<table>
<thead>
<tr>
<th><strong>Statistica Sinica Preprint No:</strong> SS-13-175R2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong> Estimation of a groupwise additive multiple-index model and its applications</td>
</tr>
<tr>
<td><strong>Manuscript ID</strong> SS-13-175R2</td>
</tr>
<tr>
<td><strong>URL</strong> <a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a></td>
</tr>
<tr>
<td><strong>DOI</strong> 10.5705/ss.2013.175</td>
</tr>
<tr>
<td><strong>Complete List of Authors</strong> Tao Wang, Jun Zhang, Hua Liang and Lixing Zhu</td>
</tr>
<tr>
<td><strong>Corresponding Author</strong> Lixing Zhu</td>
</tr>
<tr>
<td><strong>E-mail</strong> <a href="mailto:lzhu@hkbu.edu.hk">lzhu@hkbu.edu.hk</a></td>
</tr>
</tbody>
</table>

Notice: Accepted version subject to English editing.
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 real world 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 an important research topic in statistics. In many practical situations, parametric models, such as the linear model and the generalized linear model, are among the most convenient and frequently used models. However, parametric models are not flexible enough to capture the underlying relationship between the response variable and its associated predictors. Further, it remains theoretically difficult, if not impossible, to check the fit of a parametric model with a large number of predictors.
To achieve a greater realism, semi-parametric models (Ruppert, Wand and Carroll 2003) are being increasingly used to balance between the modeling bias and the “curse of dimensionality”. As their names suggest, a semi-parametric model has the flexibility and good fit of a nonparametric model and retains the parsimony and ease of interpretation of a parametric model. In this category, however, only a limited amount of work has been done on estimation with high-dimensional data. For example, 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; however, they commented that the approach developed in their paper does not offer any guarantee on the viewpoint of variable selection. Wang, Xu and Zhu (2012) studied the theoretical properties of a regularized linear least squares method for general single-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 of partially linear single-index models is that the predictors under investigation fall into two groups to affect 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 the problem of estimation in partially linear single-index models when the number of predictors can be larger than the sample size. As an application, the estimation procedures developed in this paper are applicable to partially linear single-index models and their extensions as follows.

Consider the regression of a response variable $Y \in \mathbb{R}$ on a random vector of predictors $V \in \mathbb{R}^d$. Suppose that $V = (V_1^\top, V_2^\top, \ldots, V_K^\top)^\top$ can be naturally divided into $K$ non-overlapping groups $V_k \in \mathbb{R}^{p_k}, k = 1, \ldots, K$. We consider the following groupwise additive multiple-index model:

$$Y = \sum_{k=1}^{K} g_k(\beta_k^\top V_k) + \epsilon,$$  \hspace{1cm} (1.1)

where $g_k(\cdot)$ is an unknown component function and $\beta_k \in \mathbb{R}^{p_k}$ is a single-index
vector of interest corresponding to $V_k$, and the random error $\epsilon$ is independent of $V$. Model (1.1) includes many existing models as special cases. If $K = 1$, then it is simply the well-known single-index model (Powell, Stock and Stoker 1989). If we let $K = 2$ and take $g_1(t) = t$, then model (1.1) becomes the partially linear single-index model. If we further take $p_2 = 1$, then it reduces to the partially linear model (Heckman 1986). Therefore, model (1.1) is a very flexible semiparametric model. For further discussions, see Naik and Tsai (2001) and Lin and Kulasekera (2007).

Parameter estimation for model (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; that is, iteratively estimating the nonparametric component and the parametric component.

When the total number of the predictors, $d = p_1 + \cdots + p_K$, is large, it is more realistic to regard it growing with the sample size, $n$. Yet, few results are available for estimation in the context of the groupwise additive multiple-index model when $d$ diverges with $n$. As such, we propose a simple linear least squares framework to discuss estimation. From this, we can very easily 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 then derived. As a direct application, in Subsection 3.1 we establish the theoretical properties of the least squares method for partially linear single-index models. In addition, we develop a new two-stage estimation procedure. Another application, in Subsection 3.2, concerns the variable selection problem with high-dimensional predictors. By invoking the linear least squares formulation, we propose a penalized least squares method for selecting predictors in each component function. We then study the asymptotic behavior of the penalized estimator in sparse and high-dimensional settings.
numerical studies are presented in Section 4 and Section 5. All detailed proofs are provided in the supplementary material.

2. Identifiability and estimation

We first discuss the identifiability of $\beta_k$'s in model (1.1) as it serves as a basis for our new method and theory. Denote by $\theta_{m \times 1}$ an $m \times 1$ vector of 0's. Let

$$ S = \begin{pmatrix} \beta_1 & 0_{p_1 \times 1} & \cdots & 0_{p_1 \times 1} \\ 0_{p_2 \times 1} & \beta_2 & \cdots & 0_{p_2 \times 1} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{p_K \times 1} & 0_{p_K \times 1} & \cdots & \beta_K \end{pmatrix}. $$

Then, from (1.1), $Y$ and $V$ are independent conditioned on $S^T V$. In the sufficient dimension reduction community, the column space of $S$ is called the central dimension reduction subspace (Li 1991; Cook 1998) and is a well-defined population parameter.

Assume that $\Sigma_V = \text{Cov}(V)$ is positive definite. We define the least squares direction as

$$ \beta_{LS} = \Sigma_V^{-1} \text{Cov}(V, Y). \quad (2.1) $$

We shall see that $\beta_{LS}$ lies on the column space of $S$, with the following linearity condition on the design distribution:

$$ E(V|S^T V) \text{ is a linear function of } S^T V. \quad (2.2) $$

This condition is satisfied, for example, when the distribution of $V$ is elliptically symmetric. It is also worth pointing out that when the dimension of $V$ is large, this condition is not restrictive; for a discussion, see Hall and Li (1993) and Cook and Ni (2005). Several efforts have been devoted to relaxing this condition, see Li and Dong (2009) and Dong and Li (2010). We note that 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 that it is very close to the linearity condition. In other words, this condition is not very strong when the inverse regression notion is adopted. Thus, we still use it in the present paper.

The following proposition is concerned with the identifiability of index vectors.
Proposition 1. Assume the linearity condition (2.2). Then we have

\[ \beta_{LS} = (\phi_1 \beta_1^\top, \phi_2 \beta_2^\top, \ldots, \phi_K \beta_K^\top)^\top \]

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

Thus, under mild conditions on the design distribution, the \( K \) index vectors \( \beta_k \)'s can be recovered simultaneously through a single vector \( \beta_{LS} \), providing that the additive index structure of model (1.1) holds true. We further note that the random error \( \epsilon \) in model (1.1) is allowed to be dependent on \( V \) such that \( E(\epsilon | S^\top V) = 0 \). So the results developed in this paper are still valid when the model has heteroscedasticity.

Before continuing, we emphasize that, to avoid any notational confusion, \( \beta_k \)'s are redefined such that \( \beta_{LS} = (\beta_1^\top, \ldots, \beta_K^\top)^\top \). Starting with a random sample \((v_i, y_i)\), for \( i = 1, \ldots, n \), on \((V, Y)\), we propose to estimate \( \beta_{LS} \) with the vector \( \hat{\beta}_{LS} \) from the least squares fit of \( y_i \) on \( v_i \). Denote by \( y = (y_1, \ldots, y_n)^\top \) the response vector and \( V = (v_1, \ldots, v_n)^\top \) the design matrix, respectively. Assume with no loss of generality that the data are centered, so the intercept is not included in the regression function. The least squares direction estimator is given by

\[ \hat{\beta}_{LS} = (\hat{\beta}_1^\top, \ldots, \hat{\beta}_K^\top)^\top = (V^\top V)^{-1} V^\top y. \]  

(2.3)

For a vector \( u = (u_1, \ldots, u_m)^\top \in \mathbb{R}^m \), define \( \|u\|_1 = \sum_{j=1}^m |u_j| \) and \( \|u\|_2 = (\sum_{j=1}^m u_j^2)^{1/2} \). The asymptotic behavior of \( \hat{\beta}_{LS} \) is stated in the following theorem. We note that for the time being, all the predictors are relevant to the response variable and the total number of relevant predictors \( d = d_0 \) is allowed to diverge as the sample size \( n \) tends to infinity.

Theorem 1. Assume the conditions (A1)-(A7) in the supplementary document. If \( d_0 = o(n / \log n) \), then \( \hat{\beta}_{LS} \) is a consistent estimator of \( \beta_{LS} \) such that \( \|\hat{\beta}_{LS} - \beta_{LS}\|_2 = O(\sqrt{d_0/n}) \). Consequently, we have \( \|\hat{\beta}_k - \beta_k\|_2 = O(\sqrt{d_0/n}) \) for all \( k = 1, \ldots, K \).

Similar results for the linear model and for the single-index model are available in the literature. However, to our knowledge, the results for the additive index model (1.1) are novel and thus are of interest in their own right.
Remark 1. Since the new framework is much simpler than the current one with nonparametric techniques, one might expect that this gain of simplicity would be compensated by a loss of generality. Indeed, there is no guarantee that the constants $\phi_k$’s in Proposition 1 are different from zero. Let $\zeta_k = \{ \text{Cov}(V_k) \}^{-1} \text{Cov}\{ V_k, g_k(\beta_k^T V_k) \}$ for $k = 1, \ldots, K$. When the predictors are independent, one sufficient condition is that $\zeta_k \neq 0_{p_k \times 1}$ for all $k$. In the extreme case of perfect correlation, that is, $V_k = V_l$ 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 $V_k$. For partially linear single-index models, we show explicitly in Section 3.1 the conditions for the identifiability.

In practice, if $\phi_k$, or more precisely $\beta_k$, is nonzero and estimated as nonzero, then dimension reduction within $V_k$ is achieved; for those $\phi_k$’s that are zero and estimated to be zero using a penalized method to be discussed in Subsection 3.2, a more sophisticated method should be used. For example, we may combine our method for dimension reduction and the method for high-dimensional additive models by assuming a nonparametric additive model for $V_k$’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 $\phi_k$ and the within group level for a zero $\phi_k$. Thus, our framework is, to a certain extent, useful for exploratory data analysis even when some $\phi_k$’s are zero. Since in this paper we only focus on dimension reduction for $V_k$’s with a non-vanishing index, we assume without loss of generality that $\phi_k \neq 0$ for all $k$.

3. Applications

In this section, we discuss two important applications of the linear least squares framework.

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

Let us go back to the partially linear single-index model of the form

$$Y = \alpha^T X + g(\gamma^T Z) + \epsilon,$$

where $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p, Z = (Z_1, \ldots, Z_q)^T \in \mathbb{R}^q, \alpha \in \mathbb{R}^p$ is an unknown
linear parameter, $\gamma \in \mathbb{R}^q$ is an unknown single-index parameter, and $g(\cdot)$ is an unknown link function. Then, we have $K = 2$, $V_1 = X$, $V_2 = Z$ and

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

By Proposition 1, $\beta_{LS} = (\beta_1^T, \beta_2^T)^T = S\phi$ for some vector $\phi = (\phi_1, \phi_2)^T \in \mathbb{R}^2$. It follows that $\beta_1 = \phi_1\alpha$ and $\beta_2 = \phi_2\gamma$. As a consequence, if $\phi_1 \neq 0$ and $\phi_2 \neq 0$, $\alpha$ and $\gamma$ can be identified simultaneously by just one vector $\beta_{LS}$. Let $\Sigma_Z = \text{Cov}(Z)$ and $\Sigma_{ZX} = \text{Cov}(Z, X)$. Specifically, we have the following result on the identifiability.

**Proposition 2.** Under the linearity condition (2.2), there are constants $\phi_1 \neq 0$ and $\phi_2 \neq 0$ such that $\beta_{LS} = S\phi$ or, equivalently, $\beta_1 = \phi_1\alpha$ and $\beta_2 = \phi_2\gamma$, provided that

1. $\alpha \neq 0_{p \times 1}$, $\text{Cov}\{Z, g(\gamma^T Z)\} \neq 0_{q \times 1}$,

and one of the following conditions holds:

2. $X$ is independent of $Z$;
3. $V = (X^T, Z^T)^T$ has an elliptically symmetric distribution.

As a corollary to Theorem 1, the following statement demonstrates the consistency of the least squares direction estimator $\hat{\beta}_{LS} = (\hat{\beta}_1^T, \hat{\beta}_2^T)^T$ in the presence of a diverging number of relevant predictors.

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

The methodologies in the literature for model (3.1) are quite mature. As mentioned before, the parameters in (3.1) are often estimated via an algorithm which iteratively updates estimates of the nonparametric component and the parametric component. Recently, Wang et al. (2010) introduced a two-step estimation procedure and Liang et al. (2010) proposed a profile least squares procedure. The resulting estimators can be found without employing the iterative algorithm. However, the former has two bandwidths to be selected and is still unsatisfactory, and the latter involves a nonlinear optimization problem, which is iterative in nature and may run into numerical instability with high-dimensional predictors. Another relevant study, by Feng et al. (2013), proposed using partial
dimension reduction techniques to obtain estimators without iteration. When
the dimension of $\mathbf{X}$ is very large, the computation is however a great challenge
because it involves an integration over the support of $\mathbf{X}$. Below we develop a
new estimation procedure for partially linear single-index models that is distinct
from existing ones.

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

$$Y = \varphi_1 \times \beta_1^\top \mathbf{X} + \tilde{g}(\beta_2^\top \mathbf{Z}) + \epsilon,$$  \hfill (3.2)

where $\varphi_1 = 1/\phi_1$ and $\tilde{g}(\cdot) = g(\cdot/\phi_2)$ is an unknown link function. If $\beta_{LS}$ 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 $h$. We denote by $\hat{\mu}$ and
$\hat{\mu}_1$ the local linear estimates (Fan and Gijbels 1996) of $\mu(t; \beta_2) = E(Y|\beta_2^\top \mathbf{Z} = t)$
and $\mu_1(t; \beta_2) = E(\mathbf{X}|\beta_2^\top \mathbf{Z} = t)$ respectively in the following:

$$\hat{\mu}(t; \beta_2) = \sum_{i=1}^n W_{ni}(t; \beta_2) y_i$$

and

$$\hat{\mu}_1(t; \beta_2) = \sum_{i=1}^n W_{ni}(t; \beta_2) x_i,$$

where

$$W_{ni}(t; \beta_2) = \frac{U_{ni}(t; \beta_2)}{\sum_{j=1}^n U_{nj}(t; \beta_2)},$$

$$U_{ni}(t; \beta_2) = K_h(\mathbf{z}_i^\top \beta_2 - t) \{S_{n2}(t; \beta_2) - (\mathbf{z}_i^\top \beta_2 - t) S_{n1}(t; \beta_2)\}$$

and

$$S_{nl} = \frac{1}{n} \sum_{i=1}^n (\mathbf{z}_i^\top \beta_2 - t)^l K_h(\mathbf{z}_i^\top \beta_2 - t), \quad l = 1, 2.$$

The proposed estimator of $\alpha$ is then given by

$$\hat{\alpha} = \hat{\varphi}_1 \times \hat{\beta}_1 = \frac{T_{n1}}{T_{n2}} \times \hat{\beta}_1,$$  \hfill (3.3)

where

$$T_{n1} = \frac{1}{n} \sum_{i=1}^n \{y_i - \hat{\mu}(\mathbf{z}_i^\top \hat{\beta}_2; \hat{\beta}_2)\} \{x_i - \hat{\mu}_1(\mathbf{z}_i^\top \hat{\beta}_2; \hat{\beta}_2)\}^\top \hat{\beta}_1$$
and
\[
T_{n2} = \frac{1}{n} \sum_{i=1}^{n} \left\{ (x_i - \hat{\mu}_1(z_i^\top \hat{\beta}_2)) \right\} \hat{\beta}_1^\top \hat{\beta}_1^2.
\]

In summary, the two-stage estimation procedure works as follows:

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

S2. Estimate \(\alpha\) in model (3.1) by (3.3).

The new estimation procedure is iteration-free and easy-to-implement, because iterative algorithm and nonlinear numerical optimization are no longer needed. Further, there is only one bandwidth in the nonparametric smoothing which can be selected via cross-validation.

In order to study the asymptotic behavior of the new proposed estimator, we list the following technical conditions that are standard and commonly used in the literature:

(C1) The density function of \(b_2^\top Z\) is positive and satisfies the Lipschitz condition for \(b_2^\top Z\) in a neighborhood of \(\beta_2\); \(\beta_2^\top Z\) has a density function that is bounded away from 0.

(C2) The functions \(g\) and \(\mu_{1j}\) have two bounded and continuous derivatives, where \(\mu_{1j}\) is the \(j\)th component of \(\mu_1\), \(1 \leq j \leq p\).

(C3) \(E(e) = 0, E(e^2) < \infty\) and \(\sup_t E(\|X\|^2_2 | \beta_2^\top Z = t) < \infty\).

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

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

For simplicity, we assume for the time being that all the predictors are relevant to the response variable, and the number of relevant predictors \(d = d_0\) is a fixed constant. We have the following result, whose proof is given in the supplementary document.

**Theorem 2.** Assume that the conditions of Theorem 1 hold. Under the regularity conditions (C1)-(C5), \(\hat{\alpha}\) is a \(\sqrt{n}\)-consistent estimator of \(\alpha\).
Asymptotic expansion of $\hat{\alpha} - \alpha$ 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 may replace $S_1$ by $S_1'$. 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 theoretical property. We remark that the idea of this two-stage estimation is general, and thus can be incorporated into other procedures to device 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. The main reason for doing so is that parsimonious models are more easily interpreted and can often increase the accuracy of estimation and prediction.

It should be noticed that, even for the partially linear single-index model (3.1), variable selection 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). Further, how to incorporate the grouping information into the selection process is also an important issue. In the literature, there are still no promising results in high-dimensional settings.

We can successfully address these concerns, thanks mainly to the linear least squares structure of $\beta_{LS}$, which enables us to very easily apply the penalization paradigm. Specifically, we consider the penalized least squares function

$$Q_\lambda(b) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - v_i^T b)^2 + \sum_{k=1}^{K} \sum_{j=1}^{p_k} J_\lambda(|b_{kj}|),$$

(3.4)

where $b = (b_1^T, \ldots, b_K^T)^T \in \mathbb{R}^d$ with $b_k = (b_{k1}, \ldots, b_{kp_k})^T \in \mathbb{R}^{p_k}$, $J_\lambda(\cdot)$ is a penalty function and $\lambda$ is a tuning parameter.

In the literature, there has been a surge of interest in penalized methods in high-dimensional linear or generalized linear models. In a recent work, Wang, Xu and Zhu (2012) extended the results to deal with high-dimensional single-index models. In the present subsection, we further consider the variable selection
problem for additive index models by taking advantage of the special model structure.

In what follows, 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 $\beta_{LS} = (\beta_1^\top, \ldots, \beta_K^\top)^\top$ are exactly zero. Without loss of generality, assume that the nonzero components of $\beta_k$ are located at the first $p_{k0}$ coordinates. Write $\beta_k = (\beta_{k1}^\top, 0_{(p_k-p_{k0})\times 1})^\top$. Accordingly, we define $V_{k1}$ in such a way that it consists of the first $p_{k0}$ components of $V_k$. Finally, write $V = (V_1, \ldots, V_K)^\top$ with $V_k$ being the design matrix corresponding to $V_k$, and let $V_{k1}$ be a sub-matrix formed by the first $p_{k0}$ columns of $V_k$.

Let $\hat{\beta}_{k1}^o$ be an ideal vector from the least squares fit of $y$ on $(V_{11}, \ldots, V_{K1})$. Define the least squares oracle estimator to be $\hat{\beta}^o = (\hat{\beta}_{11}^o, \ldots, \hat{\beta}_{K1}^o)^\top$, where $\hat{\beta}_{k1}^o = (\hat{\beta}_{k1}^o, 0_{(p_k-p_{k0})\times 1})^\top$. We have the following result on the oracle property.

**Theorem 3.** Assume the conditions (D1)-(D5) in the supplementary document. Let $A_\lambda$ be the set of local minima of $Q_\lambda(b)$ with the SCAD penalty. Then we have

$$\lim_{n \to \infty} P(\hat{\beta}^o \in A_\lambda) = 1.$$  

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

$$Q_\lambda(b_1, b_2) = \frac{1}{2n} \sum_{i=1}^n (y_i - x_i^\top b_1 - z_i^\top b_2)^2 + \sum_{j=1}^p J_\lambda(|b_{1j}|) + \sum_{j=1}^q J_\lambda(|b_{2j}|), \quad (3.5)$$

where $b_1 = (b_{11}, \ldots, b_{1p})^\top \in \mathbb{R}^p$, $b_2 = (b_{21}, \ldots, b_{2q})^\top \in \mathbb{R}^q$, and $J_\lambda(\cdot)$ is the SCAD penalty.

Assume that the nonzero components of $\beta_1$ and $\beta_2$ are respectively located at the first $p_0$ and $q_0$ coordinates. Write $\beta_1 = (\beta_{11}^\top, 0_{(p-p_0)\times 1})^\top$ and $\beta_2 = (\beta_{21}^\top, 0_{(q-q_0)\times 1})^\top$. Accordingly, we define $X_1$ and $Z_1$ in such a way that they consist of the first $p_0$ and $q_0$ components of $X$ and $Z$, respectively. Finally, let $X = (x_1, \ldots, x_n)^\top$ and $Z = (z_1, \ldots, z_n)^\top$ be the design matrices corresponding to $X$ and $Z$, and define $X_1$ and $Z_1$ as sub-matrices formed by the first $p_0$ and $q_0$ columns of $X$ and $Z$, respectively.
Denote by $\hat{\beta}^o = (\hat{\beta}_1^o, \hat{\beta}_2^o)\top$ the least squares oracle estimator, where $\hat{\beta}_1^o = (\hat{\beta}_{11}^o, 0_{(p-p_0)\times p_1})\top$ and $\hat{\beta}_2^o = (\hat{\beta}_{21}^o, 0_{(q-q_0)\times p_1})\top$ with $(\hat{\beta}_{11}^o, \hat{\beta}_{21}^o)\top$ being an ideal vector from the least squares fit of $y$ on $(X_1, Z_1)$. Similarly, we obtain the following corollary to Theorem 3.

**Corollary 2.** Assume the conditions of Theorem 3. Let $A_\lambda$ be the set of local minima of $Q_\lambda(b_1, b_2)$ with the SCAD penalty. Then we have

$$\lim_{n \to \infty} P(\hat{\beta}^o \in A_\lambda) = 1.$$  

**4. Simulation study**

In this section, we examine the finite-sample performance of the proposed estimation and selection methods. First, we focus on the partially linear single-index model and consider the following three models:

\begin{align*}
Y &= \beta_1^\top V_1 + 2 \times \beta_2^\top V_2 \times I(\beta_2^\top V_2 < 0) + \epsilon, \quad (4.1) \\
Y &= \beta_1^\top V_1 + \exp(\beta_2^\top V_2 / 2) + \epsilon, \quad (4.2) \\
Y &= \beta_1^\top V_1 + 2 \times \sin(\beta_2^\top V_2 / 2) + \epsilon, \quad (4.3)
\end{align*}

where $I(\cdot)$ is the indicator function. We present four cases, covering different and commonly encountered scenarios.

**Case 1.** $\epsilon \sim N(0, 1)$, $V \sim N(0_{d \times 1}, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, d$, and $\epsilon$ and $V$ 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 $\beta_1 = (1.5, 1, 1, 0, \ldots, 0)\top$ and $\beta_2 = (1, 1, 0, \ldots, 0)\top$, respectively. The sample size is $n = 200$.

**Case 2.** The same as Case 1, except that the error $\epsilon$ 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 apply the penalized least squares estimation in Subsection 3.2 with the SCAD penalty, and then invoke the two-stage estimation in Subsection 3.1. We adopt the Gaussian kernel in local linear smoothing and use the least squares cross-validation (Li and Racine 2004) to select the smoothing
parameter. For comparison purpose, we also evaluate the performance of the penalized least squares estimation with the LASSO penalty, as well as the oracle least squares estimation assuming that the irrelevant predictors are known beforehand. The resulting estimators are denoted by SCAD, LASSO and Oracle, respectively. Oracle also reflects the estimation efficiency when our method is applied. For each penalized competitor, we implement the fast and efficient coordinate descent algorithm (see, e.g., Friedman, Hastie and Tibshirani 2010) and select its tuning parameter by ten-fold cross-validation.

For any vector \( \theta \), we denote \( \theta^s \) the orthonormalized version of \( \theta \). To evaluate estimation accuracy, we compute the absolute correlation coefficient, \( ACC_{\theta_k} \), between the estimated predictor and the true one: \( ACC_{\theta_k} = |\text{corr}(\hat{\theta}_k^TV_k, \theta_k^TV_k)| \) for SCAD and LASSO and \( ACC_{\theta_k} = |\text{corr}(\hat{\theta}_k^TV_{k1}, \theta_k^TV_{k1})| \) for Oracle; plus the vector correlation coefficient, \( VCC_{\theta_k} = \hat{\theta}_k^s\theta_k^s \). Here, for both SCAD and LASSO \( \theta_k = \beta_k \) and for Oracle \( \theta_k = \beta_{k1}, k = 1, \ldots, K \). For partially linear single-index models, we use \( \text{EST}_{\theta} = \|\hat{\theta} - \theta\|_2 \) to measure the performance of the two-stage estimation procedure; for both SCAD and LASSO \( \theta = \beta_1 \) and for Oracle \( \theta = \beta_{11} \).

To assess how well SCAD and LASSO select predictors, we employ three commonly used summary statistics: the number of nonzero components (\( \text{MS}_{\theta_k} \)); the true positive rate (\( \text{TPR}_{\theta_k} \)), which is defined to be the ratio of the number of correctly identified predictors to the number of truly important predictors; and the false positive rate (\( \text{FPR}_{\theta_k} \)), which is defined to be the ratio of the number of falsely identified predictors to the total number of irrelevant predictors. Ideally, we wish to have \( \text{TPR}_{\theta_k} \) to be close to 1 and \( \text{FPR}_{\theta_k} \) to be close to 0 at the same time.

The simulation results based on 200 data replications from these four cases are summarized in Tables 4.1-4.3, respectively. In Cases 1 and 2, the predictors are serially correlated, and several conclusions can be drawn as follows. First, SCAD and LASSO have very comparable performance in terms of estimation and selection. For instance, the average absolute correlation coefficient and the average vector correlation coefficient of SCAD and LASSO are close to one and are slightly lower than those of Oracle. In addition, SCAD and LASSO successfully identify 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. Second, we can see that the estimation accuracy of all the methods considered deteriorate when we replace the normal distribution of the error with the t-distribution. In Cases 3 and 4, there are constant positive correlations among the predictors, and we can observe that SCAD is generally superior to LASSO. For instance, the values of absolute/vector correlation coefficients of SCAD are higher than those of LASSO, and the estimation error of the linear component (EST\(_{\theta_k}\)) of SCAD, which is close to that of Oracle, is significantly lower than that of LASSO. Further, it can be seen that LASSO tends to select a model with many spurious predictors. These indicate that LASSO is biased and may not have the oracle property. As a result, the performance of SCAD is more robust to the correlation structure among the predictors than that of LASSO, which is well in accordance with the theoretical results established in the literature (Fan and Lv 2010). Unreported results also show that increasing sample sizes improves the performance.

Table 4.1: Summary of model (4.1). The average absolute correlation coefficient (ACC\(_{\theta_k}\)), the average vector correlation coefficient (VCC\(_{\theta_k}\)), the average estimation error of linear component (EST\(_{\theta_k}\)), the average number of nonzero components (MS\(_{\theta_k}\)), the true positive rate (TPR\(_{\theta_k}\)) and the false positive rate (FPR\(_{\theta_k}\)), based on 200 data replications, are reported. \(\theta_k = \beta_k\) for SCAD and LASSO and \(\theta_k = \beta_{k1}\) for Oracle.
Table 4.2: Summary of model (4.2). The average absolute correlation coefficient (ACC\(\theta_k\)), the average vector correlation coefficient (VCC\(\theta_k\)), the average estimation error of linear component (EST\(\theta_k\)), the average number of nonzero components (MS\(\theta_k\)), the true positive rate (TPR\(\theta_k\)) and the false positive rate (FPR\(\theta_k\)), based on 200 data replications, are reported. \(\theta_k = \beta_k\) for SCAD and LASSO and \(\theta_k = \beta_{k1}\) for Oracle.

<table>
<thead>
<tr>
<th></th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC(\theta_1)</td>
<td>0.996 0.994 0.998</td>
<td>0.989 0.991 0.998</td>
<td>0.997 0.989 0.998</td>
<td>0.995 0.983 0.998</td>
</tr>
<tr>
<td></td>
<td>(0.003) (0.005) (0.001)</td>
<td>(0.013) (0.009) (0.002)</td>
<td>(0.004) (0.006) (0.001)</td>
<td>(0.007) (0.010) (0.001)</td>
</tr>
<tr>
<td>VCC(\theta_1)</td>
<td>0.991 0.988 0.996</td>
<td>0.977 0.982 0.994</td>
<td>0.992 0.968 0.995</td>
<td>0.982 0.948 0.993</td>
</tr>
<tr>
<td></td>
<td>(0.007) (0.009) (0.003)</td>
<td>(0.033) (0.017) (0.007)</td>
<td>(0.015) (0.022) (0.004)</td>
<td>(0.032) (0.039) (0.006)</td>
</tr>
<tr>
<td>EST(\theta_1)</td>
<td>0.250 0.296 0.167</td>
<td>0.402 0.368 0.211</td>
<td>0.238 0.560 0.182</td>
<td>0.347 0.697 0.222</td>
</tr>
<tr>
<td></td>
<td>(0.105) (0.118) (0.078)</td>
<td>(0.250) (0.158) (0.098)</td>
<td>(0.130) (0.157) (0.083)</td>
<td>(0.203) (0.210) (0.099)</td>
</tr>
<tr>
<td>TPR(\theta_1)</td>
<td>1.000 1.000</td>
<td>1.000 1.000</td>
<td>1.000 1.000</td>
<td>1.000 1.000</td>
</tr>
<tr>
<td>FPR(\theta_1)</td>
<td>0.028 0.034</td>
<td>0.009 0.049</td>
<td>0.014 0.047</td>
<td>0.014 0.047</td>
</tr>
<tr>
<td>ACC(\theta_2)</td>
<td>0.984 0.985 0.997</td>
<td>0.958 0.974 0.996</td>
<td>0.985 0.970 0.997</td>
<td>0.950 0.945 0.995</td>
</tr>
<tr>
<td></td>
<td>(0.017) (0.017) (0.003)</td>
<td>(0.035) (0.028) (0.005)</td>
<td>(0.025) (0.019) (0.003)</td>
<td>(0.068) (0.051) (0.007)</td>
</tr>
<tr>
<td>VCC(\theta_2)</td>
<td>0.963 0.973 0.992</td>
<td>0.905 0.952 0.988</td>
<td>0.962 0.919 0.992</td>
<td>0.872 0.835 0.987</td>
</tr>
<tr>
<td></td>
<td>(0.047) (0.027) (0.009)</td>
<td>(0.085) (0.047) (0.015)</td>
<td>(0.060) (0.061) (0.011)</td>
<td>(0.133) (0.142) (0.020)</td>
</tr>
<tr>
<td>TPR(\theta_2)</td>
<td>1.000 1.000</td>
<td>0.990 1.000</td>
<td>0.997 1.000</td>
<td>0.960 0.987</td>
</tr>
<tr>
<td>FPR(\theta_2)</td>
<td>0.028 0.034</td>
<td>0.035 0.031</td>
<td>0.010 0.049</td>
<td>0.015 0.048</td>
</tr>
</tbody>
</table>

To gain further insight into the proposed method, we next consider a more complex model of the form

\[
Y = (1 + \beta_1^\top V_1/2)^2 + 2 \times \sin(\beta_2^\top V_2/2) + 1.5 \times \exp(\beta_3^\top V_3/2) + \epsilon. \tag{4.4}
\]

As before, we consider four cases.

**Case 5.** \(\epsilon \sim N(0, 1), V \sim N(\Theta_{d \times 1}, \Sigma)\) with \(\Sigma_{ij} = 0.5^{|i-j|}\) for \(i, j = 1, \ldots, d\), and \(\epsilon\) and \(V\) are independent. Let \(p_1 = p_2 = p_3 = 200\) and \(p_{10} = p_{20} = p_{30} = 2\). The single-index parameters are \(\beta_1 = (1, -1, 0, \ldots, 0)^\top, \beta_2 = (1, 1, 0, \ldots, 0)^\top\) and \(\beta_3 = (-1, 1, 0, \ldots, 0)^\top\), respectively. The sample size is \(n = 200\).

**Case 6.** The same as Case 5, except that the error \(\epsilon\) 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 \(i \neq j\).
Table 4.3: Summary of model (4.3). The average absolute correlation coefficient (ACC\(\theta_k\)), the average vector correlation coefficient (VCC\(\theta_k\)), the average estimation error of linear component (EST\(\theta_k\)), the average number of nonzero components (MS\(\theta_k\)), the true positive rate (TPR\(\theta_k\)) and the false positive rate (FPR\(\theta_k\)), based on 200 data replications, are reported. \(\theta_k = \beta_k\) for SCAD and LASSO and \(\theta_k = \beta_{k1}\) for Oracle.

<table>
<thead>
<tr>
<th></th>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC(\theta_1)</td>
<td>0.998, 0.996, 0.999</td>
<td>0.994, 0.993, 0.998</td>
<td>0.999, 0.992, 0.999</td>
<td>0.999, 0.992, 0.999</td>
</tr>
<tr>
<td></td>
<td>(0.001), (0.003), (0.000)</td>
<td>(0.006), (0.007), (0.001)</td>
<td>(0.000), (0.003), (0.000)</td>
<td>(0.005), (0.007), (0.001)</td>
</tr>
<tr>
<td>VCC(\theta_1)</td>
<td>0.995, 0.992, 0.997</td>
<td>0.987, 0.986, 0.995</td>
<td>0.996, 0.980, 0.997</td>
<td>0.988, 0.960, 0.995</td>
</tr>
<tr>
<td></td>
<td>(0.003), (0.007), (0.003)</td>
<td>(0.019), (0.012), (0.005)</td>
<td>(0.002), (0.013), (0.002)</td>
<td>(0.020), (0.030), (0.005)</td>
</tr>
<tr>
<td>EST(\theta_1)</td>
<td>0.188, 0.233, 0.139</td>
<td>0.310, 0.320, 0.195</td>
<td>0.171, 0.462, 0.150</td>
<td>0.298, 0.610, 0.207</td>
</tr>
<tr>
<td></td>
<td>(0.076), (0.102), (0.062)</td>
<td>(0.181), (0.136), (0.089)</td>
<td>(0.082), (0.116), (0.062)</td>
<td>(0.171), (0.184), (0.090)</td>
</tr>
<tr>
<td>MS(\theta_1)</td>
<td>10.815, 15.170</td>
<td>15.095, 15.215</td>
<td>4.515, 24.345</td>
<td>8.000, 22.100</td>
</tr>
<tr>
<td>TPR(\theta_1)</td>
<td>1.000, 1.000</td>
<td>0.998, 1.000</td>
<td>1.000, 1.000</td>
<td>0.998, 1.000</td>
</tr>
<tr>
<td>FPR(\theta_1)</td>
<td>0.019, 0.030</td>
<td>0.030, 0.030</td>
<td>0.003, 0.053</td>
<td>0.012, 0.048</td>
</tr>
<tr>
<td>ACC(\theta_2)</td>
<td>0.992, 0.990, 0.998</td>
<td>0.966, 0.979, 0.996</td>
<td>0.992, 0.977, 0.998</td>
<td>0.968, 0.959, 0.996</td>
</tr>
<tr>
<td></td>
<td>(0.010), (0.012), (0.002)</td>
<td>(0.031), (0.022), (0.004)</td>
<td>(0.016), (0.017), (0.002)</td>
<td>(0.038), (0.031), (0.004)</td>
</tr>
<tr>
<td>VCC(\theta_2)</td>
<td>0.980, 0.980, 0.993</td>
<td>0.919, 0.961, 0.989</td>
<td>0.977, 0.942, 0.993</td>
<td>0.914, 0.886, 0.990</td>
</tr>
<tr>
<td></td>
<td>(0.029), (0.021), (0.007)</td>
<td>(0.076), (0.040), (0.012)</td>
<td>(0.045), (0.043), (0.008)</td>
<td>(0.114), (0.100), (0.012)</td>
</tr>
<tr>
<td>MS(\theta_2)</td>
<td>5.775, 7.785</td>
<td>7.960, 8.165</td>
<td>2.645, 12.950</td>
<td>4.315, 11.325</td>
</tr>
<tr>
<td>TPR(\theta_2)</td>
<td>1.000, 1.000</td>
<td>0.995, 1.000</td>
<td>1.000, 1.000</td>
<td>0.990, 0.997</td>
</tr>
<tr>
<td>FPR(\theta_2)</td>
<td>0.019, 0.029</td>
<td>0.030, 0.031</td>
<td>0.003, 0.055</td>
<td>0.011, 0.047</td>
</tr>
</tbody>
</table>

The empirical results based on 200 data replications from these four cases are reported in Table 4.4. As we can see, the results are qualitatively similar to those of the partially linear single-index model.

5. Real data analysis

In this section, we apply the proposed method to a real world dataset, which represents a set of possible advertisements on Internet pages and is available at the University of California-Irvine machine 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 \(n = 2358\) observations and \(p = 1430\) predictors. Among the predictors, the first three are related to the geometry of the image and hence are continuous, and the rest are binary. Therefore, it is nat-
Table 4.4: Summary of model (4.4). The average absolute correlation coefficient (ACC\(\theta_k\)), the average vector correlation coefficient (VCC\(\theta_k\)), the average estimation error of linear component (EST\(\theta_k\)), the average number of nonzero components (MS\(\theta_k\)), the true positive rate (TPR\(\theta_k\)) and the false positive rate (FPR\(\theta_k\)), based on 200 data replications, are reported. \(\theta_k = \beta_k\) for SCAD and LASSO and \(\theta_k = \beta_{k+1}\) for Oracle.

<table>
<thead>
<tr>
<th></th>
<th>Case 5</th>
<th>Case 6</th>
<th>Case 7</th>
<th>Case 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCAD</td>
<td>LASSO</td>
<td>Oracle</td>
<td>SCAD</td>
<td>LASSO</td>
</tr>
<tr>
<td>ACC(\theta_1)</td>
<td>0.988</td>
<td>0.933</td>
<td>0.991</td>
<td>0.960</td>
</tr>
<tr>
<td>(0.013)</td>
<td>(0.037)</td>
<td>(0.006)</td>
<td>(0.074)</td>
<td>(0.095)</td>
</tr>
<tr>
<td>VCC(\theta_1)</td>
<td>0.994</td>
<td>0.970</td>
<td>0.998</td>
<td>0.979</td>
</tr>
<tr>
<td>(0.005)</td>
<td>(0.020)</td>
<td>(0.002)</td>
<td>(0.044)</td>
<td>(0.052)</td>
</tr>
<tr>
<td>TPR(\theta_1)</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>FPR(\theta_1)</td>
<td>0.027</td>
<td>0.071</td>
<td>0.011</td>
<td>0.016</td>
</tr>
<tr>
<td>ACC(\theta_2)</td>
<td>0.990</td>
<td>0.975</td>
<td>0.997</td>
<td>0.951</td>
</tr>
<tr>
<td>(0.013)</td>
<td>(0.018)</td>
<td>(0.002)</td>
<td>(0.058)</td>
<td>(0.080)</td>
</tr>
<tr>
<td>VCC(\theta_2)</td>
<td>0.977</td>
<td>0.960</td>
<td>0.993</td>
<td>0.878</td>
</tr>
<tr>
<td>(0.035)</td>
<td>(0.026)</td>
<td>(0.009)</td>
<td>(0.138)</td>
<td>(0.133)</td>
</tr>
<tr>
<td>TPR(\theta_2)</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.955</td>
</tr>
<tr>
<td>FPR(\theta_2)</td>
<td>0.027</td>
<td>0.070</td>
<td>0.010</td>
<td>0.015</td>
</tr>
<tr>
<td>ACC(\theta_3)</td>
<td>0.986</td>
<td>0.919</td>
<td>0.994</td>
<td>0.922</td>
</tr>
<tr>
<td>(0.020)</td>
<td>(0.053)</td>
<td>(0.008)</td>
<td>(0.143)</td>
<td>(0.142)</td>
</tr>
<tr>
<td>VCC(\theta_3)</td>
<td>0.993</td>
<td>0.955</td>
<td>0.998</td>
<td>0.955</td>
</tr>
<tr>
<td>(0.015)</td>
<td>(0.035)</td>
<td>(0.002)</td>
<td>(0.108)</td>
<td>(0.105)</td>
</tr>
<tr>
<td>MS(\theta_3)</td>
<td>7.055</td>
<td>16.005</td>
<td>10.120</td>
<td>5.140</td>
</tr>
<tr>
<td>TPR(\theta_3)</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.982</td>
</tr>
<tr>
<td>FPR(\theta_3)</td>
<td>0.025</td>
<td>0.070</td>
<td>0.009</td>
<td>0.016</td>
</tr>
</tbody>
</table>

ural to divide the predictors into two groups, with the first group containing the three continuous predictors and the second group containing the others. Further, it is common practice to restrict the effects of categorical predictors to be linear, indicating that the partially linear single-index model (3.1), or its generalized version with logit link (Carroll et al. 1997), will work in this setting. We note that the conditional independence implied by model (3.1), and hence Proposition 2, still hold when we consider the generalized partially linear single-index model.

We carry out the study by repeated random splitting of the whole dataset into balanced training and test sets. Specifically, we set 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, both SCAD and LASSO are applied to the training data. Since there are only three predictors in the semi-parametric component, it is convenient to simply include them in the model without shrinkage of their coefficients. The new constructed predictors are given by $\mathbf{X}\hat{\beta}_1$ and $\mathbf{Z}\hat{\beta}_2$. Then the partially linear model and the generalized partially linear model with logit link are fitted to the same data. Finally, the performance of the fitted models is evaluated by the test samples. To reduce the variability, the splitting into training and test sets was repeated 200 times, and the results are summarized in Table 5.5. We see that the methods considered here are comparable, with LASSO having slightly lower classification errors but being apparently larger in model size. It is worth mentioning that although 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 $\{X_j : j \in \hat{M}\}$ with $\hat{M} = \{j : \hat{\beta}_{1j} \neq 0\}$, instead of $\mathbf{X}\hat{\beta}_1$, leads to inferior performance. More seriously, the resulting algorithms often fail to converge due to the large size of $\hat{M}$.

Table 5.5: Advertising data. Classification errors made and the number of predictors chosen over 200 random splitting of all samples into training and test sets

<table>
<thead>
<tr>
<th>Method</th>
<th>Training error (%) (mean±sd)</th>
<th>Test error (%) (mean±sd)</th>
<th>Number of selected predictors (mean±sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO + PLM</td>
<td>1.75±0.377</td>
<td>3.68±0.489</td>
<td>206.38±55.727</td>
</tr>
<tr>
<td>LASSO + GPLM</td>
<td>1.42±0.317</td>
<td>3.35±0.495</td>
<td>206.38±55.727</td>
</tr>
<tr>
<td>SCAD + PLM</td>
<td>2.51±0.587</td>
<td>4.43±0.594</td>
<td>61.52±15.083</td>
</tr>
<tr>
<td>SCAD + GPLM</td>
<td>2.19±0.551</td>
<td>4.04±0.549</td>
<td>61.52±15.083</td>
</tr>
</tbody>
</table>

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

Supplementary Document The supplementary file describes the regularity conditions and proofs.
References


