

Variable Selection for Partial Linear Single-Index Model with M-Estimation

Yafeng Xia and Erzhang Chang

School of Science, Lanzhou University of Technology, Lanzhou, China

Abstract—In this paper, a method of variable selection for partial linear single-index model is proposed, which is based on the M-estimation and the adaptive LASSO. And its oracle property is established and proved. Unlike the existing M-estimator of the partial linear single-index model, the unknown link function is approximated by B-spline.

Keywords—partial linear single-index model; M-estimation; variable selection; B-spline

I. INTRODUCTION

The partial linear single-index model (PLSIM) is an important tool in semi-parametric regression. A common form of the PLSIM is:

$$Y = g(X^T \alpha) + Z^T \beta + \varepsilon,$$

where, $X \in R^p$ and $Z \in R^q$ are covariates of the response variable Y , ε is the model error with $E(\varepsilon|X, Z) = 0$ and $Var(\varepsilon|X, Z) = \sigma^2$, $g(\bullet)$ is an unknown differentiable link function, α and β are unknown p -dimensional and q -dimensional parameter vectors. For identifiability, we assume $\|\alpha\| = 1$ and the first nonzero element of α is positive.

Yu and Ruppert (2002) proposed the penalized spline estimation procedure. Xia and Härdle (2006) proposed the minimizing average variance estimation (MAVE) method based on local linear approach. In real data analysis, the covariates of PLSIM may include many irrelevant covariates. To identify significant predictors will enhance the prediction performance of the fitted model. Liang et al. (2010) proposed a variable selection method for mean regression of PLSIM by combining the profile likelihood method and the SCAD penalized approach. Lai and Guo (2012) proposed a two steps variable selection procedure with adaptive LASSO.

Most existing estimation procedures for PLSIM based on either least squares or likelihood approach. When there are outliers in the data or the distribution of model error has a heavy tail, the estimation performance will not be always guaranteed. In order to overcome this shortcoming, Lv et al. (2014) proposed minimizing average check loss estimation procedure to conduct quantile regression of PLSIM. Zhu et al. (2014) proposed a robust variable selection procedure for partially varying coefficient single-index model based on modal regression.

We consider combining the M-estimation and the adaptive LASSO to achieve robustness of estimation and variable selection on the PLSIM.

II. Variable Selection

Suppose that $\{(x_i, z_i, y_i), i = 1, \dots, n\}$ is a random sample generated from the PLSIM, that is

$$y_i = g(\alpha_0^T x_i) + \beta_0^T z_i + \varepsilon_i, i = 1, \dots, n.$$

Let $a = \min\{x_1^T \alpha_0, \dots, x_n^T \alpha_0\}$, $d = \max\{x_1^T \alpha_0, \dots, x_n^T \alpha_0\}$, $a = t_0 \leq t_1 \leq \dots \leq t_{k_n} = d$ is a partition of interval $[a, d]$.

Using the $t_i (i = 1, \dots, k_n)$ as knots, we have $N = k_n + l$ B-spline basis function of order $l + 1$. Denote these basis functions as a vector $B(t) = (B_1(t), \dots, B_N(t))^T$. Let $g(\alpha_0^T x)$ be approximated by $B(\alpha_0^T x)^T \delta_0$, where $\delta_0 \in R^N$ is the spline coefficient vector, Thus our regression model can be written as

$$y_i = B(\alpha_0^T x_i)^T \delta_0 + \beta_0^T z_i + \varepsilon_i - R_{ni}, i = 1, \dots, n,$$

where R_{ni} is the remainder of approximation of $g(\bullet)$. Write $\eta_0 = (\alpha_0^T, \beta_0^T, \delta_0^T)^T$ as parameter vector to be estimated.

We then consider an M-estimator for η_0 , denote $\hat{\eta}_0 = (\hat{\alpha}_0^T, \hat{\beta}_0^T, \hat{\delta}_0^T)^T$ by minimizing

$$\sum_{i=1}^n \rho(y_i - B(\alpha_0^T x_i)^T \delta_0 - \beta_0^T z_i)$$

for a suitably chosen loss function $\rho(\bullet)$. Following Zou (2014) and the approach of Zou (2014), the estimator $(\hat{\alpha}, \hat{\beta})$ is a \sqrt{n} -consistency estimator of (α_0, β_0) . Then, we use $(\hat{\alpha}, \hat{\beta})$ to construct the adaptively weighted lasso penalized target function as

$$G_n(\alpha, \beta, \delta) = \sum_{i=1}^n \rho\left(y_i - B(\alpha^T x_i)^T \delta - \beta^T z_i\right) + \lambda_1 \sum_{j=1}^p \frac{|\alpha_j|}{|\hat{\alpha}_j|^2} + \lambda_2 \sum_{m=1}^q \frac{|\beta_m|}{|\hat{\beta}_m|^2},$$

For the given tuning parameters λ_1 and λ_2 , we obtain the penalized estimators by minimizing $G_n(\alpha, \beta, \delta)$ with respect to (α, β, δ) with constraints $\|\alpha\|=1$ and the first nonzero element of α is positive. For the sake of simplicity, we denote the resulting estimators by $\hat{\alpha}^{\lambda_1}$ and $\hat{\beta}^{\lambda_2}$.

III. MAIN RESULTS

To establish asymptotic results of variable selection for PLSIM with M-estimated, the following assumptions are required.

Assumption 1 The density function $f(t)$ of $\alpha_0^T X$ is bounded and is nonzero on $\tau = \{\alpha_0^T x, x \in C\}$, where C is a tight support of X .

Assumption 2 The r th derivative of $g(\bullet)$ is bounded for some $r \geq 2$.

Assumption 3 $\rho(\bullet)$ is convex and its first derivative $\psi(\bullet)$ satisfies: as $s \rightarrow 0$, there exists a positive constant b with $b < \infty$ such that $|E\psi(\varepsilon_1 + s) - bs| = o(s^2)$.

Assumption 4 There exist some functions $h_j(\bullet)$ at least with $(r-1)$ th bounded derivative over $\tau = \{\alpha_0^T x, x \in C\}$ such that for all $1 \leq i \leq n$ and $1 \leq j \leq p$,

$$g_{ij} = h_j(x_i^T \alpha_0) + u_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, p,$$

where $g_{ij} = g'(x_i^T \alpha_0)x_{ij}$ and u_{ij} are real sequence satisfying

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n U_i U_i^T = \Sigma_1, \quad U_i = (u_{i1}, \dots, u_{ip})^T.$$

Let $A_\alpha = \{j : \alpha_j \neq 0\}$ and $A_\beta = \{m : \beta_m \neq 0\}$. Without loss of generality, it is assumed that the correct model has regression coefficients $\alpha_0 = \begin{pmatrix} \alpha_0^1 \\ \alpha_0^2 \end{pmatrix}$ and $\beta_0 = \begin{pmatrix} \beta_0^1 \\ \beta_0^2 \end{pmatrix}$, where α_0^1 and β_0^1 are p_0 and q_0 nonzero components of α_0 and β_0 respectively, and α_0^2 and β_0^2 are $p - p_0$ and $q - q_0$ vectors with zeros. Thus, $A_\alpha = \{1, \dots, p_0\}$, $A_\beta = \{1, \dots, q_0\}$. In what follows, we show the adaptive lasso

penalized M-estimators enjoy the oracle properties.

Theorem 1 (Oracle property) Under the assumption 1~4, if $\lambda_i/\sqrt{n} \rightarrow 0$ and $\lambda_i\sqrt{n} \rightarrow \infty$ for $i=1,2$, then the adaptive lasso penalized estimators $\hat{\alpha}^{\lambda_1}$ and $\hat{\beta}^{\lambda_2}$ must satisfy:

A. Consistency in selection:

$$\lim_{n \rightarrow \infty} P\left(\left\{j : \hat{\alpha}_j^{\lambda_1} \neq 0\right\} = A_\alpha\right) = 1$$

and

$$\lim_{n \rightarrow \infty} P\left(\left\{m : \hat{\beta}_m^{\lambda_2} \neq 0\right\} = A_\beta\right) = 1;$$

B. Asymptotic normality:

$$\sqrt{n} \begin{pmatrix} \hat{\alpha}^{\lambda_1} - \alpha_0^1 \\ \hat{\beta}^{\lambda_2} - \beta_0^1 \end{pmatrix} \xrightarrow{L} N(0, \sigma^2 \tilde{\Sigma}^{-1})$$

Where $\hat{\alpha}^{\lambda_1}$ and $\hat{\beta}^{\lambda_2}$ denote the first p_0 and q_0 elements of $\hat{\alpha}^{\lambda_1}$ and $\hat{\beta}^{\lambda_2}$ respectively, and

$\sigma^2 = \frac{E(\psi(\varepsilon_1))^2}{b^2}$, $\tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_1 & \tilde{\Sigma}_3^T \\ \tilde{\Sigma}_3 & \tilde{\Sigma}_2 \end{pmatrix}$, the definition of $\tilde{\Sigma}_1, \tilde{\Sigma}_2, \tilde{\Sigma}_3$ is relegated to the Appendix.

IV. APPENDIX

we define \tilde{X} and \tilde{Z} in such a way that they consist of the first p_0 and q_0 elements of X and Z respectively. Denote

$$\tilde{v}_i = B'(\tilde{x}_i^T \alpha_0^1)^T \delta_0 \tilde{x}_i, \quad i = 1, \dots, n,$$

$$\tilde{V} = (\tilde{v}_1, \dots, \tilde{v}_n)^T,$$

$$\tilde{D}_n = \left(B(\tilde{x}_1^T \alpha_0^1), \dots, B(\tilde{x}_n^T \alpha_0^1) \right)^T,$$

$$\tilde{P} = \tilde{D}_n \left(\tilde{D}_n^T \tilde{D}_n \right)^{-1} \tilde{D}_n^T,$$

$$\tilde{T} = \tilde{X}^T \alpha_0^1,$$

$$\tilde{\zeta}(\tilde{t}) = (\tilde{\zeta}_1(\tilde{t}), \dots, \tilde{\zeta}_n(\tilde{t}))^T = E(\tilde{Z} | \tilde{T} = \tilde{t}),$$

$$\tilde{\Sigma}_2 = \text{Cov}(\tilde{Z} - \tilde{\zeta}(\tilde{T})), \quad \tilde{\Sigma}_3 = \lim_{n \rightarrow \infty} \frac{1}{n} \tilde{\Lambda},$$

Where

$$\tilde{\Lambda} = \left(\tilde{Z}^T (I - \tilde{P}) \tilde{Z} \right)^{-1} \tilde{Z}^T (I - \tilde{P}) \tilde{V} \left(\tilde{V}^T (I - \tilde{P}) \tilde{V} \right)^{-1},$$

$$H = \begin{pmatrix} \tilde{V}^T (I - \tilde{P}) \tilde{V} & \tilde{\Lambda}^T \\ \tilde{\Lambda} & \tilde{Z}^T (I - \tilde{P}) \tilde{Z} \end{pmatrix}.$$

Due to H is a symmetrical matrix, exist a matrix K , such that $H = K^T K$.

Lemma 1 If the Assumption 1~2 is satisfied, then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \tilde{V}^T (I - \tilde{P}) \tilde{V} = \tilde{\Sigma}_1,$$

where $\tilde{\Sigma}_1$ is obtained by only consider P_0 components of x in assumption 4.

Lemma 2 If the assumption 1~4 is satisfied, and $k_n \approx n^{1/2r+1}$, then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \tilde{Z}^T (I - \tilde{P}) \tilde{Z} = \tilde{\Sigma}_2.$$

Lemma 1 and lemma 2 are the special case of Gao (1997) and Zou (2014).

Proof of Theorem 1: We first prove the asymptotic normality part, let $\alpha = \alpha_0 + \frac{\xi_1}{\sqrt{n}}$, $\beta = \beta_0 + \frac{\xi_2}{\sqrt{n}}$, $\xi = (\xi_1^T, \xi_2^T)^T$, and

$$\begin{aligned} \Psi_n(\xi) = & \sum_{i=1}^n \rho(y_i - B \left(\sum_{j=1}^p x_j \left(\alpha_{0_j} + \frac{\xi_{1_j}}{\sqrt{n}} \right) \right)) \delta \\ & - \sum_{m=1}^q z_m \left(\beta_{0_m} + \frac{\xi_{2_m}}{\sqrt{n}} \right) + \lambda_{\alpha_n} \sum_{j=1}^p \hat{w}_{\alpha_j} \left| \alpha_{0_j} + \frac{\xi_{1_j}}{\sqrt{n}} \right| \\ & + \lambda_{\beta_n} \sum_{m=1}^q \hat{w}_{\beta_m} \left| \beta_{0_m} + \frac{\xi_{2_m}}{\sqrt{n}} \right| \end{aligned}$$

Where $\hat{w}_{\alpha} = |\hat{\alpha}|^{-2}$, $\hat{w}_{\beta} = |\hat{\beta}|^{-2}$. Let $\hat{\xi}^{(n)} = \arg \min \Psi_n(\xi)$. Note that $\hat{\xi}^{(n)} = \arg \min \Psi_n(\xi) - \Psi_n(0) = \Omega^{(n)}(\xi)$, following the proof of Zou (2014), we write $\Omega^{(n)}(\xi)$ as follows

$$\begin{aligned} \Omega^{(n)}(\xi) = & \frac{1}{2} \xi^T \frac{1}{n} H \xi - \frac{\varepsilon^T K}{\sqrt{n}} + o_p(1) \\ & + \frac{\lambda_{\alpha_n}}{\sqrt{n}} \sum_{j=1}^p \hat{w}_{\alpha_j} \sqrt{n} \left(\left| \alpha_{0_j} + \frac{\xi_{1_j}}{\sqrt{n}} \right| - |\alpha_{0_j}| \right) \\ & + \frac{\lambda_{\beta_n}}{\sqrt{n}} \sum_{m=1}^q \hat{w}_{\beta_m} \sqrt{n} \left(\left| \beta_{0_m} + \frac{\xi_{2_m}}{\sqrt{n}} \right| - |\beta_{0_m}| \right) \end{aligned}$$

We know that $\Sigma = \lim_{n \rightarrow \infty} H$ and $\frac{\varepsilon^T K}{\sqrt{n}} \underline{L} W = N(0, \sigma^2 \Sigma)$ by proof of Zou (2014). Now consider the limiting behavior of the third term, easy to see

$$\begin{aligned} & \frac{\lambda_{\alpha_n}}{\sqrt{n}} \hat{w}_{\alpha_j} \sqrt{n} \left(\left| \alpha_{0_j} + \frac{\xi_{1_j}}{\sqrt{n}} \right| - |\alpha_{0_j}| \right) \\ \xrightarrow{P} W(\alpha_j, \xi_{1_j}) = & \begin{cases} 0 & \alpha_{0_j} \neq 0 \\ 0 & \alpha_{0_j} = 0 \quad \xi_{1_j} = 0 \\ \infty & \alpha_{0_j} = 0 \quad \xi_{1_j} \neq 0 \end{cases}, \end{aligned}$$

and

$$\begin{aligned} & \frac{\lambda_{\beta_n}}{\sqrt{n}} \hat{w}_{\beta_m} \sqrt{n} \left(\left| \beta_{0_m} + \frac{\xi_{2_m}}{\sqrt{n}} \right| - |\beta_{0_m}| \right) \\ \xrightarrow{P} W(\beta_m, \xi_{2_m}) = & \begin{cases} 0 & \beta_{0_m} \neq 0 \\ 0 & \beta_{0_m} = 0 \quad \xi_{2_m} = 0 \\ \infty & \beta_{0_m} = 0 \quad \xi_{2_m} \neq 0 \end{cases}. \end{aligned}$$

Thus, again by Slutsky's theorem and Lemma 1 and Lemma 2, we see that $\Omega^{(n)}(\xi) \xrightarrow{L} \Omega(\xi)$ for every ξ , where

$$\Omega(\xi) = \begin{cases} \frac{1}{2} \xi_A^T \frac{1}{n} \tilde{\Sigma} \xi_A - \xi_A^T W_A, & \xi_{1_j} = 0, \forall j \notin A_{\alpha}; \xi_{2_m} = 0, \forall m \notin A_{\beta} \\ \infty & \text{otherwise} \end{cases}$$

where $A = \{(j, m) : j \in A_{\alpha}, m \in A_{\beta}\}$. $\Omega^{(n)}(\xi)$ is convex, and the unique minimum of $\Omega(\xi)$ is $(\tilde{\Sigma}^{-1} W_A, 0)^T$. Following the epi-convergence results, we have

$$\hat{\xi}_A^{(n)} \xrightarrow{L} \tilde{\Sigma}^{-1} W_A \quad \alpha \vee \delta \quad \hat{\xi}_{A^c}^{(n)} \xrightarrow{L} 0.$$

Finally, we observe that $W_A = N(0, \sigma^2 \tilde{\Sigma})$, so

•

and

2

$$\begin{aligned} & \frac{x_j^T B' \left(X^T \hat{\alpha}^{(n)} \right)^T \delta \psi \left(y - B \left(X^T \hat{\alpha}^{(n)} \right)^T \delta - Z^T \hat{\beta}^{(n)} \right)}{\sqrt{n}} \\ &= \frac{x_j^T \Sigma \sqrt{n} \left(\alpha_0 - \hat{\alpha}^{(n)} \right)}{n} + \frac{x_j^T \varepsilon}{\sqrt{n}} \end{aligned}$$

.

$$x_{j'}^T \mathcal{E} / \sqrt{n} \xrightarrow{L} N\left(0, \|x_{j'}^T\|^2 \sigma^2\right)$$
$$z_m^T \varepsilon / \sqrt{n} \xrightarrow{L} N\left(0, \|z_m^T\|^2 \sigma^2\right).$$

and

•

- [1] Y. Yu, D. Ruppert, "Penalized spline estimation for partially linear single-index models," *Journal of the American Statistical Association*, vol. 97, pp. 1042-1054, 2002.
- [2] Y. C. Xia, W. H. Härdle, "Semi-parametric estimation of partially linear single-index models," *Journal of Multivariate Analysis*, vol. 97, pp. 1162-1184, 2006.
- [3] H. Liang, R. Z. Li, and X. Liu, "Estimation and testing for partially linear single-index models," *The Annals of Statistics*, vol. 38, pp. 3811-3836, 2010.
- [4] P. Lai, S. Guo, "Variable selection for partially linear single-index model," *Journal of Computational Information Systems*, vol. 8, pp. 5917-5924, 2012.
- [5] Y. Z. Lv, R. Q. Zhang, and W. H. Zhao, "Quantile regression and variable selection of partial linear single-index model," *The Institute of Statistical Mathematics*, vol. 67, pp. 375-409, 2014.
- [6] H. M. Zhu, Z. K. Lv, K. M. Yu, "Robust variable selection in partially varying coefficient single-index model," *Journal of the Korean Statistical Society*, vol. 44, pp. 45-57, 2014.
- [7] Q. M. Zou, Z. Y. Zhu, "M-estimators for single-index model using B-spline," *Metrika*, vol. 77, pp. 225-246, 2014.
- [8] J. T. Gao, H. Liang, "Statistical inference in single-index and partially nonlinear models," *Annals of the Institute of Statistical Mathematics*, vol. 49, pp. 493-517, 1997.