Tuning parameter selectors for the smoothly clipped absolute deviation method

Hansheng Wang

Guanghua School of Management, Peking University, Beijing, China, 100871

hansheng@gsm.pku.edu.cn

Runze Li

Department of Statistics, The Pennsylvania State University, University Park Pennsylvania, 16802-2111, U.S.A.

rli@stat.psu.edu

and Chih-Ling Tsai

Graduate School of Management, University of California, Davis California, 95616-8609, U.S.A.

cltsai@ucdavis.edu

Summary

The penalised least squares approach with smoothly clipped absolute deviation penalty has been consistently demonstrated to be an attractive regression shrinkage and selection method. It not only automatically and consistently selects the important variables, but also produces estimators which are as efficient as the oracle estimator. However, these attractive features depend on appropriately choosing the tuning parameter. We show that the commonly used the generalised crossvalidation cannot select the tuning parameter satisfactorily, with a nonignorable overfitting effect in the resulting model. In addition, we propose a BIC tuning parameter selector, which is shown to be able to identify the true model consistently. Simulation studies are presented to support theoretical findings, and an empirical example is given to illustrate its use in the Female Labor Supply data.

Some key words: AIC; BIC; Generalised crossvalidation; Least absolute shrinkage and selection operator; Smoothly clipped absolute deviation

1. Introduction

In regression analysis, an underfitted model can lead to severely biased estimation and prediction. In contrast, an overfitted model can seriously degrade the efficiency of the resulting parameter estimates and predictions. Hence, obtaining a sparsely parsimonious and effectively predictive model is essential.

Traditional model selection criteria, such as AIC (Akaike, 1973) and BIC (Schwarz, 1978), suffer from a number of limitations. Their major drawback arises because parameter estimation and model selection are two different processes, which can result in instability (Breiman, 1996) and complicated stochastic properties (Fan & Li, 2001). Moreover, the total number of candidate models increases exponentially as the number of covariates increases.

To overcome the deficiency of traditional methods, Fan & Li (2001) proposed the smoothly clipped absolute deviation or SCAD method, which estimates parameters while simultaneously selecting important variables. As compared with another popular regression shrinkage and selection method, the least absolute shrinkage and selection operator or Lasso of Tibshirani (1996), the smoothly clipped absolute deviation method not only selects important variables consistently, but also produces parameter estimators as efficient as if the true model were known, i.e., the oracle estimator, a property not enjoyed by the Lasso. The above features of the smoothly clipped absolute deviation method rely on the proper choice of tuning parameter, or regularisation parameter, which is usually selected by generalised crossvalidation (Craven & Wahba, 1979).

We show that the optimal tuning parameter selected by generalised crossvalidation has a nonignorable overfitting effect even as the sample size goes to infinity. Moreover, we propose a BIC-based tuning parameter selector for the smoothly clipped absolute deviation method, and prove that the proposed procedure identifies the true model

consistently.

2. The Smoothly Clipped Absolute Deviation Method

Consider the linear regression model,

$$y_i = x_i'\beta + \epsilon_i, \tag{2.1}$$

where y_i is the response from the ith subject, $x_i = (x_{i1}, \cdots, x_{id})'$ is the associated d-dimensional explanatory covariate, $\beta = (\beta_1, \cdots, \beta_d)'$, and ϵ_i is the random error with mean 0 and variance σ_ϵ^2. Let (x_i, y_i), $i = 1, \cdots, n$, be a random sample from (2.1). To select simultaneously variables and estimate parameters, the smoothly clipped absolute deviation method of Fan & Li (2001) estimates β by minimising the penalised least squares function

$$\frac{1}{2n}||Y - X\beta||^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.2}$$

where $Y = (y_1, \cdots, y_n)'$, $X = (x_1, \cdots, x_n)'$, $||\cdot||$ stands for the Euclidean norm, and $p_\lambda(\cdot)$ is the smoothly clipped absolute deviation penalty with a tuning parameter λ to be selected by a data-driven method. The penalty $p_\lambda(\cdot)$ satisfies $p_\lambda(0) = 0$, and its first-order derivative is

$$p_\lambda'(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\},$$

where a is some constant usually taken to be $a = 3.7$ (Fan & Li, 2001), and $(t)_+ = tI\{t > 0\}$ is the hinge loss function. For a given tuning parameter, we denote the estimator obtained by minimising (2.2) by $\hat{\beta}_\lambda = (\hat{\beta}_{\lambda 1}, \cdots, \hat{\beta}_{\lambda d})'$.

Fan & Li (2001) showed that if $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ as $n \to \infty$, the method

consistently identifies irrelevant variables by producing zero solutions for their associated regression coefficients. In addition, the method estimates the coefficients of the relevant variables with the same efficiency as if the true model were known, which is referred to as the oracle property (Fan & Li, 2001). Hence, the choice of λ is critical. In practice, λ is usually selected by minimising the generalised crossvalidation criterion

$$\mathrm{GCV}_\lambda = \frac{||Y - X\hat{\beta}_\lambda||^2}{n(1 - \mathrm{DF}_\lambda/n)^2} = \frac{\hat{\sigma}^2_\lambda}{(1 - \mathrm{DF}_\lambda/n)^2}, \tag{2.3}$$

where $\hat{\sigma}^2_\lambda = n^{-1}||Y - X\hat{\beta}_\lambda||^2$, DF_λ is the generalised degrees of freedom (Fan & Li, 2001) given by

$$\mathrm{DF}_\lambda = \mathrm{tr}\Big\{X\Big(X'X + n\Sigma_\lambda\Big)^{-1}X'\Big\},$$

and $\Sigma_\lambda = \mathrm{diag}\{p'_\lambda(|\hat{\beta}_{\lambda 1}|)/|\hat{\beta}_{\lambda 1}|, \cdots, p'_\lambda(|\hat{\beta}_{\lambda d}|)/|\hat{\beta}_{\lambda d}|\}$. The diagonal elements of Σ_λ are coefficients of quadratic terms in the local quadratic approximation to the smoothly clipped absolute deviation penalty function $p_\lambda(\cdot)$ (Fan & Li, 2001). Since some coefficients of the estimator of β are exactly equal to zero, DF_λ is calculated by replacing X with its submatrix corresponding to the selected covariates, and by replacing Σ_λ with its corresponding submatrix. The resulting optimal tuning parameter is $\hat{\lambda}_{\mathrm{GCV}} = \mathrm{argmin}_\lambda \mathrm{GCV}_\lambda$.

The log-transformation of GCV_λ can be approximated by

$$\log \mathrm{GCV}_\lambda = \log \hat{\sigma}^2_\lambda - 2\log(1 - \mathrm{DF}_\lambda/n) \backsimeq \log \hat{\sigma}^2_\lambda + 2\mathrm{DF}_\lambda/n \triangleq \mathrm{AIC}_\lambda.$$

Hence, $\log \mathrm{GCV}_\lambda$ is very similar to the traditional model selection criterion AIC , which is an efficient selection criterion in that it selects the best finite-dimensional candidate model in terms of prediction accuracy when the true model is of infinite dimension. However, AIC is not a consistent selection criterion, since it does not select the correct

model with probability approaching 1 in large samples when the true model is of finite dimension. For further discussion of efficiency and consistency in model selection, see Shao (1997), McQuarrie & Tsai (1998) and Yang (2005). Consequently, the model selected by $\hat{\lambda}_{\mathrm{GCV}}$ may not identify the finite-dimensional true model consistently. This motivated us to employ a variable selection criterion known to be consistent, BIC (Schwarz, 1978), as the tuning parameter selector. We select the optimal λ by minimising

$$\mathrm{BIC}_\lambda = \log \hat{\sigma}^2_\lambda + \mathrm{DF}_\lambda \log(n)/n. \tag{2.4}$$

The resulting optimal regularisation parameter is denoted by $\hat{\lambda}_{\mathrm{BIC}}$.

3. Theoretical Results

3.1. Notation and conditions

Suppose that there is an integer $0 \leq d_0 \leq d$, such that $\beta_{j_k} \neq 0$ for $1 \leq k \leq d_0$, with the other β_j's equal to 0. Thus, the true model only contains the j_1th, $\cdots$, j_{d_0}th covariates as significant variables. Furthermore, in order to define the underfitted and overfitted models, we denote by $\mathcal{S}_F = \{1, \cdots, d\}$ and $\mathcal{S}_T = \{j_1, \cdots, j_{d_0}\}$ the full and true parsimonious submodels, respectively. Then any candidate model $\mathcal{S} \not\supset \mathcal{S}_T$, is referred to as an underfitted model in the sense that it misses at least one important variable. In contrast, any $\mathcal{S} \supset \mathcal{S}_T$ other than $\mathcal{S}_T$ itself, is referred to as an overfitted model in the sense that it contains all significant variables, but also at least one insignificant variable.

For an arbitrary model $\mathcal{S} = \{j_1, \cdots, j_{d^*}\} \subset \mathcal{S}_F$, we denote its associated covariate matrix by $X_{\mathcal{S}}$, which is an $n \times d^*$ matrix with the ith row given by $(x_{ij_1}, \cdots, x_{ij_{d^*}})$. After fitting the data with model $\mathcal{S}$ by least squares, we denote the resulting ordinary least squares estimator, the residual sum of squares, the variance estimator and the

generalised crossvalidation value by

$$\begin{aligned}
\hat{\beta}_{\mathcal{S}} &= (X_{\mathcal{S}}' X_{\mathcal{S}})^{-1}(X_{\mathcal{S}}' Y), && (3.1)\\
\text{SSE}_{\mathcal{S}} &= ||Y - X_{\mathcal{S}}\hat{\beta}_{\mathcal{S}}||^2, && (3.2)\\
\hat{\sigma}^2_{\mathcal{S}} &= \text{SSE}_{\mathcal{S}}/n, && (3.3)\\
\text{GCV}_{\mathcal{S}} &= n^{-1}\text{SSE}_{\mathcal{S}}/(1 - d^*/n)^2, && (3.4)
\end{aligned}$$

respectively. The smoothly clipped absolute deviation estimator $\hat{\beta}_\lambda$, obtained by minimising the objective function in (2.2), naturally identifies the model $\mathcal{S}_\lambda = \{j : \hat{\beta}_{\lambda j} \neq 0\}$, for which the ordinary least squares estimator is $\hat{\beta}_{\mathcal{S}_\lambda}$. By the definition of the ordinary least squares estimator, we have

$$\text{SSE}_\lambda = ||Y - X\hat{\beta}_\lambda||^2 \geq ||Y - X_{\mathcal{S}_\lambda}\hat{\beta}_{\mathcal{S}_\lambda}||^2 = \text{SSE}_{\mathcal{S}_\lambda}. \quad (3.5)$$

Furthermore, the $\hat{\sigma}^2_\lambda$ defined in (2.3) can be simply expressed as $\hat{\sigma}^2_\lambda = \text{SSE}_\lambda/n$. If $\lambda = 0$, then the penalty term in (2.2) is 0, and $\hat{\beta}_0$, is exactly the same as the full-model's ordinary least squares estimator, $\hat{\beta}_{\mathcal{S}_F}$. Moreover, $\text{SSE}_0 = \text{SSE}_{\mathcal{S}_F}$, $\hat{\sigma}^2_0 = \hat{\sigma}^2_{\mathcal{S}_F}$ and $\text{GCV}_0 = \text{GCV}_{\mathcal{S}_F}$.

In practice, however, λ is unknown, and it is necessary to search for the optimal λ from the positive real line $\mathcal{R}^+$, or, within the bounded interval $\Omega = [0, \lambda_{\max}]$, for some upper limit $\lambda_{\max}$. We now present the technical conditions that are needed for studying the theoretical properties of the tuning parameter selectors.

Condition 1. For any $\mathcal{S} \subset \mathcal{S}_F$, there is a $\sigma^2_{\mathcal{S}} > 0$ such that $\hat{\sigma}^2_{\mathcal{S}} \to \sigma^2_{\mathcal{S}}$ in probability.

Condition 2. For any $\mathcal{S} \not\supset \mathcal{S}_T$, we have $\sigma^2_{\mathcal{S}} > \sigma^2_{\mathcal{S}_T}$, where $\sigma^2_{\mathcal{S}_T}$ is a positive value such that $\hat{\sigma}^2_{\mathcal{S}_T} \to \sigma^2_{\mathcal{S}_T}$, in probability.

Condition 3. The ϵ_i's are independent and identically distributed as $N(0, \sigma^2_\epsilon)$.

Condition 4. The upper limit $\lambda_{\max} \to 0$ as $n \to \infty$.

Condition 5. The matrix $\mathrm{cov}(x_i) = \Sigma_x$ is finite and positive definite.

Condition 1 facilitates the proof of the asymptotic results, while Condition 2 elucidates the underfitting effect. Both conditions are satisfied if (x_i, ϵ_i) are jointly non-degenerate multivariate normal distribution. Similar conditions can be found in Shi & Tsai (2002, 2004) and Huang & Yang (2004). Condition 3 is needed only for evaluating generalised crossvalildation's overfitting effect in §3.2, and is not necessary for establishing the consistency of the proposed BIC criterion. Condition 4 implies that the search region for λ shrinks towards 0 as the sample size goes to infinity. This condition is used to simplify the proof of the consistency of BIC in §3.3. Note that the rate at which $\lambda_{\max}$ converges to 0 is not specified. Finally, Condition 5 ensures the root-n consistency of an unpenalised estimator.

3.2. The overfitting effect of generalised crossvalidation

We define $\Omega_- = \{\lambda \in \Omega : \mathcal{S}_\lambda \not\supset \mathcal{S}_T\}$, $\Omega_0 = \{\lambda \in \Omega : \mathcal{S}_\lambda = \mathcal{S}_T\}$, and $\Omega_+ = \{\lambda \in \Omega : \mathcal{S}_\lambda \supset \mathcal{S}_T \text{ and } \mathcal{S}_\lambda \neq \mathcal{S}_T\}$. In other words, Ω_0, Ω_- and Ω_+ are three subsets of Ω, in which the true, under, and overfitted models can be produced. We first show that the smoothly clipped absolute deviation method with generalised crossvalidation is conservative in the sense that it does not miss any important variables as long as the sample size is sufficiently large.

Lemma 1. *Under Conditions 1 and 2, we have*

$$\mathrm{pr}\Big(\inf_{\lambda \in \Omega_-} \mathrm{GCV}_\lambda > \mathrm{GCV}_{\mathcal{S}_F} = \mathrm{GCV}_0\Big) \to 1.$$

All proofs are given in the Appendix. According to this lemma, the generalised crossvalidation evaluated at the tuning parameter which produces the underfitted

model, is consistently larger than $\text{GCV}_{\mathcal{S}_F} = \text{GCV}_0$. As a result, the optimal model selected by minimising the generalised crossvalidation values, i.e., $\mathcal{S}_{\hat{\lambda}_{\text{GCV}}}$, must contain all significant variables with probability tending to one. However, this does not necessarily imply that $\mathcal{S}_{\hat{\lambda}_{\text{GCV}}}$ is the true model $\mathcal{S}_T$. In the next lemma, we show that the optimal model selected by generalised crossvalidation overfits the true model with a positive probability.

Lemma 2. *Under Conditions 1—3 there exists a nonzero probability* $\alpha > 0$ *such that* $\liminf_n \text{pr}(\inf_{\lambda \in \Omega_0} \text{GCV}_\lambda > \text{GCV}_{\mathcal{S}_F} = \text{GCV}_0) \geq \alpha$.

According to this lemma, there is a nonzero probability that the smallest value of generalised crossvalidation associated with the true model is larger than that of the full-model. Hence, there is a positive probability that any λ associated with the true model cannot be selected by generalised crossvalidation as the optimal tuning parameter. Combining the results from Lemmas 1 and 2, we obtain the following theorem.

Theorem 1. *If Conditions 1—3 hold, there is a nonzero probability* $\alpha > 0$ *such that* $\text{pr}(\mathcal{S}_{\hat{\lambda}_{\text{GCV}}} \supset \mathcal{S}_T) \to 1$ *and*

$$\liminf_n \text{pr}(\mathcal{S}_{\hat{\lambda}_{\text{GCV}}} \supset \mathcal{S}_T \text{ and } \mathcal{S}_{\hat{\lambda}_{\text{GCV}}} \neq \mathcal{S}_T) > \alpha.$$

Theorem 1 indicates that, with probability tending to 1, the model $\mathcal{S}_{\hat{\lambda}_{\text{GCV}}}$ contains all significant variables, but with nonzero probability includes superfluous variables, thereby leading to overfitting.

3.3. Consistency of BIC

To establish the consistency of BIC, we first construct a sequence of reference tuning parameters, $\lambda_n = \log(n)/\sqrt{n}$. Thus, $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$. According

to Theorem 2 of Fan & Li (2001), $\mathrm{pr}(\mathcal{S}_{\lambda_n} = \mathcal{S}_T) \to 1$ under appropriate regularity conditions. This implies that the model identified by the reference tuning parameter converges to the true model as the sample size gets large.

Lemma 3. *Under Condition 5,* $\mathrm{pr}(\mathrm{BIC}_{\lambda_n} = \mathrm{BIC}_{\mathcal{S}_T}) \to 1$.

According to this lemma, with probability tending to 1,

$$\mathrm{BIC}_{\lambda_n} = \mathrm{BIC}_{\mathcal{S}_T} = \log \hat{\sigma}^2_{\mathcal{S}_T} + d_0 \log(n)/n.$$

Applying this result, we finally show that, for any λ which cannot identify the true model, the BIC value is consistently larger than BIC_{λ_n}.

Lemma 4. *Under Conditions 1, 2, 4 and 5,*

$$\mathrm{pr}\left(\inf_{\lambda \in \Omega_- \cup \Omega_+} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}\right) \to 1.$$

Note that this lemma does not necessarily imply that $\lambda_n = \hat{\lambda}_{\mathrm{BIC}}$. However, it does indicate that those λ's which fail to identify the true model cannot be selected by BIC asymptotically, because at least the true model identified by λ_n is a better choice. As a result, the optimal value $\hat{\lambda}_{\mathrm{BIC}}$ can only be one of those λ's whose smoothly clipped absolute deviation estimator yields the true model, i.e., $\lambda \in \Omega_0$. Hence, the subsequent theorem follows immediately.

Theorem 2. *If Conditions 1, 2, 4, and 5 hold,* $\mathrm{pr}(\mathcal{S}_{\hat{\lambda}_{BIC}} = \mathcal{S}_T) \to 1$.

In addition to generalised crossvalidation and BIC, other selection criteria, such as AIC and RIC (Shi & Tsai, 2002) can be used to select the tuning parameter for the smoothly clipped absolute deviation method. Techniques similar to those used

above show that AIC performs like generalised crossvalidation, with a potential for overfitting, while RIC consistently identifies the true model.

4. Partially Linear Model

In the context of partially linear models, Bunea (2004) and Bunea & Wegkamp (2004) proposed an information-type criterion and established a consistency property. Fan & Li (2004) extended their nonconvex penalised least squares method to partially linear models with longitudinal data, and showed that the resulting estimator performs as well as the oracle estimator. However, Fan & Li (2004) employed generalised crossvalidation for selecting the tuning parameter. In this section, we demonstrate that this results in overfitting. We further propose a BIC approach and show that it can identify the true model consistently.

Consider the partially linear model

$$y_i = \alpha(u_i) + x_i'\beta + \epsilon_i, \tag{4.1}$$

where u_i is a covariate, $\alpha(u_i)$ is nonparametric smooth function of u_i, and the remainder of the notation is the same as that for model (2.1). Various estimation procedures have been proposed in the literature (Engle et al., 1986; Heckman, 1986; Robinson, 1988; Speckman, 1988), and a comprehensive survey for the partially linear model is given by Härdle et al. (2000).

Similar to (2.2), we propose

$$\frac{1}{2n}\|Y - \Theta - X\beta\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|) \tag{4.2}$$

as a penalised least squares function for the partially linear model, where

$$\Theta = (\alpha(u_1), \cdots, \alpha(u_n))',$$

the penalty function $p_\lambda(|\beta_j|)$ is defined as in (2.2), and $\alpha(\cdot)$ is a nonparametric smoothing function. To obtain the penalised least squares estimator, we first adopt Fan & Li's (2004) profile least squares technique to eliminate the nuisance parameter Θ for a given β. As a result, we have

$$y_i^* = \alpha(u_i) + \epsilon_i, \tag{4.3}$$

where $y_i^* = y_i - x_i'\beta$. We then use Fan & Gijbel's (1996) local linear regression approach to estimate $\alpha(\cdot)$. For u in a neighbourhood of u_i, we find $(\hat{\alpha}_0,\ \hat{\alpha}_1)$ by minimising

$$\sum_{i=1}^{n} \{y_i^* - \alpha_0 - \alpha_1(u_i - u)\}^2 K_h(u_i - u),$$

where $K(\cdot)$ is a kernel function, h is a bandwidth and $K_h(\cdot) = h^{-1}K(\cdot/h)$. The local linear estimator at u is simply $\hat{\alpha}(u;\beta) = \hat{\alpha}_0$.

Since the local linear estimator is a linear smoother, $\hat{\Theta}$ has the closed-form expression

$$\hat{\Theta} = S_h(Y - X\beta), \tag{4.4}$$

where S_h is the smoothing matrix corresponding to the local linear regression and depends only on u_i and $K_h(\cdot)$. Substituting Θ in (4.2) with $\hat{\Theta}$, we obtain the penalised profile least squares function

$$\frac{1}{2n}\|(I - S_h)Y - (I - S_h)X\beta\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{4.5}$$

where I is an $n \times n$ identity matrix. Under certain regularity conditions, Fan & Li (2004) established the oracle property for the penalised profile least squares estimator. Note that the profile least squares estimator of β is closely related to Speckman's (1988) partial residual estimator, which is obtained by minimising the first term of equation (4.5). In addition, Speckman (1988) used the kernel smoothing approach to estimate α, while we employed the local linear smoothing approach.

Based on (4.5), we define GCV_λ for the penalised profile least squares problem by substituting X and Y in (2.3) with $X_h = (I - S_h)X$ and $Y_h = (I - S_h)Y$, respectively. Analogously, we define BIC_λ by replacing X and Y in the calculation of σ^2_λ in (2.4) with X_h and Y_h, respectively. Applying the selector GCV_λ or BIC_λ, we are finally able to compute the SCAD estimator of β.

As shown in Fan & Li (2004), the penalised profile least squares estimator $\hat{\beta}_\lambda$ is root-n consistent provided that $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ as $n \to \infty$. Under regularity conditions given in the Appendix, the asymptotic bias and variance of $\hat{\alpha}(u)$ are of order $O_p(h^2)$ and $O_p(1/nh)$, respectively, since the parametric convergence rate of $\hat{\beta}$ is faster than the nonparametric convergence rate of $\hat{\alpha}(u)$. Furthermore, it can be shown that

$$\sup_{u \in \mathcal{U}} |\hat{\alpha}(u) - \alpha(u)| = o_P(n^{-1/4})$$

by using results in Mack & Silverman (1982), where $\mathcal{U}$ is the support of u; see Fan & Huang (2005) for details. Thus, Conditions 1 and 2 are reasonable assumptions in the proofs of asymptotic properties of GCV and BIC in partially linear models.

Applying Theorem 3.1 of Fan & Huang (2005), we have

$$n \left\{ \frac{\text{SSE}_{\mathcal{S}_T} - \text{SSE}_{\mathcal{S}_F}}{\text{SSE}_{\mathcal{S}_F}} \right\} \to \chi^2_{d-d_0},$$

in distribution. This, together with the arguments used in the proofs of Lemmas 1 and 2, implies that Theorem 1 holds for penalised profile least squares with the smoothly clipped absolute deviation penalty. Theorem 3.1 of Fan & Huang (2005) also implies that

$$\mathrm{SSE}_{\mathcal{S}_T} - \mathrm{SSE}_{\mathcal{S}_\lambda} \to \chi^2_{d_\lambda - d_0}$$

in distribution for any overfitted model $\mathcal{S}_\lambda (\supset \mathcal{S}_T)$ including d_λ variables. As a result, equation (A7) is valid for profile least squares estimators. Applying this result, in conjunction with the same arguments as those employed in the proofs of Lemmas 3 and 4, shows that Theorem 2 is true for penalised profile least squares with the smoothly clipped absolute deviation penalty.

Remark. To facilitate choosing the bandwidth h and the tuning parameter λ, we consider the $\hat{\beta}$ to be a root-n consistent estimator of β. Thus, its convergence rate is faster than the nonparametric convergence rate of $\hat{\alpha}(u)$. This motivates us to substitute β in the expression of y_i^* with its root-n consistent estimator. By following the approach of Fan & Li (2004), we can show that the leading terms in the asymptotic bias and variance of the resulting local linear estimator $\hat{\alpha}(u)$ are the same as those obtained by replacing β with its true value. This indicates that we are able to choose the bandwidth and tuning parameter separately, which expedites the computation of h and λ. To be specific, we adapt the approach of Fan & Li (2004) to obtain the difference-based estimator of β for the full-model, which is a root-n consistent estimator (Yatchew, 1997). Subsequently, we replace β in y_i^* with the difference-based estimator so that (4.3) becomes a one-dimensional smoothing problem, and we use a smoothing selector to choose the bandwidth. Here, we use the plug-in method proposed by Ruppert et al. (1995) to choose the bandwidth, but this does not exclude the use of other known bandwidth selectors, such as generalised crossvalidation. Finally,

we apply GCV_λ or BIC_λ to choose λ.

5. Numerical Studies

5.1. Preliminaries

We examine the finite sample performance of the BIC and generalised crossvalidation tuning parameter selectors in terms of both model error, i.e., lack-of-fit, and model complexity. However, we do not compare the smoothly clipped absolute deviation method with the best-subset variable selection BIC since Fan & Li (2001, 2004) have compared them by Monte Carlo. To facilitate the computational process, we directly applied the local quadratic approximation algorithm to search the smoothly clipped absolute deviation solution. We set the threshold for shrinking $\hat{\beta}_j$ to zero at 10^{-6}, which is much smaller than half of the standard error of the unpenalised least squares estimator, the threshold used in Fan & Li (2001). Thus, the average number of zeros using the generalised crossvalidation tuning parameter selector is expected to be slightly smaller than that in Fan & Li (2001). All simulations were conducted using Matlab code, which is available from the authors.

5.2. Simulation studies

We first consider Fan & Li's (2001) model error measure. Let (u, x, y) be a new observation from a regression model with $E(y|u, x) = \mu(u, x)$, and let $\hat{\mu}(\cdot, \cdot)$ be an estimate of the regression function based on data $\{(u_i, x_i, y_i), i = 1, \cdots, n\}$. Then model error is defined to be $E\{\hat{\mu}(u, x) - \mu(u, x)\}^2$, where the expectation is the conditional expectation given the data used in caluclating $\hat{\mu}(\cdot, \cdot)$. For a partially linear model, $\mu(u, x) = \alpha(u) + x'\beta$, the model error is

$$E\{\hat{\alpha}(u) - \alpha(u)\}^2 + E(x'\hat{\beta} - x'\beta)^2 + 2E\{\hat{\alpha}(u) - \alpha(u)\}(x'\hat{\beta} - x'\beta), \tag{5.1}$$

The first term in (5.1) measures the nonparametric component fit, while the second term assesses the parametric component fit. To investigate the performance of the smoothly clipped absolute deviation method on the just parametric regression component, we chose the simulation setting so that the cross-product term in (5.1) equals 0. Note that for the linear regression model (2.1), the model error exactly equals $E(x'\hat{\beta} - x'\beta)^2$. To compare the generalised crossvalidation and BIC approaches, we define the model error as

$$\mathrm{ME}(\hat{\beta}) = E(x'\hat{\beta} - x'\beta)^2 = (\hat{\beta} - \beta)'E(xx')(\hat{\beta} - \beta),$$

and define the relative model error as $\mathrm{RME} = \mathrm{ME}/\mathrm{ME}_{S_F}$, where ME_{S_F} is the model error obtained by fitting the data with the full-model S_F in conjunction with the unpenalised least squares estimator, $\hat{\beta}_{S_F}$.

In addition to model error, we also calculate the percentages of models correctly fitted, underfitted and overfitted by generalised crossvalidation and BIC, and the average number of zero coefficients produced by the smoothly clipped absolute deviation method.

Example 1. We simulated 1000 datasets, each consisting of a random sample of size n, from the linear regression model

$$y = x'\beta + \sigma_\epsilon \epsilon,$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$, $\epsilon \sim N(0, 1)$ and the 8×1 vector $x \sim N_8(0, \Sigma_x)$, in which $(\Sigma_x)_{ij} = \rho^{|i-j|}$ for all i and j. Values chosen were $\sigma_\epsilon = 3$ and 1, $n = 50$, 100 and 200, and $\rho = 0.75$, 0.5 and 0.25.

As a benchmark, we compute the oracle estimator, which is the least squares

estimator, of the true submodel, $y = \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \epsilon$. Since the pattern of the results is the same for all three correlations, we only present the results for $\rho = 0.5$. Table 1 indicates that median of RME over 1000 realisations of the BIC approach rapidly approaches that of the oracle estimator as the sample size increases or the noise level decreases, whereas the value for the generalised crossvalidation method remains at almost the same level across different noise levels and sample sizes. Hence, the BIC approach outperforms the generalised crossvalidation approach in terms of model error measure.

The column labelled 'C' in Table 1 denotes the average number of the five true zero coefficients that were correctly set to zero, and the column labeled 'I' denotes the average number of the three truly nonzero coefficients incorrectly set to zero. Table 1 also reports the proportions of models underfitted, correctly fitted and overfitted. In the case of overfitting, the columns labelled '1', '2' and '≥ 3' are the proportions of models including 1, 2 and more than 2 irrelevant covariates, respectively. It shows that the BIC method has a much better rate of correctly identifying the true submodel than does the generalised crossvalidation method. Furthermore, among the overfitted models, the BIC method is likely to include just one irrelevant variable, whereas the generalised crossvalidation approach often includes two or more. Not surprisingly, both methods improve as the signal gets stronger, i.e., σ_ϵ decreases from 3 to 1. However, the generalised crossvalidation method still seriously overfits even if $\sigma_\epsilon = 1$ and $n = 200$. In contrast, the BIC method overfits less often. These results corroborate the theoretical findings.

Example 2. In this example, we considered the partially linear model,

$$y = \alpha(u) + x'\beta + \sigma_\epsilon \epsilon,$$

where $u \sim \mathrm{Un}(0,1)$, and $\alpha(u) = \exp\{2\sin(2\pi u)\}$. The rest of the simulation settings are the same as in Example 1. As mentioned in the remark of §4, in each simulation, we replace β in y_i^* with the difference-based estimator, and then use the plug-in method proposed by Ruppert et al. (1995) to choose a bandwidth. Table 2 presents the simulation results for $\rho = 0.5$, and shows that once more, the BIC method outperforms the generalised crossvalidation method in both identifying the true model and in reducing the model error and complexity.

5.3. Real data examples

Example 3. We consider the Female Labour Supply data collected in East Germany in about 1994. The dataset consists of 607 observations and has been analysed by Fan et al. (1998) using additive models. Here we take the response variable y to be the 'wage per hour'. The u-variable in the partially linear model is the 'woman's age'; this is because the relationship between y and u cannot be characterised by a simple functional form; see Fig. 1. There are seven explanatory variables: x_1 is the weekly number of working hours; x_2 is the 'Treiman prestige index' of the woman's job; x_3 is the monthly net income of the woman's husband; $x_4 = 1$ if the years of the woman's education is between 13 and 16, and $x_4 = 0$ otherwise; $x_5 = 1$ if the years of the woman's education is not less than 17, and $x_5 = 0$ otherwise; $x_6 = 1$ if the woman has children less than 16-years-old, and $x_6 = 0$ otherwise; and x_7 is the unemployment rate in the place where she lives. After some preliminary analysis, we consider the following partially linear model with seven linear main effects and some first-order interaction effects among x_1, x_2 and x_3:

$$y = \alpha(u) + \sum_{j=1}^{7} \beta_j x_j + \sum_{k=1}^{3} \sum_{l=k}^{3} \beta_{kl} x_k x_l + \epsilon.$$

Here the x-variables have been standardised.

Following Fan & Li's (2004) approach, we first calculate the difference-based estimator for β, and then apply the plug-in method to select 4.6249 as a bandwidth for $\hat{\alpha}(\cdot)$. With this bandwidth, we next choose the tuning parameters by minimising the generalised crossvalidation and BIC scores, resulting in $\hat{\lambda}_{\mathrm{GCV}} = 0.0896$ and $\hat{\lambda}_{\mathrm{BIC}} = 0.2655$. Subsequently, we obtain the unpenalised profile least squares estimate (Fan & Li, 2004) and the smoothly clipped absolute deviation estimate based on generalised crossvalidation and BIC, together with their standard errors (Table 3). We also consider the model selected by the unpenalised full-model profile least squares method, i.e., equation (4.5) without the penalty term, and the best-subset variable selection criterion, $\mathrm{BIC} = \log \hat{\sigma}_h^2 + d^* \log(n)/n$, where $\hat{\sigma}_h^2 = \mathrm{SSE}_{hs}/n$, SSE_{hs} is the sum of squares of the errors by fitting Y_h versus $X_{hs} = (I - S_h)X_{\mathcal{S}}$, and d^* is the dimension of $X_{\mathcal{S}}$. The first four columns of Table 3 clearly show that the unpenalised full-model profile least squares approach fits spurious variables, while the smoothly clipped absolute deviation method based on GCV tends to include variables with small, insignificant effects. In contrast, all variables selected by the smoothly clipped absolute deviation method based on BIC are significant at level 0.05. Fig. 1 shows that the four estimates of $\alpha(\cdot)$ are fairly similar, but that the estimate from the unpenalised full-model profile least squares approach is slightly different from the others. Moreover, the resulting intercept function changes with age with no particular functional form.

The fourth column of Table 3 shows that the best-subset variable selection with the BIC criterion yields a simpler model than that from the smoothly clipped absolute deviation method based on BIC. However, Breiman (1996) found that the best-subset method suffers from a lack of stability. To demonstrate this point, we exclude the last 5 observations, so leaving a dataset with $n = 602$. The last column of Table 3 shows that the best-subset variable selection with BIC criterion yields a different

model from that with $n = 607$, which corroborates Breiman's finding. The model based on the smoothly clipped absolute deviation method with BIC turns out to be unchanged, although details are not given.

We conclude that the best model in this study is that selected by the smoothly clipped absolute deviation method with BIC:

$$\hat{y} = \hat{\alpha}(u) + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{11} x_1^2 + \hat{\beta}_{12} x_1 x_2 + \hat{\beta}_{22} x_2^2 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5.$$

Hence, the hourly wage of a woman depends primarily on working hours, job prestige and years of education, while the husband's income, the local unemployment rate and the indicator of whether or not the woman has a young child seem not to affect the hourly wage significantly. Fig. 1 indicates that the hourly wage is almost constant before the age of 50, but decreases rapidly thereafter.

6. Discussion

One could extend the current work by adapting Fan & Li's (2001) approach to define the penalised likelihood function for the generalised linear model by replacing the first term of equation (2.2) with twice the negative of the corresponding loglikelihood function. Subsequently, one could explore the overfitting effect of generalised crossvalidation and the consistency of BIC. It is also of interest to compare the generalised crossvalidation approach to the BIC method for semiparametric models and single-index models. Research along these lines is currently under investigation.

Acknowledgement

We are grateful to the editor, the associate editor and two referees for their helpful and constructive comments. Li's research was supported by grants from the U.S. National Institute on Drug Abuse and National Science Foundation.

Appendix

Proofs

Proof of Lemma 1. When $\lambda = 0$, we have $\mathrm{DF}_0 = d$ and $\mathrm{GCV}_0 = \mathrm{GCV}_{\mathcal{S}_F}$ in (2.3). Then, applying Condition 1 together with $2\log(1 - d/n) = O(n^{-1})$, we obtain

$$\log \mathrm{GCV}_{\mathcal{S}_F} = \log\left(\frac{\mathrm{SSE}_{\mathcal{S}_F}}{n}\right) - 2\log\left(1 - \frac{d}{n}\right) = \log \hat{\sigma}^2_{\mathcal{S}_F} + O(n^{-1}). \tag{A1}$$

According to (3.5), $\mathrm{SSE}_\lambda \geq \mathrm{SSE}_{\mathcal{S}_\lambda}$, which leads to

$$\log \mathrm{GCV}_\lambda \geq \log\left(\frac{1}{n}\mathrm{SSE}_\lambda\right) \geq \log\left(\frac{1}{n}\mathrm{SSE}_{\mathcal{S}_\lambda}\right) = \log \hat{\sigma}^2_{\mathcal{S}_\lambda}.$$

As a result,

$$\inf_{\lambda \in \Omega_-} \log \mathrm{GCV}_\lambda \geq \min_{\mathcal{S} \not\supset \mathcal{S}_T} \log \hat{\sigma}^2_{\mathcal{S}}. \tag{A2}$$

In addition, Condition 2 implies that $\sigma^2_{\mathcal{S}} > \sigma^2_{\mathcal{S}_F} = \sigma^2_\epsilon$ for $\mathcal{S} \not\supset \mathcal{S}_T$. Hence, $\min_{\mathcal{S} \not\supset \mathcal{S}_T} \log \sigma^2_{\mathcal{S}} > \log \sigma^2_{\mathcal{S}_F}$. This result, in conjunction with Conditions 1 and 2 and equations (A1), and (A2), yields

$$\mathrm{pr}\Big(\inf_{\lambda \in \Omega_-} \log \mathrm{GCV}_\lambda > \log \mathrm{GCV}_{\mathcal{S}_F}\Big) \to 1$$

as $n \to \infty$, and the proof is complete. □

Proof of Lemma 2. For any $\lambda \in \Omega_0$, we have $\mathcal{S}_\lambda = \mathcal{S}_T$. Hence, $\mathrm{GCV}_\lambda \geq (1/n)\mathrm{SSE}_{\mathcal{S}_\lambda} = (1/n)\mathrm{SSE}_{\mathcal{S}_T}$. This, together with the fact $(1 - d/n)^{-2} = 1 + 2d/n + O(n^{-2})$, leads to

$$\begin{aligned} \mathrm{pr}\left(\inf_{\lambda \in \Omega_0} \mathrm{GCV}_\lambda > \mathrm{GCV}_{\mathcal{S}_F}\right) &\geq \mathrm{pr}\left\{\frac{\mathrm{SSE}_{\mathcal{S}_T}}{n} > \frac{\mathrm{SSE}_{\mathcal{S}_F}}{n(1 - d/n)^2}\right\} \\ &= \mathrm{pr}\left\{\frac{\mathrm{SSE}_{\mathcal{S}_T} - \mathrm{SSE}_{\mathcal{S}_F}}{\hat{\sigma}^2_{\mathcal{S}_F}} > 2d + O(n^{-1})\right\} \end{aligned} \tag{A3}$$

since $\hat{\sigma}^2_{\mathcal{S}_F} = \text{SSE}_{\mathcal{S}_F}/n$. According to Conditions 1 and 2, we have that $\hat{\sigma}^2_{\mathcal{S}_F} \to \sigma^2_{\mathcal{S}_F} = \sigma^2_\epsilon$ in probability. Furthermore, under Condition 3, $(\text{SSE}_{\mathcal{S}_T} - \text{SSE}_{\mathcal{S}_F})/\sigma^2_\epsilon$ follows a $\chi^2_{d-d_0}$ distribution. As a result,

$$\begin{aligned} \text{pr}\left(\inf_{\lambda\in\Omega_0} \text{GCV}_\lambda > \text{GCV}_{\mathcal{S}_F}\right) &\geq \text{pr}\Big[\chi^2_{d-d_0}\{1+o_p(1)\} > 2d + O(n^{-1})\Big] \\ &\to \text{pr}(\chi^2_{d-d_0} > 2d) \triangleq \alpha. \end{aligned}$$

This completes the proof. □

Proof of Lemma 3. Let $\beta_S = (\beta_{j_1}, \cdots, \beta_{j_{d_0}})'$ be the vector of relevant coefficients, and let β_N consist of irrelevant coefficients. Without loss of generality, we assume that $\beta_S = (\beta_1, \cdots, \beta_{d_0})'$, and $\beta_N = (\beta_{d_0+1}, \cdots, \beta_d)'$. In addition, let $\hat{\beta}_{\lambda_n} = (\hat{\beta}'_{S\lambda_n}, \hat{\beta}'_{N\lambda_n})'$, where $\hat{\beta}'_{S\lambda_n}$ and $\hat{\beta}'_{N\lambda_n}$ are the smoothly clipped absolute deviation estimators of β'_S and β'_N, respectively. Under Condition 5, we apply Theorem 2 of Fan & Li (2001) to obtain that, with probability tending to 1, $\hat{\beta}_{S\lambda_n}$ satisfies

$$\frac{1}{n} X'_{\mathcal{S}_T}(Y - X_{\mathcal{S}_T}\hat{\beta}_{S\lambda_n}) + b_n(\hat{\beta}_{S\lambda_n}) = 0, \tag{A4}$$

where $b_n(\beta_S) = (p'_{\lambda_n}(|\beta_1|)\text{sign}(\beta_1), \cdots, p'_{\lambda_n}(|\beta_{d_0}|)\text{sign}(\beta_{d_0}))'$. According to Theorem 1 of Fan & Li (2001), $\hat{\beta}_{S\lambda_n} \to \beta_S \neq 0$ in probability. In addition, because $\lambda_n = \log(n)/\sqrt{n}$, we have $a\lambda_n \to 0$. As a result, $\text{pr}(|\hat{\beta}_{S\lambda_n}| > a\lambda_n) \to 1$, which implies that $\text{pr}\{b_n(\hat{\beta}_{S\lambda_n}) = 0\} \to 1$. This, together with (A4), implies that, with probability tending to 1, the normal equation (A4) is exactly the same as

$$\frac{1}{n} X'_{\mathcal{S}_T}(Y - X_{\mathcal{S}_T}\hat{\beta}_{S\lambda_n}) = 0,$$

which is the normal equation for the ordinary least squares estimator based on the true model. As a result, with probability tending to 1, the smoothly clipped absolute

deviation estimator $\hat{\beta}_{S\lambda_n}$ is exactly the same as $\hat{\beta}_{\mathcal{S}_T} = (X'_{\mathcal{S}_T}X_{\mathcal{S}_T})^{-1}(X_{\mathcal{S}_T}Y)$, the first d_0 elements of the oracle estimator. It follows immediately that $\text{pr}(\text{SSE}_{\lambda_n} = \text{SSE}_{\mathcal{S}_T}) \to 1$, since, with probability tending to one, $\hat{\beta}_{N\lambda_n} = 0$ by the sparsity in Theorem 2 of Fan & Li (2001). Using similar arguments, we can show that, with probability tending to one, the non-vanished diagonal elements of Σ_{λ_n} converge to zero, which implies that $\text{pr}(\text{DF}_{\lambda_n} = d_0) \to 1$. As a result, with probability tending to one, we have $\text{BIC}_\lambda = \text{BIC}_{\mathcal{S}_T}$. This completes the proof. □

Proof of Lemma 4. For $\mathcal{S}_\lambda \neq \mathcal{S}_T$, i.e., $\lambda \in \Omega_- \cup \Omega_+$, we can identify two different cases, i.e. underfitting or overfitting. In each case, we show that Lemma 4 holds as given below.

Case 1: Underfitted model, i.e. $\mathcal{S}_\lambda \not\supset \mathcal{S}_T$. Applying Lemma 3 and Condition 1, we first have that

$$\text{BIC}_{\lambda_n} = \log \hat{\sigma}^2_{\mathcal{S}_T} + d_0 \log(n)/n \to \log(\sigma^2_{\mathcal{S}_T}) = \log(\sigma^2_\epsilon), \tag{A5}$$

in probability. It follows by the fact of $\mathcal{S}_\lambda \not\supset \mathcal{S}_T$ and Conditions 1 and 2 that

$$\text{BIC}_\lambda = \log\left(\frac{1}{n}\text{SSE}_\lambda\right) + \text{DF}_\lambda \frac{\log(n)}{n} \geq \log\left\{\frac{1}{n}\text{SSE}_{\mathcal{S}_\lambda}\right\}$$

$$\geq \min_{\{\mathcal{S}:\mathcal{S}\not\supset\mathcal{S}_T\}} \log \hat{\sigma}^2_{\mathcal{S}} \to \min_{\{\mathcal{S}:\mathcal{S}\not\supset\mathcal{S}_T\}} \log \sigma^2_{\mathcal{S}} > \log \sigma^2_\epsilon, \tag{A6}$$

in probability. Finally, (A5) and (A6) imply that $\text{pr}\{\inf_{\lambda\in\Omega_-} \text{BIC}_\lambda > \text{BIC}_{\lambda_n}\} \to 1$.

Case 2: Overfitted model, i.e. $\mathcal{S}_\lambda \supset \mathcal{S}_T$, *but* $\mathcal{S}_\lambda \neq \mathcal{S}_T$. According to Condition 1, $\hat{\sigma}^2_{\mathcal{S}_T} \to_p \sigma^2_{\mathcal{S}_T} = \sigma^2_\epsilon > 0$. Next, let d_λ be the number of variables included in the model $\mathcal{S}_\lambda$. Then, for the overfitted model, $d_\lambda > d_0$. Moreover, using the theory of the sum of squares decomposition, we can easily show that $(\text{SSE}_{\mathcal{S}_T} - \text{SSE}_{\mathcal{S}_\lambda})/\sigma^2_\epsilon \to \chi^2_{d_\lambda - d_0}$ in distribution as $n \to \infty$. Under the normality assumption, Condition 3, this is also

true for a finite sample. Thus, for any overfitted model $\mathcal{S}$,

$$\text{SSE}_{\mathcal{S}_T} - \text{SSE}_{\mathcal{S}} = O_p(1). \tag{A7}$$

This, together with Lemma 3 and the definition of BIC_λ, implies that, with probability tending to 1,

$$\begin{aligned} n\Big(\text{BIC}_\lambda - \text{BIC}_{\lambda_n}\Big) &\geq n \log\left(\frac{\text{SSE}_{\mathcal{S}_\lambda}}{\text{SSE}_{\mathcal{S}_T}}\right) + (\text{DF}_\lambda - d_0)\log n \\ &= \{\hat{\sigma}_{\mathcal{S}_T}^{-2}(\text{SSE}_{\mathcal{S}_\lambda} - \text{SSE}_{\mathcal{S}_T}) + o_p(1)\} + \{d_\lambda + o_p(\lambda) - d_0\}\log n, \end{aligned}$$

where the last equality follows because $\text{DF}_\lambda = d_\lambda + o_p(\lambda)$ by Conditions 4 and 5. Thus,

$$\inf_{\lambda\in\Omega_+} n\Big(\text{BIC}_\lambda - \text{BIC}_{\lambda_n}\Big) \geq \hat{\sigma}_{\mathcal{S}_T}^2 \min_{\{\mathcal{S}\supset\mathcal{S}_T\}} \Big(\text{SSE}_{\mathcal{S}} - \text{SSE}_{\mathcal{S}_T}\Big) + \Big\{1 + o_p(\lambda)\Big\}\log n + o_p(1). \tag{A8}$$

It follows from (A7) that

$$\min_{\{\mathcal{S}:\mathcal{S}\supset\mathcal{S}_T\}} \Big(\text{SSE}_{\mathcal{S}} - \text{SSE}_{\mathcal{S}_T}\Big) = O_p(1).$$

This, together with the fact $\hat{\sigma}_{\mathcal{S}_T}^2 \to \sigma_{\mathcal{S}_T}^2$ in probability, the right-hand side of (A8) diverges to $+\infty$ as $n \to \infty$, which implies that

$$\text{pr}\left\{\inf_{\lambda\in\Omega_+} n\Big(\text{BIC}_\lambda - \text{BIC}_{\lambda_n}\Big) > 0\right\} = \text{pr}\left(\inf_{\lambda\in\Omega_+} \text{BIC}_\lambda > \text{BIC}_{\lambda_n}\right) \to 1.$$

The results of Cases 1 and 2 complete the proof. □

Regularity Conditions for Partially Linear Model. Suppose that $\{(u_i, x_i, y_i), i = 1, \cdots, n\}$, is a random sample from (4.1). The following regularity conditions are imposed to facilitate the technical proofs:

(i) the kernel function is a symmetric density function with compact support;

(ii) the random variable u_1 has a bounded support $\mathcal{U}$, and its density function $f(\cdot)$ is Lipschitz continuous and bounded away from 0 on its support;

(iii) the function $\alpha(\cdot)$ has a continuous second-order derivative for $u \in \mathcal{U}$;

(iv) the conditional expectation $E(x_1|u_1 = u)$ is Lipschitz continuous for $u \in \mathcal{U}$;

(v) there is an $s > 1$ such that $E\|x_1\|^{2s} < \infty$ and for some $\eta < 2 - s^{-1}$ such that $n^{2\eta-1}h \to \infty$;

(vi) the bandwidth $h = O_P(n^{-1/5})$.

References

Akaike (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Ed. B. N. Petrov & F. Csaki. pp. 267–81. Budapest: Akademia Kiado.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350–83.

Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Ann. Statist.* **32**, 898–927.

Bunea, F. & Wegkamp, M. (2004). Two-stage model selection procedures in partially linear regression. *Can. J. Statist.* **32**, 105–18.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline function: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numer. Math.* **31**, 337–403.

Engle, R. F., Granger, C. W. J., Rice, J. & Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Am. Statist. Assoc.* **81** 310–20.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* New York: Chapman and Hall.

Fan, J., Härdle, W. & Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *Ann. Statist.* **26**, 943-71.

Fan, J. & Huang, T. (2005). Profile likelihood inference on semiparametric varying-coefficient partially linear models. *Bernoulli.* **11**, 1031-57.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalised likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fan, J. & Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Am. Statist. Assoc.* **99**, 710–723.

Härdle, W., Liang, H. & Gao, J. (2000). *Partially Linear Models.* Heidelberg: Springer Physica-Verlag.

Heckman, N. E. (1986). Spline smoothing in partly linear models. *J. R. Statist. Soc.* B. **48**, 244–8.

Huang, J. & Yang, L. (2004). Identification of non-linear additive autoregressive models. *J. R. Statist. Soc.* B. **66**, 463–77.

Mack, Y. P. & Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahr. verw. Geb.* **61**, 405–15.

McQuarrie, D. R. & Tsai, C. L. (1998). *Regression and Time Series Model Selection.* Singapore: World Scientific.

ROBINSON, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, **56**, 931–54.

RUPPERT, D., SHEATHER, S. J. AND WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Am. Statist. Assoc.* **90**, 1257–70.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

SHAO, J. (1997). An asymptotic theory for linear model selection, *Statist. Sinica.* **7**, 221–64.

SHI, P. & TSAI, C. L. (2002). Regression model selection - a residual likelihood approach. *J. R. Statist. Soc.* B. **64**, 237–52.

SHI, P. & TSAI, C. L. (2004). A joint regression variable and autoregressive order selection criterion. *J. Time Ser. Anal.* **25**, 923–41.

SPECKMAN, P. (1988). Kernel smoothing in partially linear models. *J. R. Statist. Soc.* B. **50** , 413–36.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B.* **58**, 267–88.

YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation *Biometrika* **92**, 973–50.

YATCHEW, A. (1997). An elementary estimator for the partially linear model. *Economet. Lett.* **57**, 135-43.

Table 1: Example 1. Simulation results for the linear regression model

σ_ϵ	n	Method	Under-fitted(%)	Correctly fitted(%)	Overfitted(%) 1	2	≥ 3	No. of Zeros I	C	MRME (%)
3	50	$\hat{\lambda}_{\mathrm{GCV}}$	6.4	16.9	23.0	31.6	22.1	0.064	3.279	64.17
		$\hat{\lambda}_{\mathrm{BIC}}$	10.1	30.0	31.1	20.5	8.3	0.101	3.899	62.30
		Oracle	0	100	0	0	0	0	5	30.63
	100	$\hat{\lambda}_{\mathrm{GCV}}$	0.2	24.0	20.4	29.8	25.6	0 02	3.369	57.72
		$\hat{\lambda}_{\mathrm{BIC}}$	1.0	52.5	27.0	14.6	4.9	0.100	4.275	50.43
		Oracle	0	100	0	0	0	0	5	33.05
	200	$\hat{\lambda}_{\mathrm{GCV}}$	0	25.4	35.8	25.4	13.4	0	3.300	55.18
		$\hat{\lambda}_{\mathrm{BIC}}$	0	72.7	21.9	4.5	0.9	0	4.528	42.12
		Oracle	0	100	0	0	0	0	5	34.45
1	50	$\hat{\lambda}_{\mathrm{GCV}}$	0	17.1	24.5	33.2	25.2	0	3.272	55.64
		$\hat{\lambda}_{\mathrm{BIC}}$	0	45.6	23.9	21.1	9.4	0	4.042	40.97
		Oracle	0	100	0	0	0	0	5	30.62
	100	$\hat{\lambda}_{\mathrm{GCV}}$	0	19.0	24.5	31.8	24.7	0	3.324	55.91
		$\hat{\lambda}_{\mathrm{BIC}}$	0	54.9	23.6	16.9	4.6	0	4.277	40.53
		Oracle	0	100	0	0	0	0	5	33.05
	200	$\hat{\lambda}_{\mathrm{GCV}}$	0	48.1	37.5	11.7	2.7	0	3.302	55.00
		$\hat{\lambda}_{\mathrm{BIC}}$	0	81.8	16.4	1.3	0.5	0	4.405	38.36
		Oracle	0	100	0	0	0	0	5	34.42

I, the average number of the three truly nonzero coefficients incorrectly set to zero;
C, the average number of the five true zero coefficients that were correctly set to zero;
MRME, median of relative model error.

Table 2: Example 2. Simulation results for the partially linear regression model

σ_ϵ	n	Method	Under-fitted(%)	Correctly fitted(%)	Overfitted(%) 1	2	≥ 3	No. of Zeros I	C	MRME (%)
3	50	$\hat{\lambda}_{\mathrm{GCV}}$	10.9	15.9	24.6	25.8	22.8	0.112	3.263	66.78
		$\hat{\lambda}_{\mathrm{BIC}}$	15.5	29.3	29.3	18.4	7.5	0.160	3.929	67.04
		Oracle	0	100	0	0	0	0	5	29.29
	100	$\hat{\lambda}_{\mathrm{GCV}}$	0.8	23.1	22.6	29.7	23.8	0 08	3.368	58.15
		$\hat{\lambda}_{\mathrm{BIC}}$	1.9	51.8	29.4	13.1	3.8	0 19	4.301	52.10
	100	Oracle	0	100	0	0	0	0	5	33.58
	200	$\hat{\lambda}_{\mathrm{GCV}}$	0	22.9	21.5	30.5	25.1	0	3.352	54.47
		$\hat{\lambda}_{\mathrm{BIC}}$	0	70.0	16.7	10.9	2.4	0	4.540	43.34
		Oracle	0	100	0	0	0	0	5	34.50
1	50	$\hat{\lambda}_{\mathrm{GCV}}$	0	26.0	25.7	31.0	17.3	0	3.567	51.93
		$\hat{\lambda}_{\mathrm{BIC}}$	0.1	60.3	20.6	13.9	5.1	0 01	4.356	38.31
		Oracle	0	100	0	0	0	0	5	29.30
	100	$\hat{\lambda}_{\mathrm{GCV}}$	0	26.3	27.5	27.5	18.7	0	3.567	50.90
		$\hat{\lambda}_{\mathrm{BIC}}$	0	67.9	18.9	9.9	3.3	0	4.509	39.10
		Oracle	0	100	0	0	0	0	5	33.42
	200	$\hat{\lambda}_{\mathrm{GCV}}$	0	26.5	26.9	28.9	17.7	0	3.582	49.24
		$\hat{\lambda}_{\mathrm{BIC}}$	0	75.7	15.7	7.2	1.4	0	4.656	39.01
		Oracle	0	100	0	0	0	0	5	34.77

I, the average number of the three truly nonzero coefficients incorrectly set to zero;
C, the average number of the five true zero coefficients that were correctly set to zero;
MRME, median of relative model error.

Table 3: Female Labour Supply data. Estimated coefficients and their standard errors.

Variable	Profile LSE	SCAD $\hat{\lambda}_{\text{GCV}}$	SCAD $\hat{\lambda}_{\text{BIC}}$	Best-subset BIC	Best-subset BIC(n=602)
x_1	1.244(0.637)	1.281(0.636)	1.872(0.562)	1.343(0.496)	0
x_1^2	-1.451(0.563)	-1.446(0.559)	-1.841(0.517)	-2.192(0.497)	-0.853(0.119)
x_2	1.520(0.721)	1.602(0.704)	1.357(0.681)	0	0
x_2^2	1.162(0.617)	1.281(0.599)	1.341(0.601)	1.433(0.136)	1.410(0.137)
x_3	-1.229(0.692)	-1.063(0.549)	0	0	0
x_3^2	-0.011(0.276)	0	0	0	0
x_1x_2	-1.781(0.702)	-1.885(0.684)	-1.493(0.653)	0	0
x_1x_3	0.922(0.559)	0.995(0.549)	0	0	0
x_2x_3	0.313(0.485)	0	0	0	0
x_4	0.609(0.130)	0.593(0.130)	0.249(0.055)	0.605(0.129)	0.590(0.129)
x_5	1.194(0.140)	1.183(0.140)	1.030(0.131)	1.168(0.138)	1.172(0.139)
x_6	-0.290(0.189)	-0.028(0.019)	0	0	0
x_7	0.118(0.117)	0.005(0.006)	0	0	0

LSE, least squares estimate
SCAD, smoothly clipped absolute deviation

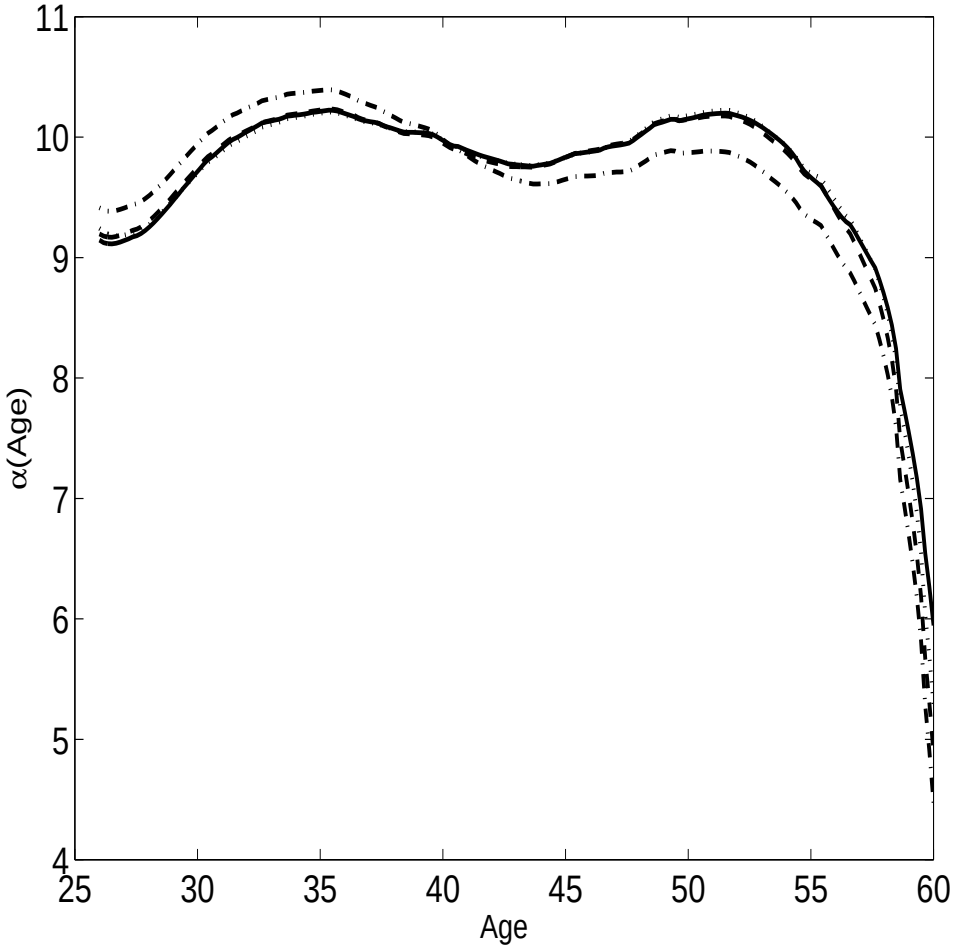

Fig. 1. Female Labour Study data. The plot of $\hat{\alpha}(u)$. The solid curve is $\hat{\alpha}(\cdot)$ based on the penalised profile least squares with $\hat{\lambda}_{\mathrm{BIC}}$, the dashed curve is $\hat{\alpha}(\cdot)$ based on the penalised profile least squares with $\hat{\lambda}_{\mathrm{GCV}}$, the dotted curve is $\hat{\alpha}(\cdot)$ obtained by smoothing partial residuals with the BIC best-subset selection, over u_i, and the dash-dotted curve is $\hat{\alpha}(\cdot)$ based on the unpenalised full-model profile least squares estimate.