Statistica Sinica **23** (2013), 543-569
doi:http://dx.doi.org/10.5705/ss.2011.275

PENALIZED MINIMUM AVERAGE VARIANCE ESTIMATION

Tao Wang[1], Peirong Xu[2] and Lixing Zhu[1]

[1]*Hong Kong Baptist University and* [2]*East China Normal University*

Abstract: For simultaneous dimension reduction and variable selection for general regression models, including the multi-index model as a special case, we propose a penalized minimum average variance estimation method, combining the ideas of minimum average variance estimation in dimension reduction and regularization in variable selection. The resulting estimator can be found in a computationally efficient manner. Under mild conditions, the new method can consistently select all relevant predictors and has the oracle property. Simulations and a data example demonstrate the effectiveness and efficiency of the proposed method.

Key words and phrases: Dimension reduction, minimum average variance estimation, oracle property, single-index model, sufficient dimension reduction, variable selection.

1. Introduction

Consider a univariate response Y and a predictor vector $X = (X_1, \ldots, X_p)^T$. The goal of regression analysis is to infer about the conditional mean function or, in general, the conditional distribution of $Y|X$. To deal with high-dimensional data that arise frequently in modern scientific research, the literature offers sufficient dimension reduction and variable selection. Sufficient dimension reduction aims to replace the predictor vector with its low-dimensional projection onto a suitable subspace of the predictor space without losing regression information, whereas variable selection is to rule out predictors that insignificantly affect the response.

More specifically, sufficient dimension reduction seeks to find a subspace $\mathcal{S}$ of minimum dimension such that $Y \perp\!\!\!\perp X | P_{\mathcal{S}} X$, where $\perp\!\!\!\perp$ indicates independence and $P_{\mathcal{S}}$ stands for the orthogonal projection onto $\mathcal{S}$; that is, Y and X are independent conditioned on $P_{\mathcal{S}} X$ or equivalently, the conditional distribution of $Y|X$ equals that of $Y|P_{\mathcal{S}} X$. Such a space, denoted $\mathcal{S}_{Y|X}$ and often called the central subspace, typically exists and is unique under mild conditions (Cook (1996)). When all projection directions are contained in the conditional mean of Y given X, $E(Y|X)$, the above independence can be written as $Y \perp\!\!\!\perp E(Y|X) | P_{\mathcal{S}} X$. The related subspace is called the central mean subspace (CMS). A number of methods have been proposed to estimate the central/central mean subspace: inverse regression methods such as sliced inverse regression method (SIR, Li (1991)) and sliced average variance estimation method (SAVE, Cook and Weisberg (1991)); direct regression methods, in particular the minimum average variance estimation method (MAVE, Xia et al. (2002)), are popular. Inverse regression methods are computationally simple and practically useful, but the probabilistic assumptions on the design of predictors, such as the linearity condition (Li (1991)), are considered to be strong (Li and Dong (2009)). Further, some of them fail in one way or another to estimate the central/central mean subspace (Cook and Weisberg (1991)). For CMS, MAVE is useful for both dimension reduction and nonparametric function estimation. It is free of the linearity condition and often has much better performance in finite samples than inverse regression methods.

The aforementioned methods assume that all predictors contain useful information. As a result, the estimated directions generally involve all of the original predictors, making their interpretation sometimes difficult. Further, if irrelevant predictors are included, quite likely in high-dimensional situations, the precision of estimation as well as the accuracy of prediction is lessened. It is thus crucial to consider variable selection along with dimension reduction, and several attempts at this have been made within the framework of sufficient dimension reduction. For instance, Li, Cook, and Nachtsheim (2005) developed test-based selection procedures, Li (2007) suggested sparse sufficient dimension reduction via integrating inverse regression estimation with a regularization paradigm, Zhou and He (2008) studied a constrained canonical correlation method, and Wang and Yin (2008) proposed sparse minimum average variance estimation. However, most methods are designed for element screening and not variable screening, producing a shrinkage solution for a basis matrix of the central/central mean subspace.

Building upon the regression-type formulation of sufficient dimension reduction methods, Chen, Zou, and Cook (2010) proposed a coordinate-independent sparse estimation that can simultaneously achieve dimension reduction and screen out irrelevant predictors. With the aid of the local quadratic approximation, the resulting estimators can be found by an application of the eigen-system analysis. Their general method is subspace-oriented and covers a number of inverse regression methods, but inherits all drawbacks of inverse regression methods, that is, the strong conditions on predictors and regression function. Also, as pointed out in Fan and Li (2001), the local quadratic approximation algorithm has to delete any small coefficient because it cannot have a sparse representation.

In this paper, by incorporating a bridge penalty (Frank and Friedman (1993)) to L_1 norms of the rows of a basis matrix, we propose a penalized minimum average variance estimation (P-MAVE) that can achieve sufficient dimension reduction and variable selection simultaneously. A fast and efficient algorithm is developed to solve the optimization problem, and the resulting estimator naturally adopts a sparse representation. A BIC-type criterion is suggested to select the tuning parameter. The new method is free of the linearity condition that is needed for inverse regression methods (see, e.g., Chen, Zou, and Cook (2010)) and has the oracle property under mild conditions, so that P-MAVE can be applied to time series data while inverse regression methods cannot.

The rest of the paper is organized as follows. Section 2 introduces the new method. Its theoretical properties are studied in Section 3. Numerical studies are reported in Section 4, and there a dataset with categorical predictors is analyzed for variable selection. Concluding remarks are included in Section 5. All technical details are relegated to the Appendix.

2. Methodology

2.1. Minimum average variance estimation revisited

Let $B_0 \in \mathbb{R}^{p\times d}$ denote a $p \times d$ orthogonal matrix with $d < p$, so $B_0^T B_0 = I_d$ where I_d is the $d \times d$ identity matrix. For dimension reduction we consider the model,

$$Y = g(B_0^T X) + \epsilon, \tag{2.1}$$

where g is an unknown smooth link function and $E(\epsilon|X) = 0$ almost surely.

At the population level, the MAVE procedure minimizes the objective function

$$E\{\mathrm{var}(Y|B^T X)\} = E\{Y - E(Y|B^T X)\}^2 \tag{2.2}$$

among all $B \in \mathbb{R}^{p\times d}$ such that $B^T B = I_d$. If $\sigma_B^2(B^T X) = E[\{Y - E(Y|B^T X)\}^2 \mid B^T X]$ is the conditional variance of Y given $B^T X$,

$$E\{Y - E(Y|B^T X)\}^2 = E\{\sigma_B^2(B^T X)\}.$$

Thus, minimizing (2.2) is equivalent to

$$\min_{B\in\mathbb{R}^{p\times d}} E\{\sigma_B^2(B^T X)\} \ \text{ subject to } B^T B = I_d.$$

Suppose $\{(y_i, \mathrm{x}_i), i = 1, \ldots, n\}$ is a random sample drawn according to (2.1). For x_i close to x_0, a local linear approximation is

$$y_i - g(B^T \mathrm{x}_i) \approx y_i - g(B^T \mathrm{x}_0) - \{\nabla g(B^T \mathrm{x}_0)\}^T B^T (\mathrm{x}_i - \mathrm{x}_0).$$

In the spirit of local linear smoothing, we might estimate $\sigma_B^2(B^T \mathrm{x}_0)$ by

$$\hat{\sigma}_B^2(B^T \mathrm{x}_0) = \min_{a\in\mathbb{R}, \mathrm{b}\in\mathbb{R}^d} \sum_{i=1}^{n} \{y_i - a - \mathrm{b}^T B^T (\mathrm{x}_i - \mathrm{x}_0)\}^2 w_{i0}.$$

Here, $w_{i0} \geq 0, i = 1, \ldots, n$, are some weights such that $\sum_{i=1}^n w_{i0} = 1$, often centering at x_0. Let $\mathrm{x}_{ij} = \mathrm{x}_i - \mathrm{x}_j$. The MAVE procedure is to minimize

$$\frac{1}{2n}\sum_{j=1}^n\sum_{i=1}^n (y_i - a_j - \mathrm{b}_j^T B^T \mathrm{x}_{ij})^2 w_{ij} \tag{2.3}$$

over $a_j \in \mathbb{R}, \mathrm{b}_j \in \mathbb{R}^d$, and $B \in \mathbb{R}^{p\times d}$ such that $B^T B = I_d$. In practice, minimizing (2.3) can be decomposed into two weighted least squares problems by fixing $(a_j, \mathrm{b}_j), j = 1, \ldots, n$, and B, alternatively. To be specific, given B we minimize, for $j = 1, \ldots, p$,

$$\frac{1}{2n}\sum_{i=1}^n (y_i - a_j - \mathrm{b}_j^T B^T \mathrm{x}_{ij})^2 w_{ij} \tag{2.4}$$

with respect to $a_j \in \mathbb{R}$ and $\mathrm{b}_j \in \mathbb{R}^d$. With $\mathrm{c}_j = (a_j, \mathrm{b}_j^T)^T$ and $\mathrm{s}_{ij} = (1, \mathrm{x}_{ij}^T B)^T$, the solution is

$$\mathrm{c}_j = \left(\sum_{i=1}^n \mathrm{s}_{ij}\mathrm{s}_{ij}^T w_{ij}\right)^{-1} \sum_{i=1}^n \mathrm{s}_{ij} y_i w_{ij}. \tag{2.5}$$

Given $(a_j, \mathrm{b}_j), j = 1, \ldots, n$, we minimize

$$\frac{1}{2n}\sum_{j=1}^n\sum_{i=1}^n (y_i - a_j - \mathrm{b}_j^T B^T \mathrm{x}_{ij})^2 w_{ij} \tag{2.6}$$

over $B \in \mathbb{R}^{p\times d}$ subject to $B^T B = I_d$. Note that

$$\frac{1}{2n}\sum_{j=1}^n\sum_{i=1}^n (y_i - a_j - \mathrm{b}_j^T B^T \mathrm{x}_{ij})^2 w_{ij} = \frac{1}{2n}\sum_{j=1}^n\sum_{i=1}^n \{y_i - a_j - (\mathrm{x}_{ij}^T \otimes \mathrm{b}_j^T)\beta\}^2 w_{ij},$$

where $\beta = \mathrm{vec}(B^T)$ and $\mathrm{vec}(\cdot)$ is a matrix operator that stacks all columns of a matrix into a vector. Again, this is a weighted least squares problem with closed-form solution

$$\beta = \left[\sum_{j=1}^n\sum_{i=1}^n \{(\mathrm{x}_{ij}\mathrm{x}_{ij}^T) \otimes (\mathrm{b}_j\mathrm{b}_j^T)\} w_{ij}\right]^{-1} \sum_{j=1}^n\sum_{i=1}^n (\mathrm{x}_{ij} \otimes \mathrm{b}_j)(y_i - a_j) w_{ij}. \tag{2.7}$$

The procedure now consists of iteration between (2.5) and (2.7). As for the weights w_{ij}, we employ the lower-dimensional kernel weights

$$w_{ij} = \frac{K_h(B^T \mathrm{x}_{ij})}{\sum_{i=1}^n K_h(B^T \mathrm{x}_{ij})}, \tag{2.8}$$

where B is the latest or current estimate of B_0, and $K_h(\cdot)$ is a d-dimensional kernel function with bandwidth h; see Xia et al. (2002) for more details.

2.2. Penalized MAVE

In practice, it is often the case that some of predictors have only marginal influence on the response. Effectively removing these irrelevant predictors could improve estimation accuracy and enhance model interpretability. This motivates us to apply regularization to simultaneously estimate parameters and select relevant predictors.

By Proposition 1 in Cook (2004), the kth predictor can be removed from the model if and only if the kth row of B_0D is a zero vector, where D denotes an arbitrary $d \times d$ orthogonal matrix. Thus, the selection of predictors can be realized through the selection of non-vanishing rows of B_0D. Building upon this observation and the iterative least squares MAVE algorithm, we propose the penalized MAVE procedure, as follows.

For some $d \times d$ orthogonal matrix D_0, suppose that $\tilde{B}$ is an initial estimator of B_0D_0. Replace B in (2.8) by $\tilde{B}$ and let $\tilde{w}_{ij}$ denote the new weights. For each j, let

$$(\tilde{a}_j, \tilde{\mathrm{b}}_j) = \underset{a_j \in \mathbb{R}, \mathrm{b}_j \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - a_j - \mathrm{b}_j^T \tilde{B}^T \mathrm{x}_{ij})^2 \tilde{w}_{ij}. \tag{2.9}$$

Write $\beta = (\beta_1^T, \ldots, \beta_p^T)^T$ and $\tilde{\beta} = (\tilde{\beta}_1^T, \ldots, \tilde{\beta}_p^T)^T$, where $\beta_k = (\beta_k^1, \ldots, \beta_k^d)^T \in \mathbb{R}^d$ is the kth row of $B \in \mathbb{R}^{p \times d}$ and $\tilde{\beta}_k = (\tilde{\beta}_k^1, \ldots, \tilde{\beta}_k^l)^T \in \mathbb{R}^d$ is the kth row of $\tilde{B} \in \mathbb{R}^{p \times d}$, respectively. Similarly, write $\tilde{\mathrm{x}}_{ij} \equiv \mathrm{x}_{ij} \otimes \tilde{\mathrm{b}}_j = (\tilde{\mathrm{x}}_{ij1}^T, \ldots, \tilde{\mathrm{x}}_{ijp}^T)^T$, where $\tilde{\mathrm{x}}_{ijk} = \mathrm{x}_{ij}^k \tilde{\mathrm{b}}_j$ and x_{ij}^k is the kth component of x_{ij}. Let $\tilde{y}_{ij} = y_i - \tilde{a}_j$. For $\mathrm{a} \in \mathbb{R}^m$ the L_r norm is written as $\|\mathrm{a}\|_r$. We consider the objective function

$$\begin{aligned} \Psi_{\lambda_n}(\beta; \tilde{\beta}) &= \frac{1}{2n} \sum_{j=1}^{n} \sum_{i=1}^{n} \{y_i - \tilde{a}_j - (\mathrm{x}_{ij}^T \otimes \tilde{\mathrm{b}}_j^T)\beta\}^2 \tilde{w}_{ij} + \lambda_n \sum_{k=1}^{p} \|\beta_k\|_1^\gamma \\ &= \frac{1}{2n} \sum_{j=1}^{n} \sum_{i=1}^{n} (\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T \beta)^2 \tilde{w}_{ij} + \lambda_n \sum_{k=1}^{p} \|\beta_k\|_1^\gamma, \end{aligned} \tag{2.10}$$

where $\lambda_n > 0$ is a tuning parameter and $\gamma \in (0, 1)$. By adopting a bridge penalty for the L_1 norms of the rows of B, it is then possible to carry out variable screening and element screening simultaneously. We call $\hat{\beta}$ minimizing (2.10) the penalized MAVE estimator (P-MAVE). In the literature, Huang et al. (2009) adopted the same penalty function to perform group variable selection in linear regression models.

2.3. Computation

Because the penalty term in (2.10) is non-convex, direct minimization of $\Psi_{\lambda_n}(\beta;\tilde{\beta})$ is difficult. However it is possible to give an equivalent form, in certain sense, of (2.10), one that is easier to compute. Take

$$\Phi_{\lambda_n}(\beta,\theta;\tilde{\beta}) = \frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij}-\tilde{\mathrm{x}}_{ij}^{T}\beta)^2\tilde{w}_{ij} + \sum_{k=1}^{p}\theta_k^{1-1/\gamma}\|\beta_k\|_1 + \tau_n\sum_{k=1}^{p}\theta_k, \quad (2.11)$$

where τ_n is a penalty parameter. The following is proved in the Appendix.

Proposition 1. *Suppose that $\gamma \in (0,1)$. If $\lambda_n = \tau_n^{1-\gamma}\gamma^{-\gamma}(1-\gamma)^{\gamma-1}$, then $\hat{\beta}$ is a minimizer of $\Psi_{\lambda_n}(\beta;\tilde{\beta})$ if and only if $(\hat{\beta},\hat{\theta})$ is a minimizer of $\Phi_{\lambda_n}(\beta,\theta;\tilde{\beta})$ subject to $\theta > 0$, where $\theta > 0$ means $\theta_k \geq 0$ for $k = 1,\ldots,p$.*

To estimate β and θ, we can use an iterative approach (see, e.g., Huang et al. (2009)). Specifically, we first fix β and estimate θ, then we fix θ and estimate β, and we iterate between these two steps to convergence. Since the value of the objective function decreases at each step, the process is guaranteed to converge. The algorithm proceeds as follows.

S1. Initialization. Set $\beta^{(0)} = \tilde{\beta}$ and $s = 1$.

S2. For $k = 1,\ldots,p$, compute

$$\theta_k^{(s)} = \left(\frac{1-\gamma}{\tau_n\gamma}\right)^{\gamma}\left\|\beta_k^{(s-1)}\right\|_1^{\gamma}.$$

S3. Compute

$$\beta^{(s)} = \operatorname*{argmin}_{\beta}\frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij}-\tilde{\mathrm{x}}_{ij}^{T}\beta)^2\tilde{w}_{ij} + \sum_{k=1}^{p}\left(\theta_k^{(s)}\right)^{1-1/\gamma}\|\beta_k\|_1.$$

This is a LASSO problem and can be solved efficiently using the LARS algorithm (Efron et al. (2004)) or the fast coordinate descent algorithm (Friedman et al. (2007)).

S4. If $\beta^{(s)}$ and $\beta^{(s-1)}$ are close enough, stop; otherwise, set $s = s+1$ and go back to Step 2.

With a good initial value $\tilde{\beta}$, the penalized MAVE estimator obtained in this way has nice statistical properties, as shown in the following section. In practice, we might invoke one further step by taking the penalized MAVE estimator as the initial estimator, updating $\tilde{a}_j$ and $\tilde{\mathrm{b}}_j$ from (2.9), and carrying out steps S1-S4 to convergence. The resulting estimator is called the one-step penalized

MAVE estimator. As seen later in numerical studies, the one-step estimator is able to stabilize the estimation and improve the penalized MAVE estimator. Furthermore, it is less sensitive to the choice of the initial estimator $\tilde{\beta}$.

3. Asymptotic Theory

3.1 Basic theoretical properties

In what follows, we assume with no loss of generality that only the first q predictors are relevant to the response, where $d \leq q < p$. Let $A_1 = \{1, \ldots, q\}$ and $A_2 = \{q+1, \ldots, p\}$. With a slight abuse of notation, write $\beta = (\beta_{A_1}^T, \beta_{A_2}^T)^T$ with $\beta_{A_1} = (\beta_k, k \in A_1)$ a sub-vector formed by the first qd components of β, and $\beta_{A_2} = (\beta_k, k \in A_2)$ formed by the last $pd - qd$ components. Let $\mu_B(\mathrm{x}) = E(X|B^T X = B^T \mathrm{x})$ and $\nu_B(\mathrm{x}) = \mu_B(\mathrm{x}) - \mathrm{x}$. We use $\nu_{B,A_1}(\mathrm{x})$ to denote a sub-vector consisting of the first q components of $\nu_B(\mathrm{x})$. Define

$$\begin{aligned} W_{g0,A_1} &= E[\{\nu_{B_0,A_1}(X)\nu_{B_0,A_1}^T(X)\} \otimes \{\nabla g(B_0^T X)\nabla^T g(B_0^T X)\}], \\ W_{g0} &= E[\{\nu_{B_0}(X)\nu_{B_0}^T(X)\} \otimes \{\nabla g(B_0^T X)\nabla^T g(B_0^T X)\}], \\ U_{g0,A_1} &= E[\{\nu_{B_0,A_1}(X)\nu_{B_0}^T(X)\} \otimes \{\nabla g(B_0^T X)\nabla^T g(B_0^T X)\}]. \end{aligned}$$

We need the following condition.

(A) $\tilde{B} = B_0 D_0 + O_P(n^{-1/2})$, equivalently $\tilde{\beta} - \beta_0 = O_P(n^{-1/2})$.

Condition (A) is mild and typically holds. In particular, the ordinary MAVE estimator can be used as the initial estimator (Xia (2006); Wang and Xia (2008)). Note that the group bridge penalty in (2.10) is added to each row of B and the penalty is not rotation free, so the sparseness depends on a special form of rotation. Under condition (A), D_0 is the underlying rotation matrix and the identifiability issue vanishes automatically.

Theorem 1. *Suppose that* $\gamma \in (0,1)$ *and* $d \leq 3$. *Assume* (A) *and conditions* (C1)−(C5) *in the Appendix.*

(i) *If* $\lambda_n n^{-1/2} = O(1)$, *then there exists a local minimizer* $\hat{\beta}$ *of* $\Psi_{\lambda_n}(\beta; \tilde{\beta})$ *such that*

$$\|\hat{\beta} - \beta_0\|_2 = O_P\left(\frac{1}{\sqrt{n}}\right).$$

(ii) *If* $\lambda_n n^{-1/2} = O(1)$ *and* $\lambda_n n^{-\gamma/2} \to \infty$, *then* $P(\hat{\beta}_{A_2} = 0) \to 1$ *as* $n \to \infty$.

(iii) *Suppose that* $\sqrt{n}(\tilde{\beta} - \beta_0) \to_L G_1$ *and*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\nu_{B_0,A_1}(\mathrm{x}_i) \otimes \nabla g(B_0^T \mathrm{x}_i)\}\epsilon_i \to_L G_2.$$

If $\lambda_n n^{-1/2} \to \lambda_0 \in [0, \infty)$ and $\lambda_n n^{-\gamma/2} \to \infty$, then

$$\sqrt{n}(\hat{\beta}_{A_1} - \beta_{0A_1}) \to_L \underset{\mathrm{u}}{\operatorname{argmin}}\{V(\mathrm{u}), \mathrm{u} = (\mathrm{u}_1^T, \ldots, \mathrm{u}_q^T)^T \in \mathbb{R}^{qd}\},$$

where

$$V(\mathrm{u}) = -\mathrm{u}^T(U_{g0}G_1 + G_2) + \mathrm{u}^T W_{g0,A1}\mathrm{u} + \lambda_0\gamma \sum_{k=1}^{q} \|\beta_{0k}\|_1^{\gamma-1} \sum_{l=1}^{d}\{u_k^l \mathrm{sign}(\beta_{0k}^l) I(\beta_{0k}^l \neq 0) + |u_k^l| I(\beta_{0k}^l = 0)\}.$$

In particular, when $\lambda_0 = 0$,

$$\sqrt{n}(\hat{\beta}_{A_1} - \beta_{0A_1}) \to_L \frac{1}{2} W_{g0,A_1}^{+}(U_{g0}G_1 + G_2),$$

where W_{g0,A_1}^{+} denotes the Moore-Penrose inverse matrix of W_{g0,A_1}.

Theorem 1 indicates that the penalized MAVE estimator is $\sqrt{n}$-root consistent, and that it is able to select all relevant predictors consistently. Part (iii) of Theorem 1 has implications: if $\lambda_0 > 0$, the limit distribution of $\hat{\beta}_{A_1}$ that corresponds to the relevant predictors puts positive probability to 0 when some components of β_{0A_1} are exactly 0, that is, element screening along with variable screening is achievable; if $\lambda_0 = 0$, the estimator $\hat{\beta}_{A_1}$ of β_{0A_1} is asymptotically normal. For general multiple-index models, the ordinary MAVE estimate is consistent, but it is hard to specify the asymptotic distribution, particularly for $d > 3$. The asymptotic distribution is available when we use inverse regression estimates, such as sliced inverse regression, as initial estimates. However, its variance has a vary complicated structure (see, Zhu and Ng (1995)). In the special case of the single-index model, we have the oracle property.

Corollary 1. *Adopt the conditions of Theorem 1 with $d = 1$. Suppose that $\tilde{\beta}$ is the ordinary MAVE estimator.*

(i) *If $\lambda_n n^{-1/2} = O(1)$ and $\lambda_n n^{-\gamma/2} \to \infty$, then $P(\hat{\beta}_{A_2} = 0) \to 1$ as $n \to \infty$.*

(ii) *If $n^{-1/2}\sum_{i=1}^{n} g'(\beta_0^T \mathrm{x}_i)\nu_{\beta_0,A_1}(\mathrm{x}_i)\epsilon_i \to_L G_2, \lambda_n n^{-1/2} \to 0$ and $\lambda_n n^{-\gamma/2} \to \infty$, then*

$$\sqrt{n}(\hat{\beta}_{A_1} - \beta_{0A_1}) \to_L W_{g0,A_1}^{+} G_2.$$

Remark 1. It is clear that B_0 in model (2.1) or B in (2.3) is not identifiable. So the initial condition $\tilde{B} = B_0 D_0 + O_P(n^{-1/2})$ for some $d \times d$ orthogonal matrix D_0 is necessary. Many estimates, including the ordinary MAVE estimate, can serve as an initial estimate because they can consistently estimate $B_0 D_0$ (D_0 may be different for different estimates). Of course, the P-MAVE depends on the initial one, and we have shown in Theorem 3.1 that it is a consistent estimator of B_0D_0. Nevertheless, the selection consistency of the P-MAVE does not depend on the choice of the initial estimator.

Note that at the population level, the (unique) solution of the MAVE can only identify B_0D_0 for some $d \times d$ orthogonal matrix D_0. This is why we say that dimension reduction methods can identify the central (mean) dimension reduction subspace. The MAVE algorithm provides a consistent estimate of B_0D_0, see Xia et al. (2002) and Wang and Xia (2008), but this does not affect the selection by our procedure. It is easy to see that, if the kth row of B_0 is a zero vector (we say B_0 is sparse in this sense), then the kth row of B_0D for any $d \times d$ orthogonal matrix D is also zero, and vice versa. Thus, the ordinary MAVE procedure estimates the sparse version B_0D_0 for some D_0, and we can apply the techniques of shrinkage and selection to produce sparse solutions, such as the P-MAVE.

3.2. Tuning parameter selection

In practice, the attractive properties of the penalized MAVE estimator rely heavily on the choice of the tuning parameter that controls the model complexity. In the literature, generalized cross-validation (GCV) has been extensively used to choose tuning parameters, see Fan and Li (2001). Recently, Wang, Li, and Tsai (2007) proved that the model selected by GCV tends to overfit, while the Bayesian information criterion (BIC) is able to identify a finite-dimensional model consistently. This motivates us to consider a BIC-type criterion for tuning parameter selection. Let

$$\mathrm{RSS}_\lambda = \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\hat{\beta}_\lambda)^2\tilde{w}_{ij}, \tag{3.1}$$

$$\mathrm{BIC}_\lambda = \log \mathrm{RSS}_\lambda + d \times df_\lambda \times \frac{\log n}{n}, \tag{3.2}$$

where $\hat{\beta}_\lambda = (\hat{\beta}_{\lambda 1}^T, \ldots, \hat{\beta}_{\lambda p}^T)^T$ is the minimizer of $\Psi_\lambda(\beta;\tilde{\beta})$, and df_λ denotes the number of nonzero sub-vectors in $\{\hat{\beta}_{\lambda 1}, \ldots, \hat{\beta}_{\lambda p}\}$.

We use the generic notation $M \subset \{1, \ldots, p\}$ to denote an arbitrary candidate model. Let $M_F = \{1, \ldots, p\}, M_T = \{1, \ldots, q\}$, and $M_\lambda = \{k : \hat{\beta}_{\lambda,k} \neq 0\}$ be the full model, the true model, and the model that is identified by $\hat{\beta}_\lambda$, respectively.

Theorem 2. *Let $\hat{\lambda}$ denote the tuning parameter selected by minimizing* BIC_λ*. Under the conditions of Corollary* 1*, $P(M_{\hat{\lambda}} = M_T) \to 1$ as $n \to \infty$.*

4. Numerical Studies

4.1. Simulations

In this subsection, we report on a study of the finite sample performance of the proposed method via simulations. Throughout, we took the ordinary MAVE estimator to be the initial estimator and used the BIC-type criterion (3.2) to select the tuning parameter. To be specific, we used the refined MAVE estimator with the multi-dimensional Gaussian kernel, and adopted the optimal bandwidth $h = \{4/(d+2)\}^{1/(d+4)}n^{-1/(d+4)}$ (Silverman (1999)). For measuring the estimation accuracy, we employed the vector correlation coefficient COR1 $= (\prod_{l=1}^{d} \phi_l^2)^{1/2}$ and the trace correlation coefficient COR2 $= (d^{-1}\sum_{l=1}^{d} \phi_l^2)^{1/2}$, where the ϕ_l^2's are the eigenvalues of the matrix $\hat{B}^T B_0 B_0^T \hat{B}$; see Ye and Weiss (2003) for more details. Furthermore, we used four summary statistics to assess how well the method selected predictors: MS, TPR, FDR, and CM. MS is the average number of identified predictors, TPR is the average ratio of the number of correctly identified predictors to the number of truly important predictors, FDR is the average ratio of the number of falsely identified predictors to the total number of identified predictors, and CM is the fraction of runs in which the correct model is selected. In each of the following example, the reported simulation results were based on 200 data replications.

Example 1. Let $d = 1, p = 10, q = 4$, and $\beta = (0.4, -0.4, 0.8, -0.2, 0, \ldots, 0)^T$. Consider the single-index model

$$Y = 1 + 2(X^T\beta + 3) \times \log(3|X^T\beta| + 1) + \epsilon,$$

where $\epsilon \sim N(0, 1), X \sim N(0, \Sigma)$ with $\Sigma_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, 10$, and ϵ and X are independent. Two values of ρ were explored, 0 and 0.5, with the sample size $n = 100$.

Example 2. The same as Example 1, except that $n = 200$ and $p = 20$.

Example 3. Let $d = 2, p = 10$, and $q = 4$. We considered the model

$$Y = \frac{(X_1 + 2X_2 + 3X_3)^2}{14} - 2X_1 + X_2 + X_4 + 0.5\epsilon,$$

where $\epsilon \sim N(0, 1), X \sim N(0, \Sigma)$ with $\Sigma_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, 10$, and ϵ and X are independent. Two values of ρ were explored, 0 and 0.5, with the sample size $n = 100$. In this example, the true dimension reduction subspace is spanned by $(1, 2, 3, 0, \ldots, 0)^T/\sqrt{14}$ and $(-2, 1, 0, 1, 0, \ldots, 0)^T/\sqrt{6}$.

Example 4. Let $d = 3, p = 10$, and $q = 3$. The regression model was

$$Y = X_1 + \frac{X_2}{0.5 + (1.5 + X_3)^2} + 0.2\epsilon,$$

where $\epsilon \sim N(0,1), X \sim N(0,\Sigma)$ with $\Sigma_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, 10$, and ϵ and X are independent. Two values of ρ were explored, 0 and 0.5, with the sample size $n = 100$. In this example, the true dimension reduction subspace is spanned by $(1, 0, \ldots, 0)^T, (0, 1, 0, \ldots, 0)^T$, and $(0, 0, 1, 0, \ldots, 0)^T$.

Example 5. We took the non-linear time series model

$$Y_t = -1 + 0.4Y_{t-1} + \cos\left(\frac{\pi}{2}Y_{t-2}\right) + 0.2\epsilon_t,$$

where the ϵ_t's are independent and identically distributed $N(0,1)$, and $X_{t-1} = (Y_{t-1}, \ldots, Y_{t-6})^T$. In this example, $d = 2, p = 6, q = 2$, and the true dimension reduction subspace is spanned by $(1, 0, \ldots, 0)^T$ and $(0, 1, 0, \ldots, 0)^T$. Our sample sizes were $n = 100$ and $n = 200$.

We evaluated the performance of the P-MAVE estimator as well as the one-step P-MAVE estimator and, in the first four examples, we also chose the refined outer product of gradients estimator (OPG, Xia et al. (2002)) as the initial estimator. We used $\gamma = 0.5$ in the bridge penalty function. For comparison purpose, we report the simulation results of the coordinate-independent sparse estimation. In particular, we applied sliced inverse regression (CISE-SIR) and principal fitted components (CISE-PFC); see Section 3 of Chen, Zou, and Cook (2010) for details. The simulation results from these five examples are summarized in Tables 1-5, respectively.

When the original MAVE estimator was used as the initial value, observations are as follows. In all examples, both the penalized MAVE estimator and its one-step counterpart improved the MAVE estimator in terms of basis estimation. This phenomenon was more evident in Example 5, where the refined MAVE estimator had a very low vector correlation coefficient (COR1) when $n = 100$ (Table 5). In general, the one-step P-MAVE estimator outperformed the penalized MAVE estimator, and the proposed method worked quite well in terms of predictor selection: the true positive rate (TPR) was very high, lowest at 97.33%; and the false discovery rate (FDR) was close to 0. Further, the CM-values in Example 1-2 were not far from 100% (Tables 1-2), which confirms the consistency of the BIC-type criterion in the context of single-index models. When there was more than one index, $d > 1$, the lowest CM-values were 67.50% and 88.50% for the P-MAVE estimator and the one-step P-MAVE estimator, respectively. We also tried two other selection criteria, generalized cross-validation and Akaike's

Table 1. Simulation results for Example 1. The means and the standard deviations (in parentheses) of the vector correlation coefficient (COR1) and the trace correlation coefficient (COR2), the average model size (MS), the true positive rate (TPR), the false discovery rate (FDR), and the proportion of selecting the correct model (CM), based on 200 data replications, are reported.

$d = 1, n = 100, p = 10, q = 4, \rho = 0$						
	COR1	COR2	MS	TPR	FDR	CM
Initial estimate: Refined MAVE	0.9979 (0.0013)	0.9979 (0.0013)				
P-MAVE	0.9993 (0.0007)	0.9993 (0.0007)	4.0950	1.0000	0.0183	0.9150
One-step P-MAVE	0.9994 (0.0006)	0.9994 (0.0006)	4.0150	1.0000	0.0030	0.9850
Initial estimate: Refined OPG	0.9976 (0.0015)	0.9976 (0.0015)				
P-MAVE	0.9992 (0.0008)	0.9992 (0.0008)	4.1150	1.0000	0.0223	0.8950
One-step P-MAVE	0.9994 (0.0006)	0.9994 (0.0006)	4.0100	1.0000	0.0020	0.9900
CISE-SIR	0.9351 (0.0593)	0.9351 (0.0593)	2.9900	0.7338	0.0130	0.2550
CISE-PFC	0.9365 (0.0575)	0.9365 (0.0575)	3.0300	0.7388	0.0173	0.2500
$d = 1, n = 100, p = 10, q = 4, \rho = 0.5$						
Initial estimate: Refined MAVE	0.9978 (0.0014)	0.9978 (0.0014)				
P-MAVE	0.9992 (0.0009)	0.9992 (0.0009)	4.1700	1.0000	0.0333	0.8400
One-step P-MAVE	0.9994 (0.0006)	0.9994 (0.0006)	4.0400	1.0000	0.0080	0.9600
Initial estimate: Refined OPG	0.9973 (0.0018)	0.9973 (0.0018)				
P-MAVE	0.9992 (0.0009)	0.9992 (0.0009)	4.1800	1.0000	0.0353	0.8300
One-step P-MAVE	0.9994 (0.0006)	0.9994 (0.0006)	4.0350	1.0000	0.0070	0.9650
CISE-SIR	0.8841 (0.0789)	0.8841 (0.0789)	2.3900	0.5537	0.0378	0.1100
CISE-PFC	0.8800 (0.0767)	0.8800 (0.0767)	2.3000	0.5325	0.0361	0.0650

information criterion, and found that they tend to over-select many irrelevant predictors. When the refined OPG estimator was used as the initial value, we can make similar conclusions. In fact, as can be seen from Tables 1-4, the P-MAVE estimator and the one-step P-MAVE estimator, especially the latter, are not very sensitive to the choice of the initial estimator.

The P-MAVE estimator and the one-step P-MAVE estimator outperformed CISE-SIR and CISE-PFC. In Example 1, CISE-SIR and CISE-PFC missed on average one or two important predictors. In Example 2, they had much better performance when the predictors were uncorrelated (Table 2). As a conclusion, the coordinate-independent sparse estimation did not work well when either the sample size of the data was low or the predictors were correlated. This is further seen in Example 4 where we found that CISE-SIR and CISE-PFC could only estimate part of the dimension reduction subspace. Unreported results showed that the performance did not improve much when the sample size went from $n = 100$ to $n = 800$. In Example 3, CISE-SIR and CISE-PFC failed as expected, as there was a symmetry component in the mean regression function.

4.2. Automobile data

As an illustration, we apply the proposed method to an automobile data set

Table 2. Simulation results for Example 2. The means and the standard deviations (in parentheses) of the vector correlation coefficient (COR1) and the trace correlation coefficient (COR2), the average model size (MS), the true positive rate (TPR), the false discovery rate (FDR), and the proportion of selecting the correct model (CM), based on 200 data replications, are reported .

$d = 1, n = 200, p = 20, q = 4, \rho = 0$						
	COR1	COR2	MS	TPR	FDR	CM
Initial estimate: Refined MAVE	0.9982 (0.0007)	0.9982 (0.0007)				
P-MAVE	0.9996 (0.0004)	0.9996 (0.0004)	4.2200	1.0000	0.0420	0.8050
One-step P-MAVE	0.9998 (0.0002)	0.9998 (0.0002)	4.0200	1.0000	0.0040	0.9800
Initial estimate: Refined OPG	0.9978 (0.0009)	0.9978 (0.0009)				
P-MAVE	0.9996 (0.0004)	0.9996 (0.0004)	4.2150	1.0000	0.0406	0.8150
One-step P-MAVE	0.9998 (0.0002)	0.9998 (0.0002)	4.0150	1.0000	0.0030	0.9850
CISE-SIR	0.9769 (0.0176)	0.9769 (0.0176)	3.7900	0.9187	0.0250	0.6050
CISE-PFC	0.9731 (0.0211)	0.9731 (0.0211)	3.6650	0.8912	0.0207	0.5050
$d = 1, n = 200, p = 20, q = 4, \rho = 0.5$						
Initial estimate: Refined MAVE	0.9979 (0.0010)	0.9979 (0.0010)				
P-MAVE	0.9996 (0.0004)	0.9996 (0.0004)	4.2900	1.0000	0.0545	0.7550
One-step P-MAVE	0.9998 (0.0002)	0.9998 (0.0002)	4.0500	1.0000	0.0097	0.9550
Initial estimate: Refined OPG	0.9972 (0.0014)	0.9972 (0.0014)				
P-MAVE	0.9996 (0.0005)	0.9996 (0.0005)	4.3200	1.0000	0.0588	0.7500
One-step P-MAVE	0.9998 (0.0002)	0.9998 (0.0002)	4.0400	1.0000	0.0077	0.9650
CISE-SIR	0.9599 (0.0482)	0.9599 (0.0482)	3.9350	0.8363	0.1203	0.2250
CISE-PFC	0.9401 (0.0614)	0.9401 (0.0614)	3.4900	0.7562	0.0960	0.1850

(available from the Machine Learning Repository at the University of California-Irvine). The objective of this analysis is to study the relationship between the car price and a set of car attributes. In the data set, there are originally 205 instances and 26 attributes, and there are also some missing values. We focus our analysis on 17 attributes: Wheel-base (X_1), Length (X_2), Width (X_3), Height (X_4), Curb-weight (X_5), Engine-size (X_6), Bore (X_7), Stroke (X_8), Compression-ratio (X_9), Horsepower (X_{10}), Peak-rpm (X_{11}), City-mpg (X_{12}), Highway-mpg (X_{13}), Fuel-type (X_{14}, $0 =$ diesel and $1 =$ gas), Aspiration (X_{15}, $0 =$ std and $1 =$ turbo), No-of-cylinders (X_{16}, $0 =$ four cylinders and 1 otherwise), and Drive-wheels (4wd, fwd and rwd). The first 13 attributes are continuous and the rest are categorical. Set $X_{17} = 0$ if a car has four-wheel drive and 2 otherwise, $X_{18} = 1$ if it has front-wheel drive and 2 otherwise. The response Y is the car price in thousands of dollars. Because the number of missing values is quite small compared to the total sample size, we simply discard the instances with missing values. The resulting data set then consists of 195 instances with complete records. For ease of explanation, all predictors, $X_1, \ldots, X_{18}$, are standardized separately. As to sufficient dimension reduction, this dataset, when the categorical predictors were removed, was analysed to identify the projection directions when the inverse regression method was applied. See Zhu and Zeng (2006). The categorical predictors were

Table 3. Simulation results for Example 3. The means and the standard deviations (in parentheses) of the vector correlation coefficient (COR1) and the trace correlation coefficient (COR2), the average model size (MS), the true positive rate (TPR), the false discovery rate (FDR), and the proportion of selecting the correct model (CM), based on 200 data replications, are reported.

$d=2, n=100, p=10, q=4, \rho=0$						
	COR1	COR2	MS	TPR	FDR	CM
Initial estimate: Refined MAVE	0.9776 (0.0098)	0.9888 (0.0049)				
P-MAVE	0.9919 (0.0077)	0.9959 (0.0038)	4.4100	1.0000	0.0732	0.7000
One-step P-MAVE	0.9933 (0.0078)	0.9967 (0.0039)	4.1200	0.9988	0.0240	0.8850
Initial estimate: Refined OPG	0.9708 (0.0136)	0.9854 (0.0068)				
P-MAVE	0.9924 (0.0064)	0.9962 (0.0032)	4.3350	1.0000	0.0626	0.7250
One-step P-MAVE	0.9941 (0.0050)	0.9971 (0.0025)	4.0850	1.0000	0.0170	0.9150
CISE-SIR	0.1985 (0.2589)	0.6834 (0.0820)	2.2850	0.4350	0.2454	0.0150
CISE-PFC	0.3310 (0.2996)	0.7289 (0.0968)	2.3950	0.5100	0.1485	0.0450
$d=2, n=100, p=10, q=4, \rho=0.5$						
Initial estimate: Refined MAVE	0.9676 (0.0155)	0.9838 (0.0077)				
P-MAVE	0.9897 (0.0140)	0.9949 (0.0068)	4.1850	0.9975	0.0365	0.8300
One-step P-MAVE	0.9911 (0.0102)	0.9956 (0.0050)	4.0600	0.9988	0.0130	0.9300
Initial estimate: Refined OPG	0.9460 (0.0408)	0.9731 (0.0197)				
P-MAVE	0.9846 (0.0267)	0.9924 (0.0129)	4.0550	0.9850	0.0227	0.8300
One-step P-MAVE	0.9905 (0.0118)	0.9952 (0.0058)	4.0450	0.9975	0.0110	0.9350
CISE-SIR	0.1916 (0.2419)	0.6707 (0.1016)	2.2700	0.4487	0.2096	0.0000
CISE-PFC	0.2637 (0.2710)	0.6879 (0.1198)	2.1900	0.4575	0.1675	0.0000

removed because inverse regression methods are limited in this regard whereas, MAVE is not. To estimate the dimension, both the cross-validation method of Xia et al. (2002) and the modified BIC-type criterion proposed by Wang and Yin (2008) suggest taking $d = 1$. The results of estimation and variable selection are summarized in Table 6.

The vector correlation coefficient, the trace correlation coefficient between the MAVE estimator and the P-MAVE estimator, is 0.9865; between the MAVE estimator and the one-step P-MAVE estimator, is 0.9762. In addition, the P-MAVE estimator and the one-step P-MAVE have similar performance in terms of predictor selection: the relevant predictors in common are Wheel-base (X_1), Length (X_2), Width (X_3), Engine-size (X_6), Stroke (X_8), Compression-ratio (X_9), Peak-rpm (X_{11}), City-mpg (X_{12}), Highway-mpg (X_{13}), and Drive-wheels (X_{18}). The plot of Y against the reduced direction $X^T\hat{\beta}$ is shown in Figure 1, displaying a strong non-linear relationship.

As seen in Figure 1, the significance of X_{18} indicates the difference between front-wheel drive cars and others. It is common knowledge that four-wheel and rear-wheel drive cars are typically more expensive to purchase than comparable front wheel drive cars.

Table 4. Simulation results for Example 4. The means and the standard deviations (in parentheses) of the vector correlation coefficient (COR1) and the trace correlation coefficient (COR2), the average model size (MS), the true positive rate (TPR), the false discovery rate (FDR), and the proportion of selecting the correct model (CM), based on 200 data replications, are reported.

$d = 3, n = 100, p = 10, , q = 3, \rho = 0$						
	COR1	COR2	MS	TPR	FDR	CM
Initial estimate: Refined MAVE	0.8207 (0.2458)	0.9476 (0.0646)				
P-MAVE	0.9546 (0.1872)	0.9927 (0.0382)	3.2900	0.9900	0.0674	0.7400
One-step P-MAVE	0.9864 (0.0997)	0.9968 (0.0211)	3.1200	0.9983	0.0255	0.9150
Initial estimate: Refined OPG	0.7769 (0.2446)	0.9341 (0.0622)				
P-MAVE	0.9193 (0.2597)	0.9893 (0.0395)	3.2250	0.9817	0.0625	0.7400
One-step P-MAVE	0.9820 (0.1262)	0.9961 (0.0265)	3.0600	0.9983	0.0149	0.9500
CISE-SIR	0.1711 (0.3594)	0.7853 (0.1362)	3.2900	0.7300	0.3317	0.1550
CISE-PFC	0.3756 (0.4585)	0.8608 (0.1202)	3.3500	0.8683	0.2150	0.3400
$d = 3, n = 100, p = 10, q = 3, \rho = 0.5$						
Initial estimate: Refined MAVE	0.6605 (0.2883)	0.9040 (0.0721)				
P-MAVE	0.8758 (0.3023)	0.9763 (0.0607)	3.3650	0.9733	0.0826	0.6750
One-step P-MAVE	0.9641 (0.1683)	0.9919 (0.0368)	3.1150	0.9933	0.0314	0.8950
Initial estimate: Refined OPG	0.5762 (0.2800)	0.8744 (0.0769)				
P-MAVE	0.8241 (0.3530)	0.9778 (0.0521)	3.3500	0.9550	0.0951	0.6100
One-step P-MAVE	0.9396 (0.2314)	0.9900 (0.0409)	3.0750	0.9833	0.0306	0.8900
CISE-SIR	0.0216 (0.1176)	0.7091 (0.1222)	3.2550	0.5933	0.4596	0.0100
CISE-PFC	0.0209 (0.1052)	0.7382 (0.1147)	3.2950	0.6483	0.4142	0.0100

Table 5. Simulation results for Example 5. The means and the standard deviations (in parentheses) of the vector correlation coefficient (COR1) and the trace correlation coefficient (COR2), the average model size (MS), the true positive rate (TPR), the false discovery rate (FDR), and the proportion of selecting the correct model (CM), based on 200 data replications, are reported.

$d = 2, n = 100, p = 6, q = 2$						
	COR1	COR2	MS	TPR	FDR	CM
Initial estimate: Refined MAVE	0.5771 (0.2414)	0.8277 (0.0805)				
P-MAVE	0.9365 (0.1841)	0.9749 (0.0680)	2.2800	0.9975	0.0795	0.8050
One-step P-MAVE	0.9683 (0.1532)	0.9889 (0.0499)	2.1000	0.9950	0.0350	0.9100
$d = 2, n = 200, p = 6, q = 2$						
Initial estimate: Refined MAVE	0.7490 (0.2079)	0.8920 (0.0739)				
P-MAVE	0.9556 (0.1458)	0.9820 (0.0526)	2.2650	1.0000	0.0772	0.8000
One-step P-MAVE	0.9920 (0.0676)	0.9969 (0.0226)	2.0650	1.0000	0.0217	0.9350

5. Concluding Remarks

In this paper, we suggest a penalized minimum average variance estimation for both dimension reduction and variable selection. This is a forward regression approach that does not depend on the eigen-decomposition of a kernel matrix. As such this method, compared with that of Chen, Zou, and Cook (2010), requires

Table 6. Estimation and variable selection for the automobile data.

Predictor	Refined MAVE	P-MAVE	One-step P-MAVE
X_1	0.1435	0.1097	0.1076
X_2	-0.2419	-0.2445	-0.2013
X_3	-0.0850	-0.0984	-0.0848
X_4	-0.0419	0	0
X_5	0.1303	0.1277	0
X_6	-0.2988	-0.2333	-0.2119
X_7	0.0134	0	0
X_8	0.0325	0.0374	0.0460
X_9	-0.2279	-0.2325	-0.2243
X_{10}	0.0644	-0.0150	0
X_{11}	-0.0548	-0.0415	-0.0644
X_{12}	0.7558	0.8079	0.8184
X_{13}	-0.3770	-0.3454	-0.3921
X_{14}	0.0679	0	0
X_{15}	-0.0608	0	0
X_{16}	0.0158	0	0
X_{17}	0.0192	0	0
X_{18}	0.1518	0.1356	0.1256

fewer conditions on the predictors and the regression function. However, as with other methods, P-MAVE seems to have difficulty handling the cases with very large p; it can be extended to the case where p diverges to infinity at certain rate as the sample size tends to infinity. As P-MAVE requires an initial nonparametric estimator with all predictors, dimension reduction and variable selection can be less accurate when p is very large. We have not yet found a good solution to this problem, it deserves further study.

Acknowledgement

The research described here was supported by a grant from the Research Council of Hong Kong, and a grant from Hong Kong Baptist University, Hong Kong. The authors thank the editor, associate editors and two referees for their constructive comments and suggestions which led to improvement of an early manuscript.

Appendix

Regularity Conditions. Here are regularity conditions needed for establishing the asymptotic properties of our estimator.

(C1) The predictor vector X is bounded and its density function $f(\cdot)$ has a bounded second order derivative; $E(X|B^TX=u)$ and $E(XX^T|B^TX=u)$

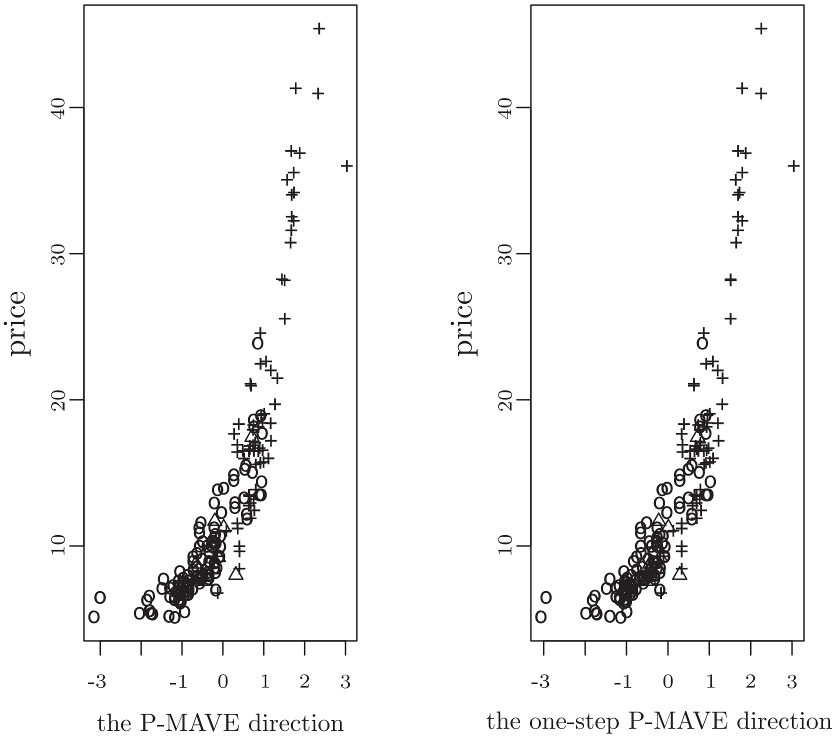

Figure 1. The automobile data. The cars are represented by "o" for front-wheel drive cars, "△" for four-wheel drive cars, and "+" for rear-wheel drive cars.

have bounded derivatives with respect to u and B in a small neighborhood of B_0.

(C2) $E|Y|^k < \infty$ for some large $k > 0$.

(C3) $E(Y|B^TX)$ has a bounded fourth order derivative in a neighborhood of B_0.

(C4) $K(\cdot)$ is a symmetric density function with bounded first order derivative. In particular, we use the Gaussian kernel with bandwidth $h \propto n^{-1/(d+4)}$.

(C5) The smallest eigenvalue of $J_0^T W_{g0} J_0$ is larger than ρ and the largest eigenvalue of W_{g0} is less than ρ^* for some positive constants ρ and ρ^*. Matrix J_0 in (C5) can be found in the proof of Theorem 1.

Conditions (C1)$-$(C4) are similar to those of Xia (2006) and Wang and Xia (2008). As for condition (C5), consider the special case of the single-index model, so W_{g0} is a square matrix of order p and $J_0^T W_{g0} J_0$ is a matrix of order $p-1$. It follows that $\beta_0^T W_{g0} = 0$ and $\beta_0^T J_0 = 0$. Note that $J_0^T W_{g0} J_0$ has the same rank as $(J_0, \beta_0)^T W_{g0} (J_0, \beta_0)$, and $(J_0, \beta_0)^T W_{g0} (J_0, \beta_0)$ has the same rank as W_{g0}, so

$J_0^T W_{g0} J_0$ has full rank $p-1$. As a result, condition (C5) is mild.

Proof of Proposition 1. This follows directly from Proposition 1 in Huang et al. (2009).

Proof of Theorem 1. There are three steps: Step I establishes the order of the minimizer $\hat\beta$ of $\Psi_{\lambda_n}(\beta;\tilde\beta)$, Step II shows that $\hat\beta$ is variable selection consistent, and Step III derives the asymptotic distribution. Following Theorem 1 and Theorem 2 of Huang et al. (2009), write

$$\Psi_{\lambda_n}(\beta;\tilde\beta) = L_n(\beta;\tilde\beta) + \lambda_n \sum_{k=1}^{p} \|\beta_k\|_1^{\gamma},$$

where

$$L_n(\beta;\tilde\beta) = \frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde y_{ij} - \tilde{\mathrm{x}}_{ij}^T\beta)^2 \tilde w_{ij}.$$

Step I. We prove that there exists a local minimizer $\hat\beta$ of $\Psi_{\lambda_n}(\beta;\tilde\beta)$ such that

$$\|\hat\beta - \beta_0\|_2 = O_P\left(\frac{1}{\sqrt{n}}\right). \tag{A.1}$$

Because $B_0^T B_0 = I_d$, the number of free parameters in B_0 equals $pd - d(d+1)/2$. In addition, there are $d(d-1)/2$ free parameters in D_0. For example, we may parameterize D_0 via the polar coordinate system. Therefore, the total number of free parameters in $B_0 D_0$ is $pd - d^2$.

For ease of presentation, we first consider the single-index model. We can simply take $D_0 = 1$. It follows that $B_0 = \beta_0 = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$, but note that $\|\beta_0\|_2 = 1$, so $g(\beta_0^T X)$ does not have a derivative at the point β_0. To address this problem, we adopt the "delete-one-component" method as follows. For any vector $\mathrm{a} = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$, let $\mathrm{a}^{(1)} = (a_2, \ldots, a_p)^T$. Then we can write $\beta_1 = \beta_1(\beta^{(1)}) = (1 - \|\beta^{(1)}\|_2^2)^{1/2}$ and $\beta = \beta(\beta^{(1)}) = (\beta_1(\beta^{(1)}), \beta^{(1)T})^T$. To exclude the trivial cases, we assume $q \geq 2$. Then the true parameter $\beta_0^{(1)}$ satisfies the constraint $\|\beta_0^{(1)}\|_2 < 1$, and β is infinitely differentiable in a neighborhood of $\beta_0^{(1)}$. We define the Jacobian matrix as $J_{\beta_0^{(1)}} = (\eta_1, \ldots, \eta_p)^T \in \mathbb{R}^{p\times(p-1)}$, where $\eta_1 = -(1 - \|\beta_0^{(1)}\|_2^2)^{-1/2}\beta_0^{(1)}$ and, for $s = 2, \ldots, p$, η_s is a $(p-1)$-dimensional unit vector with sth component being 1.

When $d > 1$, a Jacobian matrix $J_0 = J_{\beta_0^{(1)}} \in \mathbb{R}^{pd\times(pd-d^2)}$ can be defined in a similar way, where the vector $\beta_0^{(1)}$ consists of all the free parameters in $B_0 D_0$.

Let $\beta^{*(1)} = \beta_0^{(1)} + n^{-1/2}\mathrm{v}$, where $\|\mathrm{v}\|_2 = C$ for some positive constant C.

Applying a Taylor expansion, we can write

$$\begin{aligned} L_n(\beta^*;\tilde\beta)-L_n(\beta_0;\tilde\beta) &= -\sqrt{n}\mathrm{v}^T J^T_{\beta_0^{*(1)}}\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n(\tilde y_{ij}-\tilde{\mathrm{x}}_{ij}^T\beta_0)\tilde{\mathrm{x}}_{ij}\tilde w_{ij} \\ &\quad + \frac{1}{2}\mathrm{v}^T J^T_{\beta_0^{*(1)}}\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n \tilde{\mathrm{x}}_{ij}\tilde{\mathrm{x}}_{ij}^T\tilde w_{ij}J_{\beta_0^{*(1)}}\mathrm{v} \\ &\equiv I_{1n}+I_{2n}, \end{aligned}$$

where $\beta_0^{*(1)}$ lies between $\beta_0^{(1)}$ and $\beta^{*(1)}$. By Lemmas 6.4 and 6.5 in Wang and Xia (2008),

$$\begin{aligned} I_{1n} &= -\sqrt{n}\mathrm{v}^T J^T_{\beta_0^{(1)}}\Big[W_{g0}(\tilde\beta-\beta_0)+\frac{1}{n}\sum_{i=1}^n\{\nu_{B_0}(\mathrm{x}_i)\otimes\nabla g(B_0^T\mathrm{x}_i)\}\epsilon_i\Big]+o_P\left(\frac{1}{\sqrt{n}}\right), \\ I_{2n} &= \mathrm{v}^T J^T_{\beta_0^{(1)}}\{W_{g0}+o_P(1)\}J_{\beta_0^{(1)}}\mathrm{v}. \end{aligned}$$

It follows that

$$\begin{aligned} &\Psi_{\lambda_n}(\beta^*;\tilde\beta)-\Psi_{\lambda_n}(\beta_0;\tilde\beta) \\ &\quad= -\sqrt{n}\mathrm{v}^T J^T_{\beta_0^{(1)}}\Big[W_{g0}(\tilde\beta-\beta_0)+\frac{1}{n}\sum_{i=1}^n\{\nu_{B_0}(\mathrm{x}_i)\otimes\nabla g(B_0^T\mathrm{x}_i)\}\epsilon_i\Big] \\ &\qquad + \mathrm{v}^T J^T_{\beta_0^{(1)}}W_{g0}J_{\beta_0^{(1)}}\mathrm{v}+\lambda_n\sum_{k=1}^p(\|\beta_k^*\|_1^\gamma-\|\beta_{0k}\|_1^\gamma)+o_P\left(\frac{1}{\sqrt{n}}\right) \\ &\quad\equiv T_{1n}+T_{2n}+T_{3n}+o_P\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

By (C5), we have $T_{2n}\ge\rho\|\mathrm{v}\|_2^2$ with large probability. Further, by (A),

$$J^T_{\beta_0^{(1)}}W_{g0}\sqrt{n}(\tilde\beta-\beta_0)+\frac{1}{\sqrt{n}}\sum_{i=1}^n J^T_{\beta_0^{(1)}}\{\nu_{B_0}(\mathrm{x}_i)\otimes\nabla g(B_0^T\mathrm{x}_i)\}\epsilon_i = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Thus, by choosing a sufficiently large C, T_{1n} is dominated by T_{2n}.

Since $b^\gamma-a^\gamma\le 2(b-a)b^{\gamma-1}$ for $0\le a\le b$, by the Cauchy-Schwarz inequality,

$$\begin{aligned} T_{3n} &\le 2\lambda_n\sum_{k=1}^q\|\beta_{0k}\|_1^{\gamma-1}\|\beta_k^*-\beta_{0k}\|_1 \\ &\le 2\sqrt{d}\lambda_n\sum_{k=1}^q\|\beta_{0k}\|_1^{\gamma-1}\|\beta_k^*-\beta_{0k}\|_2 \\ &\le 2\sqrt{d}\sum_{k=1}^q\|\beta_{0k}\|_1^{\gamma-1}\lambda_n\Big(\sum_{k=1}^q\|\beta_k^*-\beta_{0k}\|_2^2\Big)^{1/2} \end{aligned}$$

$$= O(1)\lambda_n \frac{\|\mathrm{v}\|_2}{\sqrt{n}} = O(\|\mathrm{v}\|_2),$$

where the last equality is from $\lambda_n = O(n^{1/2})$. By choosing a sufficiently large C, T_{3n} is also dominated by T_{2n}. Consequently, it holds with large probability that $\Psi_{\lambda_n}(\beta^*;\tilde{\beta}) > \Psi_{\lambda_n}(\beta_0;\tilde{\beta})$ uniformly in $\{\mathrm{v} : \|\mathrm{v}\|_2 = C\}$. This implies that, with large probability, there exists a local minimizer of of $\Psi_{\lambda_n}(\beta;\tilde{\beta})$ in the ball $\{\mathrm{v} : \|\mathrm{v}\|_2 \leq C\}$. Therefore, there exists a local minimizer $\hat{\beta}$ of $\Psi_{\lambda_n}(\beta;\tilde{\beta})$ such that $\|\hat{\beta} - \beta_0\|_2 = O_P(n^{-1/2})$.

Step II. In this step, we establish the consistency of variable selection. Let $I(\cdot)$ denote the indicator function. Define $\bar{\beta}_k = \hat{\beta}_k I(k \in A_1)$ for $k = 1, \ldots, p$. Write $\bar{\beta} = (\bar{\beta}_1^T, \ldots, \bar{\beta}_p^T)^T$. By Proposition 1 and the Karush-Kuhn-Tucker condition, we have

$$-n\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\hat{\beta})\tilde{\mathrm{x}}_{ijk}^l\tilde{w}_{ij} = \hat{\theta}^{1-1/\gamma}\mathrm{sgn}(\hat{\beta}_k^l),\ \hat{\beta}_k^l \neq 0, \tag{A.2}$$

where $\tilde{\mathrm{x}}_{ijk}^l$ is the lth component of $\tilde{\mathrm{x}}_{ijk}$. Because $\hat{\theta}^{1-1/\gamma}\|\hat{\beta}_k\|_1 = \gamma\lambda_n\|\hat{\beta}_k\|_1^\gamma$, we obtain

$$-n\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\hat{\beta})\tilde{\mathrm{x}}_{ijk}^l\tilde{w}_{ij} = \gamma\lambda_n\|\hat{\beta}_k\|_1^{\gamma-1}\mathrm{sgn}(\hat{\beta}_k^l),\ \hat{\beta}_k^l \neq 0. \tag{A.3}$$

It follows that

$$\begin{aligned}
&-n\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\hat{\beta})\tilde{w}_{ij}\tilde{\mathrm{x}}_{ij}^T(\hat{\beta} - \bar{\beta})\\
&= -n\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\hat{\beta})\tilde{w}_{ij}\sum_{k,l:\hat{\beta}_k^l\neq 0}\tilde{\mathrm{x}}_{ijk}^l(\hat{\beta}_k^l - \bar{\beta}_k^l)\\
&= \sum_{k,l:\hat{\beta}_k^l\neq 0}\gamma\lambda_n\|\hat{\beta}_k\|_1^{\gamma-1}\mathrm{sgn}(\hat{\beta}_k^l)(\hat{\beta}_k^l - \bar{\beta}_k^l)\\
&= \sum_{k,l}\gamma\lambda_n\|\hat{\beta}_k\|_1^{\gamma-1}\mathrm{sgn}(\hat{\beta}_k^l)(\hat{\beta}_k^l - \bar{\beta}_k^l).
\end{aligned}$$

Note that from $(\hat{\beta}_k^l - \bar{\beta}_k^l)\mathrm{sgn}(\hat{\beta}_k^l) = |\hat{\beta}_k^l|I(k \in A_2)$, we obtain

$$\begin{aligned}
-n\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\hat{\beta})\tilde{w}_{ij}\tilde{\mathrm{x}}_{ij}^T(\hat{\beta} - \bar{\beta}) &= \sum_{k\in A_2}\gamma\lambda_n\|\hat{\beta}_k\|_1^{\gamma-1}\sum_{l=1}^{d}|\hat{\beta}_k^l|\\
&= \sum_{k=1}^{p}\gamma\lambda_n\|\hat{\beta}_k\|_1^{\gamma-1}(\|\hat{\beta}_k\|_1 - \|\bar{\beta}_k\|_1).
\end{aligned}$$

Since $\gamma b^{(\gamma-1)}(b-a) \le b^\gamma - a^\gamma$ for $0 \le a \le b$, we have

$$\gamma \|\hat\beta_k\|_1^{\gamma-1}(\|\hat\beta_k\|_1 - \|\bar\beta_k\|_1) \le \|\hat\beta_k\|_1^\gamma - \|\bar\beta_k\|_1^\gamma.$$

And then

$$\Big| n\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n(\tilde y_{ij} - \tilde{\mathrm{x}}_{ij}^T\hat\beta)\tilde w_{ij}\tilde{\mathrm{x}}_{ij}^T(\hat\beta - \bar\beta)\Big| \le \lambda_n\sum_{k=1}^q(\|\hat\beta_k\|_1^\gamma - \|\bar\beta_k\|_1^\gamma) + \gamma\lambda_n\sum_{k=q+1}^p\|\hat\beta_k\|_1^\gamma.$$

By the definition of $\hat\beta$, $\Psi_{\lambda_n}(\hat\beta;\tilde\beta) \le \Psi_{\lambda_n}(\bar\beta;\tilde\beta)$. It follows that

$$\begin{aligned}
&\Big| n\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n(\tilde y_{ij} - \tilde{\mathrm{x}}_{ij}^T\hat\beta)\tilde w_{ij}\tilde{\mathrm{x}}_{ij}^T(\hat\beta - \bar\beta)\Big| + (1-\gamma)\lambda_n\sum_{k=q+1}^p\|\hat\beta_k\|_1^\gamma\\
&\le \lambda_n\sum_{k=1}^p\|\hat\beta_k\|_1^\gamma - \lambda_n\sum_{k=1}^p\|\bar\beta_k\|_1^\gamma\\
&\le L_n(\bar\beta;\tilde\beta) - L_n(\hat\beta;\tilde\beta)\\
&= n\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n(\tilde y_{ij} - \tilde{\mathrm{x}}_{ij}^T\hat\beta)\tilde w_{ij}\tilde{\mathrm{x}}_{ij}^T(\hat\beta - \bar\beta)\\
&\quad + \frac{1}{2}n(\hat\beta - \bar\beta)^T\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n\tilde{\mathrm{x}}_{ij}\tilde{\mathrm{x}}_{ij}^T\tilde w_{ij}(\hat\beta - \bar\beta).
\end{aligned}$$

Then, with (C5) and Lemma 4 of Wang and Xia (2008), we have, in probability,

$$\begin{aligned}
(1-\gamma)\lambda_n\sum_{k=q+1}^p\|\hat\beta_k\|_1^\gamma &\le \frac{1}{2}n(\hat\beta - \bar\beta)^T\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n\tilde{\mathrm{x}}_{ij}\tilde{\mathrm{x}}_{ij}^T\tilde w_{ij}(\hat\beta - \bar\beta)\\
&\le n\rho^*\|\hat\beta - \bar\beta\|_2^2\\
&= n\rho^*\sum_{k=q+1}^p\|\hat\beta_k\|_2^2\\
&\le n\rho^*\|\hat\beta - \beta_0\|_2^2.
\end{aligned}$$

Using the result of Step I,

$$(1-\gamma)\lambda_n\sum_{k=q+1}^p\|\hat\beta_k\|_1^\gamma \le n\rho^*\sum_{k=q+1}^p\|\hat\beta_k\|_2^2 = O_P(1). \tag{A.4}$$

On the other hand,

$$\sum_{k=q+1}^p\|\hat\beta_k\|_1^\gamma \ge \Big(\sum_{k=q+1}^p\|\hat\beta_k\|_1\Big)^\gamma \ge \Big(\sum_{k=q+1}^p\|\hat\beta_k\|_2^2\Big)^{\gamma/2}. \tag{A.5}$$

By (A.4) and (A.5), if $\sum_{k=q+1}^{p} \|\hat{\beta}_k\|_2^2 > 0$ then

$$(1-\gamma)\lambda_n \leq n\rho^* \Big(\sum_{k=q+1}^{p} \|\hat{\beta}_k\|_2^2 \Big)^{1-\gamma/2} = n\rho^* O_P(1)(n\rho^*)^{-1+\gamma/2} = O_P(n^{\gamma/2}).$$

But, since $\lambda_n n^{-\gamma/2} \to \infty$, we obtain

$$P\Big(\sum_{k=q+1}^{p} \|\hat{\beta}_k\|_2^2 > 0 \Big) \to 0 \text{ as } n \to \infty.$$

Step III. It remains to derive the asymptotic distribution. Let $\mathbf{0}_m \in \mathbb{R}^m$ be a m-vector of zeros. Let $\mathrm{u} = (\mathrm{u}_1^T, \ldots, \mathrm{u}_q^T)^T$, where $\mathrm{u}_k = (u_k^1, \ldots, u_k^d)^T \in \mathbb{R}^d$ for $k = 1, \ldots, q$. Take

$$V_n(\mathrm{u}) = \Psi_{\lambda_n}(\beta_0 + n^{-1/2}(\mathrm{u}^T, \mathbf{0}_{(p-q)\times d}^T)^T; \tilde{\beta}) - \Psi_{\lambda_n}(\beta_0; \tilde{\beta}). \tag{A.6}$$

We have shown that, with large probability, $\hat{\beta} - \beta_0 = n^{-1/2}(\hat{\mathrm{u}}^T, \mathbf{0}_{(p-q)\times d}^T)^T$, where $\hat{\mathrm{u}}$ is a minimizer of $V_n(\mathrm{u})$. We may re-express $V_n(\mathrm{u})$ as

$$\begin{aligned}
V_n(\mathrm{u}) = & -\sqrt{n}(\mathrm{u}^T, \mathbf{0}_{(p-q)\times d}^T)\Big[W_{g0}(\tilde{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^{n}\{\nu_{B_0}(\mathrm{x}_i) \otimes \nabla g(B_0^T \mathrm{x}_i)\}\epsilon_i \Big] \\
& + (\mathrm{u}^T, \mathbf{0}_{(p-q)\times d}^T) W_{g0} (\mathrm{u}^T, \mathbf{0}_{(p-q)\times d}^T)^T \\
& + \lambda_n \sum_{k=1}^{q} (\|\beta_{0k} + n^{-1/2}\mathrm{u}_k\|_1^\gamma - \|\beta_{0k}\|_1^\gamma) + o_P(1) \\
= & -\mathrm{u}^T U_{g0}\sqrt{n}(\tilde{\beta} - \beta_0) - \mathrm{u}^T \frac{1}{\sqrt{n}} \sum_{i=1}^{n}\{\nu_{B_0,A_1}(\mathrm{x}_i) \otimes \nabla g(B_0^T \mathrm{x}_i)\}\epsilon_i \\
& + \mathrm{u}^T W_{g0,A_1}\mathrm{u} + \ \lambda_n \sum_{k=1}^{q} (\|\beta_{0k} + n^{-1/2}\mathrm{u}_k\|_1^\gamma - \|\beta_{0k}\|_1^\gamma) + o_P(1) \\
\equiv & \ V_{1n} + \lambda_n \sum_{k=1}^{q} (\|\beta_{0k} + n^{-1/2}\mathrm{u}_k\|_1^\gamma - \|\beta_{0k}\|_1^\gamma) + o_P(1) \\
\equiv & \ V_{1n} + V_{2n} + o_P(1).
\end{aligned}$$

We have

$$\begin{aligned}
V_{1n} & \to_L -\mathrm{u}^T(U_{g0,A_1}G_1 + G_2) + \mathrm{u}^T W_{g0,A_1}\mathrm{u}, \\
V_{2n} & \ \to \ \lambda_0 \gamma \sum_{k=1}^{q} \|\beta_{0k}\|_1^{\gamma-1} \sum_{l=1}^{d} \{u_k^l \mathrm{sign}(\beta_{0k}^l) I(\beta_{0k}^l \neq 0) + |u_k^l| I(\beta_{0k}^l = 0)\}.
\end{aligned}$$

Therefore,

$$\begin{aligned} V_n(\mathrm{u}) \rightarrow_L & -\mathrm{u}^T(U_{g0,A_1}G_1 + G_2) + \mathrm{u}^T W_{g0,A_1}\mathrm{u} \\ & + \lambda_0\gamma \sum_{k=1}^{q} \|\beta_{0k}\|_1^{\gamma-1} \sum_{l=1}^{d} \{u_k^l \mathrm{sign}(\beta_{0k}^l) I(\beta_{0k}^l \neq 0) + |u_k^l| I(\beta_{0k}^l = 0)\} \\ \equiv & \ V(\mathrm{u}). \end{aligned}$$

Because $\hat{\mathrm{u}} = O_P(1)$, by the argmin continuous mapping theorem of Kim and Pollard (1990),

$$\sqrt{n}(\hat{\beta}_{A_1} - \beta_{0A_1}) = \hat{\mathrm{u}} \rightarrow_L \underset{\mathrm{u}}{\mathrm{argmin}}\, V(\mathrm{u}).$$

The proof is complete.

Proof of Corollary 1. The corollary follows from observing that, by Theorem 4.2 in Xia (2006),

$$W_{g0}(\tilde{\beta} - \beta_0) = \frac{1}{n}\sum_{i=1}^{n} g'(\beta_0^T \mathrm{x}_i)\nu_{\beta_0}(\mathrm{x}_i)\epsilon_i + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof of Theorem 2. A candidate model M is said to be underfitted if it misses at least one important predictor, $M \not\supset M_T$; it is overfitted if it covers all important predictors but also contains at least one irrelevant predictor, $M \supset M_T$ but $M \neq M_T$. According to whether the model M_λ is underfitted, correctly fitted, or overfitted, we set

$$\mathbb{R}_- = \{\lambda : M_\lambda \not\supset M_T\}, \mathbb{R}_0 = \{\lambda : M_\lambda = M_T\} \text{ and } \mathbb{R}_+ = \{M_\lambda \supset M_T, M_\lambda \neq M_T\}.$$

Let λ_n be a reference tuning parameter sequence such that

$$P(M_{\lambda_n} = M_T) \rightarrow 1 \text{ as } n \rightarrow \infty. \tag{A.7}$$

To prove the theorem, it suffices to show that

$$P\left(\inf_{\lambda \in \mathbb{R}_- \cup \mathbb{R}_+} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}\right) \rightarrow 1 \text{ as } n \rightarrow \infty. \tag{A.8}$$

Clearly, we need only consider the single index model. For each M, write

$$\mathrm{RSS}_M = \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\check{\beta}_M)^2\tilde{w}_{ij},$$

where $\check{\beta}_M$ is an unpenalized estimator such that

$$\check{\beta}_M = \underset{\beta_j = 0 : \forall j \notin M}{\mathrm{argmin}} \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\beta)^2\tilde{w}_{ij}.$$

We consider two cases.

Case 1: Underfitted model. For any λ, we have $\mathrm{RSS}_\lambda \geq \mathrm{RSS}_{M_\lambda}$ and

$$\begin{aligned}\mathrm{RSS}_{M_\lambda} &= \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\check{\beta}_{M_F})^2\tilde{w}_{ij} \\ &\quad + (\check{\beta}_{M_F} - \check{\beta}_{M_\lambda})^T\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - \tilde{\mathrm{x}}_{ij}^T\check{\beta}_{M_F})\tilde{\mathrm{x}}_{ij}\tilde{w}_{ij} \\ &\quad + (\check{\beta}_{M_\lambda} - \check{\beta}_{M_F})^T\frac{1}{2n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\tilde{\mathrm{x}}_{ij}\tilde{\mathrm{x}}_{ij}^T\tilde{w}_{ij}(\check{\beta}_{M_\lambda} - \check{\beta}_{M_F}) \\ &\equiv \mathrm{RSS}_{M_F} + R_{1\lambda} + R_{2\lambda}.\end{aligned}$$

Applying techniques similar to those used in Lemma 1 of Xia et al. (2002), we obtain

$$\mathrm{RSS}_{M_F} = \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\epsilon_i^2 w_{0ij} + o_p(1) = O_P(1). \tag{A.9}$$

Let $\Pi = (\beta_0, J_{\beta_0^{(1)}})$. Because Π is full rank, there exists a vector $\phi \in \mathbb{R}^p$ such that

$$\check{\beta}_{M_\lambda} = \Pi\phi = \phi_1\beta_0 + J_{\beta_0^{(1)}}\phi^{(1)}.$$

Note that $\|\beta_0\|_2 = 1, \|\check{\beta}_{M_\lambda}\|_2 = 1$, and $\beta_0^T J_{\beta_0^{(1)}} = 0$, we have $\phi^{(1)T}J_{\beta_0^{(1)}}^T J_{\beta_0^{(1)}}\phi^{(1)} = 1 - \phi_1^2$ and

$$\|\check{\beta}_{M_\lambda} - \beta_0\|_2^2 = (\phi_1 - 1)^2 + \phi^{(1)T}J_{\beta_0^{(1)}}^T J_{\beta_0^{(1)}}\phi^{(1)} = 2(1 - \phi_1).$$

Since $\lambda \in \mathbb{R}_-$, assume without loss of generality that $\check{\beta}_{M_\lambda,2} = 0$. Then,

$$\begin{aligned}1 - \phi_1 &= \frac{1}{2}\|\check{\beta}_{M_\lambda} - \beta_0\|_2^2 \geq \frac{1}{2}\|\beta_{02}\|_2^2, \\ 1 + \phi_1 &= 1 + \check{\beta}_{M_\lambda}^T\beta_0 \geq 1 - \frac{1}{2}(1 + 1 - \|\beta_{02}\|_2^2) = \frac{1}{2}\|\beta_{02}\|_2^2.\end{aligned}$$

It follows that $\phi^{(1)T}J_{\beta_0^{(1)}}^T J_{\beta_0^{(1)}}\phi^{(1)} \geq \frac{1}{4}\|\beta_{02}\|_2^4$. Thus we have for some positive constant, $\|\phi^{(1)}\|_2^2 \geq \frac{1}{4}\|\beta_{02}\|_2^4\|\beta_{01}\|_2^2 > c$. By (C5), we have, with large probability,

$$(\check{\beta}_{M_\lambda} - \beta_0)^T\frac{1}{2n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\tilde{\mathrm{x}}_{ij}\tilde{\mathrm{x}}_{ij}^T\tilde{w}_{ij}(\check{\beta}_{M_\lambda} - \beta_0) \geq c\rho. \tag{A.10}$$

Further, we can show that

$$(\breve{\beta}_{M_F}-\beta_0)^T \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\tilde{\mathrm{x}}_{ij}\tilde{\mathrm{x}}_{ij}^T\tilde{w}_{ij}(\breve{\beta}_{M_F}-\beta_0) = O_P\left(\frac{1}{n}\right). \tag{A.11}$$

By (A.9), (A.10), and (A.11), we have $R_{1\lambda}+R_{2\lambda}\geq c\rho$ in probability. Therefore, with large probability,

$$\mathrm{RSS}_\lambda \geq \mathrm{RSS}_{M_\lambda} \geq \mathrm{RSS}_{M_F} + c\rho. \tag{A.12}$$

Note that the above results hold uniformly over all $\lambda\in\mathbb{R}_-$.

As for the reference tuning parameter λ_n, a similar decomposition yields that

$$\mathrm{RSS}_{\lambda_n} = \mathrm{RSS}_{M_F} + O_P\left(\frac{1}{n}\right). \tag{A.13}$$

This, together with (A.12) and the definition of BIC_λ, implies that

$$P\left(\inf_{\lambda\in\mathbb{R}_-}\mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}\right)\to 1 \text{ as } n\to\infty. \tag{A.14}$$

Case 2*: Overfitted model.* Consider an arbitrary $\lambda\in\mathbb{R}_+$. By definition, $\mathrm{RSS}_\lambda \geq \mathrm{RSS}_{M_\lambda}$. Further, by (C5), we can show that

$$\mathrm{RSS}_{M_\lambda} - \mathrm{RSS}_{M_F} = O_P\Big(\frac{1}{n}\Big).$$

It follows that

$$\begin{aligned}\log\mathrm{RSS}_\lambda - \log\mathrm{RSS}_{M_F} &\geq \log\mathrm{RSS}_{M_\lambda} - \log\mathrm{RSS}_{M_F} = \log\left(1+\frac{\mathrm{RSS}_{M_\lambda}-\mathrm{RSS}_{M_F}}{\mathrm{RSS}_{M_F}}\right)\\ &= \frac{\mathrm{RSS}_{M_\lambda}-\mathrm{RSS}_{M_F}}{\mathrm{RSS}_{M_F}}\{1+o_P(1)\} = O_P\Big(\frac{1}{n}\Big).\end{aligned}$$

Similarly, we obtain

$$\log\mathrm{RSS}_{\lambda_n} - \log\mathrm{RSS}_{M_F} = O_P\Big(\frac{1}{n}\Big).$$

Then we have, with large probability,

$$\begin{aligned}\inf_{\lambda\in\mathbb{R}_+}\mathrm{BIC}_\lambda - \mathrm{BIC}_{\lambda_n} &= \inf_{\lambda\in\mathbb{R}_+}\log\mathrm{RSS}_\lambda - \log\mathrm{RSS}_{\lambda_n} + (df_\lambda - df_{\lambda_n})\frac{\log n}{n}\\ &\geq \inf_{\lambda\in\mathbb{R}_+}\log\mathrm{RSS}_\lambda - \log\mathrm{RSS}_{\lambda_n} + \frac{\log n}{n} \geq O_P\Big(\frac{1}{n}\Big) + \frac{\log n}{n}.\end{aligned}$$

It thus follows that

$$P\Big(\inf_{\lambda\in\mathbb{R}_+}\mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}\Big)\to 1 \text{ as } n\to\infty. \tag{A.15}$$

The results of Cases 1 and 2 complete the proof.

References

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38**, 3696-3723.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* **91**, 983-992.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32**, 1062-1092.

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *J. Amer. Statist. Assoc.* **86**, 328-332.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

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

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-135.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302-332.

Huang, J., Ma, S., Xie, H. and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339-355.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 192-219.

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

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94**, 603-613.

Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. *J. Roy. Statist. Soc. Ser. B* **67**, 285-299.

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

Silverman, B. W. (1999). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 811-821.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Comp. Statist. Data Anal.* **52**, 4512-4520.

Xia, Y. (2006). Asymptotic distributions of two estimators of the single-index model. *Econometric Theory* **22**, 1112-1137.

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

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98**, 968-979.

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36**, 1649-1668.

Zhu, L. X. and Ng, K. W. (1995). Asymptotics for sliced inverse regression. *Statist. Sinica* **5**, 727-736.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Amer. Statist. Assoc.* **101**, 1638-1651.

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, P. R. China.
E-mail: 10466029@hkbu.edu.hk

School of Finance and Statistics, East China Normal University, Shanghai, 200241, P. R. China.
E-mail: xupeirong1108@hotmail.com

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, P. R. China.
E-mail: lzhu@hkbu.edu.hk

(Received November 2011; accepted May 2012)