Ann Inst Stat Math (2017) 69:169–187
DOI 10.1007/s10463-015-0544-7

Penalized estimation equation for an extended single-index model

Yongjin Li[1] · **Qingzhao Zhang**[2] · **Qihua Wang**[1,3]

Received: 27 September 2014 / Revised: 3 May 2015 / Published online: 8 October 2015

Abstract The single-index model is a useful extension of the linear regression model. Cui et al. (Ann Stat 39:1658–1688, 2011) proposed an estimating function method for the estimation of index vector in an extended single-index model (ESIM). Nevertheless, how to conduct variable selection for ESIM has not been studied. To solve this problem, we penalize the estimating equation with some types of penalty, such as smoothly clipped absolute deviation penalty and adaptive lasso penalty. Under some regularity conditions, the oracle property is established, i.e., the resulting estimator can be as efficient as the oracle estimator, thus we improve the explanatory ability and accuracy of estimator for the ESIM. A novel algorithm is proposed to solve the penalized estimating equation by combining quasi-Fisher scoring type algorithm and MM algorithm. Simulation study and real data application demonstrate the excellent performance of the proposed estimators.

Keywords Single-index model · Penalized estimating equations · Variable selection · Oracle property · Smoothly clipped absolute deviation · Adaptive lasso

1 Introduction

Consider the regression of a univariate response Y on a d-dimensional covariate vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$, where $\top$ denotes the transpose operator. When the dimension d

✉ Qihua Wang
qhwang@amss.ac.cn

[1] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[2] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[3] Institute of Statistical Science, Shenzhen University, Shenzhen 518006, China

is large, it suffers "curse of dimensionality" in nonparametric statistics. To address this problem, Härdle et al. (1993) proposed single-index model (SIM) $Y = g(\boldsymbol{\beta}^\top \mathbf{X}) + \epsilon$, where $\boldsymbol{\beta}$ is an unknown index parameter vector of interest, $g(\cdot)$ is an unknown link function and $E(\epsilon|\mathbf{X}) = 0$. It combines flexibility of modeling with interpretability of (linear) coefficients. From then on, single-index model has been applied to a variety of fields, such as econometrics, biostatistics, finance and so on, where high-dimensional regression models are often employed.

Most existing research about SIM focuses on efficient estimation of $\boldsymbol{\beta}$, see Härdle and Stoker (1989), Powell et al. (1989), Carroll et al. (1997), Hristache et al. (2001) and Xia et al. (2002), etc. However, much less has been done about its variable selection, which is important for any regression problems because ignoring any important predictor can lead to seriously biased results, whereas including spurious covariates can degrade the estimation efficiency substantially (Wang and Xia 2009). In linear regression model, many methods have been proposed to select variables and estimate their regression coefficients simultaneously, including least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li 2001), elastic net (EN) (Zou and Hastie 2005) and adaptive lasso (ALASSO) (Zou 2006). Donoho and Johnstone (1994) proposed the oracle property to measure the goodness of a variable selection scheme: if the method works asymptotically equivalent to the case as if the correct model was exactly known. Among these methods, SCAD and ALASSO enjoy oracle property, whereas LASSO and EN do not.

For single-index model, Kong and Xia (2007) proposed separated cross validation to exclude irrelevant covariates from single-index model. However, the method needs to compare all subsets of covariates, so it is computationally intensive and unstable. Zhu and Zhu (2009) and Zhu et al. (2011) followed the idea of sufficient dimension reduction and selected important variable in a class of single-index models via penalized least square, which requires that the covariate vector $\mathbf{X}$ satisfies the linearity condition, see Li (1991). Recently, Zeng et al. (2012) proposed a Lasso-type approach for estimation and variable selection in SIM by combining MAVE (Xia et al. 2002) and LASSO.

In this article, we consider the variable selection for an extended single-index model (ESIM), which only assumes the mean function and variance function of the response. Let $(Y_j, \mathbf{X}_j)$, $j = 1, \ldots, n$, denote the observed values with Y_j being the response variable and $\mathbf{X}_j$ being the d-dimensional explanatory variable. The mean function and variance function of Y_j are specified as follows:

$$E(Y_j|\mathbf{X}_j) = \mu\left\{g(\boldsymbol{\beta}^\top \mathbf{X}_j)\right\}, \; Var(Y_j|\mathbf{X}_j) = \sigma^2 V\left\{g(\boldsymbol{\beta}^\top \mathbf{X}_j)\right\}, \tag{1}$$

where $\mu(\cdot)$ is a known monotonic function, $V(\cdot)$ is a known covariance function, $g(\cdot)$ is an unknown univariate link function and $\boldsymbol{\beta}$ is an unknown index vector which belongs to the parameter space $\Theta = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top : \|\boldsymbol{\beta}\| = 1, \beta_1 > 0, \boldsymbol{\beta} \in \mathbb{R}^d\}$, where $\|\cdot\|$ denotes the l^2-norm. Cui et al. (2011) proposed an estimating function method (EFM) to estimate $\boldsymbol{\beta}$ in the ESIM (1) , and developed profile quasi-likelihood ratio test to test the significance of certain variables in the linear index. However, to the best of

our knowledge, no theoretical result or computational algorithm for variable selection in the ESIM has been studied.

In this paper, inspired by Johnson et al. (2008), we penalize the estimating equation of $\boldsymbol{\beta}$ to obtain a sparse estimator for ESIM, thus simultaneously select important predictors and estimate their regression coefficients. The main contributions of our work are as follows. First, we propose a penalized estimating equation approach. Second, we prove the oracle property of the proposed estimator and conduct a BIC-type criterion to select the regularization parameter, which can identify the true model consistently. Finally, a novel algorithm is proposed to solve the penalized estimating equation by combining quasi-Fisher scoring type algorithm and majorize–minimize (MM) algorithm.

The rest of this article is organized as follows. The variable selection procedures are proposed in Sect. 2. In Sect. 3, we study the asymptotic results of the method, mainly including the oracle property and selection of regularization parameter. The algorithm to solve the penalized estimating equation is presented in Sect. 4. In Sect. 5, some simulation studies are conducted to show the finite sample performance of the proposed methods. A real data case is analyzed in Sect. 6. All technical details are deferred to Appendix 7.

2 Methodology

The parameter space Θ requires that $\|\boldsymbol{\beta}\| = 1$ for the sake of identifiability. This assumption means that the true value of $\boldsymbol{\beta}$ is a boundary point on the unit sphere, and hence $g(\boldsymbol{\beta}^\top \mathbf{X})$ does not have a derivative at the point $\boldsymbol{\beta}$. By eliminating β_1, the parameter space Θ can be rearranged to a form:

$$\Theta = \left\{ \left((1 - \sum_{r=2}^{d} \beta_r^2)^{1/2}, \beta_2 \ldots, \beta_d \right)^\top : \sum_{r=2}^{d} \beta_r^2 < 1 \right\}.$$

Thus, we transform the boundary of a unit ball in $\mathbb{R}^d$ to the interior of a unit ball in $\mathbb{R}^{d-1}$. Denote $\boldsymbol{\beta}^{(1)} = (\beta_2, \ldots, \beta_d)^\top$ with the true value $\boldsymbol{\beta}_0^{(1)} = (\beta_{02}, \ldots, \beta_{0d})^\top$, then $g(\boldsymbol{\beta}^\top \mathbf{X})$ is infinitely differentiable with respect to $\boldsymbol{\beta}^{(1)}$ in a neighborhood of true parameter value $\boldsymbol{\beta}_0^{(1)}$.

Cui et al. (2011) proposed an EFM procedure to estimate $\boldsymbol{\beta}$, which can be regarded as a two-step estimation. Given $\boldsymbol{\beta}$, the estimators $\hat{g}(\cdot)$ and $\hat{g}'(\cdot)$ are obtained by solving the following kernel estimating equations with respect to α_0 and α_1:

$$\begin{aligned} &\sum_{j=1}^{n} K_{b_n}(\boldsymbol{\beta}^\top \mathbf{X}_j - t)\mu' \left\{ g_0(\boldsymbol{\beta}^\top \mathbf{X}_j) \right\} V^{-1} \left\{ g_0(\boldsymbol{\beta}^\top \mathbf{X}_j) \right\} \\ &\quad \times \left[Y_j - \mu \left\{ g_0(\boldsymbol{\beta}^\top \mathbf{X}_j) \right\} \right] = 0, \\ &\sum_{j=1}^{n} (\boldsymbol{\beta}^\top \mathbf{X}_j - t) K_{b_n}(\boldsymbol{\beta}^\top \mathbf{X}_j - t)\mu' \left\{ g_0(\boldsymbol{\beta}^\top \mathbf{X}_j) \right\} V^{-1} \left\{ g_0(\boldsymbol{\beta}^\top \mathbf{X}_j) \right\} \\ &\quad \times \left[Y_j - \mu \left\{ g_0(\boldsymbol{\beta}^\top \mathbf{X}_j) \right\} \right] = 0, \end{aligned} \tag{2}$$

where $g_0(\boldsymbol{\beta}^\top X) = \alpha_0 + \alpha_1(\boldsymbol{\beta}^\top \mathbf{X} - t)$ is local linear approximation for $g(\boldsymbol{\beta}^\top \mathbf{X})$ in a neighborhood of t (see Fan and Gijbels 1996), α_0 and α_1 are the estimators of $g(\cdot)$ and $g^{'}(\cdot)$ evaluating at t, respectively, $K_{b_n}(\cdot) = \frac{1}{b_n} K(\cdot/b_n)$ with $K(\cdot)$ being a symmetric kernel function and b_n being a bandwidth.

After obtaining the estimates $\hat{g}(\cdot)$ and $\hat{g}^{'}(\cdot)$, $\boldsymbol{\beta}$ can be estimated by the following estimating equation:

$$\hat{\mathbf{G}}(\boldsymbol{\beta}) = \sum_{j=1}^{n} \mathbf{J}^\top \hat{g}^{'}(\boldsymbol{\beta}^\top \mathbf{X}_j) \left\{ \mathbf{X}_j - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top \mathbf{X}_j) \right\} \rho_1 \left\{ \hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_j) \right\} \left[Y_j - \mu \left\{ \hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_j) \right\} \right] = 0, \tag{3}$$

where $\mathbf{J} = \partial \boldsymbol{\beta} / \partial \boldsymbol{\beta}^{(1)}$ is the Jacobian matrix of size $d \times (d-1)$ with

$$\mathbf{J} = \begin{pmatrix} -\boldsymbol{\beta}^{(1)\top} / \sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2} \\ \mathbf{I}_{d-1} \end{pmatrix},$$

and $\hat{\mathbf{h}}(t)$ is the local linear estimator for $\mathbf{h}(t) = E(\mathbf{X} | \boldsymbol{\beta}^\top \mathbf{X} = t)$,

$$\hat{\mathbf{h}}(t) = \sum_{i=1}^{n} b_i(t) \mathbf{X}_i / \sum_{i=1}^{n} b_i(t),$$

where $b_i(t) = K_{b_n}(\boldsymbol{\beta}^\top \mathbf{X}_i - t) \left\{ S_{n,2}(t) - (\boldsymbol{\beta}^\top \mathbf{X}_i - t) S_{n,1}(t) \right\}$, $S_{n,k} = \sum_{i=1}^{n} K_h (\boldsymbol{\beta}^\top \mathbf{X}_i - t)(\boldsymbol{\beta}^\top \mathbf{X}_i - t)^k$, $k = 1, 2$, and $\rho_l(z) = \left\{ \mu^{'}(z) \right\}^l V^{-1}(z)$, $l = 1, 2$.

Next, we consider variable selection for the extended single-index model. Without loss of generality, we suppose that the first explanatory variable is important, i.e., $\beta_1 \neq 0$. Otherwise, we can always obtain a root-n-consistent estimate using EFM method, and treat the covariate whose absolute value of estimated coefficient is the largest as the first explanatory variable. A penalized estimating equation is then defined as:

$$\hat{\mathbf{G}}^P(\boldsymbol{\beta}) = \hat{\mathbf{G}}(\boldsymbol{\beta}) - n \mathbf{p}_\lambda^{'} \left(|\boldsymbol{\beta}^{(1)}| \right) sgn(\boldsymbol{\beta}^{(1)}), \tag{4}$$

where $\mathbf{p}_\lambda^{'} \left(|\boldsymbol{\beta}^{(1)}| \right) = \left(p_{\lambda,2}^{'}(|\beta_2|), \ldots, p_{\lambda,d}^{'}(|\beta_d|) \right)^\top$, λ is the regularization parameter, $sgn(\cdot)$ is the sign function, and the second term of (4) is the componentwise product of $\mathbf{p}_\lambda^{'} \left(|\boldsymbol{\beta}^{(1)}| \right)$ and $sgn(\boldsymbol{\beta}^{(1)})$.

In this article, two penalty functions are considered: (1) the SCAD penalty (Fan and Li 2001), defined by

$$p_{\lambda,j}^{'}(|\theta|) = \lambda \left\{ I(|\theta| < \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| \geq \lambda) \right\}$$

for $a > 2$; (2) the ALASSO penalty (Zou 2006), $p_{\lambda,j}(|\theta|) = \lambda |\theta| \omega_j$, for a known data-driven weight ω_j. Zou (2006) used the $\omega_j = 1/|\hat{\beta}_j|^\gamma$ for some $\gamma > 0$ where $\hat{\beta}_j$ is the jth component of a root-n-consistent estimate of β. In this article, we use

the weight $\omega_j = 1/|\hat{\beta}_j|$, $j = 2, \ldots, d$, where $\hat{\boldsymbol{\beta}}^{(1)} = (\hat{\beta}_2, \ldots, \hat{\beta}_d)^\top$ refers to the $(d-1)$-dimensional vector of EFM estimator.

If the penalized estimating function (4) is continuous, the exact solution exists and we would obtain the sparse estimator of $\boldsymbol{\beta}^{(1)}$, denoted by $\hat{\boldsymbol{\beta}}_\lambda^{(1)} = (\hat{\beta}_{\lambda,2}, \ldots, \hat{\beta}_{\lambda,d})^\top$, then the estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_\lambda = \left(\sqrt{1 - ||\hat{\boldsymbol{\beta}}_\lambda^{(1)}||^2}, \hat{\boldsymbol{\beta}}_\lambda^{(1)\top}\right)^\top$. However, the penalty functions may not be continuous, hence the penalized estimating function (4) may be a discrete estimating function. Similar to Johnson et al. (2008), we introduce a zero-crossing estimating function to accommodate the discrete estimating function. Let $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$ be a zero-crossing to the penalized estimating function (4) if, for $j = 2, \ldots, d$,

$$\overline{\lim_{\epsilon \to 0^+}} \frac{1}{n} \hat{G}_j^P \left(\hat{\boldsymbol{\beta}}_\lambda^{(1)} + \epsilon \mathbf{e}_j\right) \hat{G}_j^P \left(\hat{\boldsymbol{\beta}}_\lambda^{(1)} - \epsilon \mathbf{e}_j\right) \leq 0,$$

where $\mathbf{e}_j$ is the jth canonical unit vector, ϵ is a small number and $\hat{G}_j^P(\cdot)$ is the jth component of $\hat{\mathbf{G}}^P(\cdot)$.

3 Theoretical properties

3.1 Basic theoretical properties

Recall that $\boldsymbol{\beta}_0^{(1)} = (\beta_{02}, \ldots, \beta_{0d})^\top$ denotes the true value of $\boldsymbol{\beta}^{(1)}$. Without loss of generality, suppose that $\beta_{0j} \neq 0$ for $j \leq s$ and $\beta_{0j} = 0$ for $j > s$. The set of nonzero entries in $\boldsymbol{\beta}_0^{(1)}$ is labeled as $\mathcal{A} = \{j : \beta_{0j} \neq 0, j = 2, \ldots, d\} = \{2, 3, \ldots, s\}$.

Theorem 1 *Assume that the estimating function* $\hat{\mathbf{G}}(\boldsymbol{\beta}) = 0$ *has a unique solution. Suppose the regularity conditions* C1–C8 *in Appendix* 7 *hold. If* $nh^6 \to 0$ *and* $nh^4 \to \infty$, *then there exists a root-n-consistent approximate zero-crossing of* $\hat{\mathbf{G}}^P(\boldsymbol{\beta})$, *i.e.,* $\hat{\boldsymbol{\beta}}^{(1)} = \boldsymbol{\beta}_0^{(1)} + O_p(n^{-1/2})$, *such that* $\hat{\boldsymbol{\beta}}^{(1)}$ *is an approximate zero-crossing of* $\hat{\mathbf{G}}^P(\boldsymbol{\beta})$.

Theorem 2 (Oracle property) *Assume that the estimating function* $\hat{\mathbf{G}}(\boldsymbol{\beta}) = 0$ *has a unique solution. Suppose the regularity conditions* C1–C6 *and* C8 *in the Appendix* 7 *hold. For any root-n-consistent approximate zero-crossing of* $\hat{\mathbf{G}}^P(\boldsymbol{\beta})$, *denoted by* $\hat{\boldsymbol{\beta}}_\lambda^{(1)} = (\hat{\beta}_{\lambda,2}, \ldots, \hat{\beta}_{\lambda,d})^\top$, *let* $\hat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}} = (\hat{\beta}_{\lambda,2}, \ldots, \hat{\beta}_{\lambda,s})^\top$ *and* $\boldsymbol{\beta}_{0,\mathcal{A}} = (\beta_{02}, \ldots, \beta_{0s})^\top$. *If* $nh^6 \to 0$ *and* $nh^4 \to \infty$, *we then have*

(a) *Sparsity:* $\lim_n P(\hat{\beta}_{\lambda,j} = 0 \text{ for } j > s) = 1$.
(b) *Asymptotic normality*:

$$\sqrt{n}(\mathbf{I}_{11} + \Sigma_{11}) \left\{\hat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}} - \boldsymbol{\beta}_{0,\mathcal{A}} + (\mathbf{I}_{11} + \Sigma_{11})^{-1}\mathbf{b}\right\} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}_{11}),$$

where $\mathbf{I}_{11}, \Sigma_{11}$ *are the first* $(s-1) \times (s-1)$ *submatrices of* $\mathbf{I} = \mathbf{J}^\top \Omega \mathbf{J}|_{\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}_0^{(1)}}$ *and* $diag\left\{-\mathbf{p}_\lambda''(|\boldsymbol{\beta}_0^{(1)}|)sgn(\boldsymbol{\beta}_0^{(1)})\right\}$, *and* $\mathbf{b} = -\Big(p_\lambda'(|\beta_{02}|)sgn(\beta_{02}), \ldots, p_\lambda'(|\beta_{0s}|)sgn$

$(\beta_{0s})\Big)^{\top}$. $\mathbf{J}$ *is the Jacobian matrix defined before, and*

$$\Omega = E\left[\left\{\mathbf{X}\mathbf{X}^{\top} - E(\mathbf{X}|\boldsymbol{\beta}^{\top}\mathbf{X})E(\mathbf{X}^{\top}|\boldsymbol{\beta}^{\top}\mathbf{X})\right\}\rho_2\left\{g(\boldsymbol{\beta}^{\top}\mathbf{X})\right\}\left\{g^{'}(\boldsymbol{\beta}^{\top}\mathbf{X})\right\}^2/\sigma^2\right].$$

Denote $\tilde{\mathcal{A}} = \{1\} \cup \mathcal{A}$. Let $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} = (\hat{\beta}_{\lambda,1} \quad \hat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}}^{\top})^{\top}$, where $\hat{\beta}_{\lambda,1} = \sqrt{1 - \sum_{i=2}^{d}\hat{\beta}_{\lambda,i}^2}$, and $\boldsymbol{\beta}_{\tilde{\mathcal{A}}} = (\beta_{01}, \quad \boldsymbol{\beta}_{0,\mathcal{A}}^{\top})^{\top} = (\beta_{01}, \quad \beta_{02}, \ldots, \beta_{0s})^{\top}$. Partition the matrix $\mathbf{J}$ into block matrix as follows:

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{pmatrix},$$

where $\mathbf{J}_{11}$ is a $s \times (s-1)$ submatrix. Using the multivariate delta method, we obtain the asymptotic normality of $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}$ as Corollary 1.

Corollary 1 *Under the conditions of Theorem 2, if* $\lim_n p_{\lambda}^{'}(|\theta|) = \lim_n p_{\lambda}^{''}(|\theta|) = 0$ *for* $\theta \neq 0$, *we have*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} - \boldsymbol{\beta}_{\tilde{\mathcal{A}}}\right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{J_{11}}\mathbf{I}_{11}^{-1}\mathbf{J_{11}}^{\top}).$$

Remark 1 The asymptotic variance of $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}$ is $\frac{1}{n}\mathbf{J_{11}}\mathbf{I}_{11}^{-1}\mathbf{J_{11}}^{\top}$, which is the same as that of oracle estimator. The results of Corollary 1 and Theorem 2 (a) show the oracle property of the proposed estimator.

3.2 Regularization parameter selection

We need to choose (a, λ) for SCAD penalty and λ for ALASSO penalty. Fan and Li (2001) showed that the choice of $a = 3.7$ performs well in a variety of situations and we use their suggestion throughout our numerical analysis. Hence, only the regularization parameter λ should be appropriately selected. Some selection criterions such as generalized cross-validation (GCV) (Tibshirani 1996; Fan and Li 2001), the Akaike information criterion, and the Bayes information criterion (BIC) are used to choose regularization parameter. In our practical implementation, we construct a BIC-type criterion similar to Wang and Leng (2007) as follows:

$$\mathbf{BIC}_{\lambda} = \left(\hat{\boldsymbol{\beta}}_{\lambda}^{(1)} - \tilde{\boldsymbol{\beta}}^{(1)}\right)^{\top}\hat{\mathbf{I}}^{-1}\left(\hat{\boldsymbol{\beta}}_{\lambda}^{(1)} - \tilde{\boldsymbol{\beta}}^{(1)}\right) + df_{\lambda}\log n/n, \tag{5}$$

where $\tilde{\boldsymbol{\beta}}^{(1)}$ is the solution of estimating Eq. (3), df_{λ} is the number of non-zero coefficients of $\hat{\boldsymbol{\beta}}_{\lambda}^{(1)}$, a simple estimator for degrees of freedom, and $\hat{\mathbf{I}}$ is the plug in estimator

of $\mathbf{I}$, i.e., $\hat{\mathbf{I}} = \hat{\mathbf{J}}^\top \hat{\Omega} \hat{\mathbf{J}}$, where $\hat{\mathbf{J}} = \mathbf{J}|_{\boldsymbol{\beta}^{(1)}=\hat{\boldsymbol{\beta}}_\lambda^{(1)}}$ and

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{X}_i\mathbf{X}_i^\top - \hat{h}(\hat{\boldsymbol{\beta}}_\lambda^\top \mathbf{X}_i)\hat{h}^\top(\hat{\boldsymbol{\beta}}_\lambda^\top \mathbf{X}_i)\right]\rho_2\left\{\hat{g}(\hat{\boldsymbol{\beta}}_\lambda^\top \mathbf{X}_i)\right\}\left\{\hat{g}'(\hat{\boldsymbol{\beta}}_\lambda^\top \mathbf{X}_i)\right\}^2/\hat{\sigma}^2,$$

with $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\xi}_i - \frac{1}{n}\sum_{j=1}^{n}\hat{\xi}_j\right)^2$ and $\hat{\xi}_i = \left(Y_i - \mu\left\{\hat{g}(\hat{\beta}_\lambda^\top \mathbf{X}_i)\right\}\right)/\sqrt{V\left(\hat{g}(\hat{\beta}_\lambda^\top \mathbf{X}_i)\right)}$ for $i = 1, \ldots, n$.

Minimizing $\mathbf{BIC}_\lambda$, we obtain the optimal regularization parameter. Similar to Wang and Leng (2007), it is easy to conclude that this BIC-type criterion can identify the true model consistently.

4 Algorithm

Solving the joint estimating Eqs. (2) and (4) poses some interesting challenges. Treating $\boldsymbol{\beta}^\top \mathbf{X}$ as a new predictor (with given $\boldsymbol{\beta}$), (2) gives us $\hat{g}$, $\hat{g}'$ as in Fan et al. (1995). Thus, we focus on the estimating Eq. (4). Cui et al. (2011) proposed a fixed-point iterative algorithm to solve the unpenalized estimating Eq. (3). However, their algorithm is hard to be extended to the penalized estimating Eq. (4). Inspired by Xu and Zhu (2012), we propose a quasi-Fisher scoring type algorithm to solve the Eq. (3), which is given as follows:

Set initial $\boldsymbol{\beta}_0 = (1/\sqrt{d}, 1/\sqrt{d}, \ldots, 1/\sqrt{d})^\top$ such that $\|\boldsymbol{\beta}_0\| = 1$,
Repeat for $k = 0, 1, 2, \ldots$

1. Obtain $\hat{g}(\boldsymbol{\beta}_k^\top \mathbf{X}_i)$, $\hat{g}'(\boldsymbol{\beta}_k^\top \mathbf{X}_i)$ for $i = 1, \ldots, n$ from (2).
2. Update $\tilde{\boldsymbol{\beta}}_{k+1} = \boldsymbol{\beta}_k + \mathbf{J}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_k}\left[\dot{\hat{\boldsymbol{G}}}(\boldsymbol{\beta}_k)\right]^{-1}\hat{\boldsymbol{G}}(\boldsymbol{\beta}_k)$, where

$$\dot{\hat{\boldsymbol{G}}}(\boldsymbol{\beta}) = \sum_{j=1}^{n}\mathbf{J}^\top \hat{g}'(\boldsymbol{\beta}^\top \mathbf{X}_j)\left\{\mathbf{X}_j - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top \mathbf{X}_j)\right\}\rho_2\left\{\hat{g}(\boldsymbol{\beta}^\top \mathbf{X}_j)\right\}\left\{\mathbf{X}_j - \hat{\mathbf{h}}(\boldsymbol{\beta}^\top \mathbf{X}_j)\right\}^\top \times \hat{g}'(\boldsymbol{\beta}^\top \mathbf{X}_j)\mathbf{J}.$$

3. Set $\boldsymbol{\beta}_{k+1} = \tilde{\boldsymbol{\beta}}_{k+1}/\|\tilde{\boldsymbol{\beta}}_{k+1}\|$.

Until $\max_{1\le l\le d}|\beta_{k,l} - \beta_{k-1,l}| < tol$, where $\beta_{k,l}$ is the lth component of $\boldsymbol{\beta}_k$ and tol is a given tolerance.

We can easily extend the above algorithm to the penalized estimating Eq. (4). To solve (4), we combine the above algorithm with a majorize–minimize (MM) algorithm (Hunter and Li 2005) as follows:

Set initial $\boldsymbol{\beta}_0 = (1/\sqrt{d}, 1/\sqrt{d}, \ldots, 1/\sqrt{d})^\top$ such that $\|\boldsymbol{\beta}_0\| = 1$,
Repeat for $k = 0, 1, 2, \ldots$

1. Obtain $\hat{g}(\boldsymbol{\beta}_k^\top \mathbf{X}_i)$, $\hat{g}'(\boldsymbol{\beta}_k^\top \mathbf{X}_i)$ for $i = 1, \ldots, n$ from (2).

2. Update $\tilde{\boldsymbol{\beta}}_{k+1} = \boldsymbol{\beta}_k + \mathbf{J}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_k} \left[\dot{\hat{\boldsymbol{G}}}(\boldsymbol{\beta}_k) + n\dot{\boldsymbol{P}}(\boldsymbol{\beta}_k)\right]^{-1} \hat{\boldsymbol{G}}^P(\boldsymbol{\beta}_k)$, where

$$\dot{\boldsymbol{P}}(\boldsymbol{\beta}) = \text{diag}\left\{p'_{\lambda,2}(|\beta_2|)/(\varepsilon + |\beta_2|), \ldots, p'_{\lambda,d}(|\beta_d|)/(\varepsilon + |\beta_d|)\right\}$$

for ε a small number ($\varepsilon = 10^{-6}$ in our example).
3. Set $\boldsymbol{\beta}_{k+1} = \tilde{\boldsymbol{\beta}}_{k+1}/\|\tilde{\boldsymbol{\beta}}_{k+1}\|$.

Until $\max_{1\le l\le d} |\beta_{k,l} - \beta_{k-1,l}| < tol$, where $\beta_{k,l}$ is the lth component of $\boldsymbol{\beta}_k$ and tol is a given tolerance.

5 Simulation studies

To demonstrate the finite sample performance of the proposed method, we consider some extended single-index models. Fore each simulated data, the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$ is used; the bandwidth is selected via generalized cross validation (GCV). To summarize the variable selection results, similar to Wang and Xia (2009), we consider three different situations. If the resulting model is exactly the same as the true model, we denote it as the correctly fitted model. Whenever the estimated model misses at least one relevant predictor, we denote it as the underfitted model. Whenever the estimated model includes at least one irrelevant predictor but does not miss any relevant one, we denote it as the overfitted model.

Furthermore, we consider another criterion of variable selection performances using G-means, $\text{G} = \sqrt{sensitivity \times specificity}$, which was also considered in Jeng and Daye (2011), where *sensitivity* is the true-positive rate and *specificity* is the true-negative rate. Denote TP to be number of true positive, i.e., nonzero coefficient correctly estimated as nonzero, and FP to be number of false positive, i.e., zero coefficient incorrectly estimated as nonzero, then $sensitivity(Sen) = \text{TP}/s$, $specificity(Spe) = (d - s - \text{FP})/(d - s)$, where s is the number of relevant predictors and d is the dimension of predictors. A value close to 1 for G indicates good selection, whereas a value close to 0 implies that few true predictor or too many irrelevant variables are selected, or both.

Denote $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_d)^\top$ one of the EFM, SCAD-EFM and ALASSO-EFM estimators (SCAD-EFM and ALASSO-EFM are the proposed estimators of the variable selection methods, depending on the penalty). To evaluate the estimation accuracy of the proposed variable selection methods, we consider the absolute bias (AB), which is defined as:

$$\text{AB} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{d}\sum_{j=1}^{d}|\hat{\beta}_j^{[i]} - \beta_{0j}|\right],$$

where N is the number of simulation replications, $\hat{\beta}_j^{[i]}$ is the jth component of $\hat{\boldsymbol{\beta}}^{[i]}$ with $\hat{\boldsymbol{\beta}}^{[i]}$ being $\hat{\boldsymbol{\beta}}$ obtained in the ith simulation. To investigate the oracle property

of SCAD and ALASSO methods, we consider the following relative estimation error (REE):

$$\text{REE} = 100\,\% \times \frac{\sum_{j=1}^{d} |\hat{\beta}_j - \beta_{0j}|}{\sum_{j=1}^{d} |\tilde{\beta}_j - \beta_{0j}|},$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_d)^\top$ is the oracle estimator. Thus, the corresponding REE value measures the estimation accuracy of each method to the oracle estimator. For each example below, the median of REE values (denote as MREE) is calculated.

In the following examples, we numerically compare the proposed variable selection methods with EFM. All simulations are conducted using MATLAB codes. We consider two cases in each simulation example. In Case 1, the dimension of covariates is taken $d = 10$, while taken $d = 20$ in Case 2. For each example, we simulated 500 data sets.

Example 1 (A simple model) We firstly consider the following simple single-index model:

$$Y = (\boldsymbol{\beta}^\top \mathbf{X})^2 + \epsilon. \tag{6}$$

The underlying coefficients are assumed to be $\boldsymbol{\beta} = (2, 1, 0, \ldots, 0)/\sqrt{5}$; $\mathbf{X}$ is generated from $N_d(2, \Sigma)$, where $\Sigma = (\sigma_{i,j})_{1 \le i,j \le d}$ with $\sigma_{i,j} = 0.8^{|i-j|}$, and $\epsilon \sim N(0, 0.2^2)$. A similar modeling setup was also used in Example 3 of Cui et al. (2011). The simulated results are given in Table 1 with sample size $n = 100, 200$.

As we can see from Table 1, the performances of SCAD-EFM and ALASSO-EFM are similar. The percentage of the correctly fitted models is 100 %, which confirms that our BIC criterion (5) can indeed identify the true model consistently. The absolute bias (AB) is much smaller than that of EFM. The MREE approaches 100 % when n is relatively large. The AB and MREE of EFM are much larger than that of SCAD-EFM and ALASSO-EFM, especially when n is small or d is large. The proposed approaches improve the explanatory ability and accuracy of estimator for the extended single-index model.

Example 2 (An oscillating function model) In this example, we consider the following single-index model with the link function $g(\cdot)$ and oscillating function:

$$Y = \sin\left(\frac{\pi}{2} \cdot \boldsymbol{\beta}^\top \mathbf{X}\right) + \epsilon,$$

where $\boldsymbol{\beta}$, $\mathbf{X}$ and ϵ are set in the same way as in Example 1. Table 2 reports the simulation results with sample size $n = 100, 200, 400$.

In this example, when n is relative smaller ($n = 100$), the percentage of the correctly fitted models is 90 % ($d = 10$) or 81 % ($d = 20$), mainly because of the missing of relevant predictors. However, the percentage steadily increases as the sample size increases, and approaches 100 % quickly. The MREE of ALASSO-EFM is larger than SCAD-EFM, especially when the sample size is small, which is due to the use of EFM estimator as the weights of penalty for ALASSO. Nevertheless, it decreases towards

Table 1 The simulation results of Example 1

n	Method	Percentage of models			G-means					Estimation accuracy	
		Correctly	Over	Under	TP	FP	Sen	Spe	G	AB (10^{-4})	MREE (%)
Case 1: $d = 10$											
100	SCAD	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	2.82	100.03
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	3.02	101.42
	EFM	–	–	–	–	–	–	–	–	15.88	744.35
200	SCAD	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	1.19	100.05
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	1.25	101.43
	EFM	–	–	–	–	–	–	–	–	5.25	543.73
Case 2: $d = 20$											
100	SCAD	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	1.65	100.05
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	2.20	137.67
	EFM	–	–	–	–	–	–	–	–	19.68	1988.77
200	SCAD	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	0.65	100.02
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	0.66	101.99
	EFM	–	–	–	–	–	–	–	–	5.74	1165.36

Table 2 The simulation results of Example 2

n	Method	Percentage of models			G-means					Estimation accuracy	
		Correctly	Over	Under	TP	FP	Sen	Spe	G	AB (10^{-4})	MREE (%)
Case 1: $d = 10$											
100	SCAD	0.90	0.00	0.10	1.90	0.00	0.95	1.00	0.97	88.69	100.03
	ALASSO	0.92	0.00	0.08	1.92	0.00	0.96	1.00	0.98	107.45	154.07
	EFM	–	–	–	–	–	–	–	–	183.64	885.24
200	SCAD	0.99	0.00	0.01	1.99	0.00	1.00	1.00	1.00	16.90	100.02
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	14.27	112.93
	EFM	–	–	–	–	–	–	–	–	85.28	843.36
400	SCAD	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	5.56	99.99
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	6.42	103.70
	EFM	–	–	–	–	–	–	–	–	39.40	795.80
Case 2: $d = 20$											
100	SCAD	0.81	0.00	0.19	1.81	0.00	0.91	1.00	0.95	70.19	100.05
	ALASSO	0.86	0.00	0.14	1.86	0.00	0.93	1.00	0.96	88.22	439.50
	EFM	–	–	–	–	–	–	–	–	221.74	2476.41
200	SCAD	0.90	0.00	0.10	1.90	0.00	0.95	1.00	0.97	40.71	100.04
	ALASSO	0.95	0.00	0.05	1.95	0.00	0.97	1.00	0.99	47.57	175.40
	EFM	–	–	–	–	–	–	–	–	97.79	1985.47
400	SCAD	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	2.80	100.00
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	3.50	105.23
	EFM	–	–	–	–	–	–	–	–	44.21	1799.46

100 % as the sample size increases. Similar to Example 1, the AB and MREE of EFM are much larger than those of SCAD-EFM and ALASSO-EFM, especially when n is small or d is large.

Example 3 (An heterogeneous error model) To illustrate the adaptivity of our algorithm to heterogeneous error model, we consider the regression model (6) in Example 1, where the true parameter is $\boldsymbol{\beta} = (2/\sqrt{5}, 1/\sqrt{5}, 0, \ldots, 0)$; $\mathbf{X}$ is generated from $N_d(2, I)$, and $\varepsilon \sim N\left(0, \exp(\frac{2X_1+X_2}{7})\right)$. The simulation results are summarized in Table 3 with simple size $n = 100, 200, 400$. Similar conclusion to the former examples can be made, which shows that our algorithm is also attractive to the heterogeneous cases.

Example 4 (A binary response model) The ESIM includes a series of common models, especially, Y is discrete, for example $Y = 0, 1$. Hence, we consider the follow binary response model:

$$P(Y = 1|\mathbf{X}) = \mu\{g(\boldsymbol{\beta}^\top \mathbf{X})\} = \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}/[1 + \exp\{g(\boldsymbol{\beta}^\top \mathbf{X})\}], \tag{7}$$

where $g(\boldsymbol{\beta}^\top \mathbf{X}) = \exp(5\boldsymbol{\beta}^\top \mathbf{X} - 2)/(1 + \exp(5\boldsymbol{\beta}^\top \mathbf{X} - 3)) - 1.5$. The true parameter is $\boldsymbol{\beta} = (2/\sqrt{5}, 1/\sqrt{5}, 0, \ldots, 0)$ and $X_1, X_2, \ldots, X_d$ are independent and identical distribution from $U(-2, 2)$. Similar designs for generalized partially linear single-index models are assumed in Kane et al. (2004), and Cui et al. (2011) also considered this model. Here, the sample size takes 500 and 1000, which is different from the other examples due to complexity of the this example. For this example, 250 replications are simulated and the results are displayed in Table 4. Similar conclusion to Example 1 can be made, which shows that the proposed method works well for the discrete response.

6 Real data analysis

Here, we consider the body fat data, which is available at http://lib.stat.cmu.edu/datasets/bodyfat. A variety of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. However, accurate measurement of body fat is inconvenient/costly and it is desirable to have easy methods of estimating body fat that are not inconvenient/costly. The data of 252 men contain thirteen baseline predictors. Body mass index (BMI) is a useful measure of body fat based on height and weight. Hence we calculate BMI for each sample and omit the predictors height and weight. Then, we have twelve baseline predictors X: age (x_1), BMI (x_2), circumference of the skinfold measurements neck (x_3), chest (x_4), 2 abdomen (x_5), hip (x_6), thigh (x_7), knee (x_8), ankle (x_9), biceps (x_{10}), forearm (x_{11}) and wrist (x_{12}). The response Y is the percentage of body fat. We aim at building a predictive model to relate the response to the predictors and meanwhile selecting important predictors. We delete possible outliers to a sample of size 244.

We compare the performance of EFM with the proposed variable selection methods SCAD-EFM and ALASSO-EFM. The estimated coefficients $\hat{\boldsymbol{\beta}}$ and adjusted R squared

Table 3 The simulation results of Example 3

n	Method	Percentage of models			G-means					Estimation accuracy	
		Correctly	Over	Under	TP	FP	Sen	Spe	G	AB (10^{-4})	MREE (%)
Case 1: $d = 10$											
100	SCAD	0.87	0.13	0.00	2.00	0.16	1.00	0.98	0.99	90.58	197.24
	ALASSO	0.97	0.03	0.00	2.00	0.03	1.00	1.00	1.00	91.15	141.27
	EFM	–	–	–	–	–	–	–	–	358.23	813.47
200	SCAD	0.98	0.02	0.00	2.00	0.03	1.00	1.00	1.00	50.28	134.29
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	45.19	131.54
	EFM	–	–	–	–	–	–	–	–	249.72	958.62
400	SCAD	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	27.61	100.01
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	25.80	108.88
	EFM	–	–	–	–	–	–	–	–	169.11	778.28
Case 2: $d = 20$											
100	SCAD	0.75	0.25	0.00	2.00	0.40	1.00	0.98	0.99	59.97	264.05
	ALASSO	0.94	0.05	0.01	1.99	0.09	1.00	0.99	1.00	65.67	215.86
	EFM	–	–	–	–	–	–	–	–	397.9	1809.44
200	SCAD	0.94	0.06	0.00	2.00	0.10	1.00	0.99	1.00	32.88	178.66
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	44.88	223.25
	EFM	–	–	–	–	–	–	–	–	257.85	1528.59
400	SCAD	0.99	0.01	0.00	2.00	0.01	1.00	1.00	1.00	14.94	100.02
	ALASSO	1.00	0.00	0.00	2.00	0.00	1.00	1.00	1.00	15.35	119.69
	EFM	–	–	–	–	–	–	–	–	163.81	1600.98

Table 4 The simulation results of Example 4

n	Method	Percentage of models			G-means					Estimation accuracy	
		Correctly	Over	Under	TP	FP	Sen	Spe	G	AB (10^{-4})	MREE (%)
Case 1: $d = 10$											
500	SCAD	0.69	0.31	0.00	2.00	0.36	1.00	0.95	0.98	136.16	120.24
	ALASSO	0.83	0.17	0.00	2.00	0.19	1.00	0.98	0.99	114.89	150.30
	EFM	–	–	–	–	–	–	–	–	431.46	1026.87
1000	SCAD	0.97	0.03	0.00	2.00	0.03	1.00	1.00	1.00	47.37	100.00
	ALASSO	0.98	0.02	0.00	2.00	0.02	1.00	1.00	1.00	54.94	111.04
	EFM	–	–	–	–	–	–	–	–	357.92	654.78
Case 2: $d = 20$											
500	SCAD	0.57	0.43	0.00	2.00	0.71	1.00	0.96	0.98	103.63	209.07
	ALASSO	0.59	0.41	0.00	2.00	0.68	1.00	0.96	0.98	100.85	199.26
	EFM	–	–	–	–	–	–	–	–	484.58	1953.63
1000	SCAD	0.89	0.11	0.00	2.00	0.11	1.00	0.99	1.00	42.01	109.93
	ALASSO	0.89	0.11	0.00	2.00	0.11	1.00	0.99	1.00	52.06	158.19
	EFM	–	–	–	–	–	–	–	–	382.12	1014.44

Table 5 The estimates from the body fat data

	SCAD-EFM $\hat{\boldsymbol{\beta}}$	ALASSO-EFM $\hat{\boldsymbol{\beta}}$	EFM $\hat{\boldsymbol{\beta}}$
x_1	0	0	0.1085
x_2	0.1929	0.1968	0.1882
x_3	0	0	−0.1232
x_4	0	0	−0.1348
x_5	0.9539	0.9522	0.9207
x_6	−0.1027	−0.1029	−0.1370
x_7	0	0	0.0433
x_8	0	0	−0.0253
x_9	0	0	0.0183
x_{10}	0	0	0.0372
x_{11}	0	0	0.0470
x_{12}	−0.2053	−0.2094	−0.215
Adjusted R^2	0.7423	0.7423	0.7403
Mean MAPE	2.9253	2.9921	3.1125

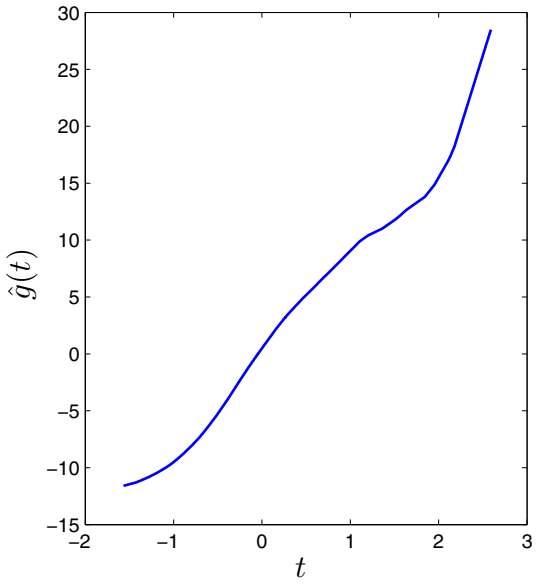

Fig. 1 The estimation of link function of SCAD-EFM

of the above three methods are listed in Table 5. The estimate of link function $g(\cdot)$ is similar among three methods, hence we only show that of SCAD-EFM in Fig. 1. The figure shows that link function is strictly increasing, approximately linear but not exactly. Both SCAD and ALASSO select parsimonious model (only select x_2, x_5, x_6 and x_{12} as the important predictors), while having competitive adjusted R squared compared with the full model (EFM).

To assess the prediction power of the proposed estimators, we use the following procedure. The data are randomly split into two separate groups with equal observations, i.e., 122 samples are used to fit the model, while the remaining observations are used to evaluate the predictive ability of the selected model. The prediction performance is measured by the median absolute prediction error (MAPE). We apply the procedure 200 times, and the mean MAPE for each methods is reported in Table 5. It appears that the model chosen by SCAD-EFM and ALASSO-EFM has lower mean MPAE than the full model. In summary, from the results of Table 5, we can see that the proposed methods improve explanatory ability and give better predictions.

Zhang and Wang (2013) also considered this body fat data and used the a semi-parameter model $Y = \exp\left(\mathbf{X}^\top \boldsymbol{\beta} + h(U)\right)\varepsilon$ to fit the data, where $h(\cdot)$ is a unknown nonlinear function, $U = x_4$ and $\mathbf{X}$ is the predictor vector containing $x_1 - x_{12}$ except for x_4. Both their method and the proposed SCAD-EFM, ALASSO-EFM select x_5 (2 abdomen) as the most important predictor, which means that it has the largest estimated coefficient. Nevertheless, the mean MAPE of their model is 2.9930, which is a little larger than the proposed SCAD-EFM (2.9253) and ALASSO-EFM (2.9921). The reason may be that the estimate of link function $g(\cdot)$ is approximately linear but not exactly, while they postulated an exponential function.

7 Appendix

Regularity conditions

Before we present the proofs of the theorems, we first introduce some regularity conditions.

(C1) $\mu(\cdot)$, $V(\cdot)$, $g(\cdot)$ and $\mathbf{h}(\cdot) = E(\mathbf{X}|\boldsymbol{\beta}^\top\mathbf{X} = \cdot)$ have bounded and continuous derivatives of order two. $V(\cdot)$ is uniformly bounded and bounded away from 0.

(C2) Let $q(z, y) = \mu'(z)V^{-1}(z)(y - \mu(z))$. Assume that $\partial q(z, y)/\partial z < 0$ for $z \in \mathbb{R}$ and y in the range of the response variable.

(C3) Define the block partition of matrix Ω as follows:

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

where Ω_{11} is a positive constant, Ω_{12} is a $(d-1)$-dimensional row vector, Ω_{21} is a $(d-1)$-dimensional column vector and Ω_{22} is a $(d-1)\times(d-1)$ nonnegative definite matrix. The largest eigenvalues of Ω_{22} are bounded away from infinity.

(C4) The density function of $\mathbf{X}$ has continuous derivative of order two on its support. The density function $f_{\boldsymbol{\beta}^\top\mathbf{X}}(\boldsymbol{\beta}^\top\mathbf{X})$ of random variable $\boldsymbol{\beta}^\top\mathbf{X}$ is bounded away from 0 on T_β and satisfies the Lipschitz condition of order 1 on T_β, where $T_\beta = \{\boldsymbol{\beta}^\top\mathbf{X} : \mathbf{X} \in T\}$ and T is the compact support set of $\mathbf{X}$.

(C5) Let $Q^*(\boldsymbol{\beta}) = \int Q\left[\mu\left\{g(\boldsymbol{\beta}^\top\mathbf{x})\right\}, y\right] f(y|\boldsymbol{\beta}_0^\top\mathbf{x}) f_{\beta^\top\mathbf{x}}(\boldsymbol{\beta}_0^\top\mathbf{x}) dy d(\boldsymbol{\beta}_0^\top\mathbf{x})$ with $\boldsymbol{\beta}_0$ denoting the true parameter value and $Q[\mu, y] = \int_u^y \frac{s-y}{V\{\mu^{-1}(s)\}} ds$. Here, $f(y|\boldsymbol{\beta}_0^\top\mathbf{x})$ is the conditional density function of Y given $\boldsymbol{\beta}_0^\top\mathbf{X} = \boldsymbol{\beta}_0^\top\mathbf{x}$. Assume

that $Q^*(\boldsymbol{\beta})$ has a unique maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$,

$$E\left[\sup_{\boldsymbol{\beta}^{(1)}}\sup_{\boldsymbol{\beta}^\top\mathbf{X}}\left|\mu'\left\{g(\boldsymbol{\beta}^\top\mathbf{X})\right\}V^{-1}\{g(\boldsymbol{\beta}^\top\mathbf{X})\}\left[Y-\mu\{g(\boldsymbol{\beta}^\top\mathbf{X})\}\right]\right|^2\right]<\infty,$$

and $E\|\mathbf{X}\|^2 < \infty$.

(C6) The kernel function $K(\cdot)$ is a bounded and symmetric density function with a bounded derivative, and satisfies

$$\int_{-\infty}^{+\infty}|t|^2K(t)\mathrm{d}t<\infty.$$

(C7) For any fixed $\theta \neq 0$, $\lim_n \sqrt{n}p_\lambda'(|\theta|) = 0$ and $\lim_n p_\lambda''(|\theta|) = 0$.

(C8) For any positive constant C, $\lim \sqrt{n}\inf_{|\theta|<Cn^{-1/2}} p_\lambda'(|\theta|) \to \infty$.

Remark 2 Conditions (C1)–(C6) are some regularity conditions for the extended single-index models, which are similar to Cui et al. (2011). Conditions (C7) and (C8) are the key for obtaining the oracle property. For SCAD penalty, if we choose appropriate regularization parameter such that $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ as $n \to \infty$, then condition (C7) holds because $\sqrt{n}p_\lambda'(|\theta|) = p_\lambda''(|\theta|) = 0$ as $n \to \infty$ for $\theta \neq 0$ and $\sqrt{n}\inf_{|\theta|<Cn^{-1/2}} p_\lambda'(|\theta|) = \sqrt{n}\lambda$. For ALASSO penalty, Conditions (C7) and (C8) are also hold by choosing the appropriate regularization parameter (see Johnson et al. 2008).

Proof of Theorem 1 Under the regularity conditions C1–C6 and $nh^6 \to 0$ and $nh^4 \to \infty$, from the proof of Theorem 2.1 in Cui et al. (2011), we have

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}^{(1)}-\boldsymbol{\beta}_0^{(1)}\right)=\frac{1}{\sqrt{n}}\mathbf{I}^+\hat{\mathbf{G}}(\boldsymbol{\beta}_0)+o_p(1),\tag{8}$$

where $\mathbf{I}_{(d-1)\times(d-1)} = \mathbf{J}^\top\Omega\mathbf{J}|_{\boldsymbol{\beta}^{(1)}=\boldsymbol{\beta}_0^{(1)}}$, $\mathbf{I}^+$ denote the Moore–Penrose inverse of the matrix $\mathbf{I}$ and $\tilde{\boldsymbol{\beta}}^{(1)}$ is the solution of estimating Eq. (3). Let $\breve{\boldsymbol{\beta}}^{(1)} = \left(\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{(1)\top} \quad \mathbf{0}_{(d-s)\times 1}^\top\right)^\top$, where $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{(1)} = \boldsymbol{\beta}_{0\mathcal{A}}^{(1)} + n^{-1}\mathbf{I}_{11}^{-1}\hat{\mathbf{G}}_{\mathcal{A}}(\boldsymbol{\beta}_0)$, with $\boldsymbol{\beta}_{0\mathcal{A}}^{(1)}$ and $\hat{\mathbf{G}}_{\mathcal{A}}(\boldsymbol{\beta}_0)$ being the first $s-1$ component of $\boldsymbol{\beta}_0^{(1)}$ and $\hat{\mathbf{G}}(\boldsymbol{\beta}_0)$, respectively, and $\mathbf{I}_{11}$ is the first $(s-1)\times(s-1)$ submatrices of $\mathbf{I}$. Thus, we have $\breve{\boldsymbol{\beta}}^{(1)} = \boldsymbol{\beta}_0^{(1)} + O_P(n^{-1/2})$. Under the condition (C7) with $\lim \sqrt{n}p_\lambda'(|\theta|) = 0$ for $j = 2, \ldots, s$, it follows that

$$\frac{1}{\sqrt{n}}\hat{G}_j^P(\breve{\boldsymbol{\beta}}^{(1)}\pm\epsilon\mathbf{e}_j)=o_p(1)-\frac{1}{\sqrt{n}}p_\lambda'(|\hat{\beta}_j\pm\epsilon|)sgn(\hat{\beta}_j\pm\epsilon)=o_p(1),$$

where $\mathbf{e}_j$ is the jth canonical unit vector and ϵ is a small number.

For $j > s$, under Condition (C8), $\frac{1}{\sqrt{n}}\hat{G}_j^P(\breve{\boldsymbol{\beta}}^{(1)} + \epsilon\mathbf{e}_j)$ and $\frac{1}{\sqrt{n}}\hat{G}_j^P(\breve{\boldsymbol{\beta}}^{(1)} - \epsilon\mathbf{e}_j)$ are dominated by $-\sqrt{n}p_\lambda'(\epsilon)$ and $\sqrt{n}p_\lambda'(\epsilon)$, which have opposite signs when $\epsilon \to 0$.

Hence $\breve{\boldsymbol{\beta}}^{(1)}$ is an approximate zero-crossing by definition. Thus, we complete the proof of Theorem 1. □

Proof of Theorem 2 Inspired by the proof of Theorem 1(b) in Johnson et al. (2008), we define probability space $B_j = \{\hat{\beta}_{\lambda,j} \neq 0\}$ for $j > s$. To prove the sparsity, we only need to show that for any $\epsilon > 0$, $P(B_j) < \epsilon$ when n is sufficiently large.

$\hat{\boldsymbol{\beta}}_{\lambda}^{(1)}$ is root-n-consistent approximate zero-crossing, hence $\hat{\beta}_{\lambda,j} = O_p(n^{-1/2})$, $j > s$, and there exists some constant $C > 0$ such that when n is large enough,

$$\begin{aligned} P(B_j) &= P\{\hat{\beta}_{\lambda,j} \neq 0, |\hat{\beta}_{\lambda,j}| \geq Cn^{-1/2}\} + P\{\hat{\beta}_{\lambda,j} \neq 0, |\hat{\beta}_{\lambda,j}| < Cn^{-1/2}\} \\ &< \epsilon/2 + P\{\hat{\beta}_{\lambda,j} \neq 0, |\hat{\beta}_{\lambda,j}| < Cn^{-1/2}\}. \end{aligned}$$

The jth penalized estimating function of (4) is

$$n^{-1/2}\hat{G}_j(\hat{\boldsymbol{\beta}}_{\lambda}^{(1)}) - \sqrt{n}p_{\lambda}^{'}(|\hat{\beta}_{\lambda,j}^{(1)}|)sgn(\hat{\beta}_{\lambda,j}^{(1)}) = o_p(1). \tag{9}$$

Equation (8) shows that the first term of (9) is $O_p(1)$, then there exists some $C^{'} > 0$ such that for large n,

$$P\left\{\hat{\beta}_{\lambda,j} \neq 0, |\hat{\beta}_{\lambda,j}| < Cn^{-1/2}, \sqrt{n}p_{\lambda}^{'}(|\hat{\beta}_{\lambda,j}^{(1)}|) > C^{'}\right\} < \epsilon/2.$$

From Condition (C8), we know that $\hat{\beta}_{\lambda,j} \neq 0$ and $|\hat{\beta}_{\lambda,j}| < Cn^{-1/2}$ imply that $\sqrt{n}p_{\lambda}^{'}(|\hat{\beta}_{\lambda,j}^{(1)}|) > C^{'}$ for large n. Therefore,

$$P(B_j) < \epsilon/2 + P\{\hat{\beta}_{\lambda,j} \neq 0, |\hat{\beta}_{\lambda,j}| < Cn^{-1/2}\} < \epsilon,$$

which proves the sparsity.

Next, we prove the asymptotic normality. After the Taylor expansion of first term of (4) at the point $\boldsymbol{\beta}_0^{(1)}$, we have

$$o_p(1) = n^{-1/2}\hat{\mathbf{G}}_{\mathcal{A}}(\boldsymbol{\beta}_0^{(1)}) + n^{-1/2}\mathbf{I}_{11}\left(\hat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}} - \boldsymbol{\beta}_{0,\mathcal{A}}\right) - \sqrt{n}\mathbf{p}_{\lambda}^{'}(|\hat{\beta}_{\lambda,\mathcal{A}}|)sgn(\hat{\beta}_{\lambda,\mathcal{A}}).$$

After the Taylor series expansion of the last term, it follows by Slutsky's theorem and the central limit theorem that

$$\sqrt{n}(\mathbf{I}_{11} + \Sigma_{11})\left\{\hat{\boldsymbol{\beta}}_{\lambda,\mathcal{A}} - \boldsymbol{\beta}_{0,\mathcal{A}} + (\mathbf{I}_{11} + \Sigma_{11})^{-1}\mathbf{b}\right\} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}_{11}).$$

Thus, we complete the proof of Theorem 2. □

Acknowledgements Wang's research was supported by the National Science Fund for Distinguished Young Scholars in China (10725106), the National Natural Science Foundation of China (General program 11171331 and Key program 11331011), a Grant from the Key Lab of Random Complex Structure and Data Science, CAS and Natural Science Foundation of SZU. Zhang's research was supported by the China Postdoctoral Science Foundation (Grant No. 2014M550799) and the National Science Foundation of China (Grant No. 11401561).

References

Carroll, R. J., Fan, J., Gijbels, I., Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*, 477–489.

Cui, X., Härdle, W., Zhu, L. X. (2011). The EFM approach for single-index models. *Annals of Statistics*, *39*, 1658–1688.

Donoho, D. L., Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*, 425–455.

Fan, J., Gijbels, I. (1996). *Local polynomial modeling and its applications*. London: Chapman Hall.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fan, J., Heckman, N. E., Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likehood functions. *Journal of the American Statistical Association*, *90*, 141–150.

Härdle, W., Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, *84*, 985–995.

Härdle, W., Hall, P., Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, *21*, 157–178.

Hristache, M., Juditsky, A., Spokoiny, V. (2001). Direct estimation of the index coeffcient in a single-index model. *Annals of Statistics*, *29*, 595–623.

Hunter, D. R., Li, R. Z. (2005). Variable selection using MM algorithms. *Annals of Statistics*, *33*, 1617–1642.

Jeng, X. J., Daye, Z. J. (2011). Sparse convariance thresholding for high-dimensional variable selection. *Statistica Sinica*, *21*, 625–657.

Johnson, B. A., Lin, D. Y., Zeng, D. L. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, *103*, 672–680.

Kane, M., Holt, J., Alen, B. (2004). Results concerning the generalized partically linear single-index model. *Journal of Statistical Computation and Simulation*, *72*, 897–912.

Kong, E., Xia, Y. C. (2007). Variable selection for the single-index model. *Biometrika*, *94*, 217–229.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, *86*, 316–342.

Powell, J. L., Stock, J. H., Stoker, T. M. (1989). Semiparametric estimation of index coeffcients. *Econometrika*, *57*, 1403–1430.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society : Series B*, *58*, 267–288.

Wang, H., Leng, C. (2007). Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association*, *102*, 1039–1048.

Wang, H. S., Xia, Y. C. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, *104*, 747–757.

Xia, Y. C., Tong, H., Li, W. K., Zhu, L. X. (2002). An adaptive estimation of optimal regression subspace. *Journal of the Royal Statistical Society*: *Series B*, *64*, 363–410.

Xu, P. R., Zhu, L. X. (2012). Estimation for a marginal generalized single-index longitudinal model. *Journal of Multivariate Analysis*, *105*, 285–299.

Zeng, P., He, T. H., Zhu, Y. (2012). A Lasso-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics*, *21*, 92–109.

Zhang, Q. Z., Wang, Q. H. (2013). Local least absolute relative error estimating approach for partially linear multiplicative model. *Statistica Sinica*, *23*, 1091–1116.

Zhu, L. P., Zhu, L. X. (2009). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *Journal of Multivariate Analysis*, *100*, 862–875.

Zhu, L. P., Qian, L. Y., Lin, J. G. (2011). Variable selection in a class of single-index models. *Annals of the Institute of Statistical Mathematics*, *63*, 1277–1293.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*: *Series B*, *67*, 310–320.