Statistica Sinica **33** (2023), 2137-2160
doi:https://doi.org/10.5705/ss.202021.0009

STRUCTURED ULTRAHIGH DIMENSIONAL MULTIPLE-INDEX MODELS WITH EFFICIENT ESTIMATION IN COMPUTATION AND THEORY

Huazhen Lin[1], Shuangxue Zhao[1], Li Liu[2] and Wenyang Zhang[3]

[1]*Southwestern University of Finance and Economics,* [2]*Wuhan University, and* [3]*The University of York*

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 special cases, including the stochastic frontier model, single-index model, and additive-index model. We estimate all of the functions and parameters based on a full likelihood-type function. As a result, the proposed estimators are shown to be semiparametrically efficient, consistent in terms of selection and estimation, and asymptotically normal. The computation is challenging owing to the combination of nonconvexity of the likelihood function, the nonsmoothness of the penalty term, and the large number of functions. To solve the computational problem, we blend spline and kernel smoothing with a majorized coordinate descendent algorithm, making the implementation easy to perform using existing packages. Intensive simulation studies show that the proposed estimation procedure outperforms alternatives for various cases. Finally, we apply the proposed SMIM and estimation procedure to a real data set from one of China's largest liquor companies, successfully identifying 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 extremely sensitive to model identification. In particular, parametric models may lead to biased estimation and selection, owing to the risk of misspecification, whereas nonparametric models may suffer from uninterpretability and instability of the resulting estimators, owing to the "curse of dimensionality." Semiparametric modeling offers a sensible

Corresponding author: Li Liu, School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China. E-mail: lliu.math@whu.edu.cn.

compromise. Multi-index models, which incorporate dimension reduction, are important semiparametric models that enjoy good asymptotic properties. However, they are not always useful in practice, because they still face the "curse of dimensionality" when the number of indices is even moderate, say three or larger. Here, a more useful approach is provided by structured multi-index models (SMIMs). Motivated by the multi-index stochastic frontier model described later, we consider the following model with a known link:

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

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

Model (1.1) is structured by specifying the link function m, which helps to incorporate information on the type of Y, and can be seen from special cases of the model. Model (1.1) includes many commonly used models, including the index heteroscedastic model (Zhu, Dong and Li (2013)) for continuous responses, $Y = f_1(\mathbf{X}'\boldsymbol{\beta}_1) + f_2(\mathbf{X}'\boldsymbol{\beta}_2)\varepsilon$, and the generalized additive-index model for various types of responses; that is, Y follows an exponential family of distribution with mean $m\{\sum_{j=1}^{d} f_j(\mathbf{X}'\boldsymbol{\beta}_j)\}$, where $f_j(\cdot)$ are unknown functions and m is a known link function determined by the type of Y, for example, a logit link for a binary response, a logarithmic function for a count response, and a linear function for a continuous response. Furthermore, generalized additive-index models take many commonly used models as special cases, such as single-index models and partial linear models. Studies on these kinds of models include the works of Carroll et al. (1997), Xia (2008), Ma and Zhu (2013), Liu, Cui and Li (2016), Guo, Box and Zhang (2017), Ke, Lian and Zhang (2020), Lian, Qiao and Zhang (2021), and the references therein. Except for the single-index models, these works focus on a fixed dimension of $\mathbf{X}$.

Model (1.1) cannot be addressed using existing methods. In particular, studies on multiple-index models focus on a fixed dimension of covariates. The methods for high-dimensional single-index models give an estimation and selection, and establish the asymptotic properties by avoiding an 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 model (1.1), which has multiple indexes and a specific structure. We pro-

vide semiparametrically efficient and computationally convenient estimators for all parameters and functions in a high-dimensional SMIM. The new estimation procedure is easy to implement, and simulation studies show that outperforms alternatives for models from the literature. We show theoretically that the estimators achieve semiparametric efficiency, in the sense of Bickel et al. (1993), which, to the best of our knowledge, has not been discussed for high-dimensional semiparametric models.

This study is motivated by the multi-index stochastic frontier model, which we use to analyze real data from one of the largest liquor companies in China. The purpose of the analysis is to investigate whether and how various factors affect the mean, frontier, inefficiency, and uncertainty of the sale 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 2,051 covariates and $n = 1941$ observations. The problem of measuring production inefficiency is important in economic, political, and social fields. One of the most satisfactory models for analyzing the problem is the stochastic frontier model introduced by Aigner, Lovell and Schmidt (1977), which is expressed as follows:

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

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

When analyzing model (1.2), a parametric functional form for f, which is usually linear in $\boldsymbol{\beta}$, has become standard practice in studies that measure efficiency. Because a misspecification in f may lead to erroneous conclusions, Fan, Li and Weersink (1996) considered model (1.2) with a completely unspecified $f(\cdot)$. Kumbhakar et al. (2007) further generalized the work of Fan, Li and Weersink (1996) by allowing the variances of the inefficiency score α_i and the measurement error ε_i to depend 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, Li and Weersink (1996) and Kumbhakar et al. (2007), even when the dimension

of the covariates is greater than three.

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

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

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; and $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$ are unknown coefficients representing the effect of $\mathbf{X}_i$ on the frontier, inefficiency, and variance functions, respectively. In particular, the covariates that affect the frontier, inefficiency, and variance may be different. By identifying the zero components in $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\beta}_3$, we can select subsets of $\mathbf{X}_i$ that are significant for the 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, which circumvents the problem of fitting high-dimensional surfaces and avoids the so-called curse of dimensionality. Model (1.3) is clearly a special case of (1.1) with $\boldsymbol{\varepsilon}_i = (\alpha_i,\ \varepsilon_i)'$.

In this study, we focus on the ultrahigh-dimensional setting for (1.1) with $p_n \gg n$, specifically, $\log(p_n) = O(n^r)$, for $0 < r < 1$. Although model (1.1) can be viewed as a unified framework accommodating some commonly used models, this study also develops a new and efficient estimation procedure that applies to any model in the unified framework.

The remainder of this paper is organized as follows. Section 2 describes 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 semiparametric efficiency. The performance of the proposed estimation procedures is assessed using simulation studies in Section 4. Here, we examine how well the proposed estimation procedures work. In Section 5, we apply the proposed SMIM and one-step estimation procedure to a real data set from one of China's largest liquor companies to explore the important factors affecting the sale of liquor. Technical proofs are relegated to the Supplementary Material. A user-friendly R package for implementing the proposed method is available at `https://github.com/LinhzLab/SMIM2.git`.

2. Estimation Procedure

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

$$L(\boldsymbol{\beta}, \boldsymbol{f}) = \frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ f_1(\mathbf{X}_i'\boldsymbol{\beta}_1), \ldots, f_d(\mathbf{X}_i'\boldsymbol{\beta}_d)\Big). \tag{2.1}$$

When the distribution of $\boldsymbol{\varepsilon}$ is given, this objective function is the conditional log-likelihood function given $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. 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 $L(\boldsymbol{\beta}, \boldsymbol{f})$ the log-likelihood function.

2.1. Kernel estimation

The proposed kernel estimation is based on back-fitting and profile likelihood estimation. The details are as follows. Pretending $\boldsymbol{\beta}_k$ are known, we apply the idea of back-fitting to estimate $f_k(\cdot)$.

Step I. We assume $f_j(\cdot) = f_j^{[\ell+1]}(\cdot)$, for $j = 1, \ldots, k-1$, $f_j(\cdot) = f_j^{[\ell]}(\cdot)$, and $j = k+1, \ldots, d$, just after the ℓth iteration. In the $\ell+1$th iteration, we update $f_k(\cdot)$ in the following way. For each given k, for $k = 1, \ldots, d$, and any given x, by Taylor's expansion, we have $f_k(\mathbf{X}_i'\boldsymbol{\beta}_k) \approx f_k(x) + \dot{f}_k(x)(\mathbf{X}_i'\boldsymbol{\beta}_k - x) \doteq \eta_{kx1} + \eta_{kx2}(\mathbf{X}_i'\boldsymbol{\beta}_k - x)$ when $\mathbf{X}_i'\boldsymbol{\beta}_k$ is in $B(x)$, a small neighborhood of x. In other words,

$$f_k(\mathbf{X}_i'\boldsymbol{\beta}_k) \approx \left\{\eta_{kx1} + \eta_{kx2}(\mathbf{X}_i'\boldsymbol{\beta}_k - x)\right\} I_{ik}(x) + f_k(\mathbf{X}_i'\boldsymbol{\beta}_k)\left\{1 - I_{ik}(x)\right\}, \tag{2.2}$$

for any $i = 1, \ldots, n$ and $k = 1, \ldots, d$, where $I_{ik}(x) = I(\mathbf{X}_i'\boldsymbol{\beta}_k \in B(x))$. Using (2.2), we extract information on $(f_k(x), \dot{f}(x))$ from all of the samples $i = 1, \ldots, n$. Substituting (2.2) into $L(\boldsymbol{\beta}, \boldsymbol{f})$, we estimate $\boldsymbol{\eta}_{kx} = (\eta_{kx1}, \eta_{kx2})'$ based on the following log-likelihood function for $\boldsymbol{\eta}_{kx}$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Q\Big(&Y_i,\ f_1(\mathbf{X}_i'\boldsymbol{\beta}_1), \ldots,\ f_{k-1}(\mathbf{X}_i'\boldsymbol{\beta}_{k-1}),\ W_{ix}(\boldsymbol{\beta}_k)'\boldsymbol{\eta}_{kx} I_{ik}(x)\\ &+ f_k(\mathbf{X}_i'\boldsymbol{\beta}_k)\left\{1 - I_{ik}(x)\right\}, f_{k+1}(\mathbf{X}_i'\boldsymbol{\beta}_{k+1}), \ldots, f_d(\mathbf{X}_i'\boldsymbol{\beta}_d)\Big),\end{aligned} \tag{2.3}$$

where $W_{ix}(\boldsymbol{\beta}_k) = (1,\ \mathbf{X}_i'\boldsymbol{\beta}_k - x)'$. Note that, with the approximation (2.2), our estimation for $\boldsymbol{\eta}_{kx}$ is based on a full likelihood function rather than a local likelihood function, which is commonly used in the nonparametric literature (Fan, Lin and Zhou (2006)). Differentiating (2.3) 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:

$$L_k(\boldsymbol{\beta}, \boldsymbol{f}; x) \hat{=} \frac{1}{n}\sum_{i=1}^{n} Q^{(01,k)}\Big(Y_i, f_1(\mathbf{X}_i'\boldsymbol{\beta}_1), \ldots, f_{k-1}(\mathbf{X}_i'\boldsymbol{\beta}_{k-1}), W_{ix}(\boldsymbol{\beta}_k)'\boldsymbol{\eta}_{kx}, f_{k+1}(\mathbf{X}_i'\boldsymbol{\beta}_{k+1}), \ldots, f_d(\mathbf{X}_i'\boldsymbol{\beta}_d)\Big) W_{ix}(\boldsymbol{\beta}_k) K_{ix}(\boldsymbol{\beta}_k) = 0, \tag{2.4}$$

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 the component k of $\partial Q(y, \mathbf{v})/\partial \mathbf{v}$. By (2.4), we obtain the updated $f_k(x)$, $f_k^{[\ell+1]}(x)$.

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

We consider the estimation of $\boldsymbol{\beta}$. The covariates are ultrahigh dimensional, and an extra task is to select the important covariates. Replacing $f_k(\cdot)$ in (2.1) with $f_k^{\hat{K}er}(\cdot; \boldsymbol{\beta})$ and applying a penalized estimation, we have the penalized likelihood

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ f_1^{\hat{K}er}(\mathbf{X}_i'\boldsymbol{\beta}_1; \boldsymbol{\beta}), \ldots,\ f_d^{\hat{K}er}(\mathbf{X}_i'\boldsymbol{\beta}_d; \boldsymbol{\beta})\Big) - \sum_{k=1}^{d}\sum_{j=1}^{p_n} \lambda_n \rho_{\lambda_n}(|\beta_{kj}|), \tag{2.5}$$

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. Next, we maximize (2.5) with respect to $\boldsymbol{\beta}_k$ subject to $\|\boldsymbol{\beta}_k\| = 1$ and $\beta_{k1} > 0$, for $k = 1, \ldots, d$. We use the resulting maximizer to estimate $\boldsymbol{\beta}_k$, and denote them by $\boldsymbol{\beta}_k^{\hat{K}er}$. Let $\boldsymbol{\beta}^{\hat{K}er}$ be $\boldsymbol{\beta}$, with each $\boldsymbol{\beta}_k$ replaced by $\boldsymbol{\beta}_k^{\hat{K}er}$. We use $\boldsymbol{f}^{\hat{K}er}(\cdot; \boldsymbol{\beta}^{\hat{K}er})$ to estimate $\boldsymbol{f}(\cdot)$, and denote it by $\boldsymbol{f}^{\hat{K}er}(\cdot)$, with $f_k^{\hat{K}er}(\cdot)$ being the kth component.

Although the kernel estimation enjoys good asymptotic properties, including consistency, asymptotic normality, and semiparametric efficiency, which are established in Section 3, it is difficult to implement. Here, we provide an algorithm that is computationally practical and has the same asymptotic properties as the kernel estimation.

2.2. Algorithm

The asymptotic theory for nonparameter estimators based on kernel smoothing or local-polynomial smoothing is better understood and established than that based on spline smoothing. Moreover, the computation based on spline smoothing is simpler than that based on kernel smoothing. Hence, the algorithm introduced in this subsection is a one-step kernel estimation based on the estimators obtained from a B-spline method.

2.2.1. B-spline estimation

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

$$f_k(x) \approx f_{k,n}(x) = \boldsymbol{B}(x)'\boldsymbol{\theta}_k, \quad k = 1, \ldots, d, \tag{2.6}$$

where $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kq_n})'$. Replacing $f_k(\cdot)$ in (2.1) by their approximations using (2.6) leads to the following penalized objection function for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_k$:

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_1)'\boldsymbol{\theta}_1, \ldots,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_d)'\boldsymbol{\theta}_d\Big) - \sum_{k=1}^{d}\sum_{j=1}^{p_n} \lambda_n \rho_{\lambda_n}(|\beta_{kj}|), \tag{2.7}$$

where β_{kj} is the jth component of $\boldsymbol{\beta}_k$. We maximize (2.7) with respect to $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_k$, subject to $\|\boldsymbol{\beta}_k\| = 1$ and $\beta_{k1} > 0$, and denote the maximizers as $\tilde{\boldsymbol{\beta}}_k$ and $\tilde{\boldsymbol{\theta}}_k$. The initial estimators of $f_k(\cdot)$ and $\boldsymbol{\beta}_k$ are taken as $\tilde{f}_k(\cdot) = \boldsymbol{B}(\cdot)'\tilde{\boldsymbol{\theta}}_k$ and $\tilde{\boldsymbol{\beta}}_k$.

2.2.2. One-step kernel estimation

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

For each k, $k = 1, \ldots, d$, and any given x, replacing $\boldsymbol{\beta}_j$ in (2.1) with $\tilde{\boldsymbol{\beta}}_j$, for $j = 1, \ldots, d$, and $f_j(\cdot)$ with $\tilde{f}_j(\cdot)$, for $j = 1, \ldots, k-1,\ k+1, \ldots, d$, and applying the local linear estimation, we obtain the local log-likelihood function for $f_k(x)$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \tilde{f}_1(\mathbf{X}_i'\tilde{\boldsymbol{\beta}}_1), \ldots,\ \tilde{f}_{k-1}(\mathbf{X}_i'\tilde{\boldsymbol{\beta}}_{k-1}),\ W_{ix}(\tilde{\boldsymbol{\beta}}_k)'\boldsymbol{\eta}_k,\\ &\tilde{f}_{k+1}(\mathbf{X}_i'\tilde{\boldsymbol{\beta}}_{k+1}), \ldots,\ \tilde{f}_d(\mathbf{X}_i'\tilde{\boldsymbol{\beta}}_d)\Big)K_{ix}(\tilde{\boldsymbol{\beta}}_k).\end{aligned} \tag{2.8}$$

We maximize (2.8) with respect to $\boldsymbol{\eta}_k$, and take the estimator of $f_k(x)$, $\hat{f}_k(x; \tilde{\boldsymbol{\beta}}_k)$,

as the first component of the maximizer.

Once the estimators $\hat{f}_k(\cdot;\boldsymbol{\beta})$ are obtained, we apply the penalized maximum likelihood estimation to estimate $\boldsymbol{\beta}$. Specifically, we maximize

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \hat{f}_1(\mathbf{X}_i'\boldsymbol{\beta}_1),\ldots,\ \hat{f}_d(\mathbf{X}_i'\boldsymbol{\beta}_d)\Big) - \sum_{k=1}^{d}\sum_{j=1}^{p_n}\lambda_n\rho_{\lambda_n}(|\beta_{kj}|) \tag{2.9}$$

with respect to $\boldsymbol{\beta}$. We use the resulting maximizer to estimate $\boldsymbol{\beta}$, and denote it as $\hat{\boldsymbol{\beta}}$. Then we define $\hat{\boldsymbol{f}}(x)$ as $\hat{\boldsymbol{f}}(x;\hat{\boldsymbol{\beta}})$, with $\hat{f}_k(x)$ being the kth component.

2.3. Computational issue and selection of the tuning parameters

When implementing the proposed estimation procedure, we have to deal with some practical issues, such as maximizing (2.7), (2.8), and (2.9), and need to select an initial estimation, bandwidth, tuning parameter, and penalty function.

We start with the initial estimation to address the maximization of (2.7). For this purpose, note that model (1.1) satisfies

$$E(Y|\mathbf{X}) = m_1\big(f_k(\mathbf{X}'\boldsymbol{\beta}_k), k\in\boldsymbol{\tau}_1\big), \tag{2.10}$$

$$var(Y|\mathbf{X}) = m_2\big(f_k(\mathbf{X}'\boldsymbol{\beta}_k), k\in\boldsymbol{\tau}_2\big), \tag{2.11}$$

where m_1 and m_2 are known link functions, and $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ are the subscript sets of the multiple indices related to the conditional mean and the conditional variance respectively. Without loss of generality, we suppose that the multiple indexes in (2.10) and (2.11) share a common part, that is, $\boldsymbol{\tau}_1\cap\boldsymbol{\tau}_2=\boldsymbol{\tau}_3$, with $\boldsymbol{\tau}_1\cup\boldsymbol{\tau}_2=\{1,\ldots,d\}$. Then, for $k\in\boldsymbol{\tau}_1$, based on (2.10), we obtain the initial estimators $\boldsymbol{\beta}_k^{(0)}$ by using the package $mave()$ in R (Xia (2008)). Note that for the ultrahigh-dimensional case, the 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 obtain $\boldsymbol{\theta}_k^{(0)}$ as the minimizer of $\sum_{i=1}^{n}(Y_i - m_1(\boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_k^{(0)})\boldsymbol{\theta}_k, k\in\boldsymbol{\tau}_1))^2$ with respect to $(\boldsymbol{\theta}_k, k\in\boldsymbol{\tau}_1)$ by using the $optim()$ function in R. The initial estimators of $f_k(\cdot)$, for $k\in\boldsymbol{\tau}_1$, 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 (2.11), for $k\in\boldsymbol{\tau}_2\backslash\boldsymbol{\tau}_3$.

Then, the maximizer of (2.7) can be obtained using the following iteration:

(I) Substituting $\boldsymbol{\theta}_k^{(0)}$ for $\boldsymbol{\theta}_k$ in (2.7), we have

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_1)'\boldsymbol{\theta}_1^{(0)},\ \ldots,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_d)'\boldsymbol{\theta}_d^{(0)}\Big) - \sum_{k=1}^{d}\sum_{j=1}^{p_n}\lambda_n\rho_{\lambda_n}(|\beta_{kj}|). \tag{2.12}$$

Maximize (2.12) with respect to $\boldsymbol{\beta}_k$ by taking $\boldsymbol{\beta}_k^{(0)}$ as the initial values, and denote the resulting maximizer by $\boldsymbol{\beta}_k^{(1)}$. This can be done using the MM principle (Lange, Hunter and Yang (2000)) and the $grpref()$ function in R.

(II) Substitute $\boldsymbol{\beta}_k^{(1)}$ for $\boldsymbol{\beta}_k$ in (2.7), and maximize (2.7) with respect to $\boldsymbol{\theta}_k$, that is, maximize

$$\frac{1}{n}\sum_{i=1}^{n} Q\Big(Y_i,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_1^{(1)})'\boldsymbol{\theta}_1,\ \ldots,\ \boldsymbol{B}(\mathbf{X}_i'\boldsymbol{\beta}_d^{(1)})'\boldsymbol{\theta}_d\Big). \tag{2.13}$$

This can be done using the $optim()$ function in R. Treat $\boldsymbol{\beta}_k^{(1)}$ and the resulting maximizer as the initial values of $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_k$, and repeat steps (I) and (II) until convergence, until we obtain the maximizer of (2.7).

The maximization of (2.8) can also be done using the $optim()$ function, and the $grpref()$ function in R can be used to maximize (2.9).

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

There are numerous studies on penalized estimation, and various penalty functions have been proposed. Examples include the LASSO of Tibshirani (1996), smoothly clipped absolute deviation (SCAD) of Fan and Li (2001), minimax concave penalty (MCP) of Zhang (2010), and elastic net of Zou and Hastie (2005). In this study, we use the MCP. The tuning parameter λ_n in the proposed estimation procedure plays a very important role. When the dimension of $\mathbf{X}$ is of a polynomial order of the sample size n, we apply the BIC to select λ_n; see (Fan and Li (2001)). When the dimension of $\mathbf{X}$ increases with an exponential order of the sample size n, the BIC does not work very well. In this case, we use the EBIC proposed in Chen and Chen (2008) to select λ_n.

3. Asymptotic Properties

Before presenting our main asymptotic results, we introduce some notation. Define $\mathcal{A}_k$ and $\mathcal{A}$ as the nonzero 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 parameter/function, for example, $\boldsymbol{\beta}^0$ 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 $\mathcal{F}_k = \{f_k : f_k \text{ has continuous } r\text{th-order derivatives}\}$ 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 this paper, C is a constant that 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 $\dot{\mathbf{S}}_{\boldsymbol{\beta}_k\boldsymbol{\beta}_{\tilde{k}}}(\boldsymbol{\beta}, \boldsymbol{f}) = \partial^2 L(\boldsymbol{\beta}, \boldsymbol{f})/\partial \boldsymbol{\beta}_k\partial \boldsymbol{\beta}'_{\tilde{k}}$, for $k, \tilde{k} = 1, \ldots, d$. We use a capital letter to denote a random variable, and its lowercase to denote its expectation, for example, $\mathbf{s}_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}, \boldsymbol{f}) = E\mathbf{S}_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}, \boldsymbol{f})$. The vector of $\{x_j, j \in \mathcal{A}\}$ is denoted as $\mathbf{x}_{\mathcal{A}}$, and the matrix $(V_{ij}, i \in \mathcal{A}, j \in \mathcal{A})$ is denoted as $\mathbf{V}_{\mathcal{A}\mathcal{A}}$. 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 the matrix A.

Before establishing the asymptotic properties of the proposed estimators $\hat{f}_k(\cdot)$ and $\hat{\boldsymbol{\beta}}_k$, 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 Conditions* (C2)–(C4) *in the Supplementary Material, 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 one 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), $\kappa(\rho_{\lambda_n}; \boldsymbol{\beta})$ is equal to $\max_{1 \leq k \leq d, 1 \leq j \leq p_n} -\rho''(|\beta_{kj}|)$, 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 derivative 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 that 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 a 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 that 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 a value in the neighborhood of $(\boldsymbol{\beta}^0,\boldsymbol{f}^0)$. Furthermore, the second-order Condition (C6) in the Supplementary Material ensures that the maximizer 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 can show the asymptotic properties of the kernel estimators, $f_k^{\hat{K}er}(\cdot)$ and $\boldsymbol{\beta}_k^{\hat{K}er}$, and then prove that the estimators $\hat{f}_k(\cdot)$ and $\hat{\boldsymbol{\beta}}_k$ based on the proposed algorithm have the same asymptotic properties.

Theorem 1. *Under regularity Conditions* (C1)–(C7) *in the Supplementary Material, if* $h_n\to0$, $nh_n/\log n\to\infty$, $\varphi_n\leq Cn^{-\gamma}$, *and*

$$\begin{aligned}
&\frac{n}{(\log p_n)^{\iota_1}}\left\{\frac{(\dot{\rho}_{\lambda_n}^{-1}(m_{\boldsymbol{\beta}})\wedge n^{\gamma})^2}{\varphi_n^2s_n^2}\wedge\frac{\dot{\rho}_{\lambda_n}^{-1}(m_{\boldsymbol{\beta}})\wedge n^{\gamma}}{\varphi_ns_n}\right\}\to\infty,\\
&\frac{n}{(\log s_n)^{\iota_1}}\left\{\frac{(\varphi_n^{-1}\wedge\mu_n)^2}{s_n^2}\wedge\frac{\varphi_n^{-1}\wedge\mu_n}{s_n}\right\}\to\infty,\quad\frac{n\lambda_n^2}{(\log p_n)^{\iota_2}}\to\infty,\\
&\left\{\frac{n^{(1-2\gamma)/2}\lambda_n}{(\log s_n)^{\iota_2}}\wedge\frac{n^{1-2\gamma}\lambda_n^2}{(\log s_n)^{\iota_2}}\right\}\to\infty,\quad m_{\boldsymbol{\beta}}\geq C\varphi_n\lambda_n\dot{\rho}(0+),\quad s_n\lambda_n\to0,\\
&\frac{\{\lambda_n/n^{\gamma}\}\wedge\{\varphi_n/s_n\}}{h_n^2+(nh_n)^{-1/2}\log^{1/2}(n)}\to\infty,\quad\varphi_n\lambda_n\leq C(h_n^2+(nh_n)^{-1/2}\log^{1/2}(n)),
\end{aligned}\tag{3.1}$$

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

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

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

(c) $\sup_{x\in\mathcal{U}}\|f_k^{\hat{K}er}(x)-f_k^0(x)\|\to0$ *in probability.*

For a bounded covariate, it can be seen that $\iota_1=1$ and $\iota_2=1$ by letting $\iota\to\infty$. Then, (3.1) holds when $n/(\varphi_n^2s_n^2\log p_n)\to\infty$, which holds if $\log p_n=o(n)$

and $s_n = o(\sqrt{n})$ when φ_n takes a constant. This means that the kernel estimation procedure is applicable to the ultrahigh-dimensional case in which the number of covariates is of an exponential order of the sample size n. The last three conditions in (3.17) guarantee that we search 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 that the kernel estimators, $\boldsymbol{\beta}^{\hat{K}er}$, enjoy selection consistency, and (b) implies the estimate consistency of $\boldsymbol{\beta}^{\hat{K}er}$, that is, $\|\boldsymbol{\beta}_{\mathcal{A}}^{\hat{K}er} - \boldsymbol{\beta}_{\mathcal{A}}^0\|_\infty \to 0$ 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 (3.1). Lastly, (c) illustrates that the estimators $f_k^{\hat{K}er}(\cdot)$ of $f_k(\cdot)$ 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 let $\boldsymbol{\Sigma}_{2n}$ be the covariance matrix of $_{\mathcal{A}}$, 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 the conditions of Theorem* 1*, if*

$$\begin{aligned} &\frac{\Lambda_{3n} n h_n^2}{s_n (\log n)^2} \to \infty, \quad \frac{n + h_n^{-4}}{(1 \vee \Lambda_{3n}) s_n^3} \to \infty, \quad \frac{\Lambda_{2n}^2 (n + h_n^{-4})}{s_n^2} \to \infty, \\ &\frac{n(\Lambda_{1n}^2 - h_n^4)}{s_n^2 (\log s_n)^{\iota_1}} \to \infty, \quad \frac{\Lambda_{1n}^4 \Lambda_{3n} (n + h_n^{-4})}{s_n^3} \to \infty, \quad \frac{n s_n \lambda_n^2 \dot{\rho}_{\lambda_n}(m_{\boldsymbol{\beta}})^2}{\Lambda_{1n}^2 \Lambda_{3n}} \to 0, \end{aligned} \tag{3.2}$$

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*

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

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

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

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 establish the asymptotic normality of the estimators $\hat{f}_k(x)$ and $\hat{\boldsymbol{\beta}}_k$ from the proposed algorithm.

Theorem 3. *Under* (3.2) *and the conditions in Theorem* 1*,*

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

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

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

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

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

When the eigenvalues Λ_j, for $j = 1,\ 2,\ 3$, are bounded away from zero, it is easy to see that 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 $h_n = O(n^{-1/3})$; furthermore, part (b) holds for the theoretical optimal bandwidth $h_n = O(n^{-1/5})$ of the 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 has the same asymptotic distribution as that of 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* $\int_{\mathcal{U}^d} \boldsymbol{\psi}_1'(\mathbf{x}) \boldsymbol{f}^{\hat{K}er}(\mathbf{x}) d\mathbf{x} + \boldsymbol{\psi}_2' \boldsymbol{\beta}_{\mathcal{A}}^{\hat{K}er}$ *and* $\int_{\mathcal{U}^d} \boldsymbol{\psi}_1'(\mathbf{x}) \hat{\boldsymbol{f}}(\mathbf{x}) d\mathbf{x} + \boldsymbol{\psi}_2' \hat{\boldsymbol{\beta}}_{\mathcal{A}}$ *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 $\boldsymbol{\beta}_{\mathcal{A}}^{\hat{K}er}$ and $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ are efficient estimators of $\boldsymbol{\beta}_{\mathcal{A}}^0$ by taking $\boldsymbol{\psi}_1(\mathbf{x}) = \mathbf{0}$, and that $\boldsymbol{f}^{\hat{K}er}(\cdot)$ and $\hat{\boldsymbol{f}}(\cdot)$ 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 by comparing existing competing procedures in terms of their bias, efficiency, predictive accuracy, and selection accuracy. To ensure the feasibility of the comparison, the settings and evaluation criteria of the first two simulations are taken from the related literature. 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 the MCP selector to select the important variables. The tuning parameter λ_n is determined using the BIC and EBIC principles for the high-dimensional and ultrahigh-dimensional cases, respectively.

Simulation 1. The setting is the same as that of Alquier and Biau (2013), in which we consider single-index models with $p_n = 10$ or 50 and a sample size of $n = 50$ 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 $\hat{f}_{Fourier}$ in Alquier and Biau (2013), the estimation $\hat{f}_{HHI}$ in Härdle, Hall and Ichimura (1993), the LASSO estimator $\hat{f}_{LASSO}$, and the standard kernel estimate $\hat{f}_{NW}$ (Nadaraya (1964); Watson (1964)).

The median, mean, and standard deviation (SD) of the MSPE based on 200 repetitions are shown in Table 1, which suggests that the proposed method has a much smaller predictive error than the competing procedures do. Compared with the LASSO estimator, this result is natural, because the LASSO estimator does not enjoy the variable selection oracle property. In addition, the MSPEs of the proposed estimators $\hat{f}_j$, for $j = 1, 2, 3$, with the smoothing parameters $q_n =$ 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 that of Case 1 in Zhu, Dong and Li (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 1,000 times with a sample size of $n = 600$.

We compare the proposed method with that of Zhu, Dong and Li (2013). Table 2 summarizes the bias, standard deviation (SD), and root mean squared error (RMSE) of the estimates for the nonzero elements of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. To evaluate the performance of the estimators for both the parameters and the functions, we also calculate the average squared error, defined as $ASE_j = n^{-1}\sum_{i=1}^{n}(\hat{f}_j(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_j) - f_j^0(\mathbf{X}_i'\boldsymbol{\beta}_j^0))^2$, for $j = 1, 2$. Table 2 shows that the proposed method is much more efficient and accurate than the method of Zhu, Dong and Li (2013), because the proposed method selects the important variables and estimate parameters and functions simultaneously, whereas the estimating equation method in Zhu, Dong and Li (2013) considers only 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 are then trimmed into the range $[-1, 1]$. The coefficients are 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 component. The functions are taken as $f_1(x) = \exp(x/2) + 2x^2$, $f_2(x) = ((x-1)^2 +$

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

$n = 50$	$p_n = 10$	$\hat{f}_{Fourier}$	$\hat{f}_{HHI}$	$\hat{f}_{LASSO}$	$\hat{f}_{NW}$	$\hat{f}^{(1)}$	$\hat{f}^{(2)}$	$\hat{f}^{(3)}$
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
$n = 100$	$p_n = 10$	$\hat{f}_{Fourier}$	$\hat{f}_{HHI}$	$\hat{f}_{LASSO}$	$\hat{f}_{NW}$	$\hat{f}^{(1)}$	$\hat{f}^{(2)}$	$\hat{f}^{(3)}$
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
$n = 50$	$p_n = 50$	$\hat{f}_{Fourier}$	$\hat{f}_{HHI}$	$\hat{f}_{LASSO}$	$\hat{f}_{NW}$	$\hat{f}^{(1)}$	$\hat{f}^{(2)}$	$\hat{f}^{(3)}$
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
$n = 100$	$p_n = 50$	$\hat{f}_{Fourier}$	$\hat{f}_{HHI}$	$\hat{f}_{LASSO}$	$\hat{f}_{NW}$	$\hat{f}^{(1)}$	$\hat{f}^{(2)}$	$\hat{f}^{(3)}$
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

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

$1)/4$, $f_3(x) = ((x + 1)^2 + 1)/4$. The simulation is repeated 1,000 times with a sample size of $n = 600$.

The simulation results are summarized in Tables 3 and 4 and Figure 1. Table 3 shows the results of the 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 is close to one and the FPR is close to zero. These results suggest that the proposed method not only

Table 2. Numerical results in Simulation 2.

Method		$\hat{\beta}_{11}\%$	$\hat{\beta}_{12}\%$	$\hat{\beta}_{13}\%$	$\hat{\beta}_{14}\%$	$\hat{\beta}_{21}\%$	$\hat{\beta}_{22}\%$	$\hat{\beta}_{28}\%$	$ASE_1\%$	$ASE_2\%$
(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, Dong and Li (2013); RMSE represents the root-mean-square error; ASE represents the average squared error, defined by $ASE_j = n^{-1}\sum_{i=1}^{n}(\hat{f}_j(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_j) - f_j^0(\mathbf{X}_i'\boldsymbol{\beta}_j^0))^2$.

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.

selects the important variables, but also rules 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 the three important covariates only. The results in Table 4 reveal that the proposed estimators are approximately unbiased, and their estimated standard errors (ESEs) agree well with the sample standard deviations (SDs). Moreover, the proposed method produces coverage percentages of the 95% confidence intervals that are close to the nominal level. Furthermore, the proposed procedure performs comparably well with the oracle estimator.

Figure 1 (a) displays the frontier function estimated using the proposed method. Neglecting the noise, the frontier function $f_1(\mathbf{x}'\boldsymbol{\beta}_1)$ is the maximum attainable output with the input $\mathbf{x}$. To see that, we further generated a validation data set with a sample size of 600 from the same model, displayed as

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

Parameter	$\hat{\boldsymbol{\beta}}$				$\boldsymbol{\beta}^{\hat{O}R}$			
	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

$\hat{\boldsymbol{\beta}}$ represents the proposed estimator; $\boldsymbol{\beta}^{\hat{O}R}$ 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.

the star point in Figure 1 (a). This plot shows that the statistical noise in the nonparametric scenario does not affect the estimation.

To further evaluate the performance of the nonparametric function estimators and to compare the prediction effect of the proposed method with that of the competing gradient boosting approach, for each repetition of 1,000 replications, we applied the fitted model to predict a newly generated data set. Figure 1 (b) and (d) display scatter plots of the true values of Y_i against the fitted values of $\hat{Y}_i = \hat{f}_1(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_1) + \hat{u}_i$ of the proposed method and the gradient boosting approach, respectively. Figure 1 (c) and (f) display scatter plots of the true simulated e_i against its predictor $\hat{e}_i = Y_i - \hat{Y}_i$ for the proposed method and the gradient boosting approach, respectively, where $\hat{u}_i = \hat{\sigma}_i\hat{\lambda}_i/(1+\hat{\lambda}_i^2)[\varphi(-\hat{\xi}_i\hat{\lambda}_i/\hat{\sigma}_i)/\Phi(-\hat{\xi}_i\hat{\lambda}_i/\hat{\sigma}_i) - \hat{\xi}_i\hat{\lambda}_i/\hat{\sigma}_i]$, $\hat{\sigma}_i^2 = \hat{f}_2^2(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_2) + \hat{f}_3^2(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_3)$, $\hat{\lambda}_i = \hat{f}_2(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_2)/\hat{f}_3(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_3)$, and $\hat{\xi}_i = Y_i - \hat{f}_1(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_1)$, following Jondrow et al. (1982). From Figure 1 (b)–(f), we can see that the predictors using the proposed method work relatively well globally, and are comparable with those of the gradient boosting approach. In fact, the MSPEs based on 1,000 newly generated data sets are 1.450 for the proposed method, with a deviation of 0.286, and 3.104 for the gradient boosting approach, with a deviation of 0.420. This shows that the proposed method possesses both high prediction ability and interpretability.

Simulation 4. The data are generated 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 the ultrahigh-dimensional case. The simulation

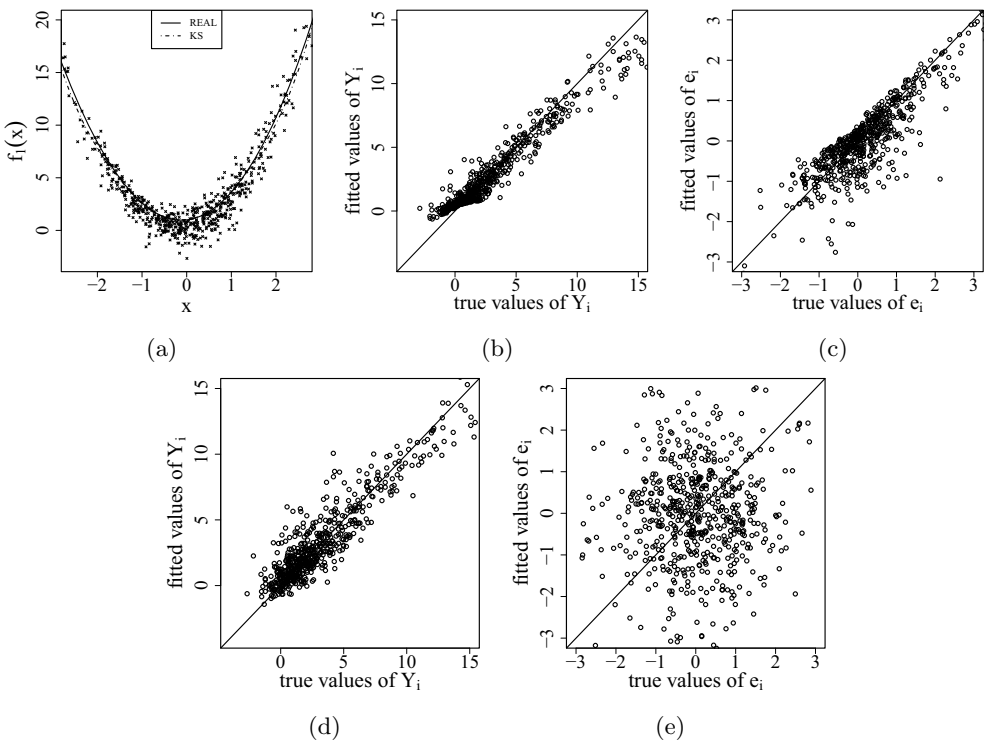

Figure 1. Plots of the nonparametric estimates in Simulation 3: true against estimated values. (a) shows the plot of true frontier and the fitted frontier function; (b) and (d) show the scatter plot of the true values and the fitted values of the response variable using the proposed method and the gradient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

results are summarized in Tables 3–4 and Figure 1 of the Supplementary Material. Moreover, we obtain MSPEs based on 1,000 newly generated data sets of 2.716 for the proposed method, with a deviation of 0.350, and 5.842 for the gradient boosting approach, with a deviation of 0.832. Therefore, we draw similar conclusions to those for Simulation 3.

5. An Application

In this section, we apply the proposed approach to analyze a data set from one of China's largest liquor companies. The purpose of the analysis is to investigate whether and how various factors affect the sale of liquor. The data set includes monthly sales (Y_i) and covariate information for $n = 1941$ observations in 31 provinces of China for the period 2011 to 2018. The covariates

comprise four parts: (1) the company's product information, including price, advertising investment, and reimbursement expense of dealers; (2) brewing industry information, including monthly liquor yields, monthly beer yields, beer imports and exports, monthly trading amounts of 12 stocks of the brewing industry, and the profit of affiliated companies; (3) economic information of related cities and towns, including per capita GDP, per capita disposable income, consumer price index, retail price index, total retail sales of consumer goods, housing sales prices, residential investment, and permanent population; and (4) geographic information, including monthly average temperature, monthly average relative humidity, geographical divisions, and the distance from the liquor producing area. Together with the lagged variables, we have 2,051 covariates. A log transformation is taken of the response variable, and all covariates and responses are standardized. Then, the multiple-index stochastic frontier model (1.3) and the proposed approach are applied to the data. The bandwidths are as described in Remark 3. The selected important variables and their regression coefficient 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 sales. Combining this with the monotone increasing function of $f_1(\cdot)$ in Figure 3 (a), we can draw the following conclusions. First, 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 a lower level of economic development buy more liquor, which is consistent with the fact that the considered product is cheap, and thus popular among low consumption groups. Second, the positive coefficients of residents and the previous year's value show that the greater the population, the larger is the demand for liquor. Third, the sales in the previous months (SL_lag1,2,4,5,6,7) have positive coefficients, which means that larger past sales result in larger sales in the current month, which is consistent with our intuition. Fourth, the price (PRICE_lag5) is statistically significant, because this is a low-end liquor product targeting price-sensitive low-consumption groups. Fifth, the coefficient of liquor_production (taking the value one if a city belongs to a province with large liquor production, and zero otherwise) is positive, which shows that people in a province with large liquor production tend to purchase more liquor. Sixth, the variable xlj_sichuan (taking the value one if a city is in Sichuan Province, where the liquor is produced, and zero otherwise) has quite a large coefficient, reflecting that the product sells well around the place of origin.

The 18 important variables selected for the inefficiency function are shown in Figure 2 (b). Combining this with the monotone increasing function of $f_2(\cdot)$ displayed in Figure 3 (b), we can draw the following conclusions. First, subsidy re-

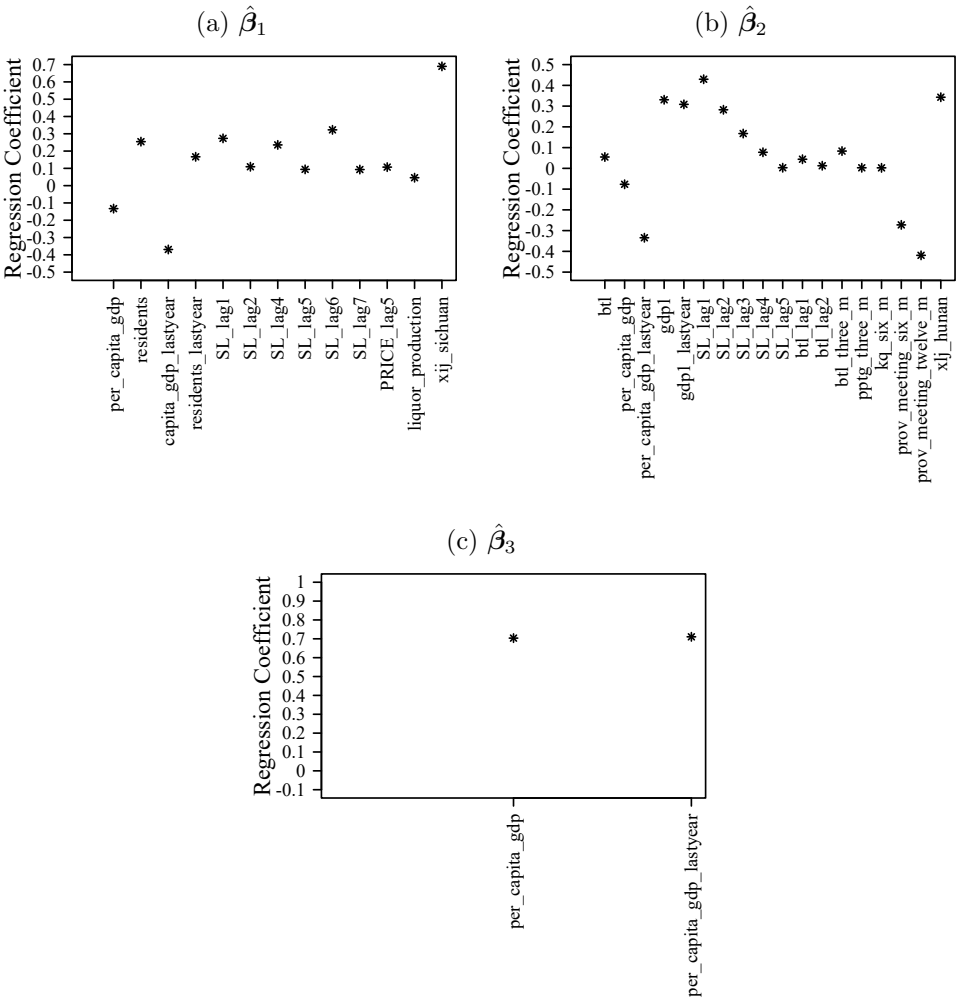

Figure 2. Selected important variables and their estimates of $\boldsymbol{\beta}_k, k = 1, 2, 3$ for liquor data.

imbursement expenses and their lagged variables (btl & btl_lag1, 2 & btl_three_m) have positive coefficients, illustrating that they are a relatively inefficient input. This is because the company subsidizes dealers based on their purchases before a particular day, and dealers usually buy much more than they can sell before that day, causing large inventories. Second, per capita GDP (per_capita_gdp) and its lagged variable (per_capita_gdp_lastyear) have negative coefficients, and thus have different effects on sales than that shown in Figure 2 (a). This may be because cities with a higher GDP are usually more efficient in terms of commercial operation. Third, the GDP from the primary industry, mainly agriculture,

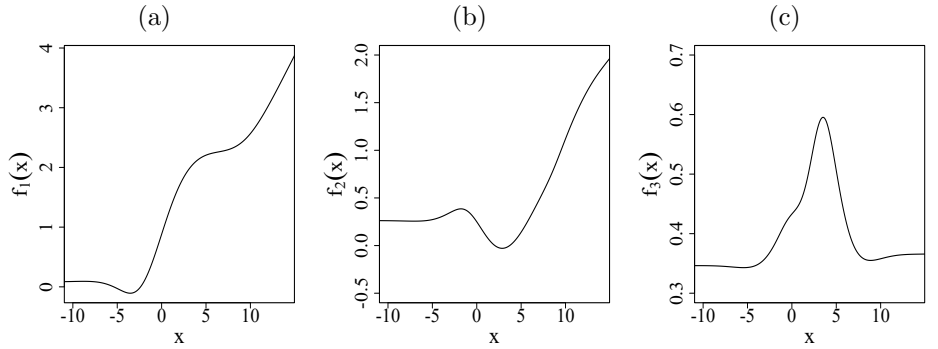

Figure 3. Plots of the kernel smoothing nonparametric estimates in liquor data analysis.

in the current and previous years (gdp1 & gdp1_lastyear) have positive regression coefficients, indicating a negative effect on sales. This may be because areas with a high agricultural output usually have a low commercial operation ability. Fourth, the positive coefficients of sales in the past one to five months (SL_lag1–5) show that there may be some waste of costs in areas with large sales in the past. Fifth, cumulative expenses on meetings and events, such as a wine expo, during the past half year and the past year (prov_meeting_six_m & prov_meeting_twelve_m, respectively) have a positive effect on sales, reflecting that effective promotional activities can increase the market share of the product. Sixth, the variable xlj_hunan (taking the value one if a city belongs to Hunan Province, which is the province with the second largest sales, and zero otherwise) has a positive regression coefficient, and thus a negative effect on sales. The dealers in this province may have some cost waste.

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

6. Conclusion

To investigate whether and how ultrahigh-dimensional factors affect various measurements, such as the mean, frontier, inefficiency, and variance, we propose an ultrahigh-dimensional structured multiple-index model. We estimate all of the functions and parameters based on a 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

nonconvexity of the likelihood function, nonsmoothness of the penalty term, and the large number of functions and ultrahigh-dimensional predictors, we blend spline and kernel smoothing with a majorized coordinate descendent algorithm, so that the computation is easily performed using existing software. Our simulation studies show that the proposed method outperforms existing methods in terms of selection and estimation for all of the cases considered, the settings of which are taken from the existing literature, if available. We apply the proposed method to a real data from one of China's largest liquor companies, finding that 31 out of 2051 factors, including price, previous sales, per capita GDP, and residents, are important for the mean, stochastic frontier, inefficiency, and variance of liquor sales.

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

Supplementary Material

The online Supplementary Material provides additional notation, conditions, the proofs of the theorems, results from Section 4, and additional simulation studies on mixed-effects models.

Acknowledgments

The authors would like to thank the associate editor and two referees for their constructive and insightful comments and suggestions. This research was supported by the National Key R&D Program of China (2022YFA1003702), the National Natural Science Foundation of China (11931014, 12171374, 11971362), and the New Cornerstone Science Foundation.

References

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

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

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and 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. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477–489.

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

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

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

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

Guo, S., Box, J. and 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. and Ichimura, H. (1993). Smooth regression analysis. *Ann. Statist.* **21**, 151–178.

Jondrow, J., Lovell, C. A. K., Materov, I. S. and 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. and 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. and Tsionas, E. G. (2007). Nonparametric stochastic frontiers: A local maximum likelihood approach. *J. Econometrics* **137**, 1–27.

Lange, K., Hunter, D. R. and 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. and Zhang, W. (2021). Homogeneity pursuit in single index models based panel data analysis. *J. Bus. Econom. Statist.* **39**, 386–401.

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

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

Ma, Y. and 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 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. and Li, R. (2013). Semiparametric estimation of conditional heteroscedasticity via single-index modeling. *Statist. Sinica* **23**, 1235–1255.

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

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

Huazhen Lin

Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China.

E-mail: linhz@swufe.edu.cn

Shuangxue Zhao

Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu 611130, China.

E-mail: zsxzsx@smail.swufe.edu.cn

Li Liu

School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China.

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

Wenyang Zhang

Department of Mathematics, University of York, York YO10 5DD, United Kingdom.

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

(Received January 2021; accepted December 2021)