Statistica Sinica Preprint No: SS-2021-0009	
Title	Structured Ultrahigh Dimensional Multiple-Index Models With Efficient Estimation in Computation and Theory
Manuscript ID	SS-2021-0009
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0009
Complete List of Authors	Huazhen Lin, Shuangxue Zhao, Li Liu and Wenyang Zhang
Corresponding Author	Li Liu
E-mail	lliu.math@whu.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Structured Ultrahigh Dimensional Multiple-Index Models with Efficient Estimation in Computation and Theory

Huazhen Lin[a], Shuangxue Zhao[a], Li Liu[b], Wenyang Zhang[c]

[a] *Center of Statistical Research and School of Statistics,*

Southwestern University of Finance and Economics, Chengdu, China

[b] *School of Mathematics and Statistics, Wuhan University, Wuhan, China*

[c] *Department of Mathematics, The University of York, York, United Kingdom*

Abstract: **In this paper, we propose a structured multiple-index model (SMIM) for ultrahigh-dimensional data analysis. The proposed model takes many commonly used semiparametric models as its special cases, such as stochastic frontier models, single-index model, additive-index models etc. We estimate all of functions and parameters based on full likelihood-type function. As a result, the proposed estimators are shown to be semiparametrically efficient in the sense of Bickel et al. (1993), as well as consistent in selection and estimation and asymptotically normal. The computation is challenging due to the combination of nonconvexity of the likelihood function, nonsmoothness of the penalty term, and the large number of functions. To solve the computational problem, we develop a technique of blending spline and kernel smoothing with a majorized coordinate descendent algorithm, so that the implementation is**

easily performed by using the existing packages. Intensive simulation studies also show that the proposed estimation procedure outperforms its alternatives for various cases. Finally, we apply the proposed SMIM together with the proposed estimation procedure to a real dataset from one of China's largest liquor companies, and successfully identify the 31, from 2051, most important factors affecting the sale of liquor.

Key words and phrases: High-dimensional covariates, Maximum likelihood estimation, Semiparametrical efficiency, Structured multiple-index models, Variable selection.

1. Introduction

Modern technologies yield abundant data with ultrahigh-dimensional risk predictors from diverse scientific fields. The estimation and variable selection of ultrahigh-dimensional risk predictors are very sensitive to model identification. Particularly, parametric models may lead to biased estimation and selection due to the risk of misspecification, whilst nonparametric models could suffer from the uninterpretability and instability of the resulting estimators due to "curse of dimensionality". Semiparametric modelling comes as a sensible compromise. The multi-index models, stimulated by dimension reduction, are important semiparametric models, and enjoy good asymptotic properties. However, when it comes to application, they are not as useful as expected, because they would still come up against "curse of dimensionality" when the number of indices is even moderate, say 3 or larger. A more useful approach would be kind of structured multi-index models (SMIM). In the paper, motivated by the multi-index stochastic frontier model described later, we consider the following model with known link:

$$Y = m\left\{f_1(\mathbf{X}'\boldsymbol{\beta}_1),\ \cdots,\ f_d(\mathbf{X}'\boldsymbol{\beta}_d),\ \boldsymbol{\varepsilon}\right\}, \tag{1.1}$$

where Y is a response variable, $\mathbf{X}$ is a p_n-dimensional vector of covariate, m is a known link function of $(d+1)$ variables, f_js are unknown functions, $\boldsymbol{\beta}_j$s are unknown vectors, and $\boldsymbol{\varepsilon}$ is a vector including random error and some latent variables. To make model (1.1) identifiable, we assume, throughout this paper, that $\|\boldsymbol{\beta}_j\| = 1$ and the first component of $\boldsymbol{\beta}_j$ is positive for $j = 1,\ \cdots,\ d$.

The model (1.1) is structured by specifying the link function m which is helpful to cooperate the information of the type of Y and can be seen from the special cases of the model (1.1). Model (1.1) includes many commonly used models. For example, the index heteroscedastic model (Zhu et al., 2013) for continuous response, $Y = f_1(\mathbf{X}'\boldsymbol{\beta}_1) + f_2(\mathbf{X}'\boldsymbol{\beta}_2)\varepsilon$; and the generalized additive-index model for various type of response, namely, Y follows the exponential family distribution with mean $m\left\{\sum_{j=1}^{d} f_j(\mathbf{X}'\boldsymbol{\beta}_j)\right\}$, where $f_j(\cdot)$s are unknown functions, m is a known link function and determined by the type of Y, e.g., a logit link for binary response, a logarithmic function for count response and a linear function for continuous response. Furthermore, the generalized additive-index models take many commonly used models as their special cases, such as single index models and partial linear models. Literatures about this kind of models include Carroll et al. (1997); Xia (2008); Ma and Zhu (2013); Liu et al. (2016); Guo et al. (2017); Ke et al. (2020); Lian et al. (2021), and the references therein. Except for the single index models, the existing works focus on fixed dimension of $\mathbf{X}$.

The model (1.1) can not been straightforward addressed by the existing methods. Particularly, the existing studies on multiple-index models focus on the case of the fixed dimension of covariates. The methods for high-dimensional single-index model give estimation and selection and establish the asymptotical properties by avoiding the estimation of the unknown link function so that the objective function involves only high-dimensional parameters. The strategy for the high-dimensional single-index model does not work for the model (1.1), which has multiple indexes and specific structure. In the paper, we provide a semiparametrically efficient and computationally convenient estimator for all of parameters and functions in high-dimensional SMIM. The new estimation procedure is easy to implement, and simulation studies show that it beats all its alternatives for the models from the existing literature. On the theoretical front, we will show the estimators achieve the semiparametric efficiency, in the sense of Bickel et al. (1993), which has not been discussed for any high-dimensional semiparametric models as we have known.

As mentioned above, the study also is motivated by the multi-index stochastic frontier model, which is used to analyze our real data from one of China's largest liquor companies in western China. The purpose of the analysis is investigating whether and how various factors affect the mean, frontier, inefficiency and uncertainty of sales of liquor. The covariates include four parts: (1) the company's product information; (2) brewing industry information; (3) economic information of related cities and towns; and (4) geographic information. Together with the lagged variables, we have 2051 covariates and $n = 1941$ observations. The problem of measuring production inefficiency has been an important issue in the economic, political and social fields. One of the most satisfactory models to analyse the problem is the stochastic frontier model introduced by Aigner et al. (1977), which is written as follows:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\beta}) + \alpha_i + \varepsilon_i, \alpha_i \leq 0, i = 1, \cdots, n, \tag{1.2}$$

where $\mathbf{X}_i$ is the covariate with fixed dimension, $\boldsymbol{\beta}$ is an unknown vector, ε_i is a noise with normal distribution and α_i is an unobservable random variable which represents firm specific technical inefficiency. α_i, ε_i and $\mathbf{X}_i$ are assumed to be independent. The density of α_i is considered to have the support $(-\infty, 0)$ and is assumed from an $N(0, 1)$ distribution truncated at 0, that is, $\alpha_i \sim -|N(0, 1)|$. This means that, neglecting the noise, $f(\mathbf{x}, \boldsymbol{\beta})$ is the maximum attainable output with input $\mathbf{x}$, called stochastic frontier function.

In the analysis of the model (1.2), a parametric functional form for f, which is usually linear in $\boldsymbol{\beta}$, has become a standard practice in efficiency measurement studies. Since the misspecification in f may lead to erroneous conclusions, Fan et al. (1996) considered model (1.2) with completely unspecified $f(\cdot)$. Kumbhakar et al. (2007) further generalized Fan et al. (1996) by allowing the variances of inefficiency score α_i and measurement error ε_i depending on $\mathbf{X}_i$ without making any assumption on the variance functions. As a result, the problem of the curse of dimensionality may arise in Fan et al. (1996) and Kumbhakar et al. (2007), even when the dimension of covariates is slightly large, say larger than 3.

As a compromise between parametric and nonparametric modeling, we hence consider the following high-dimensional multiple-index stochastic frontier model,

$$Y_i = f_1(\mathbf{X}_i'\boldsymbol{\beta}_1) + f_2(\mathbf{X}_i'\boldsymbol{\beta}_2)\alpha_i + f_3(\mathbf{X}_i'\boldsymbol{\beta}_3)\varepsilon_i, \alpha_i \leq 0, i = 1, \cdots, n, \tag{1.3}$$

where the dimension p_n of $\mathbf{X}_i$ can be much larger than n, $f_1(\cdot)$, $f_2(\cdot)$ and $f_3(\cdot)$ are unknown functions, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are unknown coefficients representing the effect of $\mathbf{X}_i$ on the frontier, inefficiency and variance functions, respectively. Particularly, the covariates which affect the frontier, inefficiency and variance may be different. By identifying the zero component in $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$, we can select the subsets of $\mathbf{X}_i$ that are significant for frontier, inefficiency and variance, respectively. It is also remarkable that all unknown functions, $f_1(\cdot)$, $f_2(\cdot)$ and $f_3(\cdot)$, are one-dimensional, so it circumvents the problem of fitting high-dimensional surfaces and avoids the so-called curse of dimensionality. It is easy to see model (1.3) is a special case of (1.1) with $\boldsymbol{\varepsilon}_i = (\alpha_i,\ \varepsilon_i)'$.

In this paper, we will focus on ultrahigh dimensional setting for (1.1) with $p_n \gg n$, specifically, $\log(p_n) = O(n^r)$, $0 < r < 1$. Although models (1.1) can be viewed as an unified framework to accommodate some commonly used models, we would like to stress the aim of this paper is not only to build a unified framework, but also to develop new and efficient estimation procedure which applies to any model in the unified framework.

The paper is organised as follows. We begin in Section 2 with a description of the proposed estimation procedures and the algorithm to implement them. In Section 3 we present the asymptotic properties of the resulting estimators, and demonstrate that the estimators achieve the semiparametric efficiency. The performance of the proposed estimation procedures is assessed by simulation studies in Section 4. Through simulation we examine how well the proposed estimation procedures work. In Section 5, we apply the proposed SMIM model together with the proposed one-step estimation procedure to the real dataset from one of China's largest liquor companies to explore the important factors affecting the sale of liquor. Technical proofs are relegated to the Appendix. A user-friendly R package to implement the proposed method has been made available at https://github.com/LinhzLab/SMIM2.git.

2. Estimation procedure

We first introduce some notations. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_d')'$ and $\boldsymbol{f} = (f_1(\cdot), \cdots, f_d(\cdot))'$. To present the proposed estimation procedure in a more generic way, we assume the objective function, based on (1.1), for the estimation is

$$L(\boldsymbol{\beta}, \boldsymbol{f}) = \frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ f_1(\mathbf{X}_i'\boldsymbol{\beta}_1),\ \cdots,\ f_d(\mathbf{X}_i'\boldsymbol{\beta}_d)\Big). \tag{2.4}$$

When the distribution of $\boldsymbol{\varepsilon}$ is given, this objective function is the conditional log-likelihood function given $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$. When the distribution of $\boldsymbol{\varepsilon}$ is unknown, it is some kind of negative loss function. For example, in model (1.3), when $\alpha_i \sim -|N(0,1)|, \varepsilon_i \sim N(0,1)$ and α_i, ε_i and $\mathbf{X}_i$ are independent, the objective function $L(\boldsymbol{\beta}, \boldsymbol{f}) = n^{-1}\sum_{i=1}^n Q(Y_i, f_1(X_i'\boldsymbol{\beta}_1), f_2(X_i'\boldsymbol{\beta}_2), f_3(X_i'\boldsymbol{\beta}_3))$, where

$$Q(y, v_1, v_2, v_3) = -\frac{1}{2}\log\left(v_2^2 + v_3^2\right) - \frac{(y-v_1)^2}{2(v_2^2+v_3^2)} + \log\left(1 - \Phi\left\{\frac{(y-v_1)\,v_2}{v_3\sqrt{v_2^2+v_3^2}}\right\}\right)$$

with Φ being the standard normal distribution function. Without any confusion, throughout this paper, we call $L(\boldsymbol{\beta}, \boldsymbol{f})$ log likelihood function.

2.1 Kernel estimation

The proposed kernel estimation is based on the ideas of back-fitting and profile likelihood estimation. The details are as follows. Pretending $\boldsymbol{\beta}_k$s are known, we apply the idea of back-fitting to estimate $f_k(\cdot)$s. Specifically, Step I. we assume $f_j(\cdot) = f_j^{[\ell+1]}(\cdot), \quad j = 1, \cdots, k-1,$ $f_j(\cdot) = f_j^{[\ell]}(\cdot), \quad j = k+1, \cdots, d$, just after the ℓth iteration. In the $\ell+1$th iteration, we update $f_k(\cdot)$ in the following way. For each given k, $k = 1, \cdots, d$, and any given x, by the Taylor's expansion, we have $f_k(\mathbf{X}_i'\boldsymbol{\beta}_k) \approx f_k(x) + \dot{f}_k(x)(\mathbf{X}_i'\boldsymbol{\beta}_k - x) \widehat{=} \eta_{kx1} + \eta_{kx2}(\mathbf{X}_i'\boldsymbol{\beta}_k - x)$, when $\mathbf{X}_i'\boldsymbol{\beta}_k$ is in $B(x)$, a small neighbourhood of x. In the other words,

$$f_k(\mathbf{X}_i'\boldsymbol{\beta}_k) \approx \{\eta_{kx1} + \eta_{kx2}(\mathbf{X}_i'\boldsymbol{\beta}_k - x)\} I_{ik}(x) + f_k(\mathbf{X}_i'\boldsymbol{\beta}_k)\{1 - I_{ik}(x)\}, \quad (2.5)$$

for any $i = 1, \cdots, n$ and $k = 1, \cdots, d$, where $I_{ik}(x) = I(\mathbf{X}_i'\boldsymbol{\beta}_k \in B(x))$. Using (2.5), we extract information on $(f_k(x), \dot{f}(x))$ from all of the samples $i = 1, \cdots, n$. Substituting (2.5) into $L(\boldsymbol{\beta}, \boldsymbol{f})$, we estimate $\boldsymbol{\eta}_{kx} = (\eta_{kx1}, \eta_{kx2})'$ based on the following log likelihood function for $\boldsymbol{\eta}_{kx}$,

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ f_1(\mathbf{X}_i'\boldsymbol{\beta}_1),\ \cdots,\ f_{k-1}(\mathbf{X}_i'\boldsymbol{\beta}_{k-1}),\ W_{ix}(\boldsymbol{\beta}_k)'\boldsymbol{\eta}_{kx} I_{ik}(x) \\ &\qquad + f_k(\mathbf{X}_i'\boldsymbol{\beta}_k)\{1 - I_{ik}(x)\}, f_{k+1}(\mathbf{X}_i'\boldsymbol{\beta}_{k+1}),\ \cdots,\ f_d(\mathbf{X}_i'\boldsymbol{\beta}_d)\Big), \end{aligned} \quad (2.6)$$

where $W_{ix}(\boldsymbol{\beta}_k) = (1,\ \mathbf{X}_i'\boldsymbol{\beta}_k - x)'$. It is worthy to mention that, with the approximation (2.5), our estimation for $\boldsymbol{\eta}_{kx}$ is based on full likelihood function, rather than local likelihood function which is commonly used in the nonparametric literature (Fan et al., 2006). Differentiating (2.6) with respect to $\boldsymbol{\eta}_{kx}$ and noting that $I_{ik}(x)(1 - I_{ik}(x)) = 0$, we estimate $\boldsymbol{\eta}_{kx}$ by solving the following equations,

$$L_k(\boldsymbol{\beta}, \boldsymbol{f}; x)\widehat{=}\frac{1}{n}\sum_{i=1}^{n} Q^{(01,k)}\Big(Y_i, f_1(\mathbf{X}_i'\boldsymbol{\beta}_1), \cdots, f_{k-1}(\mathbf{X}_i'\boldsymbol{\beta}_{k-1}), W_{ix}(\boldsymbol{\beta}_k)'\boldsymbol{\eta}_{kx}, \\ f_{k+1}(\mathbf{X}_i'\boldsymbol{\beta}_{k+1}), \cdots, f_d(\mathbf{X}_i'\boldsymbol{\beta}_d)\Big) W_{ix}(\boldsymbol{\beta}_k) K_{ix}(\boldsymbol{\beta}_k) = 0, \tag{2.7}$$

with the indicator function $I_{ik}(x)$ replaced by a kernel function $K_{ix}(\boldsymbol{\beta}_k) = K_{h_k}(\mathbf{X}_i'\boldsymbol{\beta}_k - x)$, where h_k is a bandwidth and $Q^{(01,k)}(y, \mathbf{v})$ is component k of $\partial Q(y, \mathbf{v})/\partial \mathbf{v}$. By (2.7), the updated $f_k(x)$, $f_k^{[\ell+1]}(x)$, is obtained.

Step II. Continue Step I until convergence. We denote the converged $f_k^{[\ell]}(\cdot)$ by $\widehat{f_k^{Ker}}(\cdot; \boldsymbol{\beta})$.

We consider the estimation for $\boldsymbol{\beta}$. The covariates are ultrahigh-dimensional and an extra task is to select the important covariates. Replacing $f_k(\cdot)$s in (2.4) by $\widehat{f_k^{Ker}}(\cdot; \boldsymbol{\beta})$s and applying penalised estimation, we have penalised likelihood,

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \widehat{f_1^{Ker}}(\mathbf{X}_i'\boldsymbol{\beta}_1; \boldsymbol{\beta}),\ \cdots,\ \widehat{f_d^{Ker}}(\mathbf{X}_i'\boldsymbol{\beta}_d; \boldsymbol{\beta})\Big) - \sum_{k=1}^{d}\sum_{j=1}^{p_n} \lambda_n \rho_{\lambda_n}(|\beta_{kj}|). \tag{2.8}$$

where β_{kj} is the jth component of $\boldsymbol{\beta}_k$, λ_n is a tuning parameter, and $\rho_{\lambda_n}(\cdot)$ is a penalty function. Maximise (2.8) with respect to $\boldsymbol{\beta}_k$s subject to $\|\boldsymbol{\beta}_k\| = 1$ and $\beta_{k1} > 0, \quad k = 1, \cdots, d$. We use the resulting maximiser to estimate $\boldsymbol{\beta}_k$s, and denote them by $\widehat{\boldsymbol{\beta}_k^{Ker}}$s. Let $\widehat{\boldsymbol{\beta}^{Ker}}$ be $\boldsymbol{\beta}$ with each $\boldsymbol{\beta}_k$ being replaced by $\widehat{\boldsymbol{\beta}_k^{Ker}}$. We use $\widehat{\boldsymbol{f}^{Ker}}(\cdot; \widehat{\boldsymbol{\beta}^{Ker}})$ to estimate $\boldsymbol{f}(\cdot)$, and denote it by $\widehat{\boldsymbol{f}^{Ker}}(\cdot)$, with $\widehat{f_k^{Ker}}(\cdot)$ being the kth component.

Whilst the kernel estimation enjoys good asymptotic properties, including consistency, asymptotical normality and semiparametric efficiency, which are established in Section 3, it is difficult to implement it. In the paper, we provide an algorithm that is practicable in computation and, at the same time, has the same asymptotic properties as the kernel estimation.

2.2 Algorithm

The asymptotic theories for nonparameter estimators based on kernel smoothing or local-polynomial smoothing are better understood and established than those based on spline smoothing, whilst the computation based on spline smoothing is more simple than those based on kernel smoothing, the algorithm introduced in this subsection hence is an one-step kernel estimation based on the estimators obtained from a B-spline method.

2.2.1 B-spline estimation

We denote $\mathcal{U}$ to be the bounded support set of $\mathbf{X}\boldsymbol{\beta}_k$, as defined in (C2) of Supplementary Materials. Letting $\boldsymbol{B}(\cdot) = (B_{1,m}(\cdot), \cdots, B_{q_n,m}(\cdot))'$ be the vector of B-spline basis functions on $\mathcal{U}$, we have

$$f_k(x) \approx f_{k,n}(x) = \boldsymbol{B}(x)'\boldsymbol{\theta}_k, \quad k = 1, \cdots, d, \tag{2.9}$$

where $\boldsymbol{\theta}_k = (\theta_{k1}, \cdots, \theta_{kq_n})'$. Replacing the $f_k(\cdot)$s in (2.4) by their approximations using (2.9) leads to the following penalised objection function for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_k$s

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_1)'\boldsymbol{\theta}_1,\ \cdots,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_d)'\boldsymbol{\theta}_d\Big) - \sum_{k=1}^{d}\sum_{j=1}^{p_n}\lambda_n\rho_{\lambda_n}(|\beta_{kj}|), \qquad (2.10)$$

where β_{kj} is the jth component of $\boldsymbol{\beta}_k$. Maximise (2.10) with respect to $\boldsymbol{\beta}_k$s and $\boldsymbol{\theta}_k$s subject to $\|\boldsymbol{\beta}_k\| = 1$ and $\beta_{k1} > 0$, and denote the maximiser as $\tilde{\boldsymbol{\beta}}_k$s and $\tilde{\boldsymbol{\theta}}_k$s. The initial estimators of $f_k(\cdot)$ and $\boldsymbol{\beta}_k$ are taken to be $\tilde{f}_k(\cdot) = \boldsymbol{B}(\cdot)'\tilde{\boldsymbol{\theta}}_k$ and $\tilde{\boldsymbol{\beta}}_k$.

2.2.2 One-step kernel estimation

To ensure the good asymptotic properties, we update the B-spline estimation $\tilde{f}_k(\cdot)$s and $\tilde{\boldsymbol{\beta}}_k$s by an one-step kernel estimation. We estimate $f_k(\cdot)$s first and then $\boldsymbol{\beta}_k$s.

For each k, $k = 1, \cdots, d$, and any given x, replacing $\boldsymbol{\beta}_j$ in (2.4) by $\tilde{\boldsymbol{\beta}}_j$, $j = 1, \cdots, d$, and $f_j(\cdot)$ by $\tilde{f}_j(\cdot)$, $j = 1, \cdots, k-1, k+1, \cdots, d$, and applying the local linear estimation, we obtain the local log likelihood function for $f_k(x)$

$$\begin{aligned}\tfrac{1}{n}\textstyle\sum_{i=1}^{n} Q\Big(Y_i,\ \tilde{f}_1(\mathbf{X}_i'\tilde{\boldsymbol{\beta}}_1),\ \cdots,\ \tilde{f}_{k-1}(\mathbf{X}_i'\tilde{\boldsymbol{\beta}}_{k-1}),\ W_{ix}(\tilde{\boldsymbol{\beta}}_k)'\boldsymbol{\eta}_k,\\ \tilde{f}_{k+1}(\mathbf{X}_i'\tilde{\boldsymbol{\beta}}_{k+1}),\ \cdots,\ \tilde{f}_d(\mathbf{X}_i'\tilde{\boldsymbol{\beta}}_d)\Big)K_{ix}(\tilde{\boldsymbol{\beta}}_k).\end{aligned} \qquad (2.11)$$

Maximise (2.11) with respect to $\boldsymbol{\eta}_k$, and the estimator of $f_k(x)$, $\widehat{f}_k(x; \tilde{\boldsymbol{\beta}}_k)$, is taken to be the first component of the maximiser.

Once the estimators $\widehat{f}_k(\cdot; \boldsymbol{\beta})$s are obtained, we apply the penalised maximum likelihood estimation to estimate $\boldsymbol{\beta}$. Explicitly, we maximise

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \widehat{f}_1(\mathbf{X}_i'\boldsymbol{\beta}_1),\ \cdots,\ \widehat{f}_d(\mathbf{X}_i'\boldsymbol{\beta}_d)\Big) - \sum_{k=1}^{d}\sum_{j=1}^{p_n} \lambda_n \rho_{\lambda_n}(|\beta_{kj}|). \tag{2.12}$$

with respect to $\boldsymbol{\beta}$. We use the resulting maximiser to estimate $\boldsymbol{\beta}$, and denote them by $\widehat{\boldsymbol{\beta}}$. And define $\widehat{\boldsymbol{f}}(x)$ to be $\widehat{\boldsymbol{f}}(x;\widehat{\boldsymbol{\beta}})$, with $\widehat{f}_k(x)$ being the kth component.

2.3 Computational issue and selection of tuning parameters

When implementing the proposed estimation procedure, we have to deal with some practical issues such as the maximisation of (2.10), (2.11) and (2.12), and the selection of initial estimation, bandwidth, tuning parameter and penalty function.

We start with the initial estimation to address the maximisation of (2.10). For this purpose, we note model (1.1) satisfies

$$E(Y|\mathbf{X}) = m_1\big(f_k(\mathbf{X}'\boldsymbol{\beta}_k), k \in \boldsymbol{\tau}_1\big), \tag{2.13}$$

$$var(Y|\mathbf{X}) = m_2\big(f_k(\mathbf{X}'\boldsymbol{\beta}_k), k \in \boldsymbol{\tau}_2\big), \tag{2.14}$$

where m_1 and m_2 are the known link functions, and $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ are the subscript sets of the multiple indexes related to the conditional mean and the conditional variance respectively. Without loss of generality, we suppose that the multiple indexes in (2.13) and (2.14) share a common part, i.e., $\boldsymbol{\tau}_1 \cap \boldsymbol{\tau}_2 = \boldsymbol{\tau}_3$, with $\boldsymbol{\tau}_1 \cup \boldsymbol{\tau}_2 = \{1, \ldots, d\}$. Then, for $k \in \boldsymbol{\tau}_1$, based on (2.13), we obtain the initial estimators $\boldsymbol{\beta}_k^{(0)}$'s by using package *mave*() in R software (Xia, 2008). It should be mentioned that for ultrahigh dimensional case, package *mave*() first reduces the model to a moderate scale of order $n/\log(n)$ by adapting a screening procedure (Zhu et al., 2011), and then estimates $\boldsymbol{\beta}$ based on the reduced model. After that, we get $\boldsymbol{\theta}_k^{(0)}$s as the minimiser of $\sum_{i=1}^{n}\Big(Y_i - m_1\big(\boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_k^{(0)})\boldsymbol{\theta}_k, k \in \boldsymbol{\tau}_1\big)\Big)^2$ with respect to $(\boldsymbol{\theta}_k, k \in \boldsymbol{\tau}_1)$ by using *optim*() function in R software. The initial estimators of $f_k(\cdot), k \in \boldsymbol{\tau}_1$ and $E(Y|X)$ are then taken as $f_k^{(0)}(\cdot) = \boldsymbol{B}(\cdot)'\boldsymbol{\theta}_k^{(0)}$ and $E^{(0)}(Y|\mathbf{X}) = m_1\big(f_k^{(0)}(\mathbf{X}'\boldsymbol{\beta}_k^{(0)}), k \in \boldsymbol{\tau}_1\big)$, respectively. Similarly, repeating the procedure above with Y_i replaced by $\tilde{Y}_i = (Y_i - E^{(0)}(Y|\mathbf{X}_i))^2$, we obtain $\boldsymbol{\beta}_k^{(0)}$, $\boldsymbol{\theta}_k^{(0)}$ and $f_k^{(0)}(\cdot)$ based on (2.14) for $k \in \boldsymbol{\tau}_2 \backslash \boldsymbol{\tau}_3$.

Then the maximiser of (2.10) can be obtained through the following iteration.

(I) Substituting $\boldsymbol{\theta}_k^{(0)}$s for $\boldsymbol{\theta}_k$s in (2.10), we have

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \boldsymbol{B}(\mathbf{X}_i'\beta_1)'\boldsymbol{\theta}_1^{(0)},\ \cdots,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_d)'\boldsymbol{\theta}_d^{(0)}\Big) - \sum_{k=1}^{d}\sum_{j=1}^{p_n}\lambda_n\rho_{\lambda_n}(|\beta_{kj}|). \tag{2.15}$$

Maximise (2.15) with respect to $\boldsymbol{\beta}_k$s by taking $\boldsymbol{\beta}_k^{(0)}$s as the initial values, and denote the resulting maximiser by $\boldsymbol{\beta}_k^{(1)}$s. This can be done by using the MM principle (Lange et al., 2000) and *grpref*() function in R software.

(II) Substitute $\boldsymbol{\beta}_k^{(1)}$s for $\boldsymbol{\beta}_k$s in (2.10), and maximise (2.10) with respect to $\boldsymbol{\theta}_k$s, namely, maximise

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_1^{(1)})'\boldsymbol{\theta}_1,\ \cdots,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_d^{(1)})'\boldsymbol{\theta}_d\Big) \tag{2.16}$$

This can be done by appealing *optim*() function in R software. Treat $\boldsymbol{\beta}_k^{(1)}$ and the resulting maximiser as initial values of $\boldsymbol{\beta}_k$s and $\boldsymbol{\theta}_k$s, and repeat steps (I) and (II) until convergence, we can get the maximiser of (2.10).

The maximisation of (2.11) can also be done by using *optim*() function, and *grpref*() function in R software can be used to maximise (2.12).

In the proposed estimation procedure, different $f_k(\cdot)$s are allowed to have different bandwidths. For each $f_k(\cdot)$, its bandwidth h_k can be selected by a rule of thumb, that is $h_k = b\widehat{\sigma}_k n^{-1/5}$ and $\widehat{\sigma}_k = \sqrt{var(\mathbf{X}_i'\tilde{\boldsymbol{\beta}}_k)}$, where $\tilde{\boldsymbol{\beta}}_k$ is the initial estimator of $\boldsymbol{\beta}_k$ obtained in section 2.2.1, and b is selected by $K-$fold cross validation (Fan et al., 2006). Our simulation studies show this bandwidth selection method works very well.

There is much literature about penalised estimation, and various penalty functions have been proposed. Examples include LASSO in Tibshirani (1996), SCAD in Fan and Li (2001), MCP in Zhang (2010), the Elastic net in Zou and Hastie (2005). In this paper, we use the MCP penalty. The tuning parameter λ_n in the proposed estimation procedure plays a very important role. When the dimension of the $\mathbf{X}$ is of polynomial order of sample size n, we apply BIC to select λ_n, see (Fan and Li, 2001). When the dimension of the $\mathbf{X}$ increases with an exponential order of sample size n, BIC would not work very well, we therefore appeal EBIC, proposed in Chen and Chen (2008), to select λ_n.

3. Asymptotic properties

To present our main asymptotic results, we first introduce some notations. Define $\mathcal{A}_k$ and $\mathcal{A}$ as the non-zero index set of coefficients $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}$ respectively. Let $s_n = |\mathcal{A}|$ be the cardinality of set $\mathcal{A}$. We put a superscript 0 on a parameter/function to denote the true of this parameter/function, e.g. $\boldsymbol{\beta}^0$ and $\boldsymbol{f}^0$ are the true values of $\boldsymbol{\beta}$ and $\boldsymbol{f}$, respectively. For simplicity, we also write $g(\boldsymbol{\beta}, \boldsymbol{f})$ as g when the variable $(\boldsymbol{\beta}, \boldsymbol{f})$ takes the true value $(\boldsymbol{\beta}^0, \boldsymbol{f}^0)$. Let $\mathcal{F}_k = \{f_k :$ f_k has continuous rth order derivatives$\}$ for an integer $r \geq 2$, and $\mathcal{F} = \{\boldsymbol{f} = (f_1, \ldots, f_d)' : f_k \in \mathcal{F}_k, k = 1, \ldots, d\}$. Throughout the paper, C is a constant, it may represent different values at different places.

We denote the score function by $\mathbf{S}_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}, \boldsymbol{f}) = \partial L(\boldsymbol{\beta}, \boldsymbol{f})/\partial \boldsymbol{\beta}_k$ and $\mathbf{S}_{\boldsymbol{\eta}_k}(\boldsymbol{\beta}, \boldsymbol{f}; x) = \partial L_k(\boldsymbol{\beta}, \boldsymbol{f}; x)/\partial \boldsymbol{\eta}_{kx}$. Let $\dot{\mathbf{S}}_{\boldsymbol{\eta}_k\boldsymbol{\eta}_k}(\boldsymbol{\beta}, \boldsymbol{f}; x) = \partial^2 L_k(\boldsymbol{\beta}, \boldsymbol{f}; x)/\partial \boldsymbol{\eta}_{kx}\partial \boldsymbol{\eta}'_{kx}$, $\dot{\mathbf{S}}_{\boldsymbol{\eta}_k\boldsymbol{\beta}_{\tilde{k}}}(\boldsymbol{\beta}, \boldsymbol{f}; x) = \partial^2 L_k(\boldsymbol{\beta}, \boldsymbol{f}; x)/\partial \boldsymbol{\eta}_{kx}\partial \boldsymbol{\beta}'_{\tilde{k}}$ and $\dot{\mathbf{S}}_{\boldsymbol{\beta}_k\boldsymbol{\beta}_{\tilde{k}}}(\boldsymbol{\beta}, \boldsymbol{f}) = \partial^2 L(\boldsymbol{\beta}, \boldsymbol{f})/\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}'_{\tilde{k}}$, $k, \tilde{k} = 1, \ldots, d$. We use a capital letter to denote a random variable, its lowercase to denote its expectation, e.g., $\mathbf{s}_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}, \boldsymbol{f}) = E\mathbf{S}_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}, \boldsymbol{f})$. The vector of $\{x_j, j \in \mathcal{A}\}$ is denoted by $\mathbf{x}_{\mathcal{A}}$, and the matrix $(V_{ij}, i \in \mathcal{A}, j \in \mathcal{A})$ by $\mathbf{V}_{\mathcal{A}\mathcal{A}}$. Denote

$$\kappa(\rho_{\lambda_n}; \boldsymbol{\beta}) = \lim_{\epsilon \to 0+} \max_{1 \leq k \leq d, 1 \leq j \leq p_n} \sup_{|\beta_{kj}| - \epsilon < t_1 < t_2 < |\beta_{kj}| + \epsilon} \Big\{ - \frac{\dot{\rho}_{\lambda_n}(t_2) - \dot{\rho}_{\lambda_n}(t_1)}{t_2 - t_1} \Big\},$$

$$\kappa_0 = \sup\{\kappa(\rho_{\lambda_n}; \boldsymbol{\gamma}) : \|\boldsymbol{\gamma} - \boldsymbol{\beta}^0_{\mathcal{A}}\|_\infty \leq m_{\boldsymbol{\beta}}, \boldsymbol{\gamma} \in \mathbb{R}^{s_n}\},$$

$$m_{\boldsymbol{\beta}} = \frac{1}{2} \min_{j \in \mathcal{A}} |\beta^0_j|, \quad \varphi_n = \| - \dot{\mathbf{s}}^{-1}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}}\|_\infty, \quad \mu_n = \Lambda_{\min}(-\dot{\mathbf{s}}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}}) - \lambda_n \kappa_0,$$

where $\Lambda_{\min}(A)$ is the minimum eigenvalue of matrix A.

Before establishing the asymptotic properties of the proposed estimators $\widehat{f}_k(\cdot)$ and $\widehat{\boldsymbol{\beta}}_k$, we first illustrate the local convexity of the objective function $M(\boldsymbol{\beta}, \boldsymbol{f}) = L(\boldsymbol{\beta}, \boldsymbol{f}) - \lambda_n \sum_{k=1}^{d} \sum_{j=1}^{p_n} \rho_{\lambda_n}(|\beta_{kj}|)$.

Proposition 1. *Under Condition (C2)-(C4) in the Supplementary Materials, if*

$$\frac{n}{(\log s_n)^{\iota_1}} \left\{ \frac{\mu_n^2}{s_n^2} \wedge \frac{\mu_n}{s_n} \right\} \to \infty$$

with $\iota_1 = (4+\iota)/\iota$, *then* $\Lambda_{\min}\big(- \dot{S}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}}(\boldsymbol{\beta}^0, \boldsymbol{f})\big) > \lambda_n \kappa_0$ *holds with probability tending to 1 for all* $\boldsymbol{f}$ *satisfying* $\|\boldsymbol{f} - \boldsymbol{f}^0\|_\infty = o\big(\mu_n/\{s_n(\log p_n)^{2/\iota}\}\big)$.

Remark 1. *Proposition 1 implies that* $\Lambda_{\min}(-\dot{S}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}}(\boldsymbol{\beta}^0, \boldsymbol{f})) > \lambda_n\kappa_0 \geq \lambda_n \kappa(\rho_{\lambda_n}; \boldsymbol{\beta}^0)$ *with high probability when* μ_n*, the gap between* $\Lambda_{\min}(-\dot{S}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}})$ *and* $\lambda_n\kappa_0$*, is positive and does not shrink too fast. As shown in Lv and Fan (2009),* $\kappa(\rho_{\lambda_n}; \boldsymbol{\beta})$ *equals to* $\max\limits_{1\leq k \leq d, 1 \leq j \leq p_n} -\rho''(|\beta_{kj}|)$*, provided that* ρ *has a continuous second derivative. Therefore* $\kappa(\rho_{\lambda_n}; \boldsymbol{\beta})$ *can be regarded as the local concavity of the penalty* ρ_{λ_n} *at* $\boldsymbol{\beta} = (\beta_{kj})$*. Noting that* $\dot{S}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}}(\boldsymbol{\beta}, \boldsymbol{f}))$ *is the second order derivatives of* $L(\boldsymbol{\beta}, \boldsymbol{f})$ *with respect to* $\beta_{kj} \in \mathcal{A}$*, the conclusion* $\Lambda_{\min}(-\dot{S}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}}(\boldsymbol{\beta}^0, \boldsymbol{f})) \geq \lambda_n \kappa(\rho_{\lambda_n}; \boldsymbol{\beta}^0)$ *guarantees the objective function* $M(\boldsymbol{\beta}, \boldsymbol{f}) = L(\boldsymbol{\beta}, \boldsymbol{f}) - \lambda_n \sum_{k=1}^{d}\sum_{j=1}^{p_n} \rho_{\lambda_n}(|\beta_{kj}|)$ *is strictly convex with respect to* $\boldsymbol{\beta}_{\mathcal{A}}$ *in the subspace* $\{\boldsymbol{\beta} \in \Theta : \boldsymbol{\beta}_{\mathcal{A}^c} = 0\}$ *when* $(\boldsymbol{\beta}, \boldsymbol{f})$ *takes value in the neighborhood of* $(\boldsymbol{\beta}^0, \boldsymbol{f}^0)$*. Hence,* $\Lambda_{\min}(-\dot{S}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}}(\boldsymbol{\beta}^0, \boldsymbol{f})) \geq \lambda_n \kappa(\rho_{\lambda_n}; \boldsymbol{\beta}^0)$ *guarantees the objective function* $M(\boldsymbol{\beta}, \boldsymbol{f}) = L(\boldsymbol{\beta}, \boldsymbol{f}) - \lambda_n \sum_{k=1}^{d}\sum_{j=1}^{p_n} \rho_{\lambda_n}(|\beta_{kj}|)$ *is strictly con-*

vex with respect to $\boldsymbol{\beta}_{\mathcal{A}}$ in the subspace $\{\boldsymbol{\beta} \in \Theta : \boldsymbol{\beta}_{\mathcal{A}^c} = 0\}$ when $(\boldsymbol{\beta}, \boldsymbol{f})$ takes value in the neighborhood of $(\boldsymbol{\beta}^0, \boldsymbol{f}^0)$. Furthermore, the second order Condition (C6) in the Supplementary Materials ensures the maximiser of the objective function in the subspace $\{\boldsymbol{\beta} \in \Theta : \boldsymbol{\beta}_{\mathcal{A}^c} = 0\}$ is the optimal estimator over the space $\{\boldsymbol{\beta} \in \Theta\}$ in the neighborhood of $(\boldsymbol{\beta}^0, \boldsymbol{f}^0)$.

Now, we are at the position to show the asymptotic properties of the kernel estimators, $\widehat{f_k^{Ker}}(\cdot)$ and $\widehat{\boldsymbol{\beta}_k^{Ker}}$, and then prove that the estimators $\widehat{f}_k(\cdot)$ and $\widehat{\boldsymbol{\beta}}_k$ based on the proposed algorithm have the same asymptotic properties.

Theorem 1. *Under regularity Conditions (C1)-(C7) in the Supplementary Materials, if $h_n \to 0$, $nh_n/\log n \to \infty$, $\varphi_n \leq Cn^{-\gamma}$ and*

$$
\begin{aligned}
&\frac{n}{(\log p_n)^{\iota_1}}\left\{\frac{(\dot{\rho}^{-1}_{\lambda_n}(m_{\boldsymbol{\beta}}) \wedge n^{\gamma})^2}{\varphi_n^2 s_n^2} \wedge \frac{\dot{\rho}^{-1}_{\lambda_n}(m_{\boldsymbol{\beta}}) \wedge n^{\gamma}}{\varphi_n s_n}\right\} \to \infty,\\
&\frac{n}{(\log s_n)^{\iota_1}}\left\{\frac{(\varphi_n^{-1} \wedge \mu_n)^2}{s_n^2} \wedge \frac{\varphi_n^{-1} \wedge \mu_n}{s_n}\right\} \to \infty, \quad \frac{n\lambda_n^2}{(\log p_n)^{\iota_2}} \to \infty,\\
&\left\{\frac{n^{(1-2\gamma)/2}\lambda_n}{(\log s_n)^{\iota_2}} \wedge \frac{n^{1-2\gamma}\lambda_n^2}{(\log s_n)^{\iota_2}}\right\} \to \infty, \quad m_{\boldsymbol{\beta}} \geq C\varphi_n\lambda_n\dot{\rho}(0+), \quad s_n\lambda_n \to 0,\\
&\frac{\{\lambda_n/n^{\gamma}\} \wedge \{\varphi_n/s_n\}}{h_n^2 + (nh_n)^{-1/2}\log^{1/2}(n)} \to \infty, \quad \varphi_n\lambda_n \leq C(h_n^2 + (nh_n)^{-1/2}\log^{1/2}(n)),
\end{aligned}
\tag{3.17}
$$

with $\iota_1 = (4+\iota)/\iota$ and $\iota_2 = (2+\iota)/\iota$, we have

(a) $\lim_{n\to\infty} P(\widehat{\boldsymbol{\beta}_{\mathcal{A}^c}^{Ker}} = 0) = 1.$

(b) $\lim_{n\to\infty} P(\|\widehat{\boldsymbol{\beta}_{\mathcal{A}}^{Ker}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty \leq \varphi_n\lambda_n\dot{\rho}(0+)) = 1.$

(c) $\sup_{x\in\mathcal{U}} \|\widehat{f_k^{Ker}}(x) - f_k^0(x)\| \to 0$ *in probability.*

For bounded covariate, it can be seen that $\iota_1 = 1$ and $\iota_2 = 1$ by letting $\iota \to \infty$. Then (3.17) holds when $\frac{n}{\varphi_n^2 s_n^2 \log p_n} \to \infty$, which holds if $\log p_n = o(n)$ and $s_n = o(\sqrt{n})$ when φ_n takes a constant. This means the kernel estimation procedure is applicable to ultrahigh dimensional case where the number of covariates is of exponential order of sample size n. The last three conditions in (3.17) guarantee the searching for the estimator of $\boldsymbol{\beta}$ in the neighborhood of the true parameter by choosing the appropriate order of the tuning parameter. Theorem 1 (a) shows the kernel estimators, $\widehat{\boldsymbol{\beta}^{Ker}}$, enjoy selection consistency. (b) implies the estimate consistency of $\widehat{\boldsymbol{\beta}^{Ker}}$, i.e., $\|\widehat{\boldsymbol{\beta}_{\mathcal{A}}^{Ker}} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty \to 0$ in probability, when $\sqrt{(\log p_n)^{\iota_2}/n} \wedge (\log s_n)^{\iota_2}/n^{(1-2\gamma)/2} \ll \lambda_n \ll m_{\boldsymbol{\beta}}/\{C\varphi_n\dot{\rho}(0+)\}$. Therefore, Theorem 1 guarantees the recovery of signals if $m_{\boldsymbol{\beta}} \gg \varphi_n\sqrt{(\log p_n)^{\iota_2}/n}$ under condition (3.17). (c) illustrates the estimators $\widehat{f_k^{Ker}}(\cdot)$s of $f_k(\cdot)$s are uniformly consistent.

Let $\boldsymbol{\Sigma}_{1n} = -\dot{\mathbf{s}}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}} + \dot{\mathbf{s}}_{\boldsymbol{\beta}_{\mathcal{A}}\boldsymbol{\eta}}\dot{\mathbf{s}}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}\dot{\mathbf{s}}_{\boldsymbol{\eta}\boldsymbol{\beta}_{\mathcal{A}}}$, and $\boldsymbol{\Sigma}_{2n}$ be the covariance matrix of $\mathbf{m}_{\mathcal{A}}$, the empirical efficient score for parameter $\boldsymbol{\beta}_{\mathcal{A}}$, which are defined in the Supplementary Materials. Denote $\Lambda_{1n} = \Lambda_{\min}(\boldsymbol{\Sigma}_{1n})$, $\Lambda_{2n} = \Lambda_{\min}(\boldsymbol{\Sigma}_{2n})$, and $\Lambda_{3n} = \Lambda_{\min}(\boldsymbol{\Sigma}_{1n}^{-1}\boldsymbol{\Sigma}_{2n}\boldsymbol{\Sigma}_{1n}^{-1})$. The following theorem establishes the oracle property and asymptotic normality of the kernel estimators.

Theorem 2. *Under conditions of Theorem 1, if*

$$\frac{\Lambda_{3n} n h_n^2}{s_n(\log n)^2} \to \infty, \quad \frac{n + h_n^{-4}}{(1 \vee \Lambda_{3n}) s_n^3} \to \infty, \quad \frac{\Lambda_{2n}^2 (n + h_n^{-4})}{s_n^2} \to \infty,$$
$$\frac{n(\Lambda_{1n}^2 - h_n^4)}{s_n^2 (\log s_n)^{\iota_1}} \to \infty, \quad \frac{\Lambda_{1n}^4 \Lambda_{3n}(n + h_n^{-4})}{s_n^3} \to \infty, \quad \frac{n s_n \lambda_n^2 \dot{\rho}_{\lambda_n}(m_{\boldsymbol{\beta}})^2}{\Lambda_{1n}^2 \Lambda_{3n}} \to 0, \tag{3.18}$$

then

(a) for any $\mathbf{u} \in \mathbb{R}^{s_n}$ *with* $\|\mathbf{u}\|_2 = 1$*, when* $n h_n^4 \to 0$*, we have*

$$\sqrt{n}\mathbf{u}' \boldsymbol{\Sigma}_{2n}^{-1/2} \boldsymbol{\Sigma}_{1n} (\widehat{\boldsymbol{\beta}_{\mathcal{A}}^{Ker}} - \boldsymbol{\beta}_{\mathcal{A}}^0) \xrightarrow{d} N(0, 1).$$

(b) when $s_n = o(\Lambda_{3n}^{-1}(n h_n^4 + h_n^{-1}))$*, we have*

$$\sqrt{n h_n}\Big(\widehat{f_k^{Ker}}(x) - f_k^0(x) - \frac{1}{2}\ddot{f}_k^0(x)\nu_2 h_n^2\Big) \xrightarrow{d} N(0, \sigma_k^2(x)),$$

where $\sigma_k^2(x) = \upsilon_0 e_k^{-1}(x) f_{\mathcal{X}_k}^{-1}(x)$*,* $e_k(x) = -E(Q^{(02,k)}(Y_i, f_l^0(\mathbf{X}_i'\boldsymbol{\beta}_l^0), l = 1, \ldots, d) | \mathbf{X}_i'\boldsymbol{\beta}_k^0 = x)$*,* $\nu_2 = \int_{-\infty}^{\infty} x^2 K(x) dx$*, and* $\upsilon_0 = \int_{-\infty}^{\infty} K^2(x) dx$*.*

In the following, we are going to establish the asymptotic normalities for the estimators, $\widehat{f}_k(x)$s and $\widehat{\boldsymbol{\beta}}_k$s, resulted from the proposed algorithm.

Theorem 3. *Under (3.18) and the conditions in Theorem 1,*

(a) If $s_n = o\big(\Lambda_{3n} n h_n^{-1} q_n^{2(r-1)} + \Lambda_{3n} n^{-1} h_n^{-4} q_n^{4(r-1)}\big)$*,* $s_n q_n^2 (q_n + s_n) = o\big(\Lambda_{3n} n h_n^{-1}\big)$*,* $s_n q_n^4 (q_n + s_n)^2 = o\big(\Lambda_{3n} n h_n^{-4}\big)$*,* $n h_n^4 \to 0$*,* $r \geq 2$*, we have*

$$\sqrt{n}\mathbf{u}' \boldsymbol{\Sigma}_{2n}^{-1/2} \boldsymbol{\Sigma}_{1n} (\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^0) \xrightarrow{d} N(0, 1), \ \ \textit{with} \ \|\mathbf{u}\|_2 = 1.$$

(b) If $q_n(q_n+s_n)^{1/2}=o(h_n^{-1}n^{1/2}a_n^{1/2})$, and $s_n=o(\Lambda_{3n}^{-1}(nh_n^4+h_n^{-1}))$, we have

$$\sqrt{nh_n}\Big(\widehat{f}_k(x)-f_k^0(x)-\frac{1}{2}\ddot{f}_k^0(x)\nu_2h_n^2\Big)\xrightarrow{d}N(0,\sigma_k^2(x)),$$

where $a_n=h_n^4+(nh_n)^{-1}$ and $\sigma_k^2(x)$ is defined as in Theorem 2.

When eigenvalues $\Lambda_j,\ j=1,\ 2,\ 3$ are bounded away from 0, it is easy to see part (a) in Theorem 3 holds for $s_n=o(n^{1/3})$ if we take $r=2$, $q_n=O(n^{1/3})$ and $h_n=O(n^{-1/3})$; and part (b) holds for the theoretical optimal bandwidth $h_n=O(n^{-1/5})$ of nonparametric estimation if we take $q_n=O(n^{1/5})$ and $s_n=o(n^{1/5})$. Theorem 3 implies that the proposed algorithm shares the same asymptotic distribution with the kernel estimators.

Theorem 4. *Let $\mathcal{D}_0=\{\boldsymbol{\psi}:\boldsymbol{\psi}$ has a continuous derivative on $\mathcal{U}^d$, $\int_{\mathcal{U}^d}\boldsymbol{\psi}(\mathbf{x})d\mathbf{x}=0\}$. Under the conditions for part (a) in Theorem 2, when the distribution of $\boldsymbol{\varepsilon}$ is known, both $\int_{\mathcal{U}^d}\boldsymbol{\psi}_1'(\mathbf{x})\widehat{\boldsymbol{f}^{Ker}}(\mathbf{x})d\mathbf{x}+\boldsymbol{\psi}_2'\widehat{\boldsymbol{\beta}_{\mathcal{A}}^{Ker}}$ and $\int_{\mathcal{U}^d}\boldsymbol{\psi}_1'(\mathbf{x})\widehat{\boldsymbol{f}}(\mathbf{x})d\mathbf{x}+\boldsymbol{\psi}_2'\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ are efficient estimators of $\int_{\mathcal{U}^d}\boldsymbol{\psi}_1'(\mathbf{x})\boldsymbol{f}^0(\mathbf{x})d\mathbf{x}+\boldsymbol{\psi}_2'\boldsymbol{\beta}_{\mathcal{A}}^0$, for any function $\boldsymbol{\psi}_1=(\psi_{11},\ldots,\psi_{1d})'\in\mathcal{D}_0$ and $\boldsymbol{\psi}_2\in\mathbb{R}^{s_n}$.*

Theorem 4 indicates that both $\widehat{\boldsymbol{\beta}_{\mathcal{A}}^{Ker}}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ are efficient estimators of $\boldsymbol{\beta}_{\mathcal{A}}^0$ by taking $\boldsymbol{\psi}_1(\mathbf{x})=\mathbf{0}$, and $\widehat{\boldsymbol{f}^{Ker}}(\cdot)$ and $\widehat{\boldsymbol{f}}(\cdot)$ are semiparametrically efficient estimators of $\boldsymbol{f}^0(\cdot)$, by taking $\boldsymbol{\psi}_2(\mathbf{x})=\mathbf{0}$, in the sense of Bickel et al. (1993).

4. Simulation Studies

In this section, we conduct four simulations to investigate the performance of the proposed method through the comparison between the existing competing procedures in terms of bias, efficiency, predictive accuracy and selection accuracy. For feasibility to compare, the settings, as well as evaluation criteria, of the first two simulations are taken from the related literatures. Model (1.3) in Section 1 is new and the corresponding Simulations 3 and 4 are conducted under the cases with high-dimensional and ultrahigh-dimensional covariates, respectively. We adapt M-CP selector to select important variables. Tuning parameter λ_n is determined by using BIC and EBIC principles for high dimension and ultrahigh-dimensional cases respectively.

Simulation 1. *The setting is the same as Alquier and Biau (2013) to consider the single-index models with $p_n = 10$ or 50 and the sample size $n = 50$ or 100.*

For each model, a training set of size n is generated to fit the model and the mean squared prediction error (MSPE) is evaluated on a separate validation set of the same size. We compare the results of proposed method with the Fourier estimator $\widehat{f}_{Fourier}$ in Alquier and Biau (2013), the estimation $\widehat{f}_{HHI}$ in Härdle et al. (1993), the LASSO estimator $\widehat{f}_{LASSO}$ and the standard kernel estimate $\widehat{f}_{NW}$ (Nadaraya, 1964; Watson, 1964).

The median, mean and standard deviation (SD) of MSPE based on 200 repetitions are shown in Table 1, which suggests that the proposed method has much less predictive error than all the competing procedures. Compared with the LASSO estimator, this result is natural since LASSO estimator does not enjoy variable selection oracle property. In addition, the MSPEs of the propose estimators $\widehat{f}_j, j = 1, 2, 3$ with the smoothing parameter $q_n = 4,\ 5,\ 6$ respectively are close, suggesting that the proposed one step estimation is not sensitive to the initial estimators.

Simulation 2. *The setting is the same as Case 1 in Zhu et al. (2013), considering the multi-index models* $Y = f_1(\mathbf{X}'\boldsymbol{\beta}_1) + f_2(\mathbf{X}'\boldsymbol{\beta}_2)\varepsilon$. *The simulation is repeated 1000 times with sample size* $n = 600$.

We compare the proposed method with Zhu et al. (2013). Table 2 summarizes the bias, the standard deviation (SD), the root of mean squared error (RMSE) of the estimates for non-zero elements of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. To see the performance of estimators for both parameters and functions, we also calculate the average squared errors defined as $ASE_j = n^{-1}\sum_{i=1}^{n}\left(\widehat{f}_j(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_j) - f_j^0(\mathbf{X}_i'\boldsymbol{\beta}_j^0)\right)^2,\ j = 1, 2$. From Table 2, we can see that the proposed method is much more efficient and accurate than the method in Zhu et al. (2013), which may be attributed to that the proposed method could select important variables and estimate parameters and functions simultaneously whilst the estimating equation method in Zhu et al. (2013) only considers the parameter estimation.

Simulation 3. *The data are generated from the multiple-index stochastic frontier model (1.3), where the covariates* $\mathbf{X} = (X_1, \ldots, X_{15})'$ *are generated from an* $AR(1)$ *model with* $X_1 \sim N(0,1)$ *and* $Cov(X_{j_1}, X_{j_2}) = 0.4^{|j_1-j_2|}$ *for* $j_1, j_2 = 1, \ldots, 15$ *and then are trimmed into the range* $[-1,1]$. *The coefficient is taken as* $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_3 = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, \ldots, 0)'$ *and* $\boldsymbol{\beta}_2 = (0,0,0,1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, \ldots, 0)'$ *so that there are three important covariates in each functional components. The functions are taken as* $f_1(x) = \exp(x/2) + 2x^2, \quad f_2(x) = ((x-1)^2+1)/4, \quad f_3(x) = ((x+1)^2+1)/4$. *The simulation is repeated 1000 times with sample size* $n = 600$.

The simulation results are summarized in Tables 3 and 4 and Figure 1. Table 3 shows the results of variable selection including the number of selected variables, true positive rate (TPR) and false positive rate (FPR). The numbers of selected variables are closed to the true values, the TPR close to 1 and the FPR close to 0. These suggest that the proposed method can not only select important variables but also rule out unimportant variables with high probability. Table 4 gives the estimators of the parameters using the proposed method and the oracle method, which is based on the model with only the important three covariates. The results in Table 4 reveal the proposed estimators are approximately unbiased, and their estimated standard errors (ESE) agree well with the sample standard deviations (SD). Moreover, the proposed method produces coverage percentages of the 95% confidence intervals close to the nominal level. It is evident as well that the proposed procedure performs comparably well with the oracle estimator.

Figure 1 (a) displays the estimated frontier function by the proposed method. As we have known, neglecting the noise, the frontier function $f_1(\mathbf{x}'\boldsymbol{\beta}_1)$ is the maximum attainable output with input $\mathbf{x}$. To see that, we further generated a validation data set with sample size 600 from the same model, which is displayed by star point in Figure 1 (a). This plot shows that the statistical noise encompassing in the nonparametric world does not affect the estimation.

To further evaluate the performance of nonparametric function estimators and compare the prediction effect of the propsed method with the competing gradient boosting approach, for each repetition of 1000 replications, we applied the fitted model to predict a newly generated data set. Figure 1 (b) and (d) display the scatter plot of the true values of Y_i against the fitted values of $\widehat{Y}_i = \widehat{f}_1(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_1) + \widehat{u}_i$ of the proposed method and the gradient boosting approach respectively, and Figure 1 (c) and (f) display the scatter plot of the true simulated e_i against its predictor $\widehat{e}_i = Y_i - \widehat{Y}_i$ of the proposed method and the gradient boosting approach respectively, where $\widehat{u}_i = \frac{\widehat{\sigma}_i\widehat{\lambda}_i}{1+\widehat{\lambda}_i^2}\left[\frac{\varphi(-\widehat{\xi}_i\widehat{\lambda}_i/\widehat{\sigma}_i)}{\Phi(-\widehat{\xi}_i\widehat{\lambda}_i/\widehat{\sigma}_i)} - \frac{\widehat{\xi}_i\widehat{\lambda}_i}{\widehat{\sigma}_i}\right]$, $\widehat{\sigma}_i^2 = \widehat{f}_2^2(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_2) + \widehat{f}_3^2(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_3)$, $\widehat{\lambda}_i = \widehat{f}_2(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_2)/\widehat{f}_3(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_3)$, $\widehat{\xi}_i = Y_i - \widehat{f}_1(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_1)$, following Jondrow et al. (1982). From Figure 1 (b)-(f), we can see that the predictors using proposed method work pretty well globally, and are comparable to those of the gradient boosting approach.

In fact, the MSPEs based on 1000 newly generated data sets are 1.450 for the proposed method with deviation 0.286, and 3.104 for the gradient boosting approach with deviation 0.420. This shows that the proposed method possesses both high prediction ability and interpretability at the same time.

Simulation 4. *The data are generated as the same as in Simulation 3 except that we take $p_n = 1000$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_3 = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, \ldots, 0)'$ and $\boldsymbol{\beta}_2 = (0, 0, 0, 1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}, 0, \ldots, 0)'$ to reflect ultra-high dimensional case. The simulation results are summarized in Tables 3-4 and Figure 1 of the Supplementary Materials. Moreover, we also obtain the MSPEs based on 1000 newly generated data sets being 2.716 for the proposed method with deviation 0.350, and 5.842 for the gradient boosting approach with deviation 0.832. Therefore, we can draw the similar conclusions as those in Simulation 3.*

5. An Application

The proposed approach is now applied to analyze the data set from one of China's largest liquor companies in western China. The purpose of the analysis is investigating whether and how various factors affect the sales of liquor. The data set includes monthly sales (Y_i) and covariates information for $n = 1941$ observations in 31 provinces of China from 2011 to 2018. The covariates include four parts. (1) The company's product information: price, advertising investment, reimbursement expense of dealers, etc. (2) Brewing industry information: monthly liquor yields, monthly beer yields, beer import and export, monthly trading amounts of 12 stocks of the brewing industry and profit of the affiliated companies, etc. (3) Economic information of related cities and towns: per capita GDP, per capita disposable income, consumer price index, retail price index, total retail sales of consumer goods, housing sales price, residential investment, permanent population, and so on. (4) Geographic information: monthly average temperature, monthly average relative humidity, geographical division, distance from the liquor producing area, etc. Together with the lagged variables, we have 2051 covariates. Log transformation is taken to response variable and all of covariates and response are standardized. Then, the multiple-index stochastic frontier model (1.3) and the proposed approach are applied to the data. The bandwidths are taken as those described in Remark 3. The selected important variables and their regression coefficients estimates are reported in Figure 2, and the estimated functions are displayed in Figure 3.

Figure 2 (a) displays the 13 important variables for the frontier of the sales. Combining with the monotone increasing function of $f_1(\cdot)$ displayed in Figure 3 (a), the following conclusions are drawn. Firstly, the negative coefficients of both per capita GDP (per_capita_gdp) and its lagged variable (per_capita_gdp_lastyear) indicate that consumers in cities with lower level of economic development buy more of the liquor product which is consistent with the fact that the considered product here is cheap and thus is popular among low consumption groups. Secondly, the positive coefficients of residents and its last year's value show that the greater the population, the larger demand for the liquor. Thirdly, the sales in the past months (SL_lag1,2,4,5,6,7) have positive coefficients, which means that the larger the sales in the past, the larger the sales in the current month, which is consistent with intuition. Fourthly, the price (PRICE_lag5) is statistically significant, because this is a low-end liquor product targeting at price-sensitive low consumption groups. Fifthly, the coefficient of liquor_production (taking 1 or 0, whether the city belongs to a province with large liquor production) is positive, which shows that people in province with large liquor production tend to purchase more liquor. Sixthly, the variable xlj_sichuan (taking 1 or 0, whether the city belongs to Sichuan Province where the liquor is produced) has quite large coefficient, reflecting that the product is selling well in the area around the place of origin.

The 18 important variables selected for inefficiency function are showed in Figure 2 (b). Combining with the monotone increasing function of $f_2(\cdot)$ displayed in Figure 3 (b), the following conclusions are drawn. Firstly, subsidy reimbursement expenses and its lagged variables (btl & btl_lag1, 2 & btl_three_m) have positive coefficients, illustrating that they are relatively inefficient input. This is because the company subsidizes dealers based on their purchases before specific day and dealers usually buy much more products than they can sell before that day, causing large inventories. Secondly, per capita GDP (per_capita_gdp) and its lagged variable (per_capita_gdp_lastyear) have negative coefficients, and thus have contrary influence on sales compared with Figure 2 (a). This may be attributed to that cities with higher GDP usually are more efficient in commercial operation. Thirdly, GDP from primary industry, mainly agriculture, in this year and last year (gdp1 & gdp1_lastyear) with positive regression coefficients have negative effect on sales, which may be a result of the fact that areas with high agricultural output usually have low commercial operation ability. Fourthly, positive coefficients of sales in the past 1 to 5 months (SL_lag1-5) show that there may be some waste of costs in areas with large sales in the past. Fifthly, cumulative expenses on meetings and events such as wine expo during the past half year and the past year (prov_meeting_six_m & prov_meeting_ twelve_m) have positive effect on sales, reflecting that effective promotional activities can increase the market share of the product. Sixthly, the variable xlj_hunan (taking 1 or 0), meaning whether the city belongs to Hunan Province which is the province with the second largest sales, has positive regression coefficient and thus has negative effect on sales. The dealers in this province may have some cost waste that actually has been noticed by the company.

Figure 2 (c) shows that per capita GDP (per_capita_gdp) and per capita GDP last year (per_capita_gdp_lastyear) have influence on the variance function, which is estimated as quadratic form (Figure 3 (c)). This may be attributed to that consumers in the areas with higher level of economic development have more choices of alcohol, increasing the uncertainty.

6. Concluding Remarks

To investigate whether and how ultrahigh-dimensional factors effect various of measurements, for example, mean, frontier, inefficiency and variance, we propose an ultrahigh-dimensional structured multiple-index models. We estimate all of functions and parameters based on penalized full likelihood-type function. The proposed estimators are shown to be consistent, asymptotically normal and semiparametrically efficient. To solve the computational problem caused by the combination of nonconvexity of the likelihood function, nonsmoothness of the penalty term, and the large number of functions and ultrahigh-dimensional predictors, we develop a technique of blending spline and kernel smoothing with a majorized coordinate descendent algorithm, so that the computation is easily performed by using the existing software. The simulation studies show that our method outperforms the existing methods in selection and estimation for all of the cases considered, whose settings are taken from existing literature if available. We apply the proposed method to a real data from one of China's largest liquor companies, and finds that 31 out of 2051 factors, including price, previous sales, per capita GDP, residents, are important for mean, stochastic frontier, inefficiency and variance of the liquor sales.

There are several potential extensions of the model and estimation strategy. We use the sparsity as a regularization strategy to solve the problem of ultra-high dimension. The sparse assumption implies the correlation among the high-dimensional covariates should be restricted. To handle with the correlated high-dimension covariates, other regularization strategy, for example, low rank or fusion method can be considered. Whether the procedure and associated theoretical results hold for these regularization strategies is unclear and warrants a further investigation.

Acknowledgments

The authors would like to thank the associate editor and the two referees for their constructive and insightful comments and suggestions that greatly improved the paper. This research was supported by the National Natural Science Foundation of China (11931014, 11829101, 12171374, 11971362) and Fundamental Research Funds for the Central Universities of China (JBK1806002).

Supplementary Materials

The Supplementary Materials provide additional notations, conditions, the proofs of the theorems, referred results in Section 4 and additional simulation studies on mixed effects models.

References

Aigner, D., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. J. Econometrics 6, 21–37.

Alquier, P., Biau, G., 2013. Sparse single-index model. J. Mach. Learn. Res. 14, 243–280.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., Wellner, J.A., 1993. Efficient and adaptive estimation for semiparametric models. Johns Hopkins Series in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD.

Carroll, R.J., Fan, J., Gijbels, I., Wand, M.P., 1997. Generalized partially linear single-index models. J. Amer. Statist. Assoc. 92, 477–489.

Chen, J., Chen, Z., 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95, 759–771.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Fan, J., Lin, H., Zhou, Y., 2006. Local partial-likelihood estimation for lifetime data. Ann. Statist. 34, 290–325.

Fan, Y., Li, Q., Weersink, A., 1996. Semiparametric estimation of stochastic production frontier models. J. Bus. Econom. Statist. 14, 460–468.

Guo, S., Box, J., Zhang, W., 2017. A dynamic structure for high dimensional covariance matrices and its application in portfolio allocation. J. Amer. Statist. Assoc. 112, 235–253.

Härdle, W., Hall, P., Ichimura, H., 1993. Smooth regression analysis. Ann. Statist. 21, 151–178.

Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P., 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. J. Econometrics 19, 233–238.

Ke, Y., Lian, H., Zhang, W., 2020. High-dimensional dynamic covariance matrices with homogeneous structure. J. Bus. Econom. Statist. , 1–15.

Kumbhakar, S.C., Park, B.U., Simar, L., Tsionas, E.G., 2007. Nonparametric stochastic frontiers: a local maximum likelihood approach. J. Econometrics 137, 1–27.

Lange, K., Hunter, D.R., Yang, I., 2000. Optimization transfer using surrogate objective functions. J. Comput. Graph. Statist. 9, 1–59. With discussion, and a rejoinder by Hunter and Lange.

Lian, H., Qiao, X., Zhang, W., 2021. Homogeneity pursuit in single index models based panel data analysis. J. Bus. Econom. Statist. 39, 386–401.

Liu, X., Cui, Y., Li, R., 2016. Partial linear varying multi-index coefficient model for integrative gene-environment interactions. Statist. Sinica 26, 1037–1060.

Lv, J., Fan, Y., 2009. A unified approach to model selection and sparse recovery using regularized least squares. Ann. Statist. 37, 3498–3528.

Ma, Y., Zhu, L., 2013. Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. J. R. Stat. Soc. Ser. B. Stat. Methodol. 75, 305–322.

Nadaraya, E., 1964. On estimating regression. J. Multivariate Anal. 9, 141–142.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B. Stat. Methodol. 58, 267–288.

Watson, G.S., 1964. Smooth regression analysis. Sankhyā Ser. A 26, 359–372.

Xia, Y., 2008. A multiple-index model and dimension reduction. J. Amer. Statist. Assoc. 103, 1631–1640.

Zhang, C.H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38, 894–942.

Zhu, L., Dong, Y., Li, R., 2013. Semiparametric estimation of conditional heteroscedasticity via single-index modeling. Statist. Sinica 23, 1235–1255.

Zhu, L., Li, L., Li, R., Zhu, L., 2011. Model-free feature screening for ultrahigh-dimensional data. J. Amer. Statist. Assoc. 106, 1464–1475.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 301–320.

H. Lin and S. Zhao

Center of Statistical Research and School of Statistics

Southwestern University of Finance and Economics, Chengdu, China

E-mail: (linhz@swufe.edu.cn); (zsxzsx@smail.swufe.edu.cn)

L. Liu

School of Mathematics and Statistics, Wuhan University, Wuhan, China

E-mail: (lliu.math@whu.edu.cn)

W. Zhang

Department of Mathematics, The University of York, York, United Kingdom

E-mail: (wenyang.zhang@york.ac.uk)

Table 1: Numerical results in Simulation 1, with $n = 50$ and $n = 100$

$n = 50$	$p_n = 10$	$\widehat{f}_{Fourier}$	$\widehat{f}_{HHI}$	$\widehat{f}_{LASSO}$	$\widehat{f}_{NW}$	$\widehat{f}^{(1)}$	$\widehat{f}^{(2)}$	$\widehat{f}^{(3)}$
Model 1	median	0.061	0.063	0.046	0.293	0.014	0.017	0.014
	mean	0.061	0.063	0.047	0.290	0.014	0.018	0.015
	SD	0.016	0.014	0.011	0.063	0.004	0.004	0.004
Model 2	median	0.050	0.067	0.307	0.198	0.062	0.072	0.062
	mean	0.069	0.080	0.338	0.208	0.066	0.078	0.067
	SD	0.081	0.057	0.082	0.072	0.032	0.037	0.029
$n = 100$	$p_n = 10$	$\widehat{f}_{Fourier}$	$\widehat{f}_{HHI}$	$\widehat{f}_{LASSO}$	$\widehat{f}_{NW}$	$\widehat{f}^{(1)}$	$\widehat{f}^{(2)}$	$\widehat{f}^{(3)}$
Model 1	median	0.053	0.051	0.042	0.227	0.005	0.005	0.005
	mean	0.056	0.050	0.043	0.237	0.005	0.005	0.005
	SD	0.011	0.006	0.004	0.044	0.001	0.001	0.001
Model 2	median	0.047	0.052	0.332	0.209	0.030	0.030	0.022
	mean	0.049	0.053	0.337	0.218	0.031	0.032	0.023
	SD	0.009	0.012	0.063	0.045	0.012	0.012	0.009
$n = 50$	$p_n = 50$	$\widehat{f}_{Fourier}$	$\widehat{f}_{HHI}$	$\widehat{f}_{LASSO}$	$\widehat{f}_{NW}$	$\widehat{f}^{(1)}$	$\widehat{f}^{(2)}$	$\widehat{f}^{(3)}$
Model 1	median	0.057	1.156	0.060	0.507	0.037	0.039	0.039
	mean	0.095	1.124	0.066	0.533	0.038	0.040	0.039
	SD	0.143	0.241	0.026	0.081	0.011	0.011	0.011
Model 2	median	0.150	0.502	0.795	0.308	0.114	0.118	0.114
	mean	0.151	0.539	0.776	0.326	0.127	0.125	0.127
	SD	0.111	0.200	0.208	0.109	0.053	0.053	0.058
$n = 100$	$p_n = 50$	$\widehat{f}_{Fourier}$	$\widehat{f}_{HHI}$	$\widehat{f}_{LASSO}$	$\widehat{f}_{NW}$	$\widehat{f}^{(1)}$	$\widehat{f}^{(2)}$	$\widehat{f}^{(3)}$
Model 1	median	0.053	0.092	0.050	0.519	0.007	0.006	0.008
	mean	0.054	0.100	0.050	0.508	0.007	0.006	0.008
	SD	0.007	0.026	0.006	0.026	0.002	0.002	0.002
Model 2	median	0.047	0.242	0.503	0.329	0.061	0.067	0.075
	mean	0.070	0.267	0.502	0.339	0.064	0.073	0.081
	SD	0.099	0.111	0.106	0.073	0.024	0.025	0.029

$\widehat{f}_{Fourier}$, $\widehat{f}_{HHI}$, $\widehat{f}_{LASSO}$ and $\widehat{f}_{NW}$ are the estimates suggested in Alquier and Biau (2013); $\widehat{f}^{(j)}$s, $j = 1, 2, 3$ represent the proposed estimate with the smoothing parameter $q_n = 4, 5, 6$ respectively.

Table 2: Numerical results in Simulation 2

Method		$\widehat{\beta}_{11}\%$	$\widehat{\beta}_{12}\%$	$\widehat{\beta}_{13}\%$	$\widehat{\beta}_{14}\%$	$\widehat{\beta}_{21}\%$	$\widehat{\beta}_{22}\%$	$\widehat{\beta}_{28}\%$	$ASE_1\%$	$ASE_2\%$
(Z3.1)	Bias	-0.04	0.03	0.06	-0.06	7.68	0.07	-13.56	3.83	98.56
	SD	0.92	1.19	0.99	1.14	17.14	19.84	11.89		
	RMSE	0.92	1.19	0.99	1.14	18.78	19.84	18.03		
(Z3.2)	Bias	-0.04	0.04	0.05	-0.06	2.62	0.62	-4.16	3.83	43.64
	SD	0.91	1.18	0.98	1.13	9.67	11.45	5.88		
	RMSE	0.91	1.18	0.98	1.13	10.02	11.47	7.20		
(Z3.3)	Bias	-0.05	0.04	0.05	-0.06	2.48	0.59	-4.04	3.84	43.24
	SD	0.91	1.18	0.98	1.13	9.35	11.19	5.30		
	RMSE	0.91	1.18	0.98	1.13	9.67	11.21	6.66		
Prop.	Bias	0.06	-0.02	-0.09	-0.48	3.80	-0.14	0.41	1.08	2.03
	SD	0.71	1.03	1.04	1.13	11.47	4.97	5.63		
	RMSE	0.71	1.03	1.05	1.23	12.09	4.97	5.64		

(Z3.1)-(Z3.3) represent the estimating equation methods (3.1)-(3.3) in Zhu et al. (2013); RMSE represents the root-mean-square error; ASE represents the average squared error, defined by $ASE_j = n^{-1}\sum_{i=1}^{n}\left(\widehat{f}_j(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_j) - f_j^0(\mathbf{X}_i'\boldsymbol{\beta}_j^0)\right)^2$.

Table 3: Selection results for regression coefficients in Simulation 3.

Parameter	#S	TPR	FPR
$\boldsymbol{\beta}_1$	3.009(0.151)	0.998	0.001
$\boldsymbol{\beta}_2$	3.055(0.908)	0.942	0.019
$\boldsymbol{\beta}_3$	3.334(0.987)	0.980	0.033
TRUE	3	1	0

#S means to the number of selected variables; selected standard errors are summarized in parentheses; TPR (True positive rate) means the rate that the important variables are selected; FPR (False positive rate) means the rate that the unimportant variables are selected.

Table 4: Estimate results for regression coefficients in Simulation 3.

Parameter	$\widehat{\boldsymbol{\beta}}$				$\widehat{\boldsymbol{\beta}^{OR}}$			
	Bias	SD	ESE	CP	Bias	SD	ESE	CP
$\boldsymbol{\beta}_1$	0.000	0.009	0.010	0.953	-0.000	0.010	0.011	0.949
	0.000	0.011	0.012	0.957	0.001	0.012	0.012	0.952
	-0.001	0.010	0.010	0.948	-0.001	0.012	0.011	0.949
$\boldsymbol{\beta}_2$	-0.028	0.127	0.117	0.936	-0.008	0.109	0.100	0.927
	-0.006	0.133	0.114	0.931	-0.026	0.118	0.107	0.939
	-0.014	0.123	0.119	0.947	-0.002	0.109	0.099	0.933
$\boldsymbol{\beta}_3$	-0.015	0.110	0.114	0.945	-0.004	0.096	0.090	0.951
	-0.012	0.124	0.122	0.932	-0.015	0.117	0.103	0.933
	-0.011	0.107	0.098	0.941	-0.013	0.097	0.092	0.954

$\widehat{\boldsymbol{\beta}}$ represents the proposed estimator; $\widehat{\boldsymbol{\beta}^{OR}}$ represents the oracle estimator; SD represents the sample standard deviation of the estimates; ESE represents the sample mean of the estimated standard errors; CP represents the empirical 95% coverage probability.

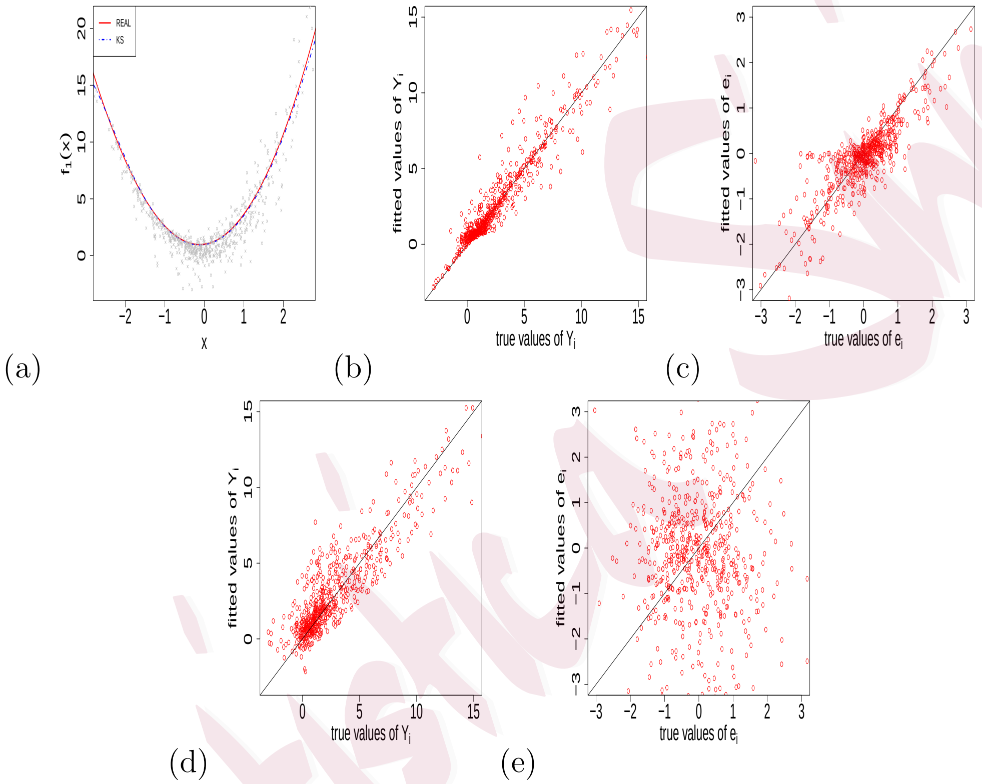
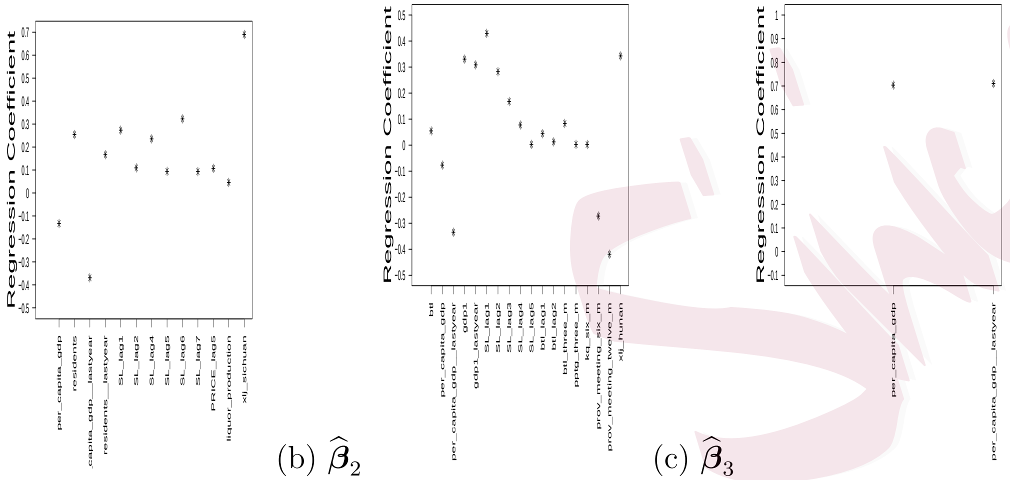
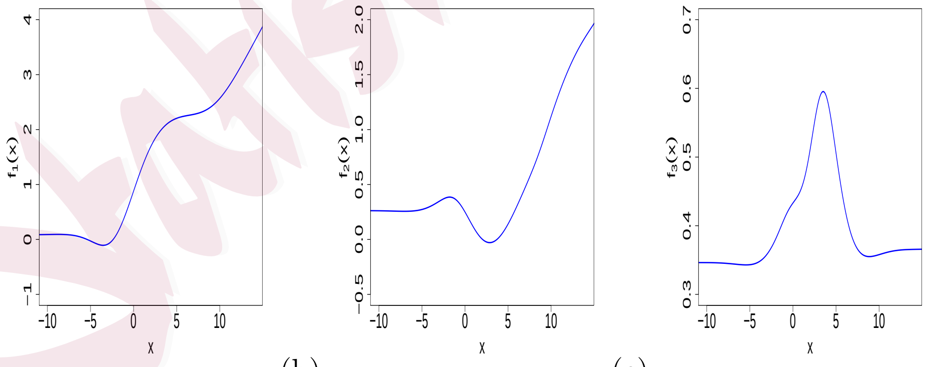


Figure 1: Plots of the nonparametric estimates in Simulation 3: true against estimated values. (a) shows the plot of true frontier and the fitted frontier function; (b) and (d) show the scatter plot of the true values and the fitted values of the response variable using the proposed method and the gradient boosting approach, respectively; (c) and (e) show the scatter plot of the true values and fitted values of the residual using the proposed method and the gradient boosting approach, respectively.



(a) $\widehat{\boldsymbol{\beta}}_1$ (b) $\widehat{\boldsymbol{\beta}}_2$ (c) $\widehat{\boldsymbol{\beta}}_3$

Figure 2: Selected important variables and their estimates of $\boldsymbol{\beta}_k, k = 1, 2, 3$ for liquor data.



(a) (b) (c)

Figure 3: Plots of the kernel smoothing nonparametric estimates in liquor data analysis.