AN ADAPTIVE ESTIMATION METHOD FOR SEMIPARAMETRIC MODELS AND DIMENSION REDUCTION

CHENLEI LENG, YINGCUN XIA AND JINFENG XU

Department of Statistics and Applied Probability
National University of Singapore
stalc@stat.nus.edu.sg, staxyc@stat.nus.edu.sg and staxj@stat.nus.edu.sg

Xia, Tong, Li and Zhu (2002) proposed a general estimation method termed minimum average variance estimation (MAVE) for semiparametric models. The method has been found very useful in estimating complicated semiparametric models (Xia, Zhang and Tong, 2004; Xia and Härdle, 2006) and general dimension reduction (Xia, 2008; Wang and Xia, 2008). The method is also convenient to combine with other methods in order to incorporate additional statistical requirements (Wang and Yin, 2007). In this paper, we give a general review on the method and discuss some issues arising in estimating semiparametric models and dimension reduction (Li, 1991 and Cook, 1998) when complicated statistical requirements are imposed, including quantile regression, sparsity of variables and censored data.

Keywords: censored data; dimension reduction; Lasso; minimum average variance estimation; nonparametric quantile regression; semiparametric models.

1. Introduction

Suppose $Y$ is a response and $X = (x_1, ..., x_p)$ are covariates. One of the basic statistical goals is to estimate the conditional mean function $m(x) = \mathbb{E}(Y | X = x)$ or model $Y = m(X) + \varepsilon$ with $\mathbb{E}(\varepsilon | X) = 0$ almost surely. More generally, statisticians are interested in function $m(x) = \sup_{m \in \mathcal{C}} E\{\rho(Y - m(x)) | X = x\}$, where $\rho(.)$ is a loss function and $\mathcal{C}$ is a class of (smooth) function. As it is well known, without any information about the structure of function $m(.)$, it is difficult to estimate the function well when the dimension $p$ is greater than 1, due to the so called “curse of dimensionality”.

As a consequence, many parametric and semiparametric models were proposed in the last decades by imposing structure or special functional forms. Write $X^T = (X^T_{[1]}, X^T_{[2]})$ where $X^T_{[1]} = (x_1, ..., x_q)$ and $X^T_{[2]} = (x_{q+1}, ..., x_p)$ with $q < p$. Here are some examples of the popular semiparametric models.

- Partially linear model (e.g. Speckman, 1988): $Y = \alpha^T X_{[1]} + g(X_{[2]}) + \varepsilon$.
- Single-index model (e.g. Ichimura, 1993): $Y = g(\beta_0^T X) + \varepsilon$.
- Semi-varying coefficient model (e.g. Zhang et al, 1999; Xia, Zhang and Tong, 2004): $Y = g_1(x_1)x_1 + g_2(x_1)x_2 + ... + g_q(x_1)x_q + a_{q+1}x_{q+1} + ... + a_p x_p + \varepsilon$.
- Single-indexing varying coefficient model (Xia and Li, 1999; Fan, Yao and Cai, 2003): $Y = g_0(\beta_0^T X) + g_1(\beta_0^T X)x_1 + ... + g_p(\beta_0^T X)x_p + \varepsilon$.
- Dimension reduction model (Li, 1991): $Y = g(\beta_1^T X, ..., \beta_d^T X, \varepsilon)$, where $d < p$.

Developing estimation method for the parameters in the semiparametric models has a long history. It was well understood from the very beginning that the root-$n$ consistency
for the estimator of parameters can be achieved, but was generally believed that the under-smoothing is necessary. The under-smoothing approach utilizes a smaller bandwidth (than the optimal bandwidth for the estimation of nonparametric functions in the model); see Robinson (1988). It was found later that many semiparametric models do not need the under-smoothing (Speckman, 1988; Härdle, Hall and Ichimura, 1993). Applying MAVE, it was well demonstrated that for a large set of semiparametric models, under-smoothing is also unnecessary for the parameters to achieve the root-$n$ consistency rate (Xia, Zhang and Tong, 2004; Xia, 2007).

Another important feature for MAVE is its easy implementation and availability of algorithms. The estimation of the semiparametric models, especially those that contain single indices, needs to solve a complicated nonlinear minimization problem, which can be difficult. A naive approach is the Newton-Raphson method, in which evaluations of the derivatives or the Hessian matrix of the unknown link function are needed. However, it is well know that the estimation of the derivatives and the Hessian matrix can be complicated. As a consequence, the Newton-Raphson method does not work well. Instead, MAVE provides a very simple way for the calculation by a local linear approximation such that the calculation is eventually converted to problems of “linear minimization”. For the latter the calculation is much easier and many efficient algorithms are available. In this review, we will give a few examples for which MAVE can be conveniently used in calculation.

Lastly, we will discuss the application of the method in dimension reduction (Li, 1991 and Cook, 1998). Xia et al (2002) proposed the MAVE method for dimension reduction in the conditional mean function. Xia (2007) and Wang and Xia (2008) generalized the idea to general dimension reduction problem. In this review, we consider the dimension reduction problem for survival data, which are often subject to censoring. When censoring occurs, the incompleteness of the observed data may induce a substantial bias in the sample. A number of approaches have been suggested to overcome the associated difficulties in regression with pre-specified model assumptions, including the censored linear regression model, the Cox proportional hazard model and many others. It is interesting to consider the problem without model specification, leading to dimension reduction in censored data. We shall show how censored data can still be easily analyzed by applying MAVE.

2. Estimating semiparametric regression models with nonsmooth loss functions

In this section, we consider the estimation of the semiparametric models with nonsmooth loss functions, including the quantile regression and estimation with $L_1$ penalty. The latter is now a popular choice used for variable selection, known as Lasso (Tibshirani, 1996). For simplicity, let us focus on the single-index model. Its extension to other models is not difficult. In the last decade or so, a series of papers (e.g. Powell, Stock and Stoker, 1989; Härdle and Stoker, 1989; Ichimura, 1993; Klein and Spady, 1993; Härdle, Hall and Ichimura, 1993; Horowitz and Härdle, 1996; Hristache, Juditski and Spokoiny, 2001; Xia, Tong, Li and Zhu, 2001) have considered the estimation of the parametric index and the nonparametric link function (i.e. the function $g$), focusing on the root-$n$ consistency of the former; efficiency issues have also been studied. Amongst the various methods of estimation, the more popular ones are the average derivative estimation (ADE) method investigated by Härdle and Stoker (1989), the sliced inverse regression (SIR) method proposed by Li (1989),

The basic algorithm for estimating the parameters in the single-index model is based on observing that

$$\theta_0 = \arg \min_{\theta} E[y - g(\theta^T X)]^2$$

subject to $\theta^T \theta = 1$. By conditioning on $\xi = \theta^T X$, we see that (1) equals $E_\xi \sigma^2_\theta(\xi)$ where

$$\sigma^2_\theta(\xi) = E\left[\left(y - g(\xi)\right)^2\bigg|\theta^T X = \xi\right].$$

It follows that

$$E[y - g(\theta^T X)]^2 = E_\xi \sigma^2_\theta(\theta^T X).$$

Therefore, minimization (1) is equivalent to,

$$\theta_0 = \arg \min_{\theta} E_\xi \sigma^2_\theta(\xi)$$

subject to $\theta^T \theta = 1$. Let $\{(X_i, y_i) : i = 1, 2, \ldots, n\}$ be a sample from $(X, y)$. The conditional expectation in (2) is now approximated by the sample analogue. For $X_i$ close to $x$, we have the following local linear approximation

$$y_i - g(\theta_0^T X_i) \approx y_i - g(\theta_0^T x) - g'(\theta_0^T x) X_i^T \theta_0,$$

where $X_{ix} = X_i - x$. Following the idea of local linear smoothing, we may estimate $\sigma^2_\theta(\theta^T x)$ by

$$\hat{\sigma}^2_\theta(\theta^T x) = \min_{a,d} \sum_{i=1}^{n} \left\{y_i - a - dX_{ix}^T \theta\right\}^2 w_{i0}.$$  

(3)

Here, $w_{i0} \geq 0, i = 1, 2, \ldots, n$, are some weights with $\sum_{i=1}^{n} w_{i0} = 1$, typically centering at $x$. Let $X_{ix} = X_i - X_j$. By (2) and (3), our estimation procedure is to minimize

$$\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \left\{y_i - a_j - d_j X_{ij}^T \theta\right\}^2 w_{ij}$$

(4)

with respect to $(a_j, d_j)$ and $\theta$. If the kernel smoothing is used with kernel function $H(\cdot)$ and bandwidth $h$, then the weight functions $w_{ij} = H_h(X_{ij})$, where $H_h(\cdot) = H(\cdot/h)/h^d$. We call the estimation procedure the minimum average (conditional) variance estimation (MAVE) method; see Xia et al (2002) for more discussions.

### 2.1. Semiparametric estimation with Lasso

When $p$ is large, the coefficients are usually sparse. As a consequence, many of the coefficients are zeros. To automatically select the variables with nonzero coefficients and to estimate the model, Tibshirani (1996) proposed to use the $L_1$ penalty. For the single-index model, we can also implement the variable selection and model estimation simultaneously by imposing the $L_1$ penalty. Following the MAVE idea and Lasso, we need to estimate the single-index by

$$\hat{\theta} = \arg \min_{\theta, \delta} \sum_{i=1=1}^{n} \sum_{i=1}^{n} \left\{Y_i - a_j - d_j X_{ij}^T \theta\right\}^2 w_{ij} + \lambda |\theta|,$$  

(5)
where $||.||$ stands for Euclidean norm and $|\theta| = |\theta_1| + ... + |\theta_p|$ is the $L_1$ norm, $\theta = (\theta_1, ..., \theta_p)^\top$. The calculation of the above minimization problem can be decomposed into two minimization problems as follows.

- Fixing $\vartheta = \theta$ and $w_{\vartheta_{ij}} = K_h(\vartheta^\top X_{ij})$, the solution to (5) of $a_j$ and $d_j$ are

$$
\begin{align*}
\begin{pmatrix}
a_j^\vartheta \\
d_j^\vartheta
\end{pmatrix} = \left\{ \sum_{i=1}^n w_{ij}^\vartheta \left( \frac{1}{X_{ij}} \right) \left( \frac{1}{\theta^\top X_{ij}} \right)^\top \right\}^{-1} \sum_{i=1}^n w_{ij}^\vartheta \left( \frac{1}{\theta^\top X_{ij}} \right) Y_i,
\end{align*}
$$

- Fixing $a_j$ and $d_j$, the minimization in (5) with respect to $\theta$ can be done as follows. Let

$$Y_{ij}^\vartheta = Y_i(w_{ij}^\vartheta)^{1/2} - a_j(w_{ij}^\vartheta)^{1/2}, \quad X_{ij}^\vartheta = d_j X_{ij}(w_{ij}^\vartheta)^{1/2}.$$

Then the problem becomes that of minimizing

$$
\sum_{i,j=1}^n \left\{ Y_{ij}^\vartheta - \theta^\top X_{ij}^\vartheta \right\}^2 + \lambda|\theta|.
$$

Repeat the two steps with $\vartheta := \theta/||\theta||$ until convergence. Similar to the linear regression model, the model estimation and variable selection can be implemented simultaneously. Using a similar idea, a more general model was investigated in Wang and Yin (2007). Note that the above algorithm is based on Lasso. One can also use lars algorithm (Efron et al., 2004).

Two pilot parameters, bandwidth $h$ and penalty $\lambda$, need to be selected. For the bandwidth, Xia (2002) gave a discussion. With $\theta$ fixed, the bandwidth is actually selected for a univariate local linear smoothing. Many existing methods can be used. In practice, the simple rule-of-thumb works well. More discussion is given later in section 3.2.

For the penalty parameter, we use the BIC criterion which is defined as

$$BIC(\lambda) = \log(RSS(\lambda)) + \frac{d_\lambda \log n}{n},$$

where $d_\lambda$ is the number of nonzero entries in the estimator for the tuning parameter $\lambda$ and $RSS(\lambda)$ is the residual sum of squares. It is remarkable that the BIC here is the one used for parametric models instead of nonparametric ones, since we are selecting number of parameters rather than nonparametric functions though it is under a semiparametric setting. We found that $BIC(\lambda)$ works well in simulations though rigorous justification is needed.

2.2. Quantile regression

Regression quantiles, along with the dual methods of regression rank scores, can be considered one of the major statistical breakthroughs of the past decades. Its advantages over the other estimation methods have been well investigated. Regression quantile methods provide a much more complete statistical analysis of the stochastic relationships among variables; in addition, they are more robust against possible outliers or extremely values, and can be computed via traditional linear programming methods. Although median regression ideas go back to the 18th century and the work of Laplace, regression quantile methods were first introduced by Koenker and Bassett (1978).
For a general loss function \( \rho(.) \), we are interested in the following minimization problem

\[
\min \mathbb{E}\{ \rho(Y - m(\theta^T X)) | X = x \}
\]  

(7)

with respect to \( ||\theta|| = 1 \) and \( m \in C \), continuous functions. Suppose the minima is achieved by \( \{\theta_0, m(.)\} \) and \( \rho(.) \) is piece-wise differentiable with derivative \( \varphi(.) \), then (7) leads to a single-index M-regression model

\[
Y = m(\theta_0^T X) + \varepsilon, \quad \mathbb{E}(\varphi(\varepsilon) | X) = 0
\]  

(8)

almost surely. An important special case for the loss function is

\[
\rho(v) = \tau I(v > 0)v + (\tau - 1)I(v \leq 0)v,
\]  

where \( 0 < \tau < 1 \) and \( I(.) \) is the indicator function, leading to the quantile regression, see Koenker and Bassett (1978).

Our main focus is the estimation of \( \theta_0 \). Suppose \( \{X_i, Y_i\}_{i=1}^n \) are observations from underlying model (8). Following MAVE, we propose to estimate the index parameter \( \theta_0 \) by

\[
\hat{\theta} = \arg \min_{\theta \in \mathbb{R}^p, \|\theta\| = 1} \frac{1}{n} \sum_{i,j=1}^{n} w_{ij} \rho\{Y_i - a_j - d_j \theta^T X_{ij}\}.
\]  

(9)

Again, the calculation of the above minimization problem can be decomposed into two minimization problems.

- Fixing \( \theta = \vartheta \) and \( w_{ij}^\vartheta = K_h(\vartheta^T X_{ij}) \), the estimation of \( a_j \) and \( d_j \) are

\[
\frac{1}{n} \sum_{i=1}^{n} \rho\{Y_i - a_j - d_j \vartheta^T X_{ij}\} w_{ij}^\vartheta.
\]

- Fixing \( a_j \) and \( d_j \), the minimization respect to \( \theta \) can be done as follows. Again, let

\[
Y_{ij}^\vartheta = Y_i(w_{ij}^\vartheta)^{1/2} - a_j(w_{ij}^\vartheta)^{1/2}, \quad X_{ij}^\vartheta = d_j X_{ij}(w_{ij}^\vartheta)^{1/2}.
\]

Then the problem becomes

\[
\min_{\theta} \frac{1}{n} \sum_{i,j=1}^{n} \rho\{Y_{ij}^\vartheta - \theta^T X_{ij}^\vartheta\}
\]

Suppose the solution to the above problem is \( \theta \). Standardize it to \( \bar{\theta} := \theta / ||\theta|| \).

Set \( \vartheta = \bar{\theta} \) and repeat the two steps until convergence. Note that both steps are simple linear quantile regression problems and that several efficient algorithms are available, see Koenker (2005). Kong and Xia (2008) proved that the above iteration converges, i.e. there exists a constant \( 0 < c < 1 \) such that

\[
||\theta - \theta_0|| = c||\vartheta - \theta_0|| + O_p(n^{-1/2}).
\]
3. Dimension Reduction with Censored data

Censoring occurs when the value of an observation is only partially known. This type of data happens quite often in the investigation of epidemiology. Let $Y^o$ be the true but sometimes unobservable lifetime, and $X = (x_1, ..., x_p)\top$ be the covariates. Let

\[ C = \text{the censoring time}, \]
\[ \delta = \text{the censoring indicator}; \delta = 1, \text{if } Y^o \text{ is not censored}; \delta = 0, \text{otherwise}, \]
\[ Y = \begin{cases} Y^o, & \text{if } Y^o \text{ is not censored;} \\ \min\{Y^o, C\}, & \text{otherwise.} \end{cases} \]

There are two kinds of censoring

Type I: $C$ is independent of $X$ and $Y^o$.

Type II: Conditional on $X$, $C$ is independent of $Y^o$.

Suppose $S_{Y^o|X}$ is the central space (CS, Cook, 1998) of $Y$ on $X$. That is for a base $B = (\beta_1, ..., \beta_q)$ of $S_{Y|X}$ with $B\top B = I_q$, we have

\[ Y^o \mathbb{1} X | B\top X, \]

which indicates that given $B\top X$, $X$ and $Y^o$ are independent. Similarly, we can define $S_{Y|X}$ and $S_{C|X}$. See Li et al (1999).

For Type I censoring, the estimation is relatively easy since the censoring does not affect the CS of $Y$ on $X$. Let $S_{Y|X}$ be the CS of $Y$ on $X$ with type I censoring. For censoring type II, besides the dimension reduction for $Y^o$ we also need to consider the dimension reduction for the censoring $C$. Let $S_{C|X}$ be the CS for $C$, i.e. $C \mathbb{1} X | B_c\top X$, where $B_c$ is a base of $S_{C|X}$. They have the following relationship.

**Proposition 3.1.** For type I censoring, we have $S_{Y|X} = S_{Y^o|X}$. For type II censoring, we have $S_{Y|X} \subseteq S_{Y^o|X} \oplus S_{C|X}$.

Next, we shall only develop estimation methods for the first type.

3.1. Dimension reduction of complete data

We first give a brief review on how MAVE method estimates the dimension reduction space when the data is complete.

**Theorem 3.2.** For any matrix $B$, $Y \mathbb{1} X | B\top X$ is equivalent to $P(Y \leq y|X = x) = P(Y \leq y|B\top X = B\top x)$ for all $y \in \mathbb{R}^1$ and $x \in \mathbb{R}^p$.

The proof of Proposition 3.2 can be found in Cook (1998). Since $P(Y \leq y|X) = E[I(Y \leq y)|X]$, Proposition 3.2 implies that the CS of $Y$ is closely related to the CMS of $I(Y \leq y)$. Consequently, as long as the CMS of $I(Y \leq y)$ can be estimated for all $y \in \mathbb{R}^1$, the CS of $Y$ should be able to be recovered. Let $M(x|y) = E[I(Y \leq y)|X = x]$ and $G(u|y) = E(I(Y < y)|B_0^\top X = u)$. Therefore, we need to consider the following regression model. Let $m(x|y) = E[I(Y \leq y)|X = x]$. Based on the discussion, we have $m(x|y) = E(I(Y \leq y)|B_0^\top X = B^\top x) := G(B^\top x|y)$, leading to model

\[ I(Y \leq y) = G(B_0^\top X|y) + \epsilon_y, \quad (10) \]
where $c_y = I(Y \leq y) - E(I(Y \leq y)|X) = I(Y \leq y) - G(B_0^\top X)$.

Consider the gradients $\nabla M(x|y) = (\partial M(x|y)/\partial x_1, \ldots, \partial M(x|y)/\partial x_p)^\top$ and $\nabla G(u|y) = (\partial G(u|y)/\partial u_1, \ldots, \partial G(u|y)/\partial u_d)^\top$, where $u = (u_1, \ldots, u_d)^\top$. It follows that

$$\frac{\partial M(x|y)}{\partial x} = B_0 \nabla G(B_0^\top x|y).$$

Moreover, the following results indicate that the CS cannot be missed as long as $y$ runs over its whole sample space.

**Proposition 3.3.** Let $\Omega(y) = E\{\nabla M(X|y) \nabla^\top M(X|y)\}$ and $\Lambda(y) = E\{\nabla G(B_0^\top X|y) \nabla^\top G(B_0^\top X|y)\}$. If $B_0$ is a basis of the CS and that $\nabla M(x|y)$ is continuous in $x$, then (1) $E\Omega(Y) = B_0 E\{\Lambda(Y)\} B_0^\top$ and (2) $E\{\Lambda(Y)\}$ is of full rank.

A proof of Proposition 3.3 can be found in Wang and Xia (2007).

### 3.2. Dimension reduction for the Censored data

Let us first consider the distribution estimation of censored data. Suppose $Y_1, \ldots, Y_n$ are the randomly observed life times, amongst them $Y_{(i)}, \ldots, Y_{(nc)}$ are censored and $Y_{(1)} \leq Y_{(c)} \leq \cdots \leq Y_{(nc)}$ are not censored, where $n_c + n'_c = n$. We propose to estimate the empirical distribution by

$$\hat{G}(y) = \frac{\#\{Y_{(i)} \leq y\}}{n'_c + \#\{Y_{(i)} > y\}}$$

It is easy to see that $\hat{G}(y)$ is an increasing function, because $\#\{Y_{(i)} \leq y\}$ is increasing and $\#\{Y_{(i)} > y\}$ is deceasing. Note that if the censored data is not used, a naive estimator of the empirical distribution is $\tilde{F}(y) = \#\{Y_{(i)} \leq y\}/n$. As a comparison we have the following results.

**Proposition 3.4.** Suppose $Y_{(1)}, \ldots, Y_{(nc)}$ are randomly censored, i.e. $Y_{(k)} = \min\{Y_{(k)}, C_k\}$, $k = 1, \ldots, n_c$. Then $E\{\hat{G}(y)\} = F(y)$ and $E\{\hat{G}(y) - F(y)\}^2 \leq E\{\tilde{F}(y) - F(y)\}^2$. The equality holds only at $y$ with $\#\{Y_{(i)} > y\} = 0$.

Proposition 3.4 indicates that $\hat{G}(y)$ is a more efficient estimator of the distribution and thus the censored data is used more efficiently than discarding the incomplete data.

**Initial estimator.** Based on Proposition 3.3, we immediately have the following estimation method. Suppose that $\{(X_i, Y_i, \Delta_i), i = 1, 2, \ldots, n\}$ is a random sample from $(X, Y)$, where $\Delta_i = 1$ or 0 respectively for complete data or censored data. To estimate the gradient $\partial M(x|y)/\partial x$, we can use the nonparametric kernel smoothing methods. For simplicity, we adopt the following notation scheme. Let $K_0(v^2)$ be a univariate symmetric density function and define $K(v_1, \ldots, v_d) = K_0(v_1^2 + \cdots + v_d^2)$ for any integer $d$ and $K_h(u) = h^{-d}K(u/h)$, where $d$ is the dimension of $u$ and $h > 0$ is a bandwidth. For any $(x, y)$, the principle of the local linear smoother suggests minimizing

$$n^{-1} \sum_{\Delta_i \geq 0} \left\{ I(Y_i \leq y) - a - b^\top (X_i - x) \right\}^2 K_h(X_{ix})$$

with respect to $a$ and $b$ to estimate $M(x|y)$ and $\partial M(x|y)/\partial x$ respectively, where $X_{ix} = X_i - x$. See Fan and Gijbels (1996) for more details. For each pair of $(X_j, Y_k)$, we consider
the following minimization problem
\[
(\hat{a}_{jk}, \hat{b}_{jk}) = \arg \min_{a_{jk}, b_{jk}} \sum_{\Delta_{y_1, y_2}} \left[I(Y_i \leq Y_k) - a_{jk} - b_{jk} X_{ij}\right]^2 w_{ij}, \tag{13}
\]
where \(X_{ij} = X_i - X_j\) and \(w_{ij} = K_h(X_{ij})\). We consider an average of their outer products
\[
\hat{\Sigma} = n^{-2} \sum_{k=1}^n \sum_{j=1}^n \rho_{jk} \hat{b}_{jk} \hat{b}_{jk}^\top,
\]
where \(\rho_{jk}\) is a trimming function introduced for technical purpose to handle the notorious boundary points. In this paper, we adopt the following trimming scheme. For any given point \((x, y)\), we use all observations to estimate its function value and its gradient as in (12). We then consider the estimates in a compact region of \((x, y)\). Moreover, for those points with too few observations around, their estimates might be unreliable. They should not be used in the estimation of the CS directions and should be trimmed off. Let \(\rho(\cdot)\) be any bounded function with bounded second order derivatives on \(\mathbb{R}\) such that \(\rho(v) > 0\) if \(v > \omega_0\); \(\rho(v) = 0\) if \(v \leq \omega_0\) for some small \(\omega_0\). We take \(\rho_{jk} = \rho(\hat{f}(X_{ij}))\), where \(\hat{f}(x)\) is estimator of the density functions of \(X\). The CS directions can be estimated by the first \(q\) eigenvectors of \(\hat{\Sigma}\). We call this estimator the outer product of gradient estimator (OPG); see also Xia et al (2002).

**MAVE for censored data** Note that if (10) holds, then the gradients \(\partial M(x|y)/\partial x\) at all \((x, y)\) are in a common \(q\)-dimensional subspace as shown in equation (11). To use this observation, we can replace \(b\) in (12), which is an estimate of the gradient, by \(Bd(x, y)\) and have the following local linear approximation
\[
n^{-1} \sum_{\Delta_{y_1, y_2}} \{I(Y_i \leq y) - a - d^\top B^\top (X_i - x)\}^2 K_h(B^\top X_{ij}),
\]
where \(d = d(x, y)\) is introduced to take the role of \(\nabla G(B_0^\top x|y)\) in (11). Note that the above weighted mean of squares is the local approximation errors of \(I(Y_i \leq y)\) by a hyperplane with the normal vectors in a common space spanned by \(B\). Since \(B\) is common for all \(x\) and \(y\), it should be estimated with aims to minimize the approximation errors for all possible \(X_i\) and \(Y_k\). As a consequence, we propose to estimate \(B_0\) by minimizing
\[
n^{-3} \sum_{k=1}^n \sum_{j=1}^n \rho_{jk} \sum_{\Delta_{y_1, y_2}} \{I(Y_i \leq Y_k) - a_{jk} - d_{jk}^\top B^\top X_{ij}\}^2 w_{ij} \tag{14}
\]
with respect to \(a_{jk}, d_{jk} = (d_{jk1}, \cdots, d_{jkq})^\top, j, k = 1, \ldots, n\) and \(B : B^\top B = I_d\), where \(\rho_{jk}\) is defined above. Because the method is proposed for censored data using MAVE, we call it the minimum average (conditional) variance estimation for censored data (cMAVE).

The minimization problem in (14) can be solved by fixing \((a_{jk}, d_{jk}), j, k = 1, \ldots, n,\) and \(B\) alternatively. As a consequence, we need to solve two quadratic programming problems which have simple analytic solutions. For any matrix \(B = (\beta_1, \cdots, \beta_q)\), we define operators \(\ell(.)\) and \(\mathcal{M}(.)\) respectively as
\[
\ell(B) = (\beta_1^\top, \cdots, \beta_q^\top)^\top \quad \text{and} \quad \mathcal{M}(\ell(B)) = B.
\]
We propose the following cMAVE algorithm to implement the estimation.
Step 2: Let \( \rho(\frac{1}{h(t)}B_{(t)}X_{ij}) \left( \begin{array}{c} 1 \\ B_{(t)}^\top X_{ij} \end{array} \right) \left( \begin{array}{c} 1 \\ B_{(t)}^\top X_{ij} \end{array} \right)^\top \right)^{-1} \times \sum_{\Delta_i=1 \text{ or } \gamma_i \geq Y_k} K_{hi}(\frac{1}{h(t)}B_{(t)}^\top X_{ij}) \left( \begin{array}{c} 1 \\ B_{(t)}^\top X_{ij} \end{array} \right) I(Y_i \leq Y_k),

where \( h_t \) and \( b_t \) are two bandwidths (details are discussed below).

Step 3: Calculate \( \Lambda_{(t+1)} = \{ \mathcal{M}(b_{(t+1)}) \}^\top \mathcal{M}(b_{(t+1)}) \) and \( B_{(t+1)} = \mathcal{M}(b_{(t+1)})\Lambda_{(t+1)}^{-1/2} \). Set \( t := t + 1 \) and go to Step 1.

Step 4: Repeat steps 1–3 until convergence. The final value of \( B_{(t)} \) is our estimator of the direction, denoted by \( \hat{B} \).

The cMAVE algorithm needs a consistent initial estimator in Step 0 to guarantee its theoretical justification. The OPG estimator above can be served as the initial estimator. In practice, we may need to standardize \( X_i = (X_{i1}, \cdots, X_{ip})^\top \) by setting \( X_i := S_x^{-1/2}(X_i - \bar{X}) \) and standardize \( Y_i \) by setting \( Y_i := (Y_i - \bar{Y})/\sqrt{s_y} \), where \( \bar{X} = n^{-1}\sum_{i=1}^n X_i \) and \( S_x = n^{-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top \), \( \bar{Y} = n^{-1}\sum_{i=1}^n Y_i \) and \( s_y = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2 \). Then the estimated CS directions are the first \( q \) columns of \( S_x^{-1/2} \hat{B} \).

Note that the estimation in the procedure is related with nonparametric estimations of conditional density functions. Several bandwidth selection methods are available for the estimation. See, e.g. Silverman (1986), Scott (1992) and Fan et al (1996). Our theoretical verification of the convergence for the algorithms requires some constraints on the bandwidths although we believe these constraints can be removed with more complicated technical proofs. To ensure the requirements on bandwidths can be satisfied, after standardizing the variables we use the following bandwidths in our calculations. In the first iteration, we use slightly larger bandwidths than the optimal ones in terms of MISE as

\[
h_0 = c_0 n^{-\frac{1}{\rho_0+6}}, \quad b_0 = c_0 n^{-\frac{1}{\rho_0+3}},
\]

where \( \rho_0 = \max(p, 3) \). Then we reduce the bandwidths in each iteration as

\[
h_{t+1} = \max\{r_n h_t, c_0 n^{-\frac{1}{\rho_0+t}}\}, \quad b_{t+1} = \max\{r_n b_t, c_0 n^{-\frac{1}{\rho_0+t}}, c_0 n^{-\frac{1}{t}}\}
\]
for \( t \geq 0 \), where \( r_n = n^{-1/(2(p_0+6))} \), \( c_0 = 2.34 \) as suggested by Silverman (1986) if the Epanechnikov kernel is used. Here, the bandwidth \( b \) is selected smaller than \( h \) based on simulation comparisons.

### 3.3. Theoretical Results

Assume that both \( B_0 \) and \( \hat{B} \) have been standardized such that \( B_0^\top B_0 = I_q \) and \( \hat{B}^\top \hat{B} = I_q \). Furthermore, we use \( ||A|| \) to denote the maximum singular value of an arbitrary matrix \( A \), which is the Euclidean norm if \( A \) is a vector. Following Wang and Xia (2007), we have the following asymptotic results for the estimation.

**Theorem 3.5.** Suppose conditions (C1)-(C5) in the appendix hold and the final bandwidth is \( h \), then there is a rotation matrix \( Q \) such that the estimator \( \hat{B} \) is consistent with

\[
||\hat{B} - B_0Q|| = O_p\{h^4 + \log n/(nh^4) + n^{-1/2}\}.
\]

If we use higher order polynomial smoothing, it is possible to show that the root-\( n \) consistency can be achieved for any dimension \( q \). See, e.g. Härdle and Stoker (1989) and Samarov (1993), where the higher order kernel, a counterpart of the higher order polynomial smoother, was used. However, using higher order polynomial smoothers increases the difficulty of calculations while the improvement of finite sample performance is not substantial.

### 4. Some numerical results

In this section we discuss some simulation results to check the numerical performance of the proposed methods.

**Example 4.1 (Variable selection Single-index via Lasso).** We consider in this example a linear regression model

\[
y = \theta_0^\top X + 0.5\varepsilon, \tag{17}
\]

where \( \varepsilon \sim N(0,1) \), \( \theta_0 \in \mathbb{R}^{18} \) with \( \theta_{0,1} = \theta_{0,7} = \theta_{0,13} = 1/\sqrt{3} \) and the rest entries are zero. We compare the proposed estimator with the unpenalized MAVE. We refer to our method as sparse MAVE (sMAVE), where we choose \( \lambda \) in (6) by minimizing the BIC criterion. The covariate vector are generated as \( X \sim \Sigma_0^{1/2} X_0 \) with \( X_0 \sim N(0, I_5) \) and \( \Sigma_0 = (\rho^{i-j})_{0 \leq i,j \leq 5} \), where \( \rho = 0, 0.5, 0.9 \). We consider two sample sizes \( n = 60, 120 \) and repeat the simulation for 100 times. The results are summarized by two measures for an estimator \( \hat{\theta}_0 \): the correlation (CORR) between \( X^\top \hat{\theta}_0 \) and \( X^\top \theta_0 \) and the mean square error \( \text{MSE} = E(\theta_0 - \hat{\theta}_0)^\top X X^\top (\theta_0 - \hat{\theta}_0) \). Furthermore, we record the percentage of models which are correctly identified (CM). Table 1 clearly shows that sMAVE outperforms MAVE. At the same time, sMAVE can select the variables efficiently.

**Example 4.2 (Median regression of Single-index).** In this example we consider the following model

\[
y = \exp\{-5(\theta_0^\top X)^2\} + \varepsilon, \tag{18}
\]

where \( X \sim \Sigma_0^{1/2} X_0 \) with \( X_0 \sim N(0, I_5) \) and \( \Sigma_0 = (0.5^{i-j})_{0 \leq i,j \leq 5} \). For the noise term, we consider several distributions with both heavy tail and thin tails as well. For simplicity, we
consider the median regression only. As a comparison, we also run the MAVE where a least square type estimation is used. With different sample sizes $n = 100, 200$, we carried out 100 replications. The calculation results are listed in Table 2.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>size</th>
<th>method</th>
<th>CORR</th>
<th>MSE × 100</th>
<th>CM (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>60</td>
<td>MAVE</td>
<td>0.960 (0.013)</td>
<td>10.87 (3.98)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sMAVE</td>
<td>0.989 (0.010)</td>
<td>2.481 (2.35)</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAVE</td>
<td>0.982 (0.006)</td>
<td>4.093 (1.25)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sMAVE</td>
<td>0.997 (0.003)</td>
<td>0.618 (0.63)</td>
<td>78</td>
</tr>
<tr>
<td>0.5</td>
<td>60</td>
<td>MAVE</td>
<td>0.962 (0.014)</td>
<td>10.77 (4.73)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sMAVE</td>
<td>0.991 (0.008)</td>
<td>2.118 (2.12)</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAVE</td>
<td>0.982 (0.007)</td>
<td>4.270 (1.44)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sMAVE</td>
<td>0.997 (0.004)</td>
<td>0.724 (0.76)</td>
<td>79</td>
</tr>
<tr>
<td>0.9</td>
<td>60</td>
<td>MAVE</td>
<td>0.979 (0.008)</td>
<td>23.56 (10.24)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sMAVE</td>
<td>0.990 (0.009)</td>
<td>6.053 (5.69)</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MAVE</td>
<td>0.990 (0.003)</td>
<td>9.531 (5.59)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sMAVE</td>
<td>0.998 (0.002)</td>
<td>1.597 (2.88)</td>
<td>74</td>
</tr>
</tbody>
</table>

the MAVE method with quadratic loss function has very bad performance when the noise has heavy tail (e.g. $t(1)$) or is highly asymmetric (e.g. $N(0,1)^4$). With the absolute value loss function, the performance is much better. Even in the situation when the noise has thin tail and symmetric, $q$MAVE still performance reasonably well.

**Example 4.3 (Dimension reduction for censored data).** In this example, we consider the following censored regression model

$$Y^o = 4 - |\theta_0^T X - 1| + 0.1\varepsilon_1,$$

$$C = a + 0.1\varepsilon_2$$

where $X \sim N(0, I_6), \varepsilon_1 \sim N(0,1), \varepsilon_2 \sim N(0,1), C \perp (X, Y^o), \theta_0 = (1,0,0,0,0,0)^T$, and $a$ is set to 3.6 or 3.2 resulting in 20% or 40% censoring respectively. The estimation error between the estimator $\hat{\theta}$ and $\theta_0$ is measured by both MSVD and SQABD, where MSVD = $\max(|\text{svd}(\hat{\theta}^T - \theta_0\theta_0^T)|)$ and SQABD = $\sqrt{1 - |\hat{\theta}^T\theta_0|}$. With different sample sizes, 100 replications are drawn from the model and the comparison is made between a naive MAVE method discarding the censored observations (MAVE) and the one we proposed (cMAVE). The simulation results are summarized in Table 3. As expected, cMAVE performs more favorably than MAVE by both measures of the estimation error, especially in the heavier censoring (40%) case.
Example 4.4 (Primary biliary cirrhosis data). The Mayo clinic has established a database of 424 patients having primary biliary cirrhosis (PBC). 276 of them have complete information of 18 variables. Considering all 18 variables, we analyze it using the cMAVE and uMAVE methods and the estimated effective dimension reductions (with the number of dimension reduction set to 1) are (0.0043, -0.6152, -0.0001, 0.3802, 0.1674, -0.0001, -0.2972, 0.5421, 0.0215, 0.0001, -0.1882, -0.0015, 0.0639, 0.1565, -0.0466, -0.0009, 0.0027) and (0.0317, -0.7340, -0.0001, 0.2852, 0.0437, 0.0003 0.1475, 0.2563, 0.2799, -0.0009, -0.0387 , -0.1907 , 0.0025, 0.2966, 0.1804, 0.2991, -0.0014 , 0.0044) which are plotted against the response variable in Figure 1 and Figure 2 respectively. Based on the plots, we can see that the proposed estimation that uses the censored observations can give clear pattern than the method that ignore the censored observations.

<table>
<thead>
<tr>
<th>censoring (%)</th>
<th>size</th>
<th>method</th>
<th>MSVD</th>
<th>SQABD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>100</td>
<td>MAVE</td>
<td>0.0439(0.0161)</td>
<td>0.0310(0.0114)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cMAVE</td>
<td>0.0419(0.0185)</td>
<td>0.0296(0.0131)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>MAVE</td>
<td>0.0277(0.0088)</td>
<td>0.0196(0.0062)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cMAVE</td>
<td>0.0257(0.0080)</td>
<td>0.0182(0.0056)</td>
</tr>
<tr>
<td>40</td>
<td>100</td>
<td>MAVE</td>
<td>0.0827(0.0318)</td>
<td>0.0586(0.0225)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cMAVE</td>
<td>0.0445(0.0174)</td>
<td>0.0315(0.0123)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>MAVE</td>
<td>0.0527(0.0231)</td>
<td>0.0373(0.0164)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cMAVE</td>
<td>0.0280(0.0106)</td>
<td>0.0198(0.0075)</td>
</tr>
</tbody>
</table>

5. Conclusion

In this review, we present some problems that MAVE can be applied easily. Discussions and simulations suggested that it is promising to apply the MAVE to complicated problems.

This work is a brief review of the the possible application of MAVE. More details related to the implementation, including how to select the pilot parameters $h$ and $\lambda$, and how to determine the dimension of the central space, need to presented. Asymptotic theory is another topic which needs to be addressed in the future.

Acknowledgement The authors thank two referees for their very valuable comments.

References

Fig. 1. The first panel is the plot of response against the dimension reduction direction estimated by cMAVE and the second that against the dimension reduction direction estimated by uMAVE.
