Int. J. Nonlinear Anal. Appl. 13 (2022) 1, 2609-2616
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2022.6063

Estimating partial linear single index model by MAVE and two stage estimation procedures

Huda Yahya Ahmed[a], Munaf Yousif Hmood[a,*]

[a] *Department of Statistics, College of Administration and Economics, University of Baghdad, Iraq*

(Communicated by Madjid Eshaghi Gordji)

Abstract

This research aims to estimate the partial linear single index model by using minimum average variance estimation (MAVE) and two stage estimation Procedures. We analyzed the creatinine ratio in blood data in order to determine the effects of different variables on renal failure. Mean squared error criterion (MSE) was used to compare between these methods, with reference to the use of the rule of thumb (ROT) method to estimate the smoothing parameter.

Keywords: PLSIM, MAVE, Two Stage Procedure, Local Linear Smoother, ROT.

1. Introduction

The use of semiparametric methods has appeared in many fields of research, especially in medical studies because of its flexibility in combining linear models and nonparametric regression models. Although there are many advantages to both models, whether linear or non-parametric, but we note that the non-parametric model suffering from the problem of the curse of dimensionality. So in order to get rid of this problem, the single index model can be used to reduce the dimensions by assuming the effect of the explanatory variables X's can be combined into a single index through using an unknown link function **g**. The term of semi-parametric appeared for the first time by Santner, Cail & Brown, 1980 in the field of Biometric and also by Whitehead, 1981 and called it as partially parametric [14]. While Millimet, List & Stengos, 2003 studied the problem of air pollution in the USA assuming there are no restrictions on function formation and showing the impact of air pollution on the economy of the USA, they made a comparison between a parametric and semiparametric models, so that the results proved the semiparametric model is the best and most flexible in representing the

*Corresponding author

Email addresses: hudayahya26@gmail.com (Huda Yahya Ahmed), munaf.yousif@coadec.uobaghdad.edu.iq (Munaf Yousif Hmood)

Received: May 2021 *Accepted:* October 2021

data [2]. Wang et al., in [13] studied partial linear single index model estimation and they proposed two-stage estimation to estimate the link function of the single index and the parameters in the single index. Su and Zhang in [11] highlighted the recent developments on estimate the variable selection for nonparametric and semiparametric regression models, they explained **SCAD** and **LASSO** methods. Hmood, in [9] studied characteristics of single index semiparametric model (SSIM). He was used both of local linear regression and Nadaraya-Watson estimators for estimating nonparametric part.

Hmood and Tariq in [10] proposed SCAD-MAVE and ALASSO-MAVE methods beside to use some semiparametric methods that based on different function penalties for estimating and selecting variables in a single index model including SCAD-NPLS method. Although the single index model plays an important role in data analysis and dimension reduction, it may not be sufficient to explain the variation of responses through the covariates (X), therefore, the partial linear single-index model can be used instead for the purpose of covering those differences and taking them into account and then representing the phenomenon better. The PLSIM first studied by Carroll et al., in [17] as in the following form:

$$Y{=}\alpha_0^T \mathrm{Z}{+}g\left(\beta_0^T \mathrm{X}\right){+}\varepsilon, \tag{1.1}$$

where:

$X \in R^P and\ Z \in R^q$ are covariates with dimensions p and q respectively.

$g(.)$: an unknown link function for the single index.

ε: is the error term with $E\left(\varepsilon\right)=0$ and $0\ \ < Var\left(\varepsilon\right)\sigma 2<\infty$.

α_0: Unknown parameters vector of degree $(q\times 1)$ for the parametric part.

β_0: Unknown parameters vector of degree $(p\times 1)$ for the nonparametric part.

We further assume that $\|\beta\|=1$ and $\beta_1>0$ for model identification [9].

2. The aim of the article

Nonparametric methods have a curse of dimensional problem that occurs when the number of explanatory variables increases, which leads to low accuracy of estimates, which forces to resort to use semiparametric methods to preserve the characteristics of some variables with parametric behavior and take into account the nonlinear behavior of other variables and then reduce the dimensions.

3. Estimation methods

3.1. Minimum Average Variance Estimation (MAVE)

Xiaa and Härdle in [17] proposed a general method for estimating partial linear single-index models called (MAVE) including the single index model for estimating the parameter vector and the link function at the same time. This method has an important advantage, which is easy to implement. Estimation of semiparametric models and especially those with a single index requires a complex solution to the problem of nonlinear reduction and method (MAVE) provides a simple method for computation through local linear approximation [4, 15, 16].

The basic algorithm for estimating parameters in [17] is based on:

$$Y{=}\alpha_0^T \mathrm{Z}{+}g\left(\beta_0^T \mathrm{X}\right){+}\varepsilon.$$

By satisfying $\beta^T\beta=1$ and conditioning on $U=\beta^T X$, we have

$$(\alpha_0,\beta_0)=\arg\underbrace{\min}_{\alpha,\beta} E[\ Y-\left(\alpha^T Z+g\ \ \left(\beta^T X\ \right)\right)]^2=\arg\underbrace{\min}_{\alpha,\beta} E_U\sigma^2_{\alpha,\beta}(U), \tag{3.1}$$

where

$$\sigma^2_{\alpha,\beta}(\mathrm{U}) = E[\ Y - \left(\alpha^T Z + g\ \ (U)\right) \mid \beta^{\mathrm{T}}\mathrm{X} =\ \ \mathrm{U}\ ]^2.$$

It follows that

$$E\ [Y\ \ --(\alpha^T Z + \ g\ (\beta^T X))\ ]^2 = \mathrm{E}_U\ \sigma^2_{\alpha,\beta}\ (\beta^T X).$$

Let $\{(X_i, Z_i,\ Y_i)$, i = 1, 2,. . . , n$\}$ be a sample from (X, Z, Y), equation (3.1) can be approximated by the Taylor's expansion, so for X_i close to x, we have the following local linear approximation from [12]:

$$Y_i - \alpha_0^T Z - g\left(\beta_0^T X_i\right) \approx Y_i - \alpha_0^T Z_i - g\left(\beta_0^T X\right) - \acute{g}\left(\beta_0^T X\right) X_{i0}^T \beta_0$$

$$X_{i0} = X_i - x_0.$$

Following the idea of local linear smoothing [3], we may estimate $\sigma^2_{\alpha,\beta}\ (\beta^T X)$ by:

$$\widehat{\sigma}^2_{\alpha,\beta}\ \left(\beta^T X\right) = \underbrace{\min}_{(a,b)} \sum_{i=1}^{n} (\{Y_i - \alpha^T Z_i - a - bX_{i0}^T \beta\}^2 W_{i0}) \tag{3.2}$$

$a = g\left(\beta_0^T X\right), b = g'\left(\beta_0^T X\right)$ and $W_{i0} = 0$, i $= 1, 2, \ldots$ n are some weights such that $\sum_{i=1}^n W_{i0} = 1$ often centering at x_0. The MAVE procedure is to minimize:

$$\frac{1}{n}\sum_{j=1}^{n} (G(\beta^T X_j) I_n(X_j) \sum_{i=1}^{n} (\{Y_i - \alpha^T Z_i - a_j - b_j X_{ij}^T \beta\}^2 W_{ij})) \tag{3.3}$$

$$X_{ij} = X_i - X_j$$

W_{ij}: represents the weight dependent on the distance between X_i, X_j .

$$W_{ij} = K_1(X_j) / \sum_{l=1}^{n} K_1(X_j)$$

$$K_1\left(X_j\right) = h^{-1} K_1\left(\frac{X_{ij}}{h}\right),$$

h: denote to the bandwidth and can be selected by rule of the thumb method [5] according to the following formula:

$$\hat{h}_{opt} = C\left(K\right)\left[\frac{\sigma^2 \int W\left(x\right) d(x)}{\sum_{j=1}^n \left(\hat{m}''(X_J)\right)^2 W\left(x\right)}\right]^{1/5}.$$

$C\left(K\right)$: a constant value that depends on the type of Kernel used.

m″(x) : The second derivative of the regression function.

$W \geq 0$: Weight function, such that $W_{ij} \rightarrow 0\ \ if\ X_i - X_j \rightarrow 0.$

With respect to ($\mathrm{a_j}$, $\mathrm{d_j}$) and (β, α), G() is another weight function that controls the contribution of ($\mathrm{X_i}, \mathrm{Z_i}$, $\mathrm{Y_i}$) to the estimation of (β, α) .

I_n (.): refers to the index function that smooth and trims data from outliers [8], such that:

$$I_n(x) = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{l=1}^n k_1(x) > c_0 \\ 0 & \text{otherwise} \end{cases},\quad c_0 > 0\ \text{ is constant} \tag{3.4}$$

Then

$$\left(\begin{array}{c} a_j \\ b_j \end{array}\right)=\left\{\sum_{i=1}^{n} W_{ij}\left(\begin{array}{c} 1 \\ X_{ij}^{T}\beta \end{array}\right)\left(\begin{array}{c} 1 \\ X_{ij}^{T}\beta \end{array}\right)^{T}\right\}^{-1}\sum_{i=1}^{n} W_{ij}\left(\begin{array}{c} 1 \\ X_{ij}^{T}\beta \end{array}\right)\left(Y_i-\alpha^{T}Z_i\right). \tag{3.5}$$

$$\begin{aligned}\left(\begin{array}{c} \alpha \\ \beta \end{array}\right)=&\left\{\sum_{j=1}^{n} G\left(\beta^{T}X_j\right) I_n\left(X_j\right)\sum_{i=1}^{n} W_{ij}\left(\begin{array}{c} Z_i \\ b_jX_{ij} \end{array}\right)\left(\begin{array}{c} Z_i \\ b_jX_{ij} \end{array}\right)^{T}\right\}^{-1} \\ &\times\sum_{j=1}^{n} G\left(\beta^{T}X_j\right) I_n\left(X_j\right)\sum_{i=1}^{n} W_{ij}\left(\begin{array}{c} Z_i \\ b_jX_{ij} \end{array}\right)(Y_i-a_j).\end{aligned} \tag{3.6}$$

We need to use two kernel functions $K_1(.)$, $K_2(.)$ that may be equal or different, $K_1(.) = K_2(.)$ or $K_1(.) \neq K_2(.)$ [7]. $W_{ij}^{\beta} = \frac{K_2\left(\beta^T X_j\right)}{\sum_{l=1}^{n} K_2\left(\beta^T X_j\right)}$ is the kernel weights for single index

$$I_n(x)=\left\{\begin{array}{l} 1,\ \ \hat{f}_{\beta}(\beta^{T}X)>c_0 \\ 0,\ \ otherwise \end{array}\right. \tag{3.7}$$

Given (a_j, b_j) we have from minimizing (3.3), where it is a solution to the problem of weighted least squares by fixing (a_j, b_j), $j = 1, 2, \ldots, n$. Add to (β, α) and by iterating between equationsv(3.5),(3.6) given (β, α) [12]. It can also be estimated $g\left(\beta^T X_j\right)$ by the solution of a_j in (3.5) to get an estimate $\hat{g}\left(\beta^T X_j\right)$. This method estimates all parameters and the nonparametric function by minimizing the loss function. [17]

3.2. A two- stage estimation for a partial linear single-index model

This method was suggested by Wang et al. in [13] to estimate the link function and parameter vector of the single index model. Constrained estimating equation leads to an asymptotically more active estimator than found estimators in the sense that it is of a smaller limiting variance, the estimator of the nonparametric link function realizes best convergence rates and the structural error variance is obtained. In addition, the results ease the construction of confidence regions and hypothesis testing for the unknown parameters. This method does not require any repetition and some indicators are based on X to explain Z.

$\{(\mathrm{X}_i,\ \mathrm{Y}_i,\ \mathrm{Z}_i);\ \mathrm{i} = 1, 2, \ldots, \mathrm{n}\ \}$ denote to the independent and symmetrically distributed variables. In equation (1.1), the estimation process takes place in two stages, that is, Z can be obtained from one indicator of X.

$$Z=\varphi\left(X^{T}\beta_{z}\right)+\eta \tag{3.8}$$

φ : is an unknown function from $\mathrm{q} \times 1$.
β_{z} : is an orthogonal matrix, $\|\beta_{\mathrm{z}}\| = 1$,β_{z} positive first component for model identification.
η : has mean zero and is independent of X with the resulting errors (residuals):

$$\eta_i = Z_i - \varphi(X_i^T\ \beta_z).$$

To estimate the link function we need firstly to estimate β_{z} and then estimate φ to get the residuals, β_{z} is estimated using general least squares by $\beta_Z = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} Z$, V refers to variance matrix.

Also, the unknown link function φ in (3.8) can be estimated by using local linear smoother in [9], so that the resulting estimator is defined as:

$$\widehat{\varphi}\left(X^T\widehat{\beta}_z\right) = \sum_{i=1}^{n} W_{ni}\left(X^T\widehat{\beta}_z\right) Z_i$$

the residual η hence becomes, $\widehat{\eta}_i = Z_i - \widehat{\emptyset}(X_i^T\ \widehat{\beta}_z)$. It is possible to use a least squares approach to estimate α_0. Then

$$\widehat{\alpha} = (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T\tilde{Y},$$

where $\tilde{Y} = Z - \widehat{\varphi}(X^T\ \widehat{\beta}_z)$ using the definition of residuals. The conditional expectation functions are as follows:

$$g_1(t) = E\ \left(Y \mid X^T\beta_0 = t\right), g_2(t) = E\ \left(Z \mid X^T\beta_0 = t\right),$$

so that

$$\hat{g}_1\left(t;\ \widehat{\beta}_0\right) = \sum_{i=1}^{n} W_{ni}\left(t;\ \widehat{\beta}_0\right) Y_i$$

$$\hat{g}_2\left(t;\ \widehat{\beta}_0\right) = \sum_{i=1}^{n} W_{ni}\left(t;\ \widehat{\beta}_0\right) Z_i.$$

We suppose that $(\hat{a})$ is a solution to the weighted least square problem [1, 6].

$$\widehat{\mathrm{g}}\,(\mathrm{x}) = \widehat{\mathrm{a}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} W_i \mathrm{Y}_i}{\sum_{\mathrm{i}=1}^{\mathrm{n}} W_i}. \tag{3.9}$$

Assuming that the parameter vector B is known, the nonparametric estimator for the function $W(t;\beta)$ is

$$W_{ni}(t;\beta) = \frac{K_h\left(X_i^T\beta - t\right)\left[S_{n,2}(t;\beta,h) - \left(X_i^T\beta - t\right)S_{n,1}(t;\beta,h)\right]}{S_{n,0}(t;\beta,h)\,S_{n,2}(t;\beta,h) - S_{n,1}^2(t;\beta,h)}. \tag{3.10}$$

$$\tilde{W}_{ni}(t;\beta) = \frac{K_{h1}\left(X_i^T\beta - t\right)\left[\left(X_i^T\beta - t\right)S_{n,0}(t;\beta,h_1) - S_{n,1}(t;\beta,h_1)\right]}{S_{n,0}(t;\beta,h_1)\,S_{n,2}(t;\beta,h_1) - S_{n,1}^2(t;\beta,h_1)} \tag{3.11}$$

$$S_{n,l}(t;\beta,h) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^T\beta - t\right)^l K_h\left(X_i^T\beta - t\right), \qquad l = 0,1,2$$

The idea of local linear smoothing through smoothing $\mathrm{Y_i} - \mathrm{Z_i^T}\widehat{\acute{\mathrm{a}}}_0$ versus $\mathrm{X_i^T}\ \hat{\mathrm{a}}_0$, $\mathrm{g}(.)$, $\mathrm{g}'(.)$, respectively, are estimated according to the following formulas

$$\hat{g}(t;\beta,\alpha) = \sum_{i=1}^{n} W_{ni}(t,\beta)\left(Y_i - Z_i^T\alpha\right), \tag{3.12}$$

and

$$\widehat{g}'(t;\beta,\alpha) = \sum_{i=1}^{n} \tilde{W}_{ni}(t,\beta)\left(Y_i - Z_i^T\alpha\right). \tag{3.13}$$

The idea of local linear smoothing to reduce the sum of the squares of error

$$\sum_{i=1}^{n}[Y_i - Z_i^T\alpha - \hat{g}(X_i^T\ \widehat{\beta}_0;\widehat{\beta}_0,\alpha)]^2. \tag{3.14}$$

The estimate for $\widehat{\text{â}}_0$ is used to update the estimate of $\widehat{\text{á}}_0$* and has been repeated more than one time to reach out to the desired extent. The resulting partial regression estimator is

$$\widehat{\text{á}}_0^{*}= (\widetilde{\mathrm{Z}}^{\mathrm{T}}\widetilde{\mathrm{Z}})^{-1}\widetilde{\mathrm{Z}}^{\mathrm{T}}\mathrm{Y}^{**} \tag{3.15}$$

$$\mathrm{Y}_{\mathrm{i}}^{**}{=}\mathrm{Y}_{\mathrm{i}}{-}\widehat{\mathrm{g}}_1(\mathrm{X}_{\mathrm{i}}^{\mathrm{T}}\ \widehat{\text{â}}_0;\ \widehat{\text{â}}_0)$$

$$\widetilde{\mathrm{Z}}_{\mathrm{i}}{=}\mathrm{Z}_{\mathrm{i}}{-}\widehat{\mathrm{g}}_2(\mathrm{X}_{\mathrm{i}}^{\mathrm{T}}\ \widehat{\text{â}}_0;\ \widehat{\text{â}}_0).$$

After updating the estimated value of $\widehat{\text{á}}_0$* and calculating the new residuals $(\mathrm{Y}{-}\mathrm{Z}^{\mathrm{T}}\widehat{\text{á}}^{*})$ the estimated value of â_0^{*} is updated. The obtained estimations $\widehat{\text{â}}_0^{*}$, $\widehat{\text{á}}_0^{*}$ are used to update the estimate of the link function g. According to the following equation:

$$\hat{g}^{*}\left(t\right) = \sum_{i=1}^{n} W_{ni}\left(t;\widehat{\beta}\right)\left(\mathrm{Y}_i\ -\ \mathrm{Z}_i^T\widehat{\text{á}}\right). \tag{3.16}$$

4. Application

In this part, we conduct an applied study of the factors affecting renal failure for a sample of 100 patients of different ages, the data collected from the Department of Planning in Iraqi Ministry of Health. Our aim is to identify the most important variables affecting on renal disease through the semiparametric estimation methods. We will deal with the calculation of the parametric component and the nonparametric component of the single index model, as the parametric component includes only the non-zero significant parameters of the parameter vector â, á with dimensions $(\mathrm{p} \times 1)$ and $(q{\times}1)$ respectively. The proposed model for the renal failure based on the most important factors or variables that be used, as follows:

$$Y_i = \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \alpha_3 Z_{i3} + g\left(\beta_1 X_{i1}\ + \beta_2 X_{i2}\ + \beta_3\ X_{i3}\right) + \varepsilon_i \tag{4.1}$$

The variables that used in equation (4.1) refers to the following:
Y_i: Creatinine.
Z_1: Cumulative sugar.
Z_2: Uric Acid.
Z_3: blood Urea Level.
X_1: Age of a person who suffers from kidney disease.
X_2: Blood percentage of a sick person.
X_3: triglycerides.

Through the real data and the adoption of the above model, the parameter vectors estimates for the estimation methods were obtained, as shown in the below:

Table 1: The estimated values of the parameter vector (β, α) and Mean Square error (MSE) using MAVE and Two-Stage procedures

Method	**MAVE**	**Two-Stage**
$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$	0.3805	0.0299
$\widehat{\boldsymbol{\beta}}_{\mathbf{2}}$	0.3813	0.0567
$\widehat{\boldsymbol{\beta}}_{\mathbf{3}}$	0.3803	-0.0010
$\widehat{\boldsymbol{\alpha}}_{\mathbf{1}}$	0.2563	0.0215
$\widehat{\boldsymbol{\alpha}}_{\mathbf{2}}$	0.2531	0.0040
$\widehat{\boldsymbol{\alpha}}_{\mathbf{3}}$	0.2531	0.0001
MSE	**0.0429**	**0.7966**

The results from table 1 show that MAVE method have the lowest MSE, hence is the best method, in representing the function of renal failure, the following figures refers to a curve with real values compared with the estimated values of the response variable by using MAVE and Two-stage procedures respectively.

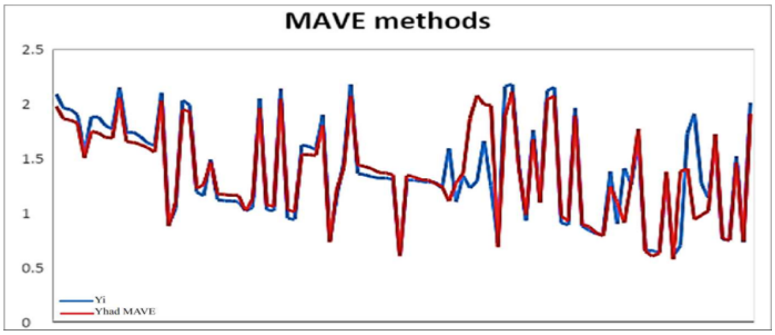

Figure 1: The real and estimated curves of the response variable by using MAVE method

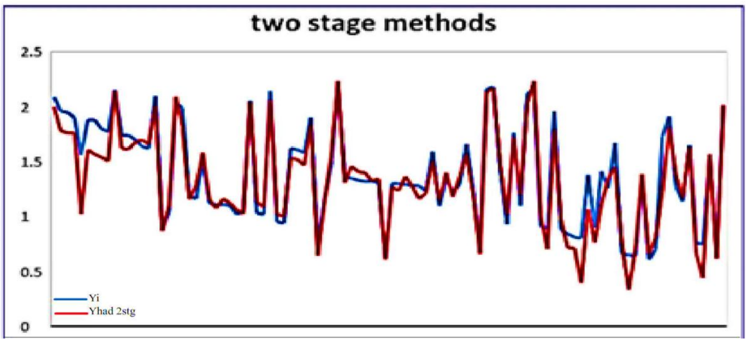

Figure 2: The real and estimated curves of the response variable by using Two-stage method

5. Conclusions

The results of the real data showed that the MAVE method is the best compared to the two-stage estimation method because it gives the lowest MSE in the model estimation, and this means that this method is able to represent the renal failure function.

It is also noted from the results that the values $\hat{a}_1, \hat{a}_2, \hat{a}_3$ for the first method have an equal effect as $\acute{a}_1, \acute{a}_2, \acute{a}_3$ accordingly we conclude that there are similar effects of the explanatory variables on the dependent variable (renal failure), which means an increase in the rate of the proportion of variables which leads to an increase in the proportion of Creatinine in blood.

References

[1] S.X. Chen, *Local linear smoothers using asymmetric kernels*, Ann. Inst. Stat. Math. 54(2) (2002) 312–323.

[2] L.M. Daniel, J.A. List and T. Stengos, *The environmental Kuznets curve: Real progress or misspecified models*, Rev.Econ. Statist. 85(4) (2003) 1038–1047.

[3] E. Kong and Y. Xia, *Variable selection for the single-index model*, Biometrika 94(1) (2007) 217–229.

[4] C. Leng, Y. Xia and J. Xu,*An adaptive estimation method for semiparametric models and dimension reduction*, Explor. Nonlinear World, An Appreciation of Howell Tong's Contributions Statist. (2009) 347–360.

[5] M.Y. Hmood and U. Stadtmuller, *A new version of local linear estimators*, Chilean J. Statist. 4(2) (2013) 61–74.

[6] M.Y. Hmood, *Comparing Nonparametric Kernel estimators for Estimating Regression Function*, MSC Thesis, Department Stat., College of Administration and Economics, University of Baghdad, 2000.

[7] M.Y. Hmood and M.M. Katee, *A comparison of the semiparametric estimators model using different smoothing methods*, J. Econ. Administ. Sci. 20(75) (2014) 376–394.

[8] M.Y. Hmood, Q.N. Nayef and M. Abbas Tahani, *Nonparametric estimation of a multivariate probability density function*, Al-Nahrain J. Sci. 11(2) (2008) 55–63.

[9] M.Y. Hmood, *On single index semiparametric model*, J. Statist. Sci. 6 (2015) 1–17.

[10] M.Y. Hmood and A.S. Tariq, *Compared some of penalized methods in analysis the semi-parametric single index model with practical application* , J. Econ. Administ. Sci. 22(90) (2016) 407–427.

[11] L. Su and Y.Zhang, *Variable Selection in Nonparametric and Semiparametric Regression Model*, The Oxford Handbook of Applied Nonparametric and Semiparametric Econom. Stat., Chapter 9, 2013.

[12] T. Wang, P. Xu and L. Zhu,*Penalized minimum average variance estimation*, Statist. Sin. 23 (2013) 543–569.

[13] J.L. Wang, L. Xue, L. Zhu and Y. S. Chong, *Estimation for a partial-linear single-index model* , Ann. Statist. 38(1) (2010) 246–274.

[14] J.A. Wellner, C.A. Klaassen and Y.A. Ritov, *Semi-parametric models: a review of progress since BKRW (1993)*, Frontiers Statist., (2006) 25–44.

[15] Y. Xia, *Asymptotic distribution for two estimators of the single index model*, Econ. Theory 22(6) (2006) 1112–1137.

[16] Y. Xia, W. Härdle and O.B. Linton, *Optimal Smoothing for A Computationally and Statistically Efficient Single Index Estimator* , Suntory Centre Suntory and Toyota International Centers for Economics and Related Disciplines, Econometrics Discussion Paper, 2009.

[17] Y. Xia and W. Härdle, *Semiparametric estimation of partially linear single index models*, J. Multivariate Anal. 97(5) (2006) 1162–1184.