selection and estimation, we propose minimizing the following penalized sum of squares

$$\widehat{Q}\,(\theta) = \widehat{R}\,(\theta) + \sum_{j=1}^{p_n} p_{\lambda_n}(|\theta_j|) I\{|\theta_j| \neq \max_{1\leq k\leq p_n}(|\theta_k|)\}, \tag{3.2.4}$$

which shrinks small components of estimated functions to zero. Note that the above minimization in (3.2.4) is for all $\theta \in \Theta_c$, so we don't penalize the largest element of θ. There are various of ways to specify the penalty function p_λ. Here we use the SCAD penalty function of [15], defined as

$$p_\lambda(\theta) = \begin{cases} \lambda\theta, & \text{if } 0 \leq \theta \leq \lambda \\ -\frac{(\theta^2 - 2a\lambda\theta + \lambda^2)}{2(a-1)}, & \text{if } \lambda < \theta < a\lambda \\ \frac{(a+1)\lambda^2}{2}, & \text{if } \theta > a\lambda \end{cases}$$

for some tuning parameter a, and $a = 3.7$ is used in all the simulation examples as well as in the real data application.

3.3 Asymptotic Properties of the Estimator

In this section, we study the asymptotic properties of the PSIM estimator. Without loss of generality, we assume that the first q_n single-index coefficients are nonzero and the remaining $p_n - q_n$ index coefficients are 0. For any $i = 1, \ldots, n$, let $X_i^T = (X_{i(1)}^T, X_{i(2)}^T)^T$, where $X_{i(1)}^T$ are the first q_n non-zero variables and $X_{i(2)}^T$ are the $p_n - q_n$ zero variables. Similarly we write $\theta_0^T = (\theta_{0(1)}^T, \theta_{0(2)}^T)$ with $\theta_{0(2)}^T = (0, \ldots, 0)^T$. Define the oracle estimator

$\widehat{\theta}^{oT} = (\widehat{\theta}^{oT}_{(1)}, 0^T)^T$, where

$$\begin{aligned} \widehat{\theta}^{o}_{(1)} &= \underset{\theta_{(1)} \in \{(\theta_1, \cdots, \theta_{q_n}) | \sum_{j=1}^{q_n} \theta_j^2 = 1, \theta_1 \geq c\}}{\arg\min} \widehat{R}^o(\theta_{(1)}) \qquad (3.3.1) \\ &= \underset{\theta_{(1)} \in \{(\theta_1, \cdots, \theta_{q_n}) | \sum_{j=1}^{q_n} \theta_j^2 = 1, \theta_1 \geq c\}}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} \left\{ Y_i - \widehat{g}_{\theta_{(1)}}(X_{i(1)}^T \theta_{(1)}) \right\}^2. \end{aligned}$$

The following two theorems demonstrate that the oracle estimator $\widehat{\theta}^{o}$ asymptotically becomes a local minimum under the SCAD penalty. Such a property is widely known as the oracle property [15, 35] . This result implies that we can find a good estimator among the local minima using the PSIM method proposed in Section 3.4, assuming we use the SCAD penalty.

Theorem 3.1. *Suppose Assumptions (A1)–(A6) stated in Section 3.7 hold and assume that $E(\varepsilon_i)^{2k} < \infty$ for an integer $k > 0$. Let $\mathcal{A}_n(\lambda_n)$ be the set of local minima of (3.2.4) with the SCAD penalty and a regularization parameter λ_n, we have*

$$Pr\{\widehat{\theta}^{o} \in \mathcal{A}_n(\lambda_n)\} \to 1$$

as $n \to \infty$ provided that $\lambda_n = o\{n^{-(1-(\alpha_2-\alpha_1))/2}\}$ and $p_n/(\sqrt{n}\lambda_n)^{2k} \to 0$.

The above theorem demonstrates that when ε_i has the all moments, the oracle property holds when $p_n = O(n^{\alpha})$ for any $\alpha > 0$, as $E(\varepsilon_i)^{2k} < \infty$ for all $k > 0$.

For Gaussian errors, the following theorem proves that the oracle property holds when $p_n = O\{\exp(\alpha_3 n)\}$ for some $\alpha_3 > 0$, that is, the dimension of covariates is allowed to grow exponentially fast.

Theorem 3.2. *Suppose Assumptions (A1)–(A6) stated in Section 3.7 hold and assume that the ε_i's are i.i.d. Gaussian random variables. Then*

$$Pr\{\widehat{\theta}^o \in \mathcal{A}(\lambda_n)\} \to 1,$$

as $n \to \infty$, provided that $p_n = O\{\exp(\alpha_3 n)\}$ and $\lambda_n = O(n^{-(1-\alpha_4)/2})$, where $0 < \alpha_3 < \alpha_4 < \alpha_2 - \alpha_1$.

3.4 Nonparametric Independence Screening

3.4.1 Nonparametric Independence Screening Procedure

The dimension p of the index parameter in model (3.2.1) can be very large, and here we consider the case that p increases much faster than the sample size n, especially by non-polynomial dimensionality or simply NP-dimensionality, i.e., $\log p = O(n^a)$ for some $a \in (0, 1/2)$. Therefore, we write the dimension as p_n. Let $\mathcal{M}_0 = \{1 \le j \le p_n : \theta_{0,j} \neq 0\}$ be the true sparse model with non-sparsity size $q_n = |\mathcal{M}_0|$. The other $p_n - q_n$ variables can also be correlated with the response variable via linkage to the predictors contained in the model. To expeditiously identify important variables in model (3.2.1), without facing the "curse of dimensionality", [16] first proposed sure independent screening (SIS) method to reduce the space of explanatory variables from a NP-dimensionality to moderate size. Later, [13] and [18] extended the SIS procedure into a class of nonparametric independence screening (NIS) models. Similarly, we consider the following p_n nonparametric marginal models. In this article, we refer to marginal models as fitting model with componentwise covariates:

$$\widehat{\theta}_j^{NIS} = \arg\min \frac{1}{2n} \sum_{i=1}^{n} \{Y_i - g_j(X_{i,j}\theta_j)\}^2, \ j = 1, \cdots, p_n, \tag{3.4.1}$$

where the function g_j can be approximated by some smoothing methods such as the polynomial spline smoothing method mentioned above. We rank the utility of covariates in model (3.4.1) according to, for example, magnitude of the absolute value of coefficients, or the sum of residuals, and then select a small group of covariates by thresholding. For example, we sort the p_n componentwise magnitudes of the $|\widehat{\theta}_j^{NIS}|$ in a decreasing order and define a submodel as

$$\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leq j \leq p_n : |\widehat{\theta}_j^{NIS}| \geq \gamma_n\}$$

where γ_n is a pre-defined threshold value. Such an independence learning ranks the importance of features according to their magnitude of marginal regression coefficients.

Another screening approach is to rank according to the descent order of the residual sum of squares of the componentwise nonparametric regressions, where we select a set of variables:

$$\widehat{\mathcal{N}}_{\nu_n} = \{1 \leq j \leq p_n : u_j \leq \nu_n\},$$

where $u_j = \frac{1}{2n}\sum_{i=1}^{n}\{Y_i - \widehat{g}_j(X_{i,j}\widehat{\theta}_j)\}^2$ represents the residual sum of squares of the marginal fit, while ν_n is another pre-defined threshold value. Such an independent screening ranks the importance according to the descent order of the residual sum of squares of the componentwise nonparametric regressions. This screening also can be viewed as ranking by the magnitude of the correlation of the marginal nonparametric estimate $\widehat{g}_j(X_{ij}\widehat{\theta}_j)_{i=1}^{n}$ with the response $\{Y_i\}_{i=1}^{n}$. In both of these senses, the proposed NIS procedure is related to the correlation learning proposed by [16].

With such NIS procedure, we dramatically decrease the dimension of the parameter space from p_n to a much smaller number with model size $|\widehat{\mathcal{M}}_{\gamma_n}|$ or $|\widehat{\mathcal{N}}_{\nu_n}|$. Thus, the computational burden is much more feasible. According to [21], although the interpretations and implications of the marginal models are biased from the joint model, the non-sparse information about the joint model can be passed along to the marginal model under a mild condition.

After variable screening, the next step is naturally to select the variables using more refined techniques in the single-index model. For example, the dissected cross-validation (DCV) method in [36], the profile least squares (PrLS) estimation procedure in [39], the adaptive LASSO with kernel smoothing in [63], the slice inverse regression based method in [62] and penalized single-index prediction models (PSIM) proposed by [50].

3.4.2 Data-Driven Thresholding Determination

To determine a data-driven threshold for the nonparametric independence screening method, we adopt a similar random permutation idea of [13]. We use random permutation to decouple X_i and Y_i in order to make the resulting data $(X_{\pi(i)}, Y_i)$ follow a null model. Here $\pi(1), \cdots, \pi(n)$ are a random permutation of the index $1, \cdots, n$. Specifically, the permutation algorithm works as follows:

1. For every $j \in \{1, \cdots, p_n\}$, find the local minimum in (3.4.1). Randomly permute the rows of X, yielding X^*. Let $\omega_{(q)}$ be the q-th quantile of $\{\widetilde{u}_j, j = 1, \cdots, p_n\}$, where

$$\widetilde{\theta}_j = \arg\min \frac{1}{2n} \sum_{i=1}^{n} \{Y_i - \widetilde{g}_j(X^*_{i,j}\theta_j)\}^2 \text{ and } \widetilde{u}_j = \frac{1}{2n} \sum_{i=1}^{n} \{Y_i - \widetilde{g}_j(X^*_{i,j}\widetilde{\theta}_j)\}^2.$$

   Then the NIS selects the following variables:

$$\mathcal{N}_1 = \{1 \leq j \leq p_n : u_j \leq \omega_{(q)}\}.$$

   As suggested by [13], in this work, we use $q = 1$ which means to take the maximum value of the empirical norm of the permuted estimates.

2. Apply PSIM [50] on the set $\mathcal{N}_1$ to select a subset A_1. Inside the PSIM algorithm, the tuning parameter is selected by high-dimensional BIC proposed by [48].

In Step 2, we use PSIM method. In fact, any variable selection method for single-index models mentioned previously would work once one has reduced from p_n to d_n via the NIS screening step in 1.

3.5 Simulation Studies

It has been shown in [16] and [20] that independent screening is a fast but crude method for reducing the dimensionality. Some extension of independent screening method include iterative SIS and multi-stage procedures, such as SIS-SCAD and SIS-LASSO. These methods can be applied to perform the final variable selection and estimation simultaneously. In this section, we present one simulation example with several different scenarios to evaluate the performance of NIS procedure for single-index models (NIS-PSIM).

We vary the sample size from 100 to 200 for different scenarios to gauge the difficulties of the simulation models. The following configurations with $p = 1000$ and 5000 are considered for generating the covariates $X = (X_1, \cdots, X_p)^T$. We consider a similar regression model as in [52],

$$Y_i = \sin\left(\frac{\pi}{4} X_i^T \theta_0\right) + \varepsilon_i$$

where ε_i's are independently and identically distributed as $N(0, 0.2^2)$, for all $i = 1, \cdots, n$. In this example, the true parameter is $\theta_0^T = (1, 1, 1, 1, 1, 0, \cdots, 0)/\sqrt{5}$, i.e.; the first five elements of θ_0 are non-zero while the remaining $p_n - 5$ elements are zero. Table 3.1

summarizes the variable selection results based on 200 replications under each of the settings with partial combinations of sample size $n = 100,\ 200$ and $p = 1000,\ 5000$. With these setting, we aim to illustrate the behaviors of the NIS procedure under different combinations of sample size and number of parameters. In this example, we compare our method (NIS-PSIM) with the method (HD-SCAD) proposed by [35], the sure independence screening plus penalized least square regression (SIS-SCAD) proposed by [16] as well as the oracle estimators which are obtained using the method proposed by [52].

In Table 3.1, the column labeled "TPN" presents the average number of zeroes, restricted only to the true zero coefficients, while column "FPN" shows the average number of the $q = 5$ true zero coefficients erroneously set to zero. The column labeled as "C" represents the percentage for which the correct model has been chosen among those 200 Monte Carlo replications. The root mean squared prediction error (RMSPE) is reported in the fourth column, and it is calculated by:

$$\text{MSPE} = \left\{ \frac{1}{n} \sum_{i=1}^{n} E\left\{ \widehat{g}\left(X_i^T \widehat{\theta} \right) \right\}^{1/2} - Y_i \right\}^2 .$$

Note that the "oracle" method always identifies the five non-zero coefficients and $p - 5$ zero coefficients correctly. The last column, "TIME", reports the running time per iteration in seconds using a desktop with Intel(R) Core(TM) i5-2500 CPU @ 3.30GHz and 8.00GB RAM.

In addition, Table 3.2 demonstrates the variable estimation results based on three measurements, including bias (BIAS), standard error (SD) and root mean squared error

(RMSE), which is calculated using $\mathrm{RMSE}(\theta_j) = \left\{ \frac{1}{200} \sum_{i=1}^{200} \left(\hat{\theta}_j - \theta_{0,j} \right)^2 \right\}^{1/2}$. To make it more comparable, in Table 3.2, we compare our NIS-PSIM estimator to the oracle estimator. From this table, one can see that as the sample size increases, the NIS-PSIM estimator becomes closer to the oracle estimator.

Summarizing Table 3.1, our proposed NIS-PSIM outperforms the other two competitive methods (HD-SCAD and SIS-SCAD), in terms of correctly identifying the correct model. Both HD-SCAD and SIS-SCAD methods penalize much too harshly and thus set too many values to zero. On the other hand, as the sample size increases, the performances of NIS-PSIM is getting closer to the oracle estimator and in terms of RMSPE, the NIS-PSIM is more comparable with the oracle estimator. From Table 3.1, one can also see that the NIS-PSIM estimator provides much smaller RMSPE compared with the other two methods. The results on both variable selection and estimation confirm that the NIS-PSIM method outperforms the other two methods.

[Tables 3.1 and 3.2 about here.]

3.6 Application

In this section, we implement our NIS-PSIM method to solve a high-dimensional gene microarray problems. The data set we used was reported by [44]. In this data set, 120 twelve-week-old male F2 rats were selected for tissue harvesting from the eyes and underwent microarray analysis. Such microarrays contains 31,042 different probes. The intensity values were normalized using the RMA (robust multi-chi averaging) and represented gene expression levels. Since gene TRIM 32 is widely known to cause Bardet-Biedl syndrome, the primary goal of this analysis is to identify the genes whose expression levels are most closely are related to gene TRIM32.

There are 31,042 probes in the data-set, with one of them recording the gene expression levels in each of the 120 rats for the TRIM32 gene. The question is what linear combination of the (presumably few) among the other 31,041 probes yields a single-index within a fitted spline that predicts the TRIM32 gene's expression levels well. One must remember that the entire microarray procedure is a 'kitchen sink' approach, where all genes in the complete mouse genome are examined, even though relatively few have anything to do with the eye network, and even fewer of those are likely to be involved in the process that causes Bardet-Biedl syndrome. More traditional biological analyses would avoid this complication entirely by simply excluding the genes (probes) that couldn't possibly be related to the factor of interest. In microarrays, that can usually be done relatively simply in an indirect way. In this case, the first step is to normalize across the chips (n=120 rats), by dividing the expression levels, $E(i, j)$ for probe i of chip j by the median expression level for the chip and taking the log_2-transform. That is:

$$W(i, j) = \log_2 \frac{E(i, j)}{M(j)},$$

where $M(j)$ is the median of the gene expression levels of chip j. This yields normalized expression levels that have a median of zero for each array. Such distributions tend to be fairly symmetric, although there are occasional outliers which must be handled by methods such as those given in [2] and [34]. Once this has been done, the probes are roughly comparable across chips (rats). Non-interesting probes would be those that have very low expression levels across all rats or which show too little variation across the rats. To eliminate the former, [44] declared gene i to be 'insufficiently expressed' if the maximum of the 120 $W(i,j)$ values was less than the 25th percentile of all $W(i,j)$ scores. Similarly, they declared gene i to be 'insufficiently variable' if the difference between the largest and smallest $W(i,j)$ values for fixed i was less than one. This is equivalent to a less than two-fold change, since the $W(i,j)$ values are in log_2 scale. For the rat data-set, employing these two criteria reduced the number of candidate probes from the original value of 31,041 to 18,795.

While the above screening procedure produces a 40% reduction in candidate probes, the number remaining far exceeds the number of rats (n=120), so more reduction must be made before employing most of the model selection procedures. The high-dimensional SCAD procedure of [35], in theory, can be made to work for very large numbers of probes, but it may be too slow to be practical. They proposed reducing the 18,795 to 3,000 (before beginning their HD-SCAD procedure) by simply choosing the 3,000 probes which displayed the largest variance across the 120 rats. For our PSIM procedure, whether we begin with 18,795 or 3,000 probes is a secondary concern – before we can apply PSIM, we must first reduce the number of candidate probes to some value k such that $k < n$. (This step of going from $p > n$ to $k < n$ probes is what most in the field call 'statistical screening', as opposed to some of the steps above which could more appropriately be called ‘biological screening’, as they tend to eliminate probes which really should never have been included in the experiment, if it had been designed more rigorously.) To achieve $k < n$, we could simply begin with the 119 probes which have the highest correlation with the probe of interest, but this is very inefficient. Instead, we used an idea based on the Sure Independence Screening procedure of [16]. We use two enhancements of their method. First, we use nonparametric independent screening (NIS) rather that the parametric SIS method which they initially introduced, to select those genes which yield small errors when the nonparametric splines are fit to model the TRIM32 gene. To determine how many (k) of these best-fitting genes to use before applying PSIM, we use the permutation adaptation of SIS discussed in [21] to select the genes that achieve the permutation threshold. For most microarrays, the SIS-permutation (or NIS-permutation) procedure will yield a number of potential probes, k, which is much less than n. In this example, we found $k = 20$ while $n = 120$. Once we attain this small set ($k = 20$) of genes which seem best able to predict the TRIM32 gene, we can then employ our PSIM method to find the best linear combination to yield the single index best fit by the nonparametric spline. In this example, the result was a set of $q = 12$ probes whose linear combination best estimated the single-index for the spline that predicted the TRIM32 gene.

The value of $q = 12$ probes that we identified is less than the $q = 24$ or $q = 19$ probes found by [32] using unrestricted and adaptive LASSO, respectively, but more than the $q = 6$ probes found by [16]. Since this is a real data-set, we don’t know what the true answer is. One way to compare the accuracy of the procedures is to fit them on a sample of size $n_0 = 80$ of the rats, with validation occurring on the remaining $n - n_0 = 120 - 80 = 40$ rats. This process (80-40 partitioning and validation) was repeated 100 times, comparing our (NIS-PSIM) method's results with those of [35]'s HD-SCAD and [16]'s SIS-SCAD. The mean model size (MS) and prediction error (PE) and their standard deviations based on these 100 replications are summarized in Table 3.3. The two rows in the table reflect whether the process began with the biological-screening population size of 18,795 candidate probes or the reduced set of 3,000 candidate probes. Overall, the NIS-PSIM method appears quite robust.

[Table 3.3 about here]

3.7 Proof of Theorems

In the following, let $\mathbf{X} = (X_{ij})_{i=1,j=1}^{n,\ p_n} = (\mathbf{X}_{(1)}, \mathbf{X}_{(2)})$ be the predictor matrix, where $\mathbf{X}_{(1)}$ is the first $n \times q$ submatrix and $\mathbf{X}_{(2)}$ is the last $n \times (p_n - q_n)$ submatrix of $\mathbf{X}$.

3.7.1 Assumptions

In this subsection, we state our assumptions below.

(A1) There exist positive constants M_1 and M_2 such that $E|X_{ij}|^4 \leq M_1$, for any $1 \leq i \leq n$ and $1 \leq j \leq p_n$, and $\sup_{u^T u=1} u^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} u \geq M_2$.

(A2) For any $\theta_1, \theta_2 \in \Theta_c$, the joint density function $f_{\theta_1,\theta_2}(x_{\theta_1}, x_{\theta_2})$ of X_{θ_1} and X_{θ_2} has r-th order ($r \geq 5$) continuous partial derivatives and is bounded below and above on $[a, b]^2$. The marginal density function of X_θ, $f_\theta(x_\theta) \in C^{(1)}[a, b]$, and is bounded below, for any $\theta \in \Theta_c$.

(A3) The true link function $g \in C^{(r)}[a, b]$ for some $r \geq 5$.

(A4) The number of interior knots N satisfies:

$$n^{1/\{2(r-1)\}} \ll N \ll \min\{n^{1/7} \log{(n)}^{-1/7}, n^{1/3} q_n^{-4/3} \log{(n)}^{-1/3}\}.$$

(A5) Let the values of $\theta_{0,1}, \theta_{0,2}, \cdots, \theta_{0,q_n}$ be nonzero, $\theta_{0,q_n+1}, \theta_{0,q_n+2}, \cdots, \theta_{0,p_n}$ be zero, and $q_n = O(n^{\alpha_1})$ for some $0 < \alpha_1 < 1/4 - 3/(8(r-1))$.

(A6) There exist positive constants α_2 and M_3 such that $\alpha_1 < \alpha_2 < 1$ and

$$n^{(1-\alpha_2)/2} \min_{j=1,\ldots,q_n} |\theta_{0,j}| \geq M_3.$$

Remark 6. *Assumptions (A2) and (A3) are typical in the nonparametric smoothing literature; see for instance, [52] and [51]. Assumption (A1) is similar to Condition (A1) and (A2) in [35]. Assumption (A4) specifies the requirement of the number of knots in spline approximation. Assumptions (A5) and (A6) are in parallel with the requirements stated in Conditions (A3) and (A4) in [35]. Note that the order of q_n not only depends on the sample size n, but also depends on the degree of smoothness r of the link function g. If we assume that g is infinitely differentiable or smooth, i.e., g has infinitely many* *derivatives, then only $\alpha_1 < 1/4$ is required. If the link function is less smooth, the bound is even tighter.*

3.7.2 Preliminary Results

Before we prove the theorem, we first state the following lemma.

Lemma 3.1. *If Assumptions (A2) and (A3) hold, one has that*

$$\sup_{\theta\in\Theta_c}\left\|\widehat{g}_\theta^{(k)} - g_\theta^{(k)}\right\|_\infty = O_P\left\{n^{-1/2}N^{1/2+k}(\log n)^{1/2} + N^{-(r-k)}\right\} \tag{3.7.1}$$

for any $k = 0, \ldots, r-2$.

Proof of Lemma 3.1 is similar to the proof of Proposition A.1 in [52], but replacing the approximation rate of cubic spline smoothing by the more general polynomial spline approximation results, and thus is omitted.

In the following, we define

$$\widehat{S}(\theta) = \left\{\widehat{S}_j(\theta)\right\}_{j=2}^{p_n} = \left\{\frac{\partial}{\partial\theta_j}\widehat{R}(\theta)\right\}_{j=2}^{p_n} = \left[n^{-1}\sum_{i=1}^{n}\left\{\widehat{g}_\theta\left(X_{\theta,i}\right) - Y_i\right\}\widehat{g}'_\theta\left(X_{\theta,i}\right)\widetilde{X}_{ij}(\theta)\right]_{j=2}^{p_n}, \tag{3.7.2}$$

where $\widehat{R}(\theta)$ is given in (3.2.3), and

$$\widetilde{X}_{ij}(\theta) = X_{ij} - \theta_j\theta_1^{-1}X_{i1}, \tag{3.7.3}$$

for any $1 \leq i \leq n$ and $2 \leq j \leq p_n$. Next let

$$\widetilde{S}(\theta) = \left\{\widetilde{S}_j(\theta)\right\}_{j=2}^{p_n} = \left[\frac{1}{n}\sum_{i=1}^{n}\left\{g_\theta\left(X_{\theta,i}\right) - Y_i\right\} g'_{\boldsymbol{\theta}}\left(X_{\theta,i}\right) \widetilde{X}_{ij}(\theta)\right]_{j=2}^{p_n}. \tag{3.7.4}$$

The following result provides the order of uniform difference between $\widehat{S}(\theta)$ and $\widetilde{S}(\theta)$.

Lemma 3.2. *Under Assumptions (A1)–(A4), one has*

$$\sup_{\theta\in\Theta_c}\max_{2\leq j\leq p_n}\left|\widehat{S}_j(\theta) - \widetilde{S}_j(\theta)\right| = o_P(n^{-1/2}). \tag{3.7.5}$$

Proof. By the definitions of $\widehat{S}_j(\theta)$ and $\widetilde{S}_j(\theta)$ in (3.7.2) and (3.7.4), one has for any $j = 2, 3, \ldots, p_n$

$$\widehat{S}_j(\theta) = \widetilde{S}_j(\theta) + K_{1,\theta,j} + K_{2,\theta,j} + K_{3,\theta,j},$$

where $\widetilde{X}_{ij}(\theta)$ is in (3.7.3) and

$$\begin{aligned}
K_{1,\theta,j} &= n^{-1}\sum_{i=1}^{n}\left(\widehat{g}_\theta - g_\theta\right)\left(X_{\theta,i}\right)\left(\widehat{g}'_\theta - g'_\theta\right)\left(X_{\theta,i}\right)\widetilde{X}_{ij}(\theta),\\
K_{2,\theta,j} &= n^{-1}\sum_{i=1}^{n}\left\{g_\theta\left(X_{\theta,i}\right) - Y_i\right\}\left(\widehat{g}'_\theta - g'_\theta\right)\left(X_{\theta,i}\right)\widetilde{X}_{ij}(\theta),\\
K_{3,\theta,j} &= n^{-1}\sum_{i=1}^{n}\left(\widehat{g}_\theta - g_\theta\right)\left(X_{\theta,i}\right) g'_\theta\left(X_{\theta,i}\right)\widetilde{X}_{ij}(\theta).
\end{aligned}$$

Using Lemma 3.1 and similar arguments in Lemma A.11 in the Supplement of [52], one obtains the following:

$$\begin{aligned}
\sup_{\theta\in\Theta_c}\max_{2\leq j\leq p_n}|K_{1,\theta,j}| &= O_P\left\{n^{-1}N^2\log n + N^{1-2r} + n^{-1/2}N^{3/2-r}(\log n)^{1/2}\right\}, \\
\sup_{\theta\in\Theta_c}\max_{2\leq j\leq p_n}|K_{2,\theta,j}| &= O_P\left\{N^{1-r} + n^{-1}N\log n\right\}, \\
\sup_{\theta\in\Theta_c}\max_{2\leq j\leq p_n}|K_{3,\theta,j}| &= O_P\left\{N^{-r} + n^{-1}\log n + n^{-1/2}N^{-1/2}(\log n)^{1/2}\right\}.
\end{aligned}$$

Therefore,

$$\sup_{\theta\in\Theta_c}\max_{2\leq j\leq p_n}\left|\widehat{S}_j(\theta)-\widetilde{S}_j(\theta)\right| = o_P\{N^{-(r-1)} + n^{-1}N^2\log n + (nN)^{-1/2}(\log n)^{1/2}\}.$$

Thus, (3.7.5) is established by Assumption (A4). ∎

Let

$$\widehat{\mathbf{H}}(\theta) = \left\{\widehat{H}_{j,j'}(\theta)\right\}_{j,j'=2}^{p_n} = \left\{\frac{\partial^2}{\partial\theta_j\partial\theta_{j'}}\widehat{R}(\theta)\right\}_{j,j'=2}^{p_n} \tag{3.7.6}$$

be the Hessian matrix of $\widehat{R}(\theta)$, then

$$\begin{aligned}
\widehat{H}_{j,j'}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)\{\widehat{g}'_\theta(X_i^T\theta)\}^2 \\
&+\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)\left\{\widehat{g}_\theta(X_i^T\theta)-Y_i\right\}\widehat{g}''_\theta(X_i^T\theta) \\
&-\frac{1}{n}\sum_{i=1}^{n}X_{i1}\left\{\widehat{g}_\theta(X_i^T\theta)-Y_i\right\}\widehat{g}'_\theta(X_i^T\theta)\delta_{j,j'}(\theta)
\end{aligned} \tag{3.7.7}$$

with $\widetilde{X}_{ij}(\theta)$ in (3.7.3) and

$$\delta_{j,j'}(\theta) = \theta_1^{-3}\left\{(\theta_j^2 + \theta_1^2)1_{(j=j')} + \theta_j\theta_{j'}1_{(j\neq j')}\right\}. \tag{3.7.8}$$

Substituting $\widehat{g}_\theta$ and its derivatives by g_θ and its corresponding derivatives, we define $\widetilde{\mathbf{H}}(\theta) = \left\{\widetilde{H}_{j,j'}(\theta)\right\}_{j,j'=2}^{p_n}$, where

$$\begin{aligned} \widetilde{H}_{j,j'}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)\{g'_\theta(X_i^T\theta)\}^2 \qquad (3.7.9) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)\left\{g_\theta(X_i^T\theta) - Y_i\right\}g''_\theta(X_i^T\theta) \\ &- \frac{1}{n}\sum_{i=1}^{n}X_{i1}\left\{g_\theta(X_i^T\theta) - Y_i\right\}g'_\theta(X_i^T\theta)\delta_{j,j'}(\theta). \end{aligned}$$

Lemma 3.3. *Under Assumptions (A1)–(A4), one has*

$$\sup_{\theta\in\Theta_c}\max_{2\leq j,j'\leq p_n}\left|\widehat{H}_{j,j'}(\theta) - \widetilde{H}_{j,j'}(\theta)\right| = o_P(1). \tag{3.7.10}$$

Proof. Note that for any j, j', we decompose the difference $\widehat{H}_{j,j'}(\theta) - \widetilde{H}_{j,j'}(\theta)$ as follows:

$$
\begin{aligned}
& \widehat{H}_{j,j'}(\theta) - \widetilde{H}_{j,j'}(\theta) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)\{(\widehat{g}'_\theta(X_{\theta,i}))^2 - (g'_\theta(X_{\theta,i}))^2\} \\
& + \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)[(\widehat{g}_\theta(X_{\theta,i}) - Y_i)\widehat{g}''_\theta(X_{\theta,i}) - (g_\theta(X_{\theta,i}) - Y_i)(g''_\theta(X_{\theta,i})] \\
& - \frac{1}{n}\sum_{i=1}^{n} X_{i1}\,[\{\widehat{g}_\theta(X_{\theta,i}) - Y_i\}\widehat{g}'_\theta(X_{\theta,i}) - \{g_\theta(X_{\theta,i}) - Y_i\}g'_\theta(X_{\theta,i})]\,\delta_{j,j'}(\theta) \\
= \; & A_{1,\theta,jj'} + A_{2,\theta,jj'} + A_{3,\theta,jj'},
\end{aligned}
\tag{3.7.11}
$$

where $\delta_{j,j'}(\theta)$ is given in (3.7.8), and

$$
\begin{aligned}
A_{1,\theta,jj'} &= \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)\{(\widehat{g}'_\theta - g'_\theta)(X_{\theta,i})\}^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n} \widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)(\widehat{g}'_\theta - g'_\theta)(X_{\theta,i})g'_\theta(X_{\theta,i}) \\
&= A_{11,\theta,jj'} + A_{12,\theta,jj'},
\end{aligned}
$$

$$
\begin{aligned}
A_{2,\theta,jj'} &= \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)(\widehat{g}_\theta - g_\theta)(X_{\theta,i})(\widehat{g}''_\theta - g''_\theta)(X_{\theta,i}) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)\{g_\theta(X_{\theta,i}) - Y_i\}(\widehat{g}''_\theta - g''_\theta)(X_{\theta,i}) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)(\widehat{g}_\theta - g_\theta)(X_{\theta,i})g''_\theta(X_{\theta,i}) \\
&= A_{21,\theta,jj'} + A_{22,\theta,jj'} + A_{23,\theta,jj'},
\end{aligned}
$$

and

$$\begin{aligned}
A_{3,\theta,jj'} &= \frac{1}{n}\sum_{i=1}^{n} X_{i1}(\widehat{g}_\theta - g_\theta)(X_{\theta,i})(\widehat{g}'_\theta - g'_\theta)(X_{\theta,i})\delta_{j,j'}(\theta) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} X_{i1}\{g_\theta(X_{\theta,i}) - Y_i\}(\widehat{g}'_\theta - g'_\theta)(X_{\theta,i})\delta_{j,j'}(\theta) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} X_{i1}(\widehat{g}_\theta - g_\theta)(X_{\theta,i})g'_\theta(X_{\theta,i})\delta_{j,j'}(\theta) \\
&= A_{31,\theta,jj'} + A_{32,\theta,jj'} + A_{33,\theta,jj'}.
\end{aligned}$$

Note that

$$(E|\widetilde{X}_{ij}(\theta)|^2)^{1/2} = (E|X_{ij} - \theta_j\theta_1^{-1}X_{i1}|^2)^{1/2} \le (E|X_{ij}|^2)^{1/2} + |\theta_j\theta_1^{-1}|(E|X_{i1}|^2)^{1/2}.$$

According to Assumptions (A1)–(A3), there exist positive constant c_1, c_2, and c_3 such that for any $2 \le j, j' \le p_n$,

$$\begin{aligned}
&\sup_{\theta\in\Theta_c} E|\widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)| \le \{E|\widetilde{X}_{ij}(\theta)|^2 E|\widetilde{X}_{ij'}(\theta)|^2\}^{1/2} \le c_1, \\
&\sup_{\theta\in\Theta_c} E|\widetilde{X}_{ij}(\theta)\widetilde{X}_{ij'}(\theta)|^4 \le \{E|\widetilde{X}_{ij}(\theta)|^4 E|\widetilde{X}_{ij'}(\theta)|^4\}^{1/2} \le c_2, \\
&\{E|g_\theta(X_{\theta,i}) - Y_i|^2\}^{1/2} \le \{E|g_\theta(X_{\theta,i}) - g(X_{\theta_0,i})|^2\}^{1/2} + \{E|\varepsilon_i|^2\}^{1/2} \le c_3.
\end{aligned}$$

In addition, one has $\sup_{\theta\in\Theta_c}\|g'_\theta\|_\infty \le c_4$ and $\sup_{\theta\in\Theta_c}\|g''_\theta\|_\infty \le c_5$ for some positive constants c_4 and c_5.

By Lemma 3.1, one has for $J_{11,\theta,jj'}$ and $J_{12,\theta,jj'}$,

$$\begin{aligned}
\sup_{\theta\in\Theta_c}|A_{11,\theta,jj'}| &\leq (c_1+o_P(1))\sup_{\theta\in\Theta_c}\|\widehat{g}'_\theta-g'_\theta\|_\infty^2=o_P(1),\\
\sup_{\theta\in\Theta_c}|A_{12,\theta,jj'}| &\leq 2(c_1+o_P(1))c_4\sup_{\theta\in\Theta_c}\|\widehat{g}'_\theta-g'_\theta\|_\infty=o_P(1).
\end{aligned}$$

Therefore, $\sup_{\theta\in\Theta_c}|A_{1,\theta,jj'}|=o_P(1)$. For $A_{21,\theta,jj'}$ and $A_{23,\theta,jj'}$, one has

$$\begin{aligned}
\sup_{\theta\in\Theta_c}|A_{21,\theta,jj'}| &\leq (c_1+o_P(1))\sup_{\theta\in\Theta_c}\|\widehat{g}'_\theta-g'_\theta\|_\infty=o_P(1),\\
\sup_{\theta\in\Theta_c}|A_{23,\theta,jj'}| &\leq (c_1+o_P(1))c_5\sup_{\theta\in\Theta_c}\|\widehat{g}_\theta-g_\theta\|_\infty=o_P(1).
\end{aligned}$$

For $A_{22,\theta,jj'}$, one has

$$\begin{aligned}
\sup_{\theta\in\Theta_c}|A_{22,\theta,jj'}| &\leq \sup_{\theta\in\Theta_c}\|\widehat{g}''_\theta-g''_\theta\|_\infty\times\sup_{\theta\in\Theta_c}\left\{\frac{1}{n}\sum_{i=1}^n|\widetilde{X}_{ij}(\theta)\widetilde{X}_{ij}(\theta)|^2\right\}^{1/2}\\
&\quad\times\sup_{\theta\in\Theta_c}\left\{\frac{1}{n}\sum_{i=1}^n|g_\theta(X_{\theta,i})-Y_i|^2\right\}^{1/2}\\
&\leq o_P(1)\times(c_2+o_P(1))\times(c_3+o_P(1))=o_P(1).
\end{aligned}$$

Therefore, $\sup_{\theta\in\Theta_c}|A_{2,\theta,jj'}| = o_P(1)$. Note that $\sup_{\theta\in\Theta_c}|\lambda_{j,j'(\theta)}| < c_6$ for some positive constant c_6. Similarly, for $A_{31,\theta,jj'}$, $A_{32,\theta,jj'}$ and $A_{33,\theta,jj'}$, one has

$$\begin{aligned}
\sup_{\theta\in\Theta_c}|A_{31,\theta,jj'}| &\leq (c_1 + o_P(1))c_6 \sup_{\theta\in\Theta_c}\|\widehat{g}'_\theta - g'_\theta\|_\infty = o_P(1), \\
\sup_{\theta\in\Theta_c}|A_{32,\theta,jj'}| &\leq \sup_{\theta\in\Theta_c}\frac{1}{n}\sum_{i=1}^{n}|X_{i1}\{g_\theta(X_{\theta,i}) - g(X_{\theta_0,i}) - \varepsilon_i\}| \\
&\quad \times \sup_{\theta\in\Theta_c}\|\widehat{g}''_\theta - g''_\theta\|_\infty \times c_6 = o_P(1), \\
\sup_{\theta\in\Theta_c}|A_{33,\theta,jj'}| &\leq (c_1 + o_P(1))c_4 \sup_{\theta\in\Theta_c}\|\widehat{g}_\theta - g_\theta\|_\infty = o_P(1).
\end{aligned}$$

Hence $\sup_{\theta\in\Theta_c}|A_{3,\theta,jj'}| = o_P(1)$. The desired results follows from (3.7.11) ▮

In the following for $\theta = (\theta_1, \theta_2, \cdots, \theta_{p_n})^T$, let $\theta^* = (\theta_2, \cdots, \theta_{p_n})^T$ be a $(p_n - 1)$-dimensional vector after removing the first component in θ. Similarly, we denote θ_0, $\hat{\theta}^{o*}$, $\theta^*_{0(1)}$ and $\hat{\theta}^{o*}_{(1)}$ as their corresponding vectors but without the first element. Define a Jacobian matrix of θ with respect to θ^*

$$\mathbf{J}(\theta^*) = \begin{pmatrix} \mathbf{J}_{(1)}(\theta^*) \\ \mathbf{J}_{(2)}(\theta^*) \end{pmatrix} = -(1 - \|\theta^*\|^2)^{-1/2}\theta^*. \tag{3.7.12}$$

Next let $\mathbf{X} = \{X_{ij}\}_{i=1,j=1}^{n\ \ p_n}$ and

$$\widetilde{\mathbf{X}}(\theta) \equiv \left(\widetilde{X}_{i,j}(\theta)\right)_{i=1,j=2}^{n,\ p_n} = \mathbf{X}\begin{pmatrix} \mathbf{J}^T(\theta^*) \\ \mathbf{I}_{(p_n-1)} \end{pmatrix}, \tag{3.7.13}$$

where $\widetilde{X}_{i,j}(\theta)$ are defined in (3.7.3). Let $\widetilde{\mathbf{X}} = \widetilde{\mathbf{X}}(\theta_0) = \left(\widetilde{\mathbf{X}}_{(1)}, \mathbf{X}_{(2)}\right)$.

Further we denote $\mathbf{g} = \mathrm{diag}\{g_{\theta_0}(X_{\theta_0,i}),\ i = 1, \ldots, n\}$, $\dot{\mathbf{g}} = \mathrm{diag}\{g'_{\theta_0}(X_{\theta_0,i}),\ i = 1, \ldots, n\}$, $\ddot{\mathbf{g}} = \mathrm{diag}\{g''_{\theta_0}(X_{\theta_0,i}),\ i = 1, \ldots, n\}$, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$, and $\boldsymbol{\varepsilon} = \mathrm{diag}(\varepsilon_1, \cdots, \varepsilon_n)$.

Substituting in the true parameter θ_0 into (3.7.4) and (3.7.9), one has

$$\widetilde{S}(\theta_0) = \left\{\widetilde{S}_j(\theta_0)\right\}_{j=2}^{p_n} = \begin{pmatrix} \widetilde{S}_{(1)}(\theta_0) \\ \widetilde{S}_{(2)}(\theta_0) \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\widetilde{\mathbf{X}}_{(1)}^T \dot{\mathbf{g}}\varepsilon \\ \frac{1}{n}\mathbf{X}_{(2)}^T \dot{\mathbf{g}}\varepsilon \end{pmatrix}, \tag{3.7.14}$$

$$\begin{aligned} \widetilde{\mathbf{H}}(\theta_0) &= \frac{1}{n}\left(\mathbf{J}(\theta_0^*), \mathbf{I}_{(p_n-1)}\right)\left\{\mathbf{X}^T(\dot{\mathbf{g}}^2 + \boldsymbol{\varepsilon}\ddot{\mathbf{g}})\mathbf{X} + (\theta_{0,1}^{-1}1^T\boldsymbol{\varepsilon}\dot{\mathbf{g}}X_1)\mathbf{I}_{p_n}\right\}\left(\mathbf{J}(\theta_0^*), \mathbf{I}_{(p_n-1)}\right)^T \\ &= \frac{1}{n}\widetilde{\mathbf{X}}^T(\dot{\mathbf{g}}^2 + \boldsymbol{\varepsilon}\ddot{\mathbf{g}})\widetilde{\mathbf{X}} + \frac{1}{n}\left(\mathbf{J}(\theta_0^*), \mathbf{I}_{(p_n-1)}\right)(\theta_{0,1}^{-1}1^T\boldsymbol{\varepsilon}\dot{\mathbf{g}}X_1)\mathbf{I}_{p_n}\left(\mathbf{J}(\theta_0^*), \mathbf{I}_{(p_n-1)}\right)^T. \end{aligned}$$

Applying the Law of Large Numbers, one has

$$\widetilde{\mathbf{H}}(\theta_0) = \mathbf{H}(\theta_0) + o_P(1),$$

where

$$\mathbf{H}(\theta_0) = \frac{1}{n}\widetilde{\mathbf{X}}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}} = \begin{pmatrix} \mathbf{H}_{(1,1)}(\theta_0) & \mathbf{H}_{(1,2)}(\theta_0) \\ \mathbf{H}_{(2,1)}(\theta_0) & \mathbf{H}_{(2,2)}(\theta_0) \end{pmatrix}, \tag{3.7.15}$$

with

$$\begin{aligned} \mathbf{H}_{(1,1)}(\theta_0) &= \frac{1}{n}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}, \quad \mathbf{H}_{(1,2)}(\theta_0) = \frac{1}{n}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\mathbf{X}_{(2)}, \\ \mathbf{H}_{(2,1)}(\theta_0) &= \frac{1}{n}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}, \quad \mathbf{H}_{(2,2)}(\theta_0) = \frac{1}{n}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}^2\mathbf{X}_{(2)}. \end{aligned}$$

3.7.3 Asymptotic Properties of the Oracle Estimator

Note that function $R(\theta)$ depends only on θ^*, thus, we use $R(\theta^*)$ and $\widehat{R}(\theta^*)$ instead of $R(\theta)$ and $\widehat{R}(\theta)$ in the following proof and define

$$R^o(\theta^*_{(1)}) = \frac{1}{2n}\sum_{i=1}^{n}\{Y_i - g(X^T_{i(1)}\theta_{(1)})\}^2,$$

and

$$\widehat{R}^o(\theta^*_{(1)}) = \frac{1}{2n}\sum_{i=1}^{n}\{Y_i - \widehat{g}_{\theta_{(1)}}(X^T_{i(1)}\theta_{(1)})\}^2. \tag{3.7.16}$$

Similarly, we can represent the score and Hessian matrices in (3.7.2) and (3.7.7) using θ^* as:

$$\widehat{S}(\theta^*) = \left\{\widehat{S}_j(\theta^*)\right\}_{j=2}^{p_n} = \begin{pmatrix} \widehat{S}_{(1)}(\theta^*) \\ \widehat{S}_{(2)}(\theta^*) \end{pmatrix}, \quad \widehat{\mathbf{H}}(\theta^*) = \begin{pmatrix} \widehat{\mathbf{H}}_{(1,1)}(\theta^*) & \widehat{\mathbf{H}}_{(1,2)}(\theta^*) \\ \widehat{\mathbf{H}}_{(2,1)}(\theta^*) & \widehat{\mathbf{H}}_{(2,2)}(\theta^*) \end{pmatrix}.$$

Lemma 3.4. *If Conditions (A1)–(A5) are satisfied, then there is a local minimizer* $\hat{\theta}^{o*}_{(1)}$ *of* $\hat{R}^o(\theta^*_{(1)})$ *such that* $\|\hat{\theta}^{o*}_{(1)} - \theta^*_{0(1)}\| = O_P\{(n^{-1}q_nN^3\log n)^{1/2}\}$.

Proof of Lemma 3.4. Let $\alpha_n = q_n^{1/2}n^{-1/2}N^{3/2}$ and set $\|u\| = C$, where C is a large enough constant. To show the existence of such an oracle minimizer, it is equivalent to prove that for any given ε there is a large constant C such that, for large n we have

$$P\left\{\inf_{\|u\|=C}\hat{R}^o(\theta^*_{0(1)} + \alpha_n u) > \hat{R}^o(\theta^*_{0(1)})\right\} \geq 1 - \varepsilon.$$

This implies that with probability tending to 1 there is an oracle minimizer $\hat{\theta}_{(1)}^{o*}$ in the ball $\{\theta_{0(1)}^{*}+\alpha_n u : \|u\| \leq C\}$ such that $\|\hat{\theta}_{(1)}^{o*}-\theta_{0(1)}^{*}\| = O_P(\alpha_n)$.

By Taylor's expansion, we obtain

$$\begin{aligned}
L(u) &= \hat{R}^o(\theta_{0(1)}^{*}+\alpha_n u)-\hat{R}^o(\theta_{0(1)}^{*}) \\
&= \alpha_n \left\{\frac{\partial}{\partial\theta_{(1)}^{*}}\hat{R}^o(\theta_{0(1)}^{*})\right\} u+\frac{1}{2}\alpha_n^2 u^T \left\{\frac{\partial^2}{\partial\theta_{(1)}^{*}\partial\theta_{(1)}^{*T}}\hat{R}^o(\theta_{0(1)}^{*})\right\} u \\
&\quad +\frac{1}{6}\alpha_n^3\frac{\partial}{\partial\theta_{(1)}^{*}}\left[u^T\left\{\frac{\partial^2}{\partial\theta_{(1)}^{*}\partial\theta_{(1)}^{*T}}\hat{R}^o(\bar{\theta})\right\}u\right]u \\
&= \alpha_n \hat{S}_{(1)}(\theta_0^{*})u+\frac{1}{2}\alpha_n^2 u^T\widehat{\mathbf{H}}_{(1,1)}(\theta_0^{*})u+\frac{1}{6}\alpha_n^3\frac{\partial}{\partial\theta_{(1)}^{*}}\left[u^T\left\{\frac{\partial^2}{\partial\theta_{(1)}^{*}\partial\theta_{(1)}^{*T}}\hat{R}^o(\bar{\theta})\right\}u\right]u \\
&= L_1(u)+L_2(u)+L_3(u),
\end{aligned}$$

where the vector $\overline{\theta}$ lies between $\theta_{0(1)}^{*}$ and $\theta_{0(1)}^{*}+\alpha_n u$. Note that $\frac{\partial}{\partial\theta_{(1)}^{*}}R^o(\theta_{(1)}^{*})=0$ at $\theta_{(1)}^{*}=\theta_{0(1)}^{*}$, by Assumptions (A1)–(A4) and Lemma 3.1 we have

$$\begin{aligned}
|L_1| &\leq \alpha_n\left\|\frac{\partial}{\partial\theta_{(1)}^{*}}\left\{\hat{R}^o\left(\theta_{0(1)}^{*}\right)-R^o\left(\theta_{0(1)}^{*}\right)\right\}\right\|\|u\| \\
&= \alpha_n\|u\|\times O_P\left\{(n^{-1}q_n N^3\log n)^{1/2}+q_n^{1/2}N^{-r+1}\right\} \\
&= O_P\left(\alpha_n^2\right)\|u\|. \qquad (3.7.17)
\end{aligned}$$

Next, we consider L_2,

$$\begin{aligned}
L_2 &= \frac{1}{2}u^T\left\{\frac{\partial^2}{\partial\theta_{(1)}^{*}\partial\theta_{(1)}^{*T}}\hat{R}^o(\theta_{0(1)}^{*})-\frac{\partial^2}{\partial\theta_{(1)}^{*}\partial\theta_{(1)}^{*T}}R^o(\theta_{0(1)}^{*})\right\}u\alpha_n^2+\frac{1}{2}u^T\left\{\frac{\partial^2}{\partial\theta_{(1)}^{*}\partial\theta_{(1)}^{*T}}R^o(\theta_{0(1)}^{*})\right\}u\alpha_n^2 \\
&= \frac{1}{2}u^T\left\{\widehat{\mathbf{H}}_{(1,1)}(\theta_0^{*})-\mathbf{H}_{(1,1)}(\theta_0^{*})\right\}u\alpha_n^2+\frac{1}{2}u^T\mathbf{H}_{(1,1)}(\theta_0^{*})u\alpha_n^2.
\end{aligned}$$

According to Lemma 3.1 and Assumption (A4), we have

$$\begin{aligned} |L_2| &\leq \frac{1}{2}u^T\mathbf{H}_{(1,1)}(\theta_0^*)u\alpha_n^2 + O\left\{\left\{(n^{-1}N^5\log n)^{1/2} + N^{-r+2}\right\}q_n\right\}\alpha_n^2\|u\|^2 \\ &= \frac{1}{2}u^T\mathbf{H}_{(1,1)}(\theta_0^*)u\alpha_n^2 + o_P\left(1\right)\times\alpha_n^2\|u\|^2. \end{aligned} \tag{3.7.18}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} L_3 &= \frac{1}{6}\alpha_n^3\frac{\partial}{\partial\theta_{(1)}^*}\left[u^T\left\{\frac{\partial^2}{\partial\theta_{(1)}^*\partial\theta_{(1)}^{*T}}\hat{R}^o(\theta_{(1)}^*)\right\}u\right]u \\ &\leq \frac{1}{6}\alpha_n^3\frac{\partial}{\partial\theta_{(1)}^*}\left[u^T\frac{\partial^2}{\partial\theta_{(1)}^*\partial\theta_{(1)}^{*T}}\left\{\hat{R}^o(\theta_{(1)}^*) - R^o(\theta_{(1)}^*)\right\}u\right]u \\ &\quad+\frac{1}{6}\alpha_n^3\frac{\partial}{\partial\theta_{(1)}^*}\left[u^T\left\{\frac{\partial^2}{\partial\theta_{(1)}^*\partial\theta_{(1)}^{*T}}R^o(\theta_{(1)}^*)\right\}u\right]u. \end{aligned}$$

Using the result in Lemma 3.1 again, together with Assumption (A4), implies that

$$\begin{aligned} |L_3| &\leq O_P\left(q_n^{3/2}\alpha_n\right)\alpha_n^2\|u\|^3 + O_P\left\{\left((n^{-1}N^7\log n)^{1/2} + N^{-r+3}\right)q_n^{3/2}\alpha_n\right\}\alpha_n^2\|u\|^3 \\ &= o_P\left(1\right)\times\alpha_n^2\|u\|^2. \end{aligned} \tag{3.7.19}$$

By equations (3.7.17)-(3.7.19), when $\|u\|$ is large enough, all terms L_1 and L_3 are dominated by a positive term L_2. Hence, Lemma 3.4 holds. ∎

3.7.4 Proof of Theorems 3.1 and 3.2

Proof of Theorem 3.1. By the second-order sufficiency of the Karush-Kuhn-Tucker condition in Bertsekas (1999), any θ satisfying

(C.1) $\widehat{S}_j(\theta^*) = 0$ and $|\theta^*_j| \geq a\lambda$ for $j = 1, \ldots, q_n - 1$,

(C.2) $|\widehat{S}_j(\theta^*)| \leq \lambda$ and $|\theta^*_j| < \lambda$ for $j = q_n, \ldots, p_n - 1$,

is an element of $A(\lambda)$. Thus, it suffices to show that the oracle estimator $\widehat{\theta}^{o*} = (\widehat{\theta}^{o*T}_{(1)}, 0^T)$ defined in (3.3.1) satisfies (C.1) and (C.2) with $\lambda = \lambda_n$.

Now, we consider the first condition in (C.1). By (3.3.1), $\widehat{\theta}^{*o}_{(1)}$ is the minimizer of (3.7.16) over all $\theta_{(1)} \in \{(\theta_1, \cdots, \theta_{q_n}) | \sum_{j=1}^{q_n} \theta_j^2 = 1, \theta_1 > c\}$, which implies that $\widehat{S}^o_j(\widehat{\theta}^{o*}) = \frac{\partial}{\partial \theta_j} \widehat{R}^o(\theta^*_{(1)}) = 0$ for any $j = 1, \ldots, q_n - 1$. Denote that $\dot{\mathbf{P}}_j = (\mathbf{I} - \mathbf{P}_\theta)\, \dot{\mathbf{B}}_j \left(\mathbf{B}_\theta^T \mathbf{B}_\theta\right)^{-1} \mathbf{B}_\theta^T$, then, we have for any $j = 1, \ldots, q_n - 1$

$$\widehat{S}_j(\widehat{\theta}^{o*}) = -\frac{1}{n}\left\{Y^T \dot{\mathbf{P}}_{\widehat{\theta}^o, j} Y - \widehat{\theta}^o_j \widehat{\theta}^{o-1}_1 Y^T \dot{\mathbf{P}}_{\widehat{\theta}^o, 1} Y\right\} = \widehat{S}^o_j(\widehat{\theta}^{o*}) = 0.$$

As a result, the first condition in (C.1) holds.

For the second condition in (C.1), so it suffices to show that as $n \to \infty$,

$$Pr(|\widehat{\theta}^{o*}_j| \geq a\lambda_n, \text{ for } j = 1, \ldots, q_n - 1) \to 1.$$

Note that $|\widehat{\theta}_j^{o*}| \geq |\theta_{0,j}^*| - |\widehat{\theta}_j^{o*} - \theta_{0,j}^*|$. According to Assumption (A6), $\min_{j=1,\ldots,q_n-1} |\theta_{0,j}^*| = o(n^{-(1-\alpha_2)/2})$ and $\lambda_n = o(n^{-(1+\alpha_1-\alpha_2)/2})$, it suffices to show that

$$\max_{j=1,\cdots,q_n-1} |\widehat{\theta}_j^{o*} - \theta_{0,j}^*| = o_P(n^{-(1-\alpha_2)/2}). \tag{3.7.20}$$

Let $\xi_j = \sqrt{n}(\widehat{\theta}_j^{o*} - \theta_{0,j}^*)$, then it is equivalent to show that

$$\max_{j=1,\ldots,q_n-1} |\xi_j| = o_P(n^{\alpha_2/2}).$$

Let $\xi = (\xi_1, \xi_2, \cdots, \xi_{q_n-1})^T$, then $\xi = -\sqrt{n}\widehat{\mathbf{H}}_{(1,1)}^{-1}(\bar{\theta}^*)\widehat{S}_{(1)}(\theta_0^*)$, where $\bar{\theta}^* = (\bar{\theta}_{(1)}^{*T}, 0^T)^T$ and $\bar{\theta}_{(1)}^*$ is between $\theta_{0(1)}^*$ and $\widehat{\theta}_{(1)}^{o*}$. By Lemmas 3.2 and 3.3,

$$\begin{aligned} \xi &= -\sqrt{n}\widehat{\mathbf{H}}_{(1,1)}^{-1}(\bar{\theta}^*)\left\{\widetilde{S}_{(1)}(\theta_0) + o_P(n^{-1/2})\right\} \\ &= -\sqrt{n}\{\widetilde{\mathbf{H}}_{(1,1)}(\bar{\theta}^*) + o_P(1)\}^{-1}\left\{\widetilde{S}_{(1)}(\theta_0) + o(n^{-1/2})\right\}. \end{aligned}$$

According to Lemma 3.4, $\|\widehat{\theta}_{(1)}^{o*} - \theta_{0(1)}^*\| = O_P\{(n^{-1}q_nN^3\log n)^{1/2}\}$. Let

$$\mathcal{C}_{\theta^*} = \{\theta^* = (\theta_{(1)}^{*T}, 0^T)^T : \|\theta_{(1)}^* - \theta_{0(1)}^*\| = O\{(n^{-1}q_nN^3\log n)^{1/2}\}\}.$$

Note that for any $1 \leq j, j' \leq q_n - 1$,

$$\sup_{\theta^*\in\mathcal{C}_{\theta^*}} |\widetilde{H}_{j,j'}(\theta^*) - \widetilde{H}_{j,j'}(\theta_0^*)| = o_P(1),$$

so $\widetilde{\mathbf{H}}_{(1,1)}(\theta^*) = \widetilde{\mathbf{H}}_{(1,1)}(\theta_0^*) + o_P(1)$. Thus,

$$\begin{aligned}
\xi &= -\sqrt{n}\{\widetilde{\mathbf{H}}_{(1,1)}(\theta_0^*) + o_P(1)\}^{-1}\left\{\widetilde{S}_{(1)}(\theta_0^*) + o(n^{-1/2})\right\} \\
&= -\sqrt{n}\{\mathbf{H}_{(1,1)}(\theta_0^*) + o_P(1)\}^{-1}\left\{\widetilde{S}_{(1)}(\theta_0^*) + o(n^{-1/2})\right\}.
\end{aligned}$$

Using the notations in (3.7.14) and (3.7.15), one has

$$\begin{aligned}
\xi &= -\frac{1}{\sqrt{n}}(n^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)})^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}\varepsilon + o_P(1) \\
&= \mathbf{D}_{(1)}^T\varepsilon + o_P(1),
\end{aligned}$$

where $\mathbf{D}_{(1)}^T \equiv (D_{(1),1}, \cdots, D_{(1),q_n-1})^T = -\sqrt{n}(\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)})^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}$. Note that

$$n^{-1}\mathbf{D}_{(1)}^T\mathbf{D}_{(1)} = (\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)})^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}(\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)})^{-1} = (\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)})^{-1}.$$

By Assumption (A1), one has $\|D_{(1),j}\|_2^2 \le M_2^{-1}$ for all $j = 1, \ldots, q_n - 1$. Hence, $E\{\sqrt{n}\widehat{S}_j(\widehat{\theta}^o)\}^{2k} < \infty$ and $E(\xi_j)^{2k} < \infty$ for all $j = 1, \ldots, q_n - 1$ because $E(\varepsilon_i)^{2k} < \infty$. Thus,

$$Pr(|\xi_j| > t) = O(t^{-2k}). \tag{3.7.21}$$

Therefore, for any $\delta > 0$, we can write

$$\begin{aligned}
&Pr\left(|\xi_j| > \delta n^{\alpha_2/2} \text{ for some } j = 1, \ldots, q_n - 1\right) \\
&\leq \sum_{j=1}^{q_n-1} Pr(|\xi_j| > \delta n^{\alpha_2/2}) \leq \sum_{j=1}^{q_n-1} \delta^{-1} n^{-\alpha_2 k} \\
&< \delta^{-1} q_n n^{-\alpha_2 k} \leq \delta^{-1} n^{-(\alpha_2 - \alpha_1)k} \to 0.
\end{aligned}$$

Hence, condition (C.1) holds for the oracle estimator.

We now proceed to show that the oracle estimator $\widehat{\theta}^{o*}$ also satisfies Condition (C.2). For the second part of (C.2), by the definition of $\widehat{\theta}^{o*}_j = 0$ for $j = q_n, \cdots, p_n - 1$, we have $|\widehat{\theta}^{o*}_j| \leq \lambda_n$.

For the first part, it suffices to show that

$$Pr\left\{|\widehat{S}_j(\widehat{\theta}^{o*})| > \lambda_n \text{ for some } j = q_n, \cdots, p_n - 1\right\} \to 0.$$

According to (3.7.20) and Assumptions (A4) and (A5), one has

$$\sqrt{n}\|\widehat{\theta}^{o*}_{(1)} - \theta^*_{0(1)}\|^2 = O_P(n^{-1/2} q_n N^3 \log n) = o_P(1).$$

In addition, by Lemma 3.2 and (3.7.4)

$$
\begin{aligned}
&\sqrt{n}\widehat{S}_{(2)}(\widehat{\theta}^{o*}) = \sqrt{n}\widetilde{S}_{(2)}(\widehat{\theta}^{o*}) + o_P(1) \\
&= \frac{1}{\sqrt{n}}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}\varepsilon - \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\left\{g'_{\widehat{\theta}^o}(X_i^T\widehat{\theta}^o) - g'(X_i^T\theta_0)\right\}\varepsilon_i\right]_{j=q_n+1}^{p_n} \\
&+ \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\left\{g_{\widehat{\theta}^o}(X_i^T\widehat{\theta}^o) - g(X_i^T\theta_0)\right\}\left\{g'_{\widehat{\theta}^o}(X_i^T\widehat{\theta}^o) - g'(X_i^T\theta_0)\right\}\right]_{j=q_n+1}^{p_n} \\
&+ \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\left\{g_{\widehat{\theta}^o}(X_i^T\widehat{\theta}^o) - g(X_i^T\theta_0)\right\}g'(X_i^T\theta_0)\right]_{j=q_n+1}^{p_n} + o_P(1) \\
&= \frac{1}{\sqrt{n}}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}\varepsilon + \left[\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\left\{g_{\widehat{\theta}^o}(X_i^T\widehat{\theta}^o) - g(X_i^T\theta_0)\right\}g'(X_i^T\theta_0)\right]_{j=q_n+1}^{p_n} + o_P(1).
\end{aligned}
$$

Note that

$$
\begin{aligned}
g_{\widehat{\theta}^o}(X_i^T\widehat{\theta}^o) - g(X_i^T\theta_0) \quad = \quad & g'(X_i^T\theta_0)X_{i(1)}^T\left(\mathbf{J}_{(1)}(\theta_0^*), \mathbf{I}_{(q_n-1)}\right)^T(\widehat{\theta}^o_{(1)} - \theta_{0(1)}) \\
& + \frac{1}{2}g''_{\bar{\theta}}(X_i^T\bar{\theta})(\widehat{\theta}^o_{(1)} - \theta_{0(1)})^T\left(\mathbf{J}_{(1)}(\theta_0^*), \mathbf{I}_{(q_n-1)}\right)X_{i(1)} \\
& \times X_{i(1)}^T\left(\mathbf{J}_{(1)}(\theta_0^*), \mathbf{I}_{(q_n-1)}\right)^T(\widehat{\theta}^o_{(1)} - \theta_{0(1)})
\end{aligned}
$$

for some $\bar{\theta} = (\bar{\theta}_{(1)}^T, 0^T)^T$, where $\bar{\theta}_{(1)}$ is between $\widehat{\theta}^o_{(1)}$ and $\theta_{0(1)}$. Thus,

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}g'(X_i^T\theta_0)\left\{g_{\widehat{\theta}^o}(X_i^T\widehat{\theta}^o) - g(X_i^T\theta_0)\right\} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\{g'(X_i^T\theta_0)\}^2 X_{i(1)}^T\left(\mathbf{J}_{(1)}(\theta_0^*), \mathbf{I}_{(q_n-1)}\right)^T(\widehat{\theta}^o_{(1)} - \theta_{0(1)}) + o_P(1)
\end{aligned}
$$

Using similar arguments as in the proof of (C.1), one has

$$\sqrt{n}(\widehat{\theta}^{o}_{(1)} - \theta_{0(1)}) = -\sqrt{n}\mathbf{H}^{-1}_{(1,1)}(\theta_0)\widetilde{S}_{(1)}(\theta_0) + o_P(1).$$

Thus,

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{ij} g'(X_i^T\theta_0)\left\{g_{\widehat{\theta}^o}(X_i^T\widehat{\theta}^{o}) - g(X_i^T\theta_0)\right\} \\ & = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{ij}\{g'(X_i^T\theta_0)\}^2 X_{i(1)}^T \left(\mathbf{J}_{(1)}(\theta_0^*), \mathbf{I}_{(q_n-1)}\right)^T \mathbf{H}^{-1}_{(1,1)}(\theta_0)\widetilde{S}_{(1)}(\theta_0) + o_P(1). \end{aligned}$$

Therefore, one has

$$\begin{aligned} \sqrt{n}\widehat{S}_{(2)}(\widehat{\theta}^{o}) &= \frac{1}{\sqrt{n}}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}\varepsilon - \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{ij}\{g'(X_i^T\theta_0)\}^2 X_{i(1)}^T \left(\mathbf{J}_{(1)}(\theta_0^*), \mathbf{I}_{(q_n-1)}\right)^T\right]_{j=q_n+1}^{p_n} \\ & \quad \times\mathbf{H}^{-1}_{(1,1)}(\theta_0)\widetilde{S}_{(1)}(\theta_0) + o_P(1) \\ &= \frac{1}{\sqrt{n}}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}\varepsilon - \frac{1}{\sqrt{n}}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}\left\{\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}\right\}^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}\varepsilon + o_P(1) \\ &= \frac{1}{\sqrt{n}}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}\left[\mathbf{I} - \dot{\mathbf{g}}\widetilde{\mathbf{X}}_{(1)}\left\{\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}\right\}^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}\right]\varepsilon + o_P(1). \end{aligned}$$

Next we write

$$\widehat{S}_j(\widehat{\theta}^{o}) = D_{(2),j}^T\varepsilon, \text{ for } j = q_n+1, \cdots, p_n,$$

with $D_{(2),j}^T$ being the j-th column of $\mathbf{D}_{(2)}^T$, where

$$\mathbf{D}_{(2)}^T = \frac{1}{\sqrt{n}}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}\left[\mathbf{I} - \dot{\mathbf{g}}\widetilde{\mathbf{X}}_{(1)}\left\{\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}\right\}^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}\right]$$

Note that

$$\begin{aligned}\mathbf{D}_{(2)}^T\mathbf{D}_{(2)} &= \frac{1}{n}\mathbf{X}_{(2)}^T\dot{\mathbf{g}}\left[\mathbf{I}-\dot{\mathbf{g}}\widetilde{\mathbf{X}}_{(1)}\left\{\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}\right\}^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}\right] \\ &\quad \times\left[\mathbf{I}-\dot{\mathbf{g}}\widetilde{\mathbf{X}}_{(1)}\left\{\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}\right\}^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}\right]\dot{\mathbf{g}}\mathbf{X}_{(2)}.\end{aligned}$$

If we let $\mathbf{W}=\dot{\mathbf{g}}\widetilde{\mathbf{X}}_{(1)}^T$, it is trivial to see that all of the eigenvalues of $\dot{\mathbf{g}}\widetilde{\mathbf{X}}_{(1)}\left\{\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}\right\}^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}=\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T$ and $\mathbf{I}-\dot{\mathbf{g}}\widetilde{\mathbf{X}}_{(1)}\left\{\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}^2\widetilde{\mathbf{X}}_{(1)}\right\}^{-1}\widetilde{\mathbf{X}}_{(1)}^T\dot{\mathbf{g}}=\mathbf{I}-\mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T$ are either 0 or 1. Thus one has $\|D_{(2),j}\|_2^2\le M_2$ for all $j=q_n+1,\cdots,p_n$. Hence, $E\{\sqrt{n}\widehat{S}_j(\widehat{\theta}^o)\}^{2k}<\infty$ and

$$Pr(|\sqrt{n}\widehat{S}_j(\widehat{\theta}^o)|>t)=O(t^{-2k}). \qquad (3.7.22)$$

Therefore,

$$\begin{aligned}& Pr(|\widehat{S}_j(\widehat{\theta}^o)|>\lambda_n \text{ for some } j=q_n+1,\cdots,p_n) \\ =\;& Pr(|\sqrt{n}\widehat{S}_j(\widehat{\theta}^o)|>\sqrt{n}\lambda_n \text{ for some } j=q_n+1,\cdots,p_n) \\ \le\;& \sum_{j=q_n}^{p_n} Pr(|\sqrt{n}\widehat{S}_j(\widehat{\theta}^o)|>\sqrt{n}\lambda_n) \\ =\;& (p_n-q_n)O\{(\sqrt{n}\lambda_n)^{-2k}\}=O\{p_n(\sqrt{n}\lambda_n)^{-2k}\}\to 0.\end{aligned}$$

Thus, this completes the proof. ∎

Proof of Theorem 3.2. For a Gaussian random variable Z with mean 0 and variance σ^2, we have

$$P(|Z|>t)<\sqrt{\frac{2}{\pi}}\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

for $t \geq \sigma$. Similar to [35], Theorem 3.2 can be simply proved by replacing the tail bounds of (3.7.21) and (3.7.22) with the above exponential bounds. ∎

Table 3.1: Selection results for Example 1

n	p_n	METHOD	TPN	FPN	C (%)	RMSPE	TIME (s)
100	1000	ORACLE	995.00	0.00	100.0	0.07	0.4
		NIS-PSIM	994.36	0.00	82.0	0.08	119.5
		SIS-SCAD	993.01	4.99	0.0	0.20	5.3
		HD-SCAD	994.99	4.98	0.0	0.21	975.7
200	1000	ORACLE	995.00	0.00	100.0	0.06	0.9
		NIS-PSIM	995.00	0.00	100.0	0.06	303.0
		SIS-SCAD	994.00	4.00	0.0	0.19	4.2
		HD-SCAD	995.00	3.91	0.0	0.21	3536.3
200	5000	ORACLE	4995.00	0.00	100.0	0.06	1.0
		NIS-PSIM	4994.91	0.00	65.0	0.06	1837.3
		SIS-SCAD	4992.97	5.00	0.0	0.18	250.2
		HD-SCAD	–	–	–	–	–

Table 3.2: Estimation results for Example 1

Est.	n	p_n	ORACLE			NIS-PSIM		
			BIAS	SD	RMSE	BIAS	SD	RMSE
θ_1	100	1000	-0.0014	0.009	0.009	-0.0265	0.111	0.114
θ_2			-0.0126	0.009	0.015	-0.0200	0.110	0.110
θ_3			0.0092	0.008	0.012	-0.0136	0.117	0.118
θ_4			-0.0057	0.009	0.011	-0.0211	0.101	0.105
θ_5			0.0061	0.009	0.011	-0.0054	0.092	0.095
θ_1	200	1000	0.0001	0.006	0.006	-0.0024	0.006	0.006
θ_2			-0.0007	0.006	0.006	0.0024	0.008	0.008
θ_3			0.0004	0.007	0.007	-0.0013	0.006	0.006
θ_4			0.0003	0.006	0.006	0.0001	0.006	0.005
θ_5			-0.0004	0.006	0.006	0.0009	0.006	0.005
θ_1	200	5000	-0.0004	0.006	0.006	-0.0312	0.016	0.035
θ_2			-0.0001	0.006	0.006	-0.0138	0.014	0.019
θ_3			-0.0002	0.006	0.006	-0.0270	0.015	0.031
θ_4			0.0005	0.006	0.006	-0.0379	0.016	0.041
θ_5			-0.0003	0.006	0.006	-0.0281	0.015	0.032

Table 3.3: Performance results of 100 random partitions of the data

p	NIS-PSIM		HD-SCAD		SIS-SCAD	
	MS	PE	MS	PE	MS	PE
3,000	11.22(.231)	.393(.011)	15.70(.417)	.471(.010)	2.00(.000)	.423(.009)
18,976	11.28(.282)	.396(.018)	—*	—*	2.00(.000)	.615(.010)

* We don't have results here because the computing time is more than 600 hours on a PC with Intel(R) Core(TM) i5-2500 CPU @ 3.30GHz and 8.00GB RAM.

Chapter 4

Conclusion

In Chapter 2, we consider the model selection for high-dimensional single-index prediction models for weakly dependent data. We apply the SCAD penalty and polynomial spline basis function expansion to perform variable selection and estimation simultaneously. We provide new statistical theory in the framework of a slowly diverging number of index parameters where the divergence rate is similar to that of parametric models in [19]. The proposed method has the following advantages and properties: (1) under regularity conditions, the proposed method is shown to have the "oracle" property when the number of parameters tends to infinity as the sample size increases; (2) both the variable selection and estimation are robust against deviations from the genuine single-index models; (3) the implemented algorithm is fast and efficient because it takes advantage of global spline smoothing as well as the iterative method; (4) our method is useful for selection of significant predictors not only for independent data but also for weakly dependent time series data.

In Chapter 3, we have developed a fast and efficient variable selection process for ultra-high dimensional single-index models. We also have proved the new theoretical result for the oracle estimators of the single-index coefficients. The numerical results given in Section 3.5 show that the proposed NIS-PSIM estimator is comparable to the oracle estimator in terms of MPE. In addition, in this dissertation, we have considered not only the situation where p is not very large, but also $p \gg n$ under the sparsity assumption. In the big data era, both the sample size and the dimensionality could be extremely large. An important direction of future research is to develop a efficient way to analyze such new “big data” sets.

Bibliography

[1] Bai, Z. D., Rao, C. R. and Wu, Y. 1999. Model selection with data-oriented penalty. Journal of Statistical Planning and Inference. 77, 103–117.

[2] Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 19, 185–93.

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

[4] Carroll, R., Fan, J., Gijbles, I. and Wang, M. P. 1997. Generalized partially linear single-index models. Journal of the American Statistical Association. 92, 477–489.

[5] Chang, Z., Xue, L. and Zhu, L. 2010. On an asymptotically more efficient estimation of the single-index model. Journal of Multivariate Analysis. 101, 1898–1901.

[6] Chen, H. 1991. Estimation of a projection – pursuit type regression model. Annals of Statistics. 19, 142–157.

[7] Cui, X., Härdle, W. K. and Zhu, L. 2011. The EFM approach for single-index models. Annals of Statistics 39, 1658–1688.

[8] deBoor, C. 2001. A Practical Guide to Splines. Springer-Verlag, New York.

[9] DeVore, R. A. and Lorentz, G. G. 1993. Constructive Approximation. Springer, New York.

[10] Donoho, D. L. and Elad, M. 2003. Optimally sparse representation in general (nonorthogonal) dictionaries via L1 minimization. Proceedings of the National Academy of Science. 100, 2197–2202.

[11] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004. Least angle regression. Annals of Statistics. 32, 407–499.

[12] Fan, J. 1997. Comments on "Wavelets in statistics: A review". Journal of the Italian Statistical Society. 6, 131–138.

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

[14] Fan, J. and Gijbels, I. 1996. Local polynomial modelling and its applications. Chapman & Hall.

[15] Fan, J. and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its orable properties. Journal of the American Statistical Association. 96, 1348–1360.

[16] Fan, J. and Lv, J. 2008. Sure independence screening for ultra-high dimensional feature space. Journal of the Royal Statistical Society: Series B. 70, 849–911.

[17] Fan, J. and Lv, J. 2011. Nonconcave penalized iikelihood with NP-dimensionality. IEEE Transactions on Information Theory. 57, 5467–5484.

[18] Fan, J., Ma, Y. and Dai, W. 2014. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. Journal of the American Statistical Association. 109, 1270–1284

[19] Fan, J. and Peng, H. 2004. Nonconcave penalized likelihood with a diverging number of parameters. Annals of Statistics 32, 928–961.

[20] Fan, J., Samworth, R. and Wu, Y. 2009. Sure independence screening in generalized linear models with NP dimensionality. Annals of Statistics. 38, 3567–3604.

[21] Fan, J. and Song, R. 2010. Ultra-high dimensional feature selection: beyond the linear model. Journal of Machine Learning Research. 10, 2013–2038.

[22] Fan, J. and Yao, Q. 2003. Nonlinear Time Series: Nonparametric and Parametric Methods. New York: Springer-Verlag.

[23] Frank, I. E. and Friedman, J. H. 1993. A statistical view of some chemometrics regression tools. Technometrics. 35, 109–148.

[24] Friedman, J. H. and Stuetzle, W. 1981. Projection pursuit regression. Journal of the American Statistical Association. 76, 817–823.

[25] Hall, P. and Miller,H. 2009. Using generalized correlation to affect variable selection in very high dimensional problems. Journal of Computational and Graphical Statistics. 18, 533–550.

[26] Hall, P., Titterington, D. M. and Xue, J. H. 2009. Tilting methods for assessing the influence of components in a classifier. Journal of the Royal Statistical Society, Series B. 71, 7830803.

[27] Härdle, W. 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge.

[28] Härdle, W., Hall, P. and Ichimura, H. 1993. Optimal smoothing in single-index models. Annals of Statistics. 21, 157–178.

[29] Härdle, W. and Stoker, T. M. 1989. Investigating smooth multiple regression by the method of average derivatives. Journal of the American Statistical Association. 84, 986–995.

[30] Hristache, M., Juditski, A. and Spokoiny, V. 2001. Direct estimation of the index coefficients in a single-index model. Annals of Statistics. 29, 595–623.

[31] Huang, J., Horowitz, J. and Ma, S. 2008. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Annals of Statistics. 36, 587–613.

[32] Huang, J., Ma, S. and Zhang, C. 2008. Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica, 18, 1603–1618.

[33] Huber, P. J. 1985. Projection pursuit (with discussion). Annals of Statistics. 13, 435–525.

[34] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. 4, 249–264.

[35] Kim, Y., Choi, H. and Oh, H. S. 2008. Smoothly clipped absolute deviation on high dimensions. Journal of the American Statistical Association. 103, 1665–1673.

[36] Kong, E. and Xia, Y. 2007. Variable selection for the single-index model. Biometrika. 94, 217–229.

[37] Li, G., Peng, H., Zhang, J. and Zhu, L. 2012. Robust rank correlation based screening. Annals of Statistics. 40, 1846–1877.

[38] Li, R., Zhong, W. and Zhu, L. 2012. Feature screening via distance correlation learning. Journal of the American Statistical Association. 107, 1129–1139.

[39] Liang, H., Liu, X., Li, R. and Tsai, C. L. 2010. Estimation and testing for partially linear single-index models. Annals of Statistics. 38, 3811–3836.

[40] Liu, J., Zhang, R., Zhao, W. and Lv, Y. 2013. A robust and efficient estimation method for single index models. Journal of Multivariate Analysis. 122, 226–238.

[41] Peng, H. and Huang, T. 2011. Penalized least squares for single index models. Journal of Statistical Planning and Inference 141, 1362–1379.

[42] Pham, D. T. 1986. The mixing properties of bilinear and generalized random coefficient autoregressive models. Stochastic Analysis and Applications 23, 291–300.

[43] Powell, J. L., Stock, J. H. and Stoker, T. M. 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.

[44] Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C. and Stone, E. M. 2006. Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences. 103, 14429–14434.

[45] Shen, X. T. and Ye, J. M. 2002. Adaptive model selection. Journal of the American Statistical Association. 97, 210–221.

[46] Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 58, 267–288.

[47] Tong, H. 1990. Nonlinear Time Series: A Dynamical System Approach. Oxford University Press, Oxford, U.K.

[48] Wang, H., Li, B. and Leng, C. 2009. Shrinkage tuning parameter selection with a diverging number of parameters. Journal of the Royal Statistical Society: Series B. 71, 671–683.

[49] Wang, L., Kim, Y. and Li, R. 2013. Scalibrating nonconvex penalized regression in untra-high dimension. Annals of Statistics. 41, 2505–2536.

[50] Wang, G. and Wang, L. 2015. Spline estimation and variable selection for single-index prediction models with diverging number of index parameters. Journal of Statistical Planning and Inference. 162, 1–19.

[51] Wang, L. and Xue, L. and Qu, A. and Liang, H. 2014. Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. Annals of Statistics. 42, 592–624.

[52] Wang, L. and Yang, L. 2009. Spline estimation of single-index models. Statistica Sinica. 19, 765–783.

[53] Wang, L., Zhou, J. and Qu, A. 2012. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics. 68, 353–360.

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

[55] Xia, Y., Tong, H., Li, W. K. and Zhu, L. 2002. An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society: Series B. 64, 363–410.

[56] Xue, L. and Yang, L. 2006. Additive coefficient modelling via polynomial spline. Statistica Sinica. 16, 1423–1446.

[57] Zhao, S. D. and Li, Y. 2012. Principled sure independence screening for Cox models with ultra-high dimensional covariates. Journal of Multivariate Analysis. 105, 397–411.

[58] Zhang, C. H. 2010. Nearly unbiased variable selection under minimax concave penalty. Annals of Statististics. 38, 894–942.

[59] Zhang, R., Huang, Z. and Lv, Y. 2010. Statistical inference for the index parameter in single-index model. Journal of Multivariate Analysis. 101, 1026–1041.

[60] Zeng, P., He, T. and Zhu, Y., 2012. A lasso-type approach for estimation and variable selection in single index models. Journal of Computational Graphical Statistics. 21, 92–109.

[63] Zhu, L., Qian, L. and Lin, J. 2011. Variable selection in a class of single-index models. Annals of the Institute of Statistical Mathematics. 63, 1277–1293.

[62] Zhu, L. and Zhu, L. 2009. Nonconcave penalized inverse regression in single-index models with high dimension predictors. Journal of Multivariate Analysis. 100, 862–875.

[63] Zhu, L., Qian, L. and Lin, J. 2011. Variable selection in a class of single-index models. Annals of the Institute of Statistical Mathematics. 63, 1277–1293.

[64] Zou, H. 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 101, 1418–1429.

[65] Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 67, 768–768.

[66] Zou, H. and Li, R. 2008. One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics. 36, 1509–1533.