Statistica Sinica **25** (2015), 551-566
doi:http://dx.doi.org/10.5705/ss.2013.175

ESTIMATION OF A GROUPWISE ADDITIVE MULTIPLE-INDEX MODEL AND ITS APPLICATIONS

Tao Wang, Jun Zhang, Hua Liang and Lixing Zhu

Yale University, Shenzhen University, George Washington University and Hong Kong Baptist University

Abstract: In this paper, we propose a simple linear least squares framework to deal with estimation and selection for a groupwise additive multiple-index model, of which the partially linear single-index model is a special case, and in which each component function has a single-index structure. We show that, somewhat unexpectedly, all index vectors can be recovered through a single least squares coefficient vector. As a direct application, for partially linear single-index models we develop a new two-stage estimation procedure that is iterative-free and easily implemented. This estimation approach can also be applied to develop, for the semi-parametric model under study, a penalized least squares estimation and establish its asymptotic behavior in sparse and high-dimensional settings without any nonparametric treatment. A simulation study and a data analysis are presented.

Key words and phrases: High dimensionality, index estimation, least squares, multiple-index models, variable selection.

1. Introduction

High-dimensional and complex data characterize many contemporary statistical applications, from areas as broad ranging as genomics, genetics, finance, and economics (Fan and Li (2006)). There is little doubt that high-dimensional data analysis has become important. In many practical situations, parametric models, such as the linear model and the generalized linear model, are among the most convenient and frequently used. However, they are not flexible enough to capture the underlying relationship between the response variable and its associated predictors, and one cannot sensibly check the fit of a parametric model with a large number of predictors.

Semi-parametric models (Ruppert, Wand, and Carroll (2003)) are increasingly used to balance modeling bias and the "curse of dimensionality". Semi-parametric models have the flexibility and good fit of nonparametric models and retains the parsimony and ease of interpretation of parametric models. Here, however, little work has been done on estimation with high-dimensional data. Wei, Huang, and Li (2011) studied the estimation and selection properties of

an adaptive group LASSO approach using B-spline basis approximation in time varying coefficient models. Xue and Qu (2012) proposed a penalized polynomial spline procedure for varying coefficient models by adopting a truncated L_1 penalty and investigated the global optimality properties of the penalized estimator. Alquier and Biau (2013) considered the single-index model estimation problem from a sparsity perspective using a PAC-Bayesian approach, but their approach offers no guarantees on the issue of variable selection. Wang, Xu, and Zhu (2012) studied the theoretical properties of a regularized linear least squares method for general singe-index models.

The partially linear single-index model is an important extension of the single-index model and of the partially linear model. A nice feature here is that the predictors under investigation fall into two groups affecting the response variable, making it easy to interpret the model parameters (Carroll et al. (1997)). To the best of our knowledge, there has been no work on estimation in partially linear single-index models when the number of predictors can be larger than the sample size. The estimation procedures developed here are applicable to partially linear single-index models and their extensions.

Consider the regression of a response variable $Y \in \mathbb{R}$ on a random vector of predictors $\boldsymbol{V} \in \mathbb{R}^d$. Suppose that $\boldsymbol{V} = (\boldsymbol{V}_1^\top, \boldsymbol{V}_2^\top, \ldots, \boldsymbol{V}_K^\top)^\top$ can be naturally divided into K non-overlapping groups $\boldsymbol{V}_k \in \mathbb{R}^{p_k}, k = 1, \ldots, K$. We consider the groupwise additive multiple-index model

$$Y = \sum_{k=1}^{K} g_k(\boldsymbol{\beta}_k^\top \boldsymbol{V}_k) + \epsilon, \tag{1.1}$$

where $g_k(\cdot)$ is an unknown component function, $\boldsymbol{\beta}_k \in \mathbb{R}^{p_k}$ is a single-index vector of interest corresponding to $\boldsymbol{V}_k$, and the random error ϵ is independent of $\boldsymbol{V}$. If $K = 1$, (1.1) is the well-known single-index model (Powell, Stock, and Stoker (1989)). If $K = 2$, with $g_1(t) = t$, then (1.1) is the partially linear single-index model. If, further, $p_2 = 1$, then it reduces to the partially linear model (Heckman (1986)). For further discussion, see Naik and Tsai (2001) and Lin and Kulasekera (2007).

Parameter estimation for (1.1), or its special cases, has received a great deal of attention in the literature. See, for instance, Carroll et al. (1997), Yu and Ruppert (2002), Li, Li, and Zhu (2010), Ruan and Yuan (2010), and references therein. In particular, Li, Li, and Zhu (2010) extended the minimum average variance estimation method of Xia et al. (2002) to deal with a more general model for groupwise dimension reduction. Generally, these methods are computationally demanding since the resulting estimators need to be solved via an iterative procedure.

With $d = p_1 + \cdots + p_K$, few results are available for estimation in this context when d diverges with n. We propose a simple linear least squares framework to discuss estimation, and can deal with high-dimensional data by apply existing variable selection techniques.

The rest of the paper is organized as follows. In Section 2, we discuss the issue of identifiability and introduce a linear least squares estimation procedure for model (1.1). Large sample properties are derived. In Subsection 3.1 we establish the theoretical properties of the least squares method for partially linear single-index models, and develop a new two-stage estimation procedure. Subsection 3.2 concerns the variable selection problem with high-dimensional predictors. We propose a penalized least squares method for selecting predictors in each component function, and study the asymptotic behavior of the penalized estimator in sparse and high-dimensional settings. Numerical studies are presented in Section 4 and Section 5. Proofs are provided in the supplementary material.

2. Identifiability and Estimation

We first discuss the identifiability of $\boldsymbol{\beta}_k$'s in model (1.1). Denote by $\boldsymbol{0}_{m\times 1}$ an $m \times 1$ vector of 0's, and let

$$\mathbf{S} = \begin{pmatrix} \boldsymbol{\beta}_1 & \boldsymbol{0}_{p_1\times 1} & \cdots & \boldsymbol{0}_{p_1\times 1} \\ \boldsymbol{0}_{p_2\times 1} & \boldsymbol{\beta}_2 & \cdots & \boldsymbol{0}_{p_2\times 1} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{0}_{p_K\times 1} & \boldsymbol{0}_{p_K\times 1} & \cdots & \boldsymbol{\beta}_K \end{pmatrix}.$$

Then, from (1.1), Y and $\boldsymbol{V}$ are independent conditioned on $\mathbf{S}^\top \boldsymbol{V}$. The column space of $\mathbf{S}$ is called the central dimension reduction subspace (Li (1991)); Cook (1998)) and is a well-defined population parameter.

When $\boldsymbol{\Sigma}_{\boldsymbol{V}} = \mathrm{Cov}(\boldsymbol{V})$ is positive-definite, define the least squares direction as

$$\boldsymbol{\beta}_{LS} = \boldsymbol{\Sigma}_{\boldsymbol{V}}^{-1}\mathrm{Cov}(\boldsymbol{V}, Y). \tag{2.1}$$

Then $\boldsymbol{\beta}_{LS}$ is in the column space of $\mathbf{S}$, provided

$$E(\boldsymbol{V}|\mathbf{S}^\top \boldsymbol{V}) \text{ is a linear function of } \mathbf{S}^\top \boldsymbol{V}. \tag{2.2}$$

This condition is satisfied, for example, when the distribution of $\boldsymbol{V}$ is elliptically symmetric and when the dimension of $\boldsymbol{V}$ is large, it is not restrictive; see Hall and Li (1993) and Cook and Ni (2005). Several efforts have been devoted to relaxing the condition, see Li and Dong (2009) and Dong and Li (2010). Feng, Wang, and Zhu (2012) recently provided a necessary and sufficient condition for the least squares coefficient vector to work in a similar scenario, and found it close to the linearity condition. This condition is not very strong when the inverse regression notion is adopted, and we use it here.

Proposition 1. *If* (2.2) *hold, then*

$$\boldsymbol{\beta}_{LS} = (\phi_1 \boldsymbol{\beta}_1^\top, \phi_2 \boldsymbol{\beta}_2^\top, \ldots, \phi_K \boldsymbol{\beta}_K^\top)^\top$$

for some constants $\phi_k, k = 1, \ldots, K$.

Thus, under mild conditions on the design distribution, the K index vectors $\boldsymbol{\beta}_k$ can be recovered simultaneously through a single vector $\boldsymbol{\beta}_{LS}$, if the additive index structure of model (1.1) holds true. The random error ϵ at (1.1) is allowed to be dependent on $\boldsymbol{V}$ such that $E(\epsilon|\mathbf{S}^\top \boldsymbol{V}) = 0$, so our results are still valid under heteroscedasticity.

To avoid confusion, the $\boldsymbol{\beta}_k$'s are redefined so $\boldsymbol{\beta}_{LS} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top$. Given a random sample $(\boldsymbol{v}_i, y_i)$, $i = 1, \ldots, n$, on $(\boldsymbol{V}, Y)$, we propose to estimate $\boldsymbol{\beta}_{LS}$ with the vector $\hat{\boldsymbol{\beta}}_{LS}$ from the least squares fit of y_i on $\boldsymbol{v}_i$. Denote by $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ the response vector and $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)^\top$ the design matrix. Assuming centered data, the intercept is not included in the regression function. The least squares direction estimator is

$$\hat{\boldsymbol{\beta}}_{LS} = (\hat{\boldsymbol{\beta}}_1^\top, \ldots, \hat{\boldsymbol{\beta}}_K^\top)^\top = (\mathbf{V}^\top \mathbf{V})^{-1} \mathbf{V}^\top \boldsymbol{y}. \tag{2.3}$$

For a vector $\boldsymbol{u} = (u_1, \ldots, u_m)^\top \in \mathbb{R}^m$, take $\|\boldsymbol{u}\|_1 = \sum_{j=1}^m |u_j|$ and $\|\boldsymbol{u}\|_2 = (\sum_{j=1}^m u_j^2)^{1/2}$. For the time being, all the predictors are relevant to the response variable and their total number, $d = d_0$, is allowed to diverge as the sample size n tends to infinity.

Theorem 1. *Under the conditions* (A1)−(A7) *in the supplementary document, if* $d_0 = o(n/\log n)$ *then* $\|\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}_{LS}\|_2 = O(\sqrt{d_0/n})$. *Consequently,* $\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_2 = O(\sqrt{d_0/n})$ *for all* $k = 1, \ldots, K$.

Similar results for the linear model and for the single-index model are available in the literature. But, to our knowledge, the results for the additive index model (1.1) are novel.

Remark 1. There is no guarantee that the constants ϕ_k's in Proposition 1 are different from zero. With $\boldsymbol{\zeta}_k = \{\mathrm{Cov}(\boldsymbol{V}_k)\}^{-1} \mathrm{Cov}\{\boldsymbol{V}_k, g_k(\boldsymbol{\beta}_k^\top \boldsymbol{V}_k)\}$ for $k = 1, \ldots, K$, if the predictors are independent, a sufficient condition is that $\boldsymbol{\zeta}_k \neq \boldsymbol{0}_{p_k \times 1}$ for all k. In the case of perfect correlation, $\boldsymbol{V}_k = \boldsymbol{V}_l$ for some $k \neq l$, model (1.1) is clearly not identifiable. As a result, there must be some linear trend in each component function (Wang, Xu, and Zhu (2012)) and some regularization constraint imposed on the linear relationship among the components $\boldsymbol{V}_k$. For partially linear single-index models, we give the conditions for identifiability in Section 3.1.

If ϕ_k, or more precisely $\boldsymbol{\beta}_k$, is nonzero and estimated as nonzero, dimension reduction within $\boldsymbol{V}_k$ is achieved; for those ϕ_k's that are zero and estimated to be zero using the method in Subsection 3.2, a more sophisticated method is needed. Thus, we can combine our method for dimension reduction and the method for high-dimensional additive models by assuming a nonparametric additive model for the $\mathbf{V}_k$'s with a vanishing index. Then, after dimension reduction, the additive model can be used to estimate the component functions at both the group level for a nonzero ϕ_k and the within group level for a zero ϕ_k. Thus, our framework can be useful for exploratory data analysis even when some ϕ_k's are zero. Since we focus on dimension reduction for $\boldsymbol{V}_k$'s with a non-vanishing index, we assume without loss of generality that $\phi_k \neq 0$ for all k.

3. Applications

3.1. Partially linear single-index models: A two-stage estimation procedure

Consider the partially linear single-index model

$$Y = \boldsymbol{\alpha}^\top \boldsymbol{X} + g(\boldsymbol{\gamma}^\top \boldsymbol{Z}) + \epsilon, \tag{3.1}$$

where $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$, $\boldsymbol{Z} = (Z_1, \ldots, Z_q)^\top \in \mathbb{R}^q$, $\boldsymbol{\alpha} \in \mathbb{R}^p$ is an unknown linear parameter, $\boldsymbol{\gamma} \in \mathbb{R}^q$ is an unknown single-index parameter, and $g(\cdot)$ is an unknown link function. Then $K = 2$, $\boldsymbol{V}_1 = \boldsymbol{X}$, $\boldsymbol{V}_2 = \boldsymbol{Z}$, and

$$\mathbf{S} = \begin{pmatrix} \boldsymbol{\alpha} & \boldsymbol{0}_{p\times 1} \\ \boldsymbol{0}_{q\times 1} & \boldsymbol{\gamma} \end{pmatrix}.$$

By Proposition 1, $\boldsymbol{\beta}_{LS} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top = \mathbf{S}\boldsymbol{\phi}$ for some vector $\boldsymbol{\phi} = (\phi_1, \phi_2)^\top \in \mathbb{R}^2$. It follows that $\boldsymbol{\beta}_1 = \phi_1\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_2 = \phi_2\boldsymbol{\gamma}$. As a consequence, if $\phi_1 \neq 0$ and $\phi_2 \neq 0$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ can be identified simultaneously by just one vector $\boldsymbol{\beta}_{LS}$. Let $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \mathrm{Cov}(\boldsymbol{Z})$ and $\boldsymbol{\Sigma}_{\boldsymbol{ZX}} = \mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{X})$.

Proposition 2. *Under* (2.2)*, there are constants* $\phi_1 \neq 0$ *and* $\phi_2 \neq 0$ *such that* $\boldsymbol{\beta}_{LS} = \mathbf{S}\boldsymbol{\phi}$*,* $\boldsymbol{\beta}_1 = \phi_1\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}_2 = \phi_2\boldsymbol{\gamma}$*, provided*
(B0) $\boldsymbol{\alpha} \neq \boldsymbol{0}_{p\times 1}, \mathrm{Cov}\{\boldsymbol{Z}, g(\boldsymbol{\gamma}^\top \boldsymbol{Z})\} \neq \boldsymbol{0}_{q\times 1}$*,*
and one of the following holds:
(B1) $\boldsymbol{X}$ *is independent of* $\boldsymbol{Z}$*;*
(B2) $\boldsymbol{V} = (\boldsymbol{X}^\top, \boldsymbol{Z}^\top)^\top$ *has an elliptically symmetric distribution.*

Corollary 1. *Under the conditions of Theorem* 1*, if* $d = d_0 = o(n/\log n)$*, then* $\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_2 = O(\sqrt{d_0/n})$ *for* $k = 1, 2$.

The parameters in (3.1) are often estimated via an algorithm that iteratively updates estimates of the nonparametric component and the parametric component. Wang et al. (2010) introduced a two-step estimation procedure, and Liang et al. (2010) proposed a profile least squares procedure that involves a nonlinear optimization problem which is iterative in nature. Both estimators can be found without an iterative procedure. Feng et al. (2013) proposed using partial dimension reduction techniques to obtain estimators without iteration, but when the dimension of $\boldsymbol{X}$ is large, computation is a challenge, because it involves an integration over the support of $\boldsymbol{X}$. We develop a new estimation procedure for partially linear single-index models.

By Proposition 2, we can re-express (3.1) as

$$Y = \varphi_1 \times \boldsymbol{\beta}_1^\top \boldsymbol{X} + \tilde{g}(\boldsymbol{\beta}_2^\top \boldsymbol{Z}) + \epsilon, \tag{3.2}$$

where $\varphi_1 = 1/\phi_1$ and $\tilde{g}(\cdot) = g(\cdot/\phi_2)$ is an unknown link function. If $\boldsymbol{\beta}_{LS}$ is given, then (3.2) reduces to a partially linear model. Let $K(\cdot)$ be a kernel function and $K_h(\cdot) = h^{-1}K(\cdot/h)$ be a re-scaling of K with bandwidth h. The local linear estimates (Fan and Gijbels (1996)) of $\mu(t;\boldsymbol{\beta}_2) = E(Y|\boldsymbol{\beta}_2^\top \boldsymbol{Z} = t)$ and $\boldsymbol{\mu}_1(t;\boldsymbol{\beta}_2) = E(\boldsymbol{X}|\boldsymbol{\beta}_2^\top \boldsymbol{Z} = t)$ are, respectively,

$$\hat{\mu}(t;\boldsymbol{\beta}_2) = \sum_{i=1}^n W_{ni}(t;\boldsymbol{\beta}_2) y_i,$$

$$\hat{\boldsymbol{\mu}}_1(t;\boldsymbol{\beta}_2) = \sum_{i=1}^n W_{ni}(t;\boldsymbol{\beta}_2) \boldsymbol{x}_i,$$

where

$$W_{ni}(t;\boldsymbol{\beta}_2) = \frac{U_{ni}(t;\boldsymbol{\beta}_2)}{\sum_{j=1}^n U_{nj}(t;\boldsymbol{\beta}_2)},$$

$$U_{ni}(t;\boldsymbol{\beta}_2) = K_h(\boldsymbol{z}_i^\top \boldsymbol{\beta}_2 - t)\{S_{n2}(t;\boldsymbol{\beta}_2) - (\boldsymbol{z}_i^\top \boldsymbol{\beta}_2 - t) S_{n1}(t;\boldsymbol{\beta}_2)\},$$

$$S_{nl} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{z}_i^\top \boldsymbol{\beta}_2 - t)^l K_h(\boldsymbol{z}_i^\top \boldsymbol{\beta}_2 - t), \quad l = 1, 2.$$

The proposed estimator of $\boldsymbol{\alpha}$ is

$$\hat{\boldsymbol{\alpha}} = \hat{\varphi}_1 \times \hat{\boldsymbol{\beta}}_1 = \frac{T_{n1}}{T_{n2}} \times \hat{\boldsymbol{\beta}}_1, \tag{3.3}$$

where

$$T_{n1} = \frac{1}{n}\sum_{i=1}^n \{y_i - \hat{\mu}(\boldsymbol{z}_i^\top \hat{\boldsymbol{\beta}}_2; \hat{\boldsymbol{\beta}}_2)\}\{\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1(\boldsymbol{z}_i^\top \hat{\boldsymbol{\beta}}_2; \hat{\boldsymbol{\beta}}_2)\}^\top \hat{\boldsymbol{\beta}}_1,$$

$$T_{n2} = \frac{1}{n}\sum_{i=1}^n [\{\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1(\boldsymbol{z}_i^\top \hat{\boldsymbol{\beta}}_2; \hat{\boldsymbol{\beta}}_2)\}^\top \hat{\boldsymbol{\beta}}_1]^2.$$

The two-stage estimation procedure works as follows.

S1. Obtain the least squares estimator $\hat{\boldsymbol{\beta}}_{LS} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)^\top$.

S2. Estimate $\boldsymbol{\alpha}$ in (3.1) by (3.3).

The new estimation procedure is iteration-free and easy-to-implement. The bandwidth in the nonparametric smoothing can be selected via cross-validation.

We require some standard technical conditions.

(C1) The density function of $\boldsymbol{b}_2^\top \boldsymbol{Z}$ is positive and satisfies a Lipschitz condition for $\boldsymbol{b}_2$ in a neighborhood of $\boldsymbol{\beta}_2$; $\boldsymbol{\beta}_2^\top \boldsymbol{Z}$ has a density function that is bounded away from 0.

(C2) The functions g and μ_{1j} have two bounded and continuous derivatives, where μ_{1j} is the jth component of $\boldsymbol{\mu}_1, 1 \le j \le p$.

(C3) $E(\epsilon) = 0, E(\epsilon^2) < \infty$, and $\sup_t E(\|\boldsymbol{X}\|_2^2 | \boldsymbol{\beta}_2^\top \boldsymbol{Z} = t) < \infty$.

(C4) The kernel function K is a bounded and symmetric density function with a bounded derivative, and satisfies $0 < \int_{-\infty}^{\infty} t^2 K(t)\,\mathrm{d}x < \infty$.

(C5) The bandwidth h satisfies $\limsup_{n\to\infty} nh^5 < \infty, nh^3 \to \infty$, and $\log^2 n/(nh^2) \to 0$.

We assume now that all the predictors are relevant to the response variable, and that the number of relevant predictors $d = d_0$ is fixed.

Theorem 2. *If the conditions of Theorem* 1 *and the regularity conditions* (C1)–(C5) *hold,* $\hat{\boldsymbol{\alpha}}$ *is a* $\sqrt{n}$*-consistent estimator of* $\boldsymbol{\alpha}$.

Asymptotic expansion of $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}$ can be found in the proof in the supplementary material. Under additional conditions, we can also prove convergence results for the nonparametric link function. In the presence of many irrelevant predictors, we can replace S1 by

S1'. Obtain the penalized least squares estimator via (3.5) in Subsection 3.2.

Because of the oracle property, the resulting two-stage estimator has the same property. We remark that the idea of this two-stage estimation is general, and can be incorporated into other procedures to devise effective algorithms.

3.2. Predictor selection for large-d-small-n problems

When there are a large number of predictors, it is desired to select significant ones to the response variable. Even for the partially linear single-index model (3.1), this is challenging as it includes selection of significant predictors and estimation of the associated coefficients in the parametric component, as well

as identification of significant predictors in the nonparametric component; it is more difficult for model (1.1). How to incorporate the grouping information into the selection process is also an important issue.

We can address these concerns due mainly to the linear least squares structure of $\boldsymbol{\beta}_{LS}$ that enables us to apply the penalization paradigm. Specifically, we consider the penalized least squares function

$$Q_\lambda(\boldsymbol{b}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{v}_i^\top \boldsymbol{b})^2 + \sum_{k=1}^{K}\sum_{j=1}^{p_k} J_\lambda(|b_{kj}|), \tag{3.4}$$

where $\boldsymbol{b} = (\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_K^\top)^\top \in \mathbb{R}^d$ with $\boldsymbol{b}_k = (b_{k1}, \ldots, b_{kp_k})^\top \in \mathbb{R}^{p_k}$, $J_\lambda(\cdot)$ is a penalty function, and λ is a tuning parameter.

There has been much interest in penalized methods in high-dimensional linear or generalized linear models, and Wang, Xu, and Zhu (2012) have dealt with high-dimensional single-index models. We consider variable selection for additive index models by taking advantage of the special model structure.

We study the large sample properties of the penalized least squares estimator with the SCAD penalty (Fan and Li (2001)). The model is assumed to be sparse in the sense that many components of the regression coefficient vector $\boldsymbol{\beta}_{LS} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top$ are exactly zero. Take the nonzero components of $\boldsymbol{\beta}_k$ to be the first p_{k0} coordinates and write $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}^\top, \boldsymbol{0}_{(p_k-p_{k0})\times 1}^\top)^\top$. Accordingly, $\boldsymbol{V}_{k1}$ consists of the first p_{k0} components of $\boldsymbol{V}_k$, and we write $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_K)^\top$ with $\mathbf{V}_k$ the design matrix corresponding to $\boldsymbol{V}_k$, and $\mathbf{V}_{k1}$ the sub-matrix formed by the first p_{k0} columns of $\mathbf{V}_k$.

Let $\hat{\boldsymbol{\beta}}_{k1}^o$ be an ideal vector from the least squares fit of $\boldsymbol{y}$ on $(\mathbf{V}_{11}, \ldots, \mathbf{V}_{K1})$, and take the least squares oracle estimator to be $\hat{\boldsymbol{\beta}}^o = (\hat{\boldsymbol{\beta}}_1^{o\top}, \ldots, \hat{\boldsymbol{\beta}}_K^{o\top})^\top$, where $\hat{\boldsymbol{\beta}}_k^o = (\hat{\boldsymbol{\beta}}_{k1}^{o\top}, \boldsymbol{0}_{(p_k-p_{k0})\times 1}^\top)^\top$.

Theorem 3. *Assume the conditions* (D1)−(D5) *in the supplementary document. If $\mathcal{A}_\lambda$ is the set of local minima of $Q_\lambda(\boldsymbol{b})$, then* $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}^o \in \mathcal{A}_\lambda) = 1$.

To select predictors in the partially linear single-index model, we consider the penalized least squares function

$$Q_\lambda(\boldsymbol{b}_1, \boldsymbol{b}_2) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{b}_1 - \boldsymbol{z}_i^\top \boldsymbol{b}_2)^2 + \sum_{j=1}^{p} J_\lambda(|b_{1j}|) + \sum_{j=1}^{q} J_\lambda(|b_{2j}|), \tag{3.5}$$

where $\boldsymbol{b}_1 = (b_{11}, \ldots, b_{1p})^\top \in \mathbb{R}^p, \boldsymbol{b}_2 = (b_{21}, \ldots, b_{2q})^\top \in \mathbb{R}^q$, and $J_\lambda(\cdot)$ is the SCAD penalty.

Take the nonzero components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ to be, respectively, the first p_0 and q_0 coordinates, and write $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{11}^\top, \boldsymbol{0}_{(p-p_0)\times 1}^\top)^\top$, $\boldsymbol{\beta}_2 = (\boldsymbol{\beta}_{21}^\top, \boldsymbol{0}_{(q-q_0)\times 1}^\top)^\top$.

Accordingly, take $\boldsymbol{X}_1$ and $\boldsymbol{Z}_1$ to consist of the first p_0 and q_0 components of $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively. Let $\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top$ and $\mathbf{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)^\top$ be the design matrices corresponding to $\boldsymbol{X}$ and $\boldsymbol{Z}$, and take $\mathbf{X}_1$ and $\mathbf{Z}_1$ as the sub-matrices formed by the first p_0 and q_0 columns of $\mathbf{X}$ and $\mathbf{Z}$, respectively.

Let $\hat{\boldsymbol{\beta}}^o = (\hat{\boldsymbol{\beta}}_1^{o\top}, \hat{\boldsymbol{\beta}}_2^{o\top})^\top$ be the least squares oracle estimator, where $\hat{\boldsymbol{\beta}}_1^o = (\hat{\boldsymbol{\beta}}_{11}^{o\top}, \boldsymbol{0}_{(p-p_0)\times 1}^\top)^\top$ and $\hat{\boldsymbol{\beta}}_2^o = (\hat{\boldsymbol{\beta}}_{21}^{o\top}, \boldsymbol{0}_{(q-q_0)\times 1}^\top)^\top$ with $(\hat{\boldsymbol{\beta}}_{11}^{o\top}, \hat{\boldsymbol{\beta}}_{21}^{o\top})^\top$ an ideal vector from the least squares fit of $\boldsymbol{y}$ on $(\mathbf{X}_1, \mathbf{Z}_1)$.

Corollary 2. *Assume the conditions of Theorem* 3. *If $\mathcal{A}_\lambda$ is the set of local minima of $Q_\lambda(\boldsymbol{b}_1, \boldsymbol{b}_2)$, then* $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}^o \in \mathcal{A}_\lambda) = 1$.

4. Simulation Study

We examined the finite-sample performance of the proposed estimation and selection methods. We focused on the partially linear single-index model and considered the models

$$Y = \boldsymbol{\beta}_1^\top \boldsymbol{V}_1 + 2 \times \boldsymbol{\beta}_2^\top \boldsymbol{V}_2 \times I(\boldsymbol{\beta}_2^\top \boldsymbol{V}_2 < 0) + \epsilon, \tag{4.1}$$

$$Y = \boldsymbol{\beta}_1^\top \boldsymbol{V}_1 + \exp\Big(\frac{\boldsymbol{\beta}_2^\top \boldsymbol{V}_2}{2}\Big) + \epsilon, \tag{4.2}$$

$$Y = \boldsymbol{\beta}_1^\top \boldsymbol{V}_1 + 2 \times \sin\Big(\frac{\boldsymbol{\beta}_2^\top \boldsymbol{V}_2}{2}\Big) + \epsilon, \tag{4.3}$$

where $I(\cdot)$ is the indicator function. We covered four cases.

Case 1. $\epsilon \sim N(0,1)$, $\boldsymbol{V} \sim N(\boldsymbol{0}_{d\times 1}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, d$, and ϵ and $\boldsymbol{V}$ are independent. Let $p_1 = 400, p_2 = 200, p_{10} = 3$ and $p_{20} = 2$. The linear parameter and the single-index parameter are $\boldsymbol{\beta}_1 = (1.5, 1, 1, 0, \ldots, 0)^\top$ and $\boldsymbol{\beta}_2 = (1, 1, 0, \ldots, 0)^\top$, respectively. The sample size is $n = 200$.

Case 2. The same as Case 1, except that the error ϵ has a t-distribution with 4 degrees of freedom.

Case 3. The same as Case 1, except that $\Sigma_{ij} = 0.5$ for all $i \neq j$.

Case 4. The same as Case 2, except that $\Sigma_{ij} = 0.5$ for all $i \neq j$.

In each example, we applied the penalized least squares estimation of Subsection 3.2 with the SCAD penalty, and then invoked the two-stage estimation in Subsection 3.1. We adopted the Gaussian kernel in local linear smoothing and used the least squares cross-validation (Li and Racine (2004)) to select the smoothing parameter. For comparison, we also evaluated the performance of the penalized least squares estimation with the LASSO penalty, as well as the

oracle least squares estimation assuming the irrelevant predictors known beforehand. The resulting estimators are denoted by SCAD, LASSO, and Oracle, respectively. For each penalized competitor, we implemented the fast and efficient coordinate descent algorithm (see, e.g., Friedman, Hastie, and Tibshirani (2010)) and selected its tuning parameter by ten-fold cross-validation.

For any vector $\boldsymbol{\theta}$, $\boldsymbol{\theta}^s$ is the orthonormalized version of $\boldsymbol{\theta}$. To evaluate estimation accuracy, we computed the absolute correlation coefficient, $\mathrm{ACC}_{\boldsymbol{\theta}_k}$, between the estimated predictor and the true one: $\mathrm{ACC}_{\boldsymbol{\theta}_k} = |\mathrm{corr}(\hat{\boldsymbol{\theta}}_k^\top \boldsymbol{V}_k, \boldsymbol{\theta}_k^\top \boldsymbol{V}_k)|$ for SCAD and LASSO, and $\mathrm{ACC}_{\boldsymbol{\theta}_k} = |\mathrm{corr}(\hat{\boldsymbol{\theta}}_k^\top \boldsymbol{V}_{k1}, \boldsymbol{\theta}_k^\top \boldsymbol{V}_{k1})|$ for Oracle, with the vector correlation coefficient, $\mathrm{VCC}_{\boldsymbol{\theta}_k} = \hat{\boldsymbol{\theta}}_k^{s\top} \boldsymbol{\theta}_k^s$. Here, for both SCAD and LASSO $\boldsymbol{\theta}_k = \boldsymbol{\beta}_k$, and for Oracle $\boldsymbol{\theta}_k = \boldsymbol{\beta}_{k1}, k = 1, \ldots, K$. For partially linear single-index models, we used $\mathrm{EST}_{\boldsymbol{\theta}} = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2$ to measure the performance of the two-stage estimation procedure; for SCAD and LASSO $\boldsymbol{\theta} = \boldsymbol{\beta}_1$, and for Oracle $\boldsymbol{\theta} = \boldsymbol{\beta}_{11}$.

To assess how well SCAD and LASSO selected predictors, we employed the number of nonzero components ($\mathrm{MS}_{\boldsymbol{\theta}_k}$); the true positive rate ($\mathrm{TPR}_{\boldsymbol{\theta}_k}$), the ratio of the number of correctly identified predictors to the number of truly important predictors; and the false positive rate ($\mathrm{FPR}_{\boldsymbol{\theta}_k}$), the ratio of the number of falsely identified predictors to the total number of irrelevant predictors. Ideally, one wishes $\mathrm{TPR}_{\boldsymbol{\theta}_k}$ close to 1 and $\mathrm{FPR}_{\boldsymbol{\theta}_k}$ close to 0 at the same time.

The simulation results based on 200 data replications from these four cases are summarized in Tables 1−3, respectively. In Cases 1 and 2, the predictors are serially correlated, and several conclusions can be drawn. First, SCAD and LASSO have comparable performance in terms of estimation and selection. The average absolute correlation coefficient and the average vector correlation coefficient of SCAD and LASSO are close to one, and slightly lower than those of Oracle. SCAD and LASSO successfully identified the relevant predictors in the model: the lowest true positive rates are 99.50% and 99.00% for the linear component and the single-index component, respectively. The estimation accuracy of all the methods considered deteriorated when we replaced the normal distribution of the error with the t-distribution. In Cases 3 and 4, there were constant positive correlations among the predictors, and SCAD was generally superior to LASSO. The estimation error of the linear component ($\mathrm{EST}_{\boldsymbol{\theta}_k}$) of SCAD, close to that of Oracle, was significantly lower than that of LASSO, and LASSO tended to select a model with many spurious predictors. That SCAD is more robust to the correlation structure among the predictors than LASSO is well in accordance with results in the literature (Fan and Lv (2010)). Unreported results also show that increasing sample sizes improves the performance.

We also considered a more complex model

$$Y = \left(1 + \frac{\boldsymbol{\beta}_1^\top \boldsymbol{V}_1}{2}\right)^2 + 2 \times \sin\left(\frac{\boldsymbol{\beta}_2^\top \boldsymbol{V}_2}{2}\right) + 1.5 \times \exp\left(\frac{\boldsymbol{\beta}_3^\top \boldsymbol{V}_3}{2}\right) + \epsilon, \qquad (4.4)$$

Table 1. Summary of model (4.1). The average absolute correlation coefficient ($\mathrm{ACC}_{\boldsymbol{\theta}_k}$), the average vector correlation coefficient ($\mathrm{VCC}_{\boldsymbol{\theta}_k}$), the average estimation error of linear component ($\mathrm{EST}_{\boldsymbol{\theta}_k}$), the average number of nonzero components ($\mathrm{MS}_{\boldsymbol{\theta}_k}$), the true positive rate ($\mathrm{TPR}_{\boldsymbol{\theta}_k}$) and the false positive rate ($\mathrm{FPR}_{\boldsymbol{\theta}_k}$), based on 200 data replications, are reported. $\boldsymbol{\theta}_k = \boldsymbol{\beta}_k$ for SCAD and LASSO, and $\boldsymbol{\theta}_k = \boldsymbol{\beta}_{k1}$ for Oracle

	Case 1			Case 2			Case 3			Case 4		
	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle
$\mathrm{ACC}_{\boldsymbol{\theta}_1}$	0.995	0.993	0.998	0.990	0.990	0.997	0.997	0.987	0.998	0.994	0.982	0.997
	(0.004)	(0.006)	(0.001)	(0.008)	(0.009)	(0.002)	(0.004)	(0.006)	(0.001)	(0.007)	(0.009)	(0.002)
$\mathrm{VCC}_{\boldsymbol{\theta}_1}$	0.989	0.985	0.994	0.976	0.979	0.992	0.991	0.962	0.994	0.980	0.941	0.992
	(0.014)	(0.012)	(0.006)	(0.027)	(0.017)	(0.008)	(0.014)	(0.023)	(0.005)	(0.026)	(0.041)	(0.008)
$\mathrm{EST}_{\boldsymbol{\theta}_1}$	0.267	0.323	0.177	0.425	0.397	0.229	0.251	0.613	0.182	0.375	0.735	0.244
	(0.151)	(0.138)	(0.084)	(0.226)	(0.161)	(0.119)	(0.137)	(0.155)	(0.087)	(0.214)	(0.204)	(0.120)
$\mathrm{MS}_{\boldsymbol{\theta}_1}$	11.480	15.330		16.275	15.360		5.580	21.550		8.055	22.340	
$\mathrm{TPR}_{\boldsymbol{\theta}_1}$	1.000	1.000		0.995	1.000		1.000	1.000		1.000	1.000	
$\mathrm{FPR}_{\boldsymbol{\theta}_1}$	0.021	0.031		0.033	0.031		0.006	0.046		0.012	0.048	
$\mathrm{ACC}_{\boldsymbol{\theta}_2}$	0.995	0.991	0.998	0.986	0.987	0.997	0.996	0.982	0.998	0.987	0.971	0.997
	(0.006)	(0.008)	(0.002)	(0.017)	(0.013)	(0.003)	(0.009)	(0.012)	(0.002)	(0.019)	(0.020)	(0.003)
$\mathrm{VCC}_{\boldsymbol{\theta}_2}$	0.988	0.982	0.994	0.970	0.974	0.992	0.990	0.958	0.994	0.967	0.926	0.992
	(0.014)	(0.016)	(0.007)	(0.044)	(0.025)	(0.009)	(0.025)	(0.035)	(0.007)	(0.053)	(0.058)	(0.009)
$\mathrm{MS}_{\boldsymbol{\theta}_2}$	5.935	8.035		8.395	7.950		3.190	11.135		4.675	11.570	
$\mathrm{TPR}_{\boldsymbol{\theta}_2}$	1.000	1.000		1.000	1.000		1.000	1.000		1.000	1.000	
$\mathrm{FPR}_{\boldsymbol{\theta}_2}$	0.019	0.030		0.032	0.030		0.006	0.046		0.013	0.048	

again with four cases.

Case 5. $\epsilon \sim N(0,1)$, $\boldsymbol{V} \sim N(\boldsymbol{0}_{d\times 1}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, d$, and ϵ and $\boldsymbol{V}$ are independent. Let $p_1 = p_2 = p_3 = 200$ and $p_{10} = p_{20} = p_{30} = 2$. The single-index parameters are $\boldsymbol{\beta}_1 = (1, -1, 0, \ldots, 0)^\top, \boldsymbol{\beta}_2 = (1, 1, 0, \ldots, 0)^\top$ and $\boldsymbol{\beta}_3 = (-1, 1, 0, \ldots, 0)^\top$, respectively. The sample size is $n = 200$.

Case 6. The same as Case 5, except that the error ϵ has a t-distribution with 4 degrees of freedom.

Case 7. The same as Case 5, except that $\Sigma_{ij} = 0.5$ for all $i \neq j$.

Case 8. The same as Case 6, except that $\Sigma_{ij} = 0.5$ for all $i \neq j$.

The empirical results based on 200 data replications from these four cases are reported in Table 4. The results are qualitatively similar to those of the partially linear single-index model.

5. Data Analysis

We applied the proposed method to a dataset of possible advertisements on Internet pages that is available at the University of California-Irvine machine

Table 2. Summary of model (4.2). The average absolute correlation coefficient ($\mathrm{ACC}_{\boldsymbol{\theta}_k}$), the average vector correlation coefficient ($\mathrm{VCC}_{\boldsymbol{\theta}_k}$), the average estimation error of linear component ($\mathrm{EST}_{\boldsymbol{\theta}_k}$), the average number of nonzero components ($\mathrm{MS}_{\boldsymbol{\theta}_k}$), the true positive rate ($\mathrm{TPR}_{\boldsymbol{\theta}_k}$) and the false positive rate ($\mathrm{FPR}_{\boldsymbol{\theta}_k}$), based on 200 data replications, are reported. $\boldsymbol{\theta}_k = \boldsymbol{\beta}_k$ for SCAD and LASSO, and $\boldsymbol{\theta}_k = \boldsymbol{\beta}_{k1}$ for Oracle.

	Case 1			Case 2			Case 3			Case 4		
	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle
$\mathrm{ACC}_{\boldsymbol{\theta}_1}$	0.996	0.994	0.998	0.989	0.991	0.998	0.997	0.989	0.998	0.995	0.983	0.998
	(0.003)	(0.005)	(0.001)	(0.013)	(0.009)	(0.002)	(0.004)	(0.006)	(0.001)	(0.007)	(0.010)	(0.001)
$\mathrm{VCC}_{\boldsymbol{\theta}_1}$	0.991	0.988	0.996	0.977	0.982	0.994	0.992	0.968	0.995	0.982	0.948	0.993
	(0.007)	(0.009)	(0.003)	(0.033)	(0.017)	(0.007)	(0.015)	(0.022)	(0.004)	(0.032)	(0.039)	(0.006)
$\mathrm{EST}_{\boldsymbol{\theta}_1}$	0.250	0.296	0.167	0.402	0.368	0.211	0.238	0.560	0.182	0.347	0.697	0.222
	(0.105)	(0.118)	(0.078)	(0.250)	(0.158)	(0.098)	(0.130)	(0.157)	(0.083)	(0.203)	(0.210)	(0.099)
$\mathrm{MS}_{\boldsymbol{\theta}_1}$	14.155	16.760		17.250	15.975		6.825	22.815		8.805	21.955	
$\mathrm{TPR}_{\boldsymbol{\theta}_1}$	1.000	1.000		0.995	1.000		1.000	1.000		0.998	1.000	
$\mathrm{FPR}_{\boldsymbol{\theta}_1}$	0.028	0.034		0.035	0.032		0.009	0.049		0.014	0.047	
$\mathrm{ACC}_{\boldsymbol{\theta}_2}$	0.984	0.985	0.997	0.958	0.974	0.996	0.985	0.970	0.997	0.950	0.945	0.995
	(0.017)	(0.017)	(0.002)	(0.035)	(0.028)	(0.005)	(0.025)	(0.019)	(0.003)	(0.068)	(0.051)	(0.007)
$\mathrm{VCC}_{\boldsymbol{\theta}_2}$	0.963	0.973	0.992	0.905	0.952	0.988	0.962	0.919	0.992	0.872	0.835	0.987
	(0.047)	(0.027)	(0.009)	(0.085)	(0.047)	(0.015)	(0.060)	(0.061)	(0.011)	(0.133)	(0.142)	(0.020)
$\mathrm{MS}_{\boldsymbol{\theta}_2}$	7.675	8.740		8.970	8.230		4.020	11.805		4.940	11.650	
$\mathrm{TPR}_{\boldsymbol{\theta}_2}$	1.000	1.000		0.990	1.000		0.997	1.000		0.960	0.987	
$\mathrm{FPR}_{\boldsymbol{\theta}_2}$	0.028	0.034		0.035	0.031		0.010	0.049		0.015	0.048	

learning repository. The features or predictors encode the geometry of the image as well as phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text. The task is to predict whether an image is an advertisement or not.

After preprocessing the dataset contains $n = 2,358$ observations and $p = 1,430$ predictors. Among the predictors, the first three are related to the geometry of the image and hence are continuous, the rest are binary. We divided the predictors into groups of the three continuous predictors and of the others. We restricted the effects of categorical predictors to be linear, granting that the partially linear single-index model (3.1), or its generalized version with logit link (Carroll et al. (1997)), works in this setting. The conditional independence implied by model (3.1), and hence Proposition 2, still holds when we consider the generalized partially linear single-index model.

We carried out the study by repeated random splitting of the full dataset, one-half of the observations from the advertisement class and one-half of the observations from the non-advertisement class as training samples, and the rest as test samples. For each split, SCAD and LASSO were applied to the training data. Since there were only three predictors in the semi-parametric component, we included them in the model without shrinkage of their coefficients, with predictors

Table 3. Summary of model (4.3). The average absolute correlation coefficient ($\mathrm{ACC}_{\boldsymbol{\theta}_k}$), the average vector correlation coefficient ($\mathrm{VCC}_{\boldsymbol{\theta}_k}$), the average estimation error of linear component ($\mathrm{EST}_{\boldsymbol{\theta}_k}$), the average number of nonzero components ($\mathrm{MS}_{\boldsymbol{\theta}_k}$), the true positive rate ($\mathrm{TPR}_{\boldsymbol{\theta}_k}$) and the false positive rate ($\mathrm{FPR}_{\boldsymbol{\theta}_k}$), based on 200 data replications, are reported. $\boldsymbol{\theta}_k = \boldsymbol{\beta}_k$ for SCAD and LASSO, and $\boldsymbol{\theta}_k = \boldsymbol{\beta}_{k1}$ for Oracle

	Case 1			Case 2			Case 3			Case 4		
	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle
$\mathrm{ACC}_{\boldsymbol{\theta}_1}$	0.998	0.996	0.999	0.994	0.993	0.998	0.999	0.992	0.999	0.996	0.987	0.998
	(0.001)	(0.003)	(0.000)	(0.006)	(0.007)	(0.001)	(0.000)	(0.003)	(0.000)	(0.005)	(0.007)	(0.001)
$\mathrm{VCC}_{\boldsymbol{\theta}_1}$	0.995	0.992	0.997	0.987	0.986	0.995	0.996	0.980	0.997	0.988	0.960	0.995
	(0.003)	(0.007)	(0.003)	(0.019)	(0.012)	(0.005)	(0.002)	(0.013)	(0.002)	(0.020)	(0.030)	(0.005)
$\mathrm{EST}_{\boldsymbol{\theta}_1}$	0.188	0.233	0.139	0.310	0.320	0.195	0.171	0.462	0.150	0.298	0.610	0.207
	(0.076)	(0.102)	(0.062)	(0.181)	(0.136)	(0.089)	(0.082)	(0.116)	(0.062)	(0.171)	(0.184)	(0.090)
$\mathrm{MS}_{\boldsymbol{\theta}_1}$	10.815	15.170		15.095	15.215		4.515	24.345		8.000	22.100	
$\mathrm{TPR}_{\boldsymbol{\theta}_1}$	1.000	1.000		0.998	1.000		1.000	1.000		0.998	1.000	
$\mathrm{FPR}_{\boldsymbol{\theta}_1}$	0.019	0.030		0.030	0.030		0.003	0.053		0.012	0.048	
$\mathrm{ACC}_{\boldsymbol{\theta}_2}$	0.992	0.990	0.998	0.966	0.979	0.996	0.992	0.977	0.998	0.968	0.959	0.996
	(0.010)	(0.012)	(0.002)	(0.031)	(0.022)	(0.004)	(0.016)	(0.017)	(0.002)	(0.038)	(0.031)	(0.004)
$\mathrm{VCC}_{\boldsymbol{\theta}_2}$	0.980	0.980	0.993	0.919	0.961	0.989	0.977	0.942	0.993	0.914	0.886	0.990
	(0.029)	(0.021)	(0.007)	(0.076)	(0.040)	(0.012)	(0.045)	(0.043)	(0.008)	(0.114)	(0.100)	(0.012)
$\mathrm{MS}_{\boldsymbol{\theta}_2}$	5.775	7.785		7.960	8.165		2.645	12.950		4.315	11.325	
$\mathrm{TPR}_{\boldsymbol{\theta}_2}$	1.000	1.000		0.995	1.000		1.000	1.000		0.990	0.997	
$\mathrm{FPR}_{\boldsymbol{\theta}_2}$	0.019	0.029		0.030	0.031		0.003	0.055		0.011	0.047	

given by $\mathbf{X}\hat{\boldsymbol{\beta}}_1$ and $\mathbf{Z}\hat{\boldsymbol{\beta}}_2$. The partially linear model and the generalized partially linear model with logit link were fitted to the same data. The performance of the fitted models was evaluated by the test samples. To reduce variability, the split into training and test sets was repeated 200 times, with the results summarized in Table 5. The methods considered here were comparable, with LASSO having slightly lower classification errors but apparently larger in model size. While LASSO may not be selection consistent, it is often persistent (Greenshtein and Ritov (2004)); the concept of persistency focuses on expected prediction losses, not the accuracy of estimated parameters. Unreported results also show that using the set of selected predictors $\{X_j : j \in \hat{\mathcal{M}}\}$ with $\hat{\mathcal{M}} = \{j : \hat{\beta}_{1j} \neq 0\}$, instead of $\mathbf{X}\hat{\boldsymbol{\beta}}_1$, leads to inferior performance. More seriously, the resulting algorithms often fail to converge due to the large size of $\hat{\mathcal{M}}$.

Supplementary Material. The supplementary file covers the regularity conditions and proofs.

Acknowledgement

Zhang's research was supported by the National Science Foundation (NSF) of Shenzhen University (801, 00036112), the NSF of China (Tianyuan fund for

Table 4. Summary of model (4.4). The average absolute correlation coefficient ($\mathrm{ACC}_{\boldsymbol{\theta}_k}$), the average vector correlation coefficient ($\mathrm{VCC}_{\boldsymbol{\theta}_k}$), the average estimation error of linear component ($\mathrm{EST}_{\boldsymbol{\theta}_k}$), the average number of nonzero components ($\mathrm{MS}_{\boldsymbol{\theta}_k}$), the true positive rate ($\mathrm{TPR}_{\boldsymbol{\theta}_k}$) and the false positive rate ($\mathrm{FPR}_{\boldsymbol{\theta}_k}$), based on 200 data replications, are reported. $\boldsymbol{\theta}_k = \boldsymbol{\beta}_k$ for SCAD and LASSO, and $\boldsymbol{\theta}_k = \boldsymbol{\beta}_{k1}$ for Oracle

	Case 5			Case 6			Case 7			Case 8		
	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle	SCAD	LASSO	Oracle
$\mathrm{ACC}_{\boldsymbol{\theta}_1}$	0.988	0.943	0.995	0.974	0.895	0.993	0.983	0.933	0.991	0.960	0.887	0.986
	(0.013)	(0.037)	(0.006)	(0.050)	(0.092)	(0.009)	(0.027)	(0.061)	(0.013)	(0.074)	(0.095)	(0.019)
$\mathrm{VCC}_{\boldsymbol{\theta}_1}$	0.994	0.970	0.998	0.985	0.937	0.997	0.992	0.967	0.997	0.979	0.937	0.995
	(0.005)	(0.020)	(0.002)	(0.064)	(0.082)	(0.003)	(0.014)	(0.025)	(0.004)	(0.044)	(0.052)	(0.006)
$\mathrm{MS}_{\boldsymbol{\theta}_1}$	7.380	16.185		10.045	15.595		4.215	13.450		5.250	12.875	
$\mathrm{TPR}_{\boldsymbol{\theta}_1}$	1.000	1.000		0.997	0.995		1.000	1.000		1.000	1.000	
$\mathrm{FPR}_{\boldsymbol{\theta}_1}$	0.027	0.071		0.040	0.068		0.011	0.057		0.016	0.054	
$\mathrm{ACC}_{\boldsymbol{\theta}_2}$	0.990	0.975	0.997	0.958	0.952	0.995	0.989	0.969	0.997	0.951	0.929	0.995
	(0.013)	(0.018)	(0.002)	(0.038)	(0.032)	(0.007)	(0.021)	(0.033)	(0.002)	(0.058)	(0.080)	(0.006)
$\mathrm{VCC}_{\boldsymbol{\theta}_2}$	0.977	0.960	0.993	0.906	0.921	0.985	0.971	0.924	0.993	0.878	0.830	0.986
	(0.035)	(0.026)	(0.009)	(0.090)	(0.055)	(0.019)	(0.059)	(0.057)	(0.008)	(0.138)	(0.133)	(0.017)
$\mathrm{MS}_{\boldsymbol{\theta}_2}$	7.355	15.870		9.870	15.725		3.980	13.310		4.990	13.030	
$\mathrm{TPR}_{\boldsymbol{\theta}_2}$	1.000	1.000		0.985	1.000		1.000	1.000		0.955	0.977	
$\mathrm{FPR}_{\boldsymbol{\theta}_2}$	0.027	0.070		0.039	0.069		0.010	0.057		0.015	0.055	
$\mathrm{ACC}_{\boldsymbol{\theta}_3}$	0.986	0.919	0.994	0.939	0.814	0.991	0.979	0.907	0.989	0.922	0.828	0.980
	(0.020)	(0.053)	(0.008)	(0.141)	(0.155)	(0.013)	(0.048)	(0.078)	(0.015)	(0.143)	(0.142)	(0.027)
$\mathrm{VCC}_{\boldsymbol{\theta}_3}$	0.993	0.955	0.998	0.957	0.871	0.997	0.990	0.953	0.996	0.955	0.897	0.993
	(0.015)	(0.035)	(0.002)	(0.143)	(0.156)	(0.004)	(0.025)	(0.036)	(0.005)	(0.108)	(0.105)	(0.010)
$\mathrm{MS}_{\boldsymbol{\theta}_3}$	7.055	16.005		10.120	15.855		3.870	13.210		5.140	13.255	
$\mathrm{TPR}_{\boldsymbol{\theta}_3}$	1.000	1.000		0.982	0.967		1.000	1.000		0.982	0.985	
$\mathrm{FPR}_{\boldsymbol{\theta}_3}$	0.025	0.070		0.041	0.070		0.009	0.056		0.016	0.056	

Table 5. Advertising data. Classification errors made and the number of predictors chosen over 200 random splitting of all samples into training and test sets

Method	Training error (%) (mean±sd)	Test error (%) (mean±sd)	Number of selected predictors (mean±sd)
LASSO + PLM	1.753±0.377	3.681±0.489	206.380±55.727
LASSO + GPLM	1.423±0.317	3.353±0.495	206.380±55.727
SCAD + PLM	2.512±0.587	4.439±0.594	61.525±15.083
SCAD + GPLM	2.198±0.551	4.047±0.549	61.525±15.083

Mathematics, No. 11326179), and the NSF of China (11101157). Liang's research was partially supported by NSF grants DMS-1007167 and DMS-1207444. Zhu's research was supported by a grant from the Research Council of Hong Kong, and a grant from Hong Kong Baptist University, Hong Kong.

References

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

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.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics.* Wiley, New York.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100**, 410-428.

Dong, Y. X. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika* **97**, 279–294.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall, New York.

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. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians* (M. Sanz-Sole, J. Soria, J.L. Varona, and J. Verdera, eds.), Vol. III, 595-622.

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

Feng, Z., Wang, T., and Zhu, L. X. (2012). Transformation-based estimation. Manuscript.

Feng, Z., Wen, X., Yu, Z. and Zhu, L. X. (2013). On partial sufficient dimension reduction with applications to partially linear multi-index models. *J. Amer. Statist. Assoc.* **108**, 237-246.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1-22.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971-988.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* **21**, 867-889.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. Ser. B* **48**, 244-248.

Li, B. and Dong, Y. X. (2009). Dimension reduction for non-elliptically distributed predictors. *Ann. Statist.* **37**, 1272-1298.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, L., Li, B. and Zhu, L. X. (2010). Groupwise dimension reduction. *J. Amer. Statist. Assoc.* **105**, 1118-1201.

Li, Q. and Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statist. Sinica* **14**, 485-512.

Liang, H., Liu, X., Li, R. and Tsai, C. (2010). Estimation and testing for partially linear single-index models. *Ann. Statist.* **38**, 3811-3836.

Lin, W. and Kulasekera, K. B. (2007). Identifiability of single-index models and additive-index models. *Biometrika* **94**, 496-501.

Naik, P. and Tsai, C. L. (2001). Single-index model selections. *Biometrika* **88**, 821-832.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403-1430.

Ruan, L. and Yuan, M. (2010). Dimension reduction and parameter estimation for additive index models. *Stat. Interface* **3**, 493-499.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, New York.

Wang, J. L., Xue, L. G., Zhu, L. X. and Chong, Y. S. (2010). Estimation for a partial-linear single-index model. *Ann. Statist.* **38**, 246-274.

Wang, T., Xu, P. R. and Zhu, L. X. (2012). Non-convex penalized estimation in high-dimensional models with single-index structure. *J. Multivariate Anal.* **109**, 221–235.

Wei, F., Huang, J. and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica* **21**, 1515-1540.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *J. Mach. Learn. Res.* **13**, 1973-1998.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97**, 1042-1054.

Department of Biostatistics, Yale University, New Haven, CT 06520, U.S.A.

E-mail: tao.wang.tw376@yale.edu

Institute of Statistical Sciences at Shenzhen University.

Shenzhen-Hong Kong Joint Research Center for Applied Statistical Sciences, College of Mathematics and Computational Science, Shenzhen University, Shenzhen 518060, China.

E-mail: zhangjunstat@gmail.com

Department of Statistics, George Washington University, Washington, D.C. 20052, U.S.A.

E-mail: hliang@gwu.edu

Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.

E-mail: lzhu@hkbu.edu.hk

(Received June 2013; accepted March 2014)