Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 41: 725–741, 2014
doi: 10.1111/sjos.12057

New Robust Variable Selection Methods for Linear Regression Models

ZIQI CHEN
School of Mathematics and Statistics, Central South University

MAN-LAI TANG
Department of Mathematics and Statistics, Hang Seng Management College

WEI GAO
Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University

NING-ZHONG SHI
Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University

ABSTRACT. Motivated by an entropy inequality, we propose for the first time a penalized profile likelihood method for simultaneously selecting significant variables and estimating unknown coefficients in multiple linear regression models in this article. The new method is robust to outliers or errors with heavy tails and works well even for error with infinite variance. Our proposed approach outperforms the adaptive lasso in both theory and practice. It is observed from the simulation studies that (i) the new approach possesses higher probability of correctly selecting the exact model than the least absolute deviation lasso and the adaptively penalized composite quantile regression approach and (ii) exact model selection via our proposed approach is robust regardless of the error distribution. An application to a real dataset is also provided.

Key words: adaptive lasso, entropy inequality, oracle properties, penalized profile likelihood, profile likelihood, robustness, variable selection

1. Introduction

In a linear regression setting, biased parameter estimates and prediction results may be produced if a significant explanatory variable is omitted. On the other hand, the efficiency of the resulting estimate may be degraded, and less accurate predictions will be produced when unnecessary predictors are included (Tibshirani, 1996; Hastie *et al.*, 2009). Hence, correct selection of the true model is important (Fan & Li, 2006). Responses of a linear regression model subject to outliers or error with a heavy tail are commonly encountered in practice (Rousseeuw & Leroy, 1987; He *et al.*, 2005). Even worse, the variance of the error may not be finite (e.g. the Cauchy error; Zou & Yuan, 2008; Kai *et al.*, 2011). Hence, it is of great interest to consider the problem of robust model selection for linear regression models (Wang *et al.*, 2007; Lambert-Lacroix & Zwald, 2011).

To shrink unnecessary coefficients to 0 and estimate the significant coefficients in multiple linear models, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (lasso). Zou (2006) proposed the adaptive lasso method for linear regression models to reduce bias in the original lasso. However, the ordinary least squares (OLS) criterion used in lasso or adaptive lasso is sensitive to outliers or heavy-tailed errors and may introduce bias. Besides, the lasso and the adaptive lasso both break down when the variance of the error tends to infinity.

Wang *et al.* (2007) proposed the least absolute deviation (LAD) lasso, which is robust to heavy-tailed errors or outliers in the response. Unfortunately, the LAD lasso estimator is expected to be less efficient than the OLS estimator with adaptive lasso penalty when the error has no heavy tail (e.g. the normal error). Zou & Yuan (2008) proposed the adaptive-lasso-penalized composite quantile regression (ACQR) procedure. They showed that their method works well for the data contaminated with outliers or generated from infinite-variance errors.

Motivated by the entropy inequality (4), we propose a penalized profile likelihood method to select variables and estimate coefficients simultaneously in linear regression models. The new method has built-in robustness because it requires no specification of the error distribution. Our proposed method is robust against outliers or errors with heavy tails and performs well even for errors with infinite variances. We theoretically show by several examples that estimators for the significant coefficients obtained by the new method are more efficient than the ACQR estimators. Furthermore, our proposed method has higher probability of correctly selecting the true model than the ACQR method.

We organize our paper as follows. In Section 2, on the basis of an entropy inequality, we propose the maximum profile likelihood (MPL) estimator and the adaptively penalized MPL (AMPL) estimator. The asymptotic properties of the estimators are given in Section 3. In Section 4, simulation studies are conducted to evaluate the performances of our proposed methods. The AMPL approach is illustrated with a plasma retinol level dataset. A brief discussion is presented in Section 5. The technical conditions and lemmas are relegated to the Appendix. The detailed proofs of Lemmas 3, 4, and 5 and Theorems 1, 2, and 3 are put in the Supporting Information.

2. The penalized profile likelihood

We consider the following linear regression model:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, \cdots, n. \tag{1}$$

Let $\beta = (\beta_1, \ldots, \beta_p)'$ and β^* be the corresponding true parameter.

Define $X_i = (x_{i1}, \ldots, x_{ip})'$. Suppose $(y_i, X_i)_{i=1}^n$ are n independent samples from (1) with $0 < \mathrm{Var}(X_1) < \infty$, X_i is independent of ϵ_i $(i = 1, \cdots, n)$, and $\beta^* \in \mathrm{interior}(\Theta)$ with Θ being a compact subset of R^p. Denote $\epsilon_i(\beta) = y_i - \sum_{j=1}^p x_{ij}\beta_j$. Let $f_{\epsilon(\beta)}$ and f_0 be the probability density function of $\epsilon_1(\beta)$ and the true density of ϵ_1, respectively. It is noted that $f_{\epsilon(\beta^*)} = f_0$. Denote $\mathcal{A} = \{j : \beta_j^* \neq 0\}$. We assume that the number of elements in $\mathcal{A}$ (i.e. q) is less than p; that is, the true model depends only on a subset of the predictors.

Fan & Li (2001) showed that we could use the penalized likelihood method to shrink unnecessary coefficients to 0 and estimate the significant coefficients in (1) if one ideally knows the error distribution. The resultant penalized likelihood estimators of the significant coefficients are efficient by the oracle properties. Unfortunately, the error density in a linear regression model is practically unknown.

2.1. The penalized likelihood for the known error density

Ideally, if one knows the true density function of ϵ, that is, $f_0(u)$, β could be estimated through the maximum likelihood (ML) estimation procedure, that is,

$$\hat{\beta}^{\mathrm{ML}} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \log f_0\{\epsilon_i(\beta)\}.$$

For example, the ML estimator is essentially the OLS estimator when the error ϵ follows the standard normal distribution. $\hat{\beta}^{\mathrm{ML}}$ is reputed to be an efficient estimator. It is well known that the consistency of $\hat{\beta}^{\mathrm{ML}}$ is based on the inequality

$$\int f_{\epsilon(\beta^*)}(u) \log f_{\epsilon(\beta)}(u)\, \mathrm{d}u < \int f_{\epsilon(\beta^*)}(u) \log f_{\epsilon(\beta^*)}(u)\, \mathrm{d}u, \tag{2}$$

for any $\beta \neq \beta^*$. Note that $f_{\epsilon(\beta^*)} = f_0$.

Zou (2006) proposed the adaptively penalized ML (AML) estimator of β^*, which is given by

$$\hat{\beta}^{\mathrm{AML}} = \arg\min_{\beta} \left[-\sum_{i=1}^{n} \log f_0\{\epsilon_i(\beta)\} + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j^{\mathrm{ML}}|^2} \right], \tag{3}$$

where λ is the tuning parameter. For example, if the error follows the standard normal distribution, the AML estimator is indeed the adaptive lasso estimator. $\hat{\beta}^{\mathrm{AML}}$ is based on the true error density and is therefore an efficient estimator.

2.2. The penalized profile likelihood for the unknown error distribution

Unfortunately, one seldom knows the density of ϵ. By Lemma 1, we have

$$\int f_{\epsilon(\beta)}(u) \log f_{\epsilon(\beta)}(u)\, \mathrm{d}u < \int f_{\epsilon(\beta^*)}(u) \log f_{\epsilon(\beta^*)}(u)\, \mathrm{d}u, \tag{4}$$

for any $\beta \neq \beta^*$. We call (4) the entropy inequality based on model (1). Hence, we can write

$$\beta^* = \arg\max_{\beta \in \Theta} \int f_{\epsilon(\beta)}(u) \log f_{\epsilon(\beta)}(u)\, \mathrm{d}u, \tag{5}$$

and β^* is the unique value that satisfies (5). Let $\epsilon_i(\beta) = y_i - \sum_{j=1}^{p} x_{ij}\beta_j$. The density function of $\epsilon(\beta)$ can be estimated by

$$\hat{f}_{\epsilon(\beta)}(u) = \frac{1}{nh} \sum_{l=1}^{n} K\left(\frac{\epsilon_l(\beta) - u}{h}\right), \tag{6}$$

where K is a scalar kernel and h is any appropriate bandwidth. Thus, β^* can be estimated by the sample analogue of (5), that is,

$$\hat{\beta} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_{\epsilon(\beta)(-i)}(\epsilon_i(\beta)), \tag{7}$$

where $\hat{f}_{\epsilon(\beta)(-i)}(\cdot)$ is the estimator of the density function of $\epsilon(\beta)$ obtained by leaving out the i-th observation by (6).

The estimation procedure proposed here is analogous to the profile likelihood approach for estimating regression parameters in semiparametric models (see also, Severini & Wong, 1992; Fan *et al.*, 2007; Lombardia & Sperlich, 2008). The basic idea of profile likelihood is to replace the unknown function by its non-parametric (kernel) estimate for given parametric components. For this reason, we also call our proposed estimator $\hat{\beta}$ the MPL estimator and denote it as $\hat{\beta}^{\mathrm{MPL}}$. A similar estimation procedure also appeared in Linton *et al.* (2008) for estimating the parameters in their semiparametric transformation models.

If the true density of ϵ is unknown, we can obtain the kernel estimator of the density of $\epsilon_1(\beta)$ as (6). Thus, we construct (7) to obtain the estimator of β^*. As the sample size (i.e. n) goes to infinity, we have that $\hat{f}_{\epsilon(\beta)(-i)}(u)$ is equivalent to $f_{\epsilon(\beta)}(u)$ for any $\beta \in \mathrm{interior}(\Theta)$. By (5),

we can find that our proposed estimator is a consistent estimator of β^*. Theorem 1 shows the consistency of the MPL estimator theoretically. Therefore, our proposed method works well without the specification of the error density. In other words, our proposed estimator is robust regardless of the error density. In particular, the proposed estimate is robust to outliers or errors with heavy tails or infinite variances. We will further demonstrate this point through some real examples in Section 3.3 and some simulation studies.

According to equation (4), our proposed estimator is a consistent estimator of β^* intuitively. It is noteworthy that equation (4) is different from equation (2), on which the consistency of the ML estimator is based.

It is noted that $\hat{\beta}^{\mathrm{MPL}}$ is a consistent estimator of β^* by Theorem 1. Following the adaptive lasso idea of Zou (2006) and on the basis of the entropy inequality (4), we adopt $\hat{\beta}^{\mathrm{MPL}}$ to construct an adaptively weighted lasso penalty and define the penalized profile likelihood as follows:

$$\sum_{i=1}^{n} \log \hat{f}_{\epsilon(\beta)(-i)}\{\epsilon_i(\beta)\} - \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j^{\mathrm{MPL}}|^2}, \tag{8}$$

where λ is the tuning parameter. Maximizing the penalized profile likelihood function (8) is equivalent to minimizing

$$Q(\beta) := -\sum_{i=1}^{n} \log \hat{f}_{\epsilon(\beta)(-i)}\{\epsilon_i(\beta)\} + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j^{\mathrm{MPL}}|^2}. \tag{9}$$

By minimizing (9) with respect to β, we obtain the AMPL estimator of β^* (denoted as $\hat{\beta}^{\mathrm{AMPL}}$). It can be seen that the penalized profile likelihood combines the profile likelihood and the adaptive lasso penalty. Hence, the resulting estimator is expected to be robust regardless of the error density. In particular, the AMPL is robust against outliers or errors with heavy tails and works well even for errors with infinite variance.

Remarks 1. For any given likelihood, penalized likelihood estimators have been extensively discussed (e.g. Fan & Li, 2001; Zou, 2006; Meier *et al.*, 2008; Zou & Li, 2008). In practice, the error distribution of the linear regression model is seldom known. Even worse, if one wrongly specifies the distribution of the error, the existing methods may work badly. For example, if the true error distribution is Cauchy and we wrongly specify the error density as normal, the penalized normal likelihood with adaptive lasso penalty (i.e. the adaptive lasso) would lead to biased estimator and large model error and may not be able to select the exact model as indicated in the simulation studies. Our proposed adaptive MPL lasso estimation method works well without error distribution specification. Thus, our proposed methodology is practically more flexible and could be widely applied.

2.3. Computations and tuning

We propose the so-called iterative marginal optimization (IMO) algorithm (Wang, 2007) to obtain the AMPL estimator of β^*. We summarize the procedure as follows:

Step 0. A convenient and good initial value for $\beta^{(0)}$ is $\hat{\beta}^{\mathrm{MPL}}$.
Step $m+1$. Through the following p grid steps, we obtain $\beta^{(m+1)}$.
Grid step k ($k = 1, \ldots, p$). β_k is updated as

$$\beta_k^{(m+1)} = \arg\min_{\beta_k} Q\left(\beta_1^{(m+1)}, \ldots, \beta_{k-1}^{(m+1)}, \beta_k, \beta_{k+1}^{(m)} \ldots, \beta_p^{(m)}\right).$$

Until convergence, we obtain the estimator $\hat{\beta}^{\mathrm{AMPL}}$.

The original p-dimensional joint optimization problem (9) is rewritten as the marginal univariate optimization problems by the IMO algorithm so that the marginal optimizer could be found in an efficient way. The IMO algorithm is computationally stable because the value of $Q(\beta^{(m)})$ decreases in both step (m) and grid step (k). According to our experiences, the IMO algorithm converges quickly and is computationally efficient. Note that the IMO algorithm is identical to the idea of the coordinate descent method proposed by Friedman *et al.* (2010).

Regarding the selection of the tuning parameter λ, we choose λ such that it minimizes the following Bayesian information criterion (BIC)-like criterion (Wang *et al.*, 2009; Kai *et al.*, 2011):

$$BIC(\lambda) = \log\left[-\sum_{i=1}^{n} \log \hat{f}_{\epsilon(\beta)(-i)}\{\epsilon_i(\beta)\}|_{\beta=\hat{\beta}_{\lambda}^{\mathrm{AMPL}}}\right] + \frac{\log(n)}{n} df_{\lambda},$$

where $\hat{\beta}_{\lambda}^{\mathrm{AMPL}}$ is the penalized profile likelihood estimator of β^* with tuning parameter λ and df_{λ} is the number of non-zero coefficient in $\hat{\beta}_{\lambda}^{\mathrm{AMPL}}$.

3. Asymptotic properties of the estimators

In this section, we study the large-sample properties of the estimators given in Section 2.

3.1. The maximum profile likelihood estimator

For any function φ, we define $\dot{\varphi} := \partial\varphi/\partial\beta$ and $\dot{\hat{\varphi}} := \partial\hat{\varphi}/\partial\beta$. Similarly, we define for any function φ: $\varphi'(u) := \partial\varphi(u)/\partial u$ and $\hat{\varphi}'(u) := \partial\hat{\varphi}(u)/\partial u$. We now show that $\hat{\beta}^{\mathrm{MPL}}$ is consistent and possesses asymptotic normality. The proofs of the following two theorems are provided in the Supporting Information.

Theorem 1. *Suppose conditions* (a)–(e) *in the Appendix hold, then* $\hat{\beta}^{\mathrm{MPL}} \to_p \beta^*$.

Theorem 2. *Under conditions* (a)–(e) *in the Appendix, we have*

$$\begin{aligned}\sqrt{n}(\hat{\beta}^{\mathrm{MPL}} - \beta^*) &= -\Gamma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\dot{f}_{\epsilon(\beta^*)}(\epsilon_i(\beta^*)) + f'_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))(-X_i)}{f_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))} + o_p(1)\\ &= \Gamma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{f_0'(\epsilon_i(\beta^*))}{f_0(\epsilon_i(\beta^*))}(X_i - EX_i) + o_p(1),\end{aligned}$$

equivalently,

$$\sqrt{n}(\hat{\beta}^{\mathrm{MPL}} - \beta^*) \to_L N(\mathbf{0}, a(\mathrm{Var}X_1)^{-1}),$$

where $\Gamma = [2E\{f_0''(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\} - E\{f_0'(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\}^2]\mathrm{Var}X_1$ *and* $a = [\mathrm{Var}\{f_0'(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\}]/[2E\{f_0''(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\} - E\{f_0'(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\}^2]^2$.

Assume that $C =: \lim_{n\to\infty}\mathbf{X}\mathbf{X}'/n$ is a $p \times p$ positive definite matrix, where $\mathbf{X} = (X_1, \ldots, X_n)$ is the predictor matrix.

Let $\hat{\beta}^{\mathrm{OLS}}$ be the OLS estimate of β^*. We have

$$\sqrt{n}(\hat{\beta}^{\mathrm{OLS}} - \beta^*) \to_L N(0, \sigma^2 C^{-1}),$$

where σ^2 is the variance of ϵ. Therefore, $\hat{\beta}^{\mathrm{OLS}}$ is no longer a $\sqrt{n}$-consistent estimator if the variance of ϵ is infinite. On the other hand, according to Theorem 2, the asymptotic variance of $\hat{\beta}^{\mathrm{MPL}}$ depends on the density of ϵ, but not directly on the variance of ϵ. This intuitively suggests that $\hat{\beta}^{\mathrm{MPL}}$ enjoys a $\sqrt{n}$-consistency property even when σ^2 is infinite.

Remarks 2. If we know that the true density of $\epsilon_1(\beta^*)$ is $f_0(u)$ with support being R (e.g. the normal distribution), we have

$$\sqrt{n}(\hat{\beta}^{\mathrm{ML}} - \beta^*) = -\left[\frac{1}{n}\sum_{i=1}^{n}\{\frac{f_0^{'}(\epsilon_i(\beta^*))}{f_0(\epsilon_i(\beta^*))}\}^2 X_i X_i^{'}\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{f_0^{'}(\epsilon_i(\beta^*))}{f_0(\epsilon_i(\beta^*))}X_i + o_p(1).$$

Because $E\{f_0^{''}(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\} = 0$, we have

$$\sqrt{n}(\hat{\beta}^{\mathrm{MPL}} - \beta^*) \to_L N(\mathbf{0}, \Omega^{-1}),$$

where $\Omega = \mathrm{Var}\{f_0^{'}(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\}\mathrm{Var}X_1$. This indicates that $\hat{\beta}^{\mathrm{MPL}}$ is asymptotically distributed the same with the ML estimator $\hat{\beta}^{\mathrm{ML}}$ assuming $EX_1 = 0$ without loss of generality. In this case, $\hat{\beta}^{\mathrm{MPL}}$ is also an efficient estimator of β^*.

Remarks 3. When the true (unknown) distribution of $\epsilon_1(\beta^*)$ is the truncated normal distribution with density function $f_0(u) = \exp(-u^2/2)/\int_{-3}^{3}\exp(-t^2/2)\,\mathrm{d}t$, for $-3 \le u \le 3$ and zero elsewhere, we have $2E\{f_0^{''}(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\} - E\{f_0^{'}(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\}^2 = 2E(-1+\epsilon_1(\beta^*)^2) - E\epsilon_1(\beta^*)^2 < -1$. Hence,

$$\sqrt{n}(\hat{\beta}^{\mathrm{MPL}} - \beta^*) \to_L N(\mathbf{0}, a(\mathrm{Var}X_1)^{-1}),$$

where $a < \mathrm{Var}(\epsilon_1(\beta^*)) < 1$. Note that Γ is negative definite. In this case, we have $\sqrt{n}(\hat{\beta}^{\mathrm{OLS}} - \beta^*) \to_L N(\mathbf{0}, \mathrm{Var}(\epsilon_1(\beta^*))(\mathrm{Var}X_1)^{-1})$ assuming $EX_1 = 0$ without loss of generality. In other words, $\hat{\beta}^{\mathrm{MPL}}$ is more efficient than the OLS estimator.

3.2. The adaptively penalized maximum profile likelihood estimator

Without loss of generality, suppose that the first q coefficients of β are non-zero, that is, $\beta_j \neq 0$, for $j = 1, 2, \ldots, q$, and $\beta_j = 0$, for $j = q+1, \ldots, p$. Denote $\mathcal{A} := \{1, 2, \ldots, q\}$, $X_i^* := (x_{i1}, \cdots, x_{iq})^{'}$, $\beta_{\mathcal{A}} := (\beta_1, \cdots, \beta_q)^{'}$, and $\beta_{\mathcal{A}^c} := (\beta_{q+1}, \cdots, \beta_p)^{'}$. The oracle who knows the true subset $\mathcal{A}$ would use the model

$$y_i = X_i^{*'}\beta_{\mathcal{A}}^* + \epsilon_i, \quad i = 1, \ldots, n.$$

We define the MPL oracle estimator as

$$\hat{\beta}^{\mathrm{MPL}}_{(\mathrm{oracle})\mathcal{A}} = \arg\max_{\beta_{\mathcal{A}}} \sum_{i=1}^{n} \log \hat{f}_{\epsilon(\beta_{\mathcal{A}})(-i)}\{\epsilon_i(\beta_{\mathcal{A}})\}, \tag{10}$$

and $\hat{\beta}^{\mathrm{MPL}}_{(\mathrm{oracle})\mathcal{A}^c} = 0$, where $\hat{f}_{\epsilon(\beta_{\mathcal{A}})(-i)}(u) = \sum_{j\neq i} K\{(\epsilon_j(\beta_{\mathcal{A}}) - u)/h\}/(nh)$ and $\epsilon_j(\beta_{\mathcal{A}}) = y_j - \sum_{l=1}^{q} x_{jl}\beta_l$. According to Theorem 2, we have

$$\sqrt{n}(\hat{\beta}^{\mathrm{MPL}}_{(\mathrm{oracle})\mathcal{A}} - \beta_{\mathcal{A}}^*) \to_L N(0, a(\mathrm{Var}X_1^*)^{-1}). \tag{11}$$

We show that the AMPL estimator enjoys the oracle properties of the MPL oracle. The proof of the following theorem is presented in the Supporting Information.

Theorem 3. *(Oracle properties) Assume that conditions* (a)–(e) *in the Appendix hold. If* $\lambda/\sqrt{n} \to 0$ *and* $\sqrt{n}\lambda \to \infty$, *then* $\hat{\beta}^{\mathrm{AMPL}}$ *satisfies the following*:

(a) *Consistent selection*: $\Pr(\{j : \hat{\beta}_j^{\mathrm{AMPL}} \neq 0\} = \mathcal{A}) \to 1$.

(b) *Efficient estimation*: $\sqrt{n}(\hat{\beta}_{\mathcal{A}}^{\mathrm{AMPL}} - \beta_{\mathcal{A}}^*) \to_d N(0, a(\mathrm{Var}X_1^*)^{-1})$, *where* $a(\mathrm{Var}X_1^*)^{-1}$ *is the asymptotical covariance matrix knowing the true subset model as in* (11) (*i.e.* $\sqrt{n}(\hat{\beta}_{\mathcal{A}}^{\mathrm{AMPL}} - \beta_{\mathcal{A}}^*)$ *is asymptotically distributed, the same as* $\sqrt{n}(\hat{\beta}_{(\mathrm{oracle})\mathcal{A}}^{\mathrm{MPL}} - \beta_{\mathcal{A}}^*)$).

3.3. The asymptotic relative efficiency

We study the asymptotic relative efficiency (ARE) of our MPL estimator with respect to the composite quantile regression (CQR), OLS, and LAD estimators. Because the AMPL, ACQR, adaptive lasso, and LAD lasso estimators all possess oracle properties, the ARE of the AMPL estimators of the significant coefficients with respect to the ACQR (adaptive lasso and LAD lasso) estimators of the non-zero coefficients could be discussed similarly.

Suppose the CQR method uses quantiles τ_i, $i = 1, \ldots, K$, and the parameter associated with τ_k is b_k, which is given in equation (2.3) of Zou & Yuan (2008). From Theorem 2.1 of Zou & Yuan (2008) and Theorem 2, we obtain the ARE of the MPL with respect to the CQR as follows:

$$\mathrm{ARE}(\tau_1, \ldots, \tau_K, a, f_0) = \frac{\{\sum_{i,j=1}^{K} \min(\tau_i, \tau_j)(1 - \max(\tau_i, \tau_j))\}}{a\{\sum_{i=1}^{K} f_0(b_{\tau_i}^*)\}^2}.$$

Let $\mathrm{ARE}(K, a, f_0) := \mathrm{ARE}(\tau_1, \ldots, \tau_K, a, f_0)$ when $\tau_i = i/(K+1)$, for $i = 1, \ldots, K$. Following Zou & Yuan (2008), we consider the case when $K \to \infty$. Let $\delta(a, f_0) = \lim_{K\to\infty} \mathrm{ARE}(K, a, f_0)$. From Theorem 3.1 of Zou & Yuan (2008), we have

$$\delta(a, f_0) = \frac{1}{12a\{Ef_0(\epsilon_1(\beta^*))\}^2}.$$

We calculate the AREs under the normal error, the heavy-tailed errors (e.g. T-distribution and logistic distribution), and the error of infinity variance (e.g. Cauchy distribution).

Normal distribution. Let the true density of the error ϵ follow $N(0, \sigma^2)$. We have $Ef_0(\epsilon_1(\beta^*)) = 1/(2\sqrt{\pi}\sigma)$, $a = \sigma^2$, and $\delta(a, f_0) = 1.047$. Hence, our MPL estimator is slightly more efficient than the CQR estimator. In addition, the ARE of our MPL estimator with respect to the LAD estimator is 1.576, which means that our MPL is obviously more efficient than the LAD. It is noted that our MPL estimator is as efficient as the OLS estimator under a normality assumption and is thus the most efficient estimator. We have that the OLS estimator is more efficient than the LAD estimator under the normality assumption.

T-distribution. Suppose the true density of ϵ is the T-distribution with degrees of freedom $\nu > 2$. We have that $a = (\nu+3)/(\nu+1)$ and

$$\delta(a, f_0) = \frac{\pi\nu(\nu+1)}{12(\nu+3)}\left(\frac{\Gamma(\nu/2)}{\Gamma((\nu+1)/2)}\right)^4\left(\frac{\Gamma(\nu+1)}{\Gamma(\nu+1/2)}\right)^2.$$

We plot the ARE between CQR estimator and our proposed MPL estimator in Fig. 1(A), and we can conclude that our MPL estimator is slightly more efficient than the CQR estimator. We also calculate the ARE of our MPL estimator with respect to the OLS estimator, which is $\{\nu(\nu+1)\}/\{(\nu-2)(\nu+3)\}$. According to Fig. 1(B), we observe that our MPL estimator is more efficient than the OLS estimator for small degrees of freedom and the ARE approaches 1 for

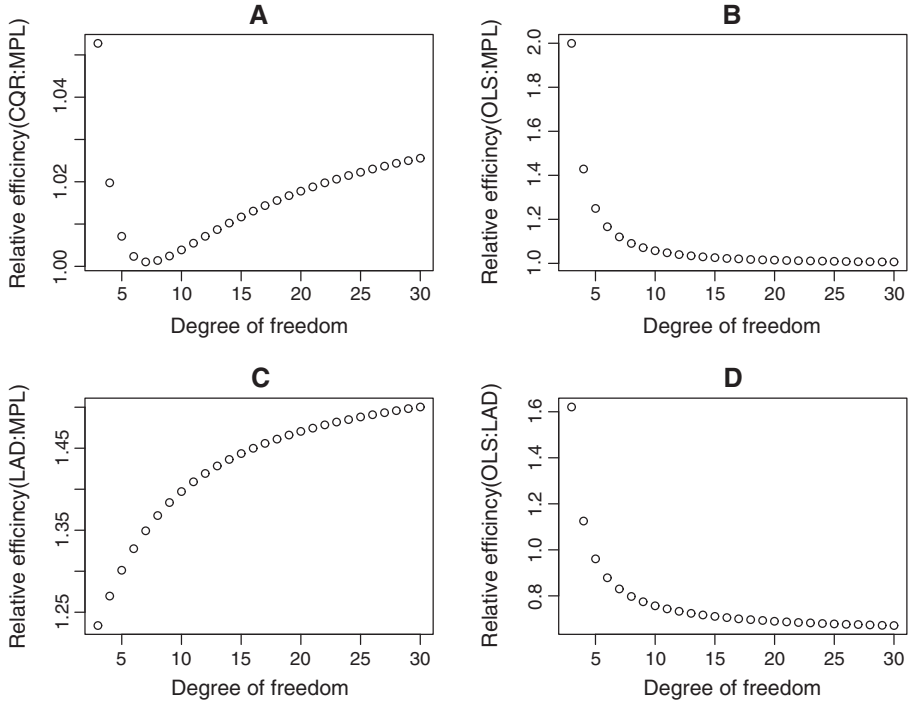

Fig. 1. The relative efficiency as a function of the degrees of freedom for the T-distribution.

large degrees of freedom. The ARE of our MPL estimator with respect to the LAD estimator is plotted in Fig. 1(C), which indicates that our MPL estimator is more efficient than the LAD. We observe from Fig. 1(D) that the LAD is more efficient than the OLS for heavy-tailed errors (i.e. T-distributions with small degrees of freedom). On the other hand, the LAD is less efficient than the OLS for large degrees of freedom (i.e. error has no heavy tail).

Logistic distribution. Suppose the true density of ϵ is the logistic distribution (i.e. $f_0(\epsilon) = e^{\epsilon}/(1+e^{\epsilon})^2$). We have $Ef_0(\epsilon_1(\beta^*)) = 1/6$, $a = 3$, and $\delta(a, f_0) = 1$. In other words, our MPL estimator is as efficient as the CQR estimator. It is noted that both the MPL and CQR estimators can asymptotically achieve the information bound for the logistic distribution error. In addition, the ARE of the MPL estimator with respect to the OLS estimator is 1.097, which suggests that the MPL estimator is slightly more efficient than the OLS estimator. Moreover, the ARE of the MPL estimator with respect to the LAD estimator is 1.333, which indicates that our MPL is more efficient than the LAD.

Standard Cauchy distribution. Suppose the density of ϵ follows the standard Cauchy distribution, whose variance is infinite. We have $\delta(a, f_0) = 1.645$, so our MPL estimator is significantly more efficient than the CQR estimator. In addition, the ARE of the MPL estimator with respect to the LAD estimator is 1.234, which tells that our MPL is more efficient than the LAD. Besides, our proposed MPL estimator can asymptotically attain the information bound when the error follows the Cauchy distribution. Unlike the asymptotic variance of the OLS estimator, the asymptotic variance of our MPL estimator is finite.

We have several conclusions after calculating the ARE. First, our MPL is based on the estimated density of the error. On the other hand, the CQR is obtained via the information of

K quantiles of the error only. That is, the MPL borrows more information of the error to estimate the unknown parameters. Intuitively, our MPL estimator generally possesses smaller asymptotic variance than the CQR estimator theoretically. Second, we observe that our MPL estimators are more efficient than the OLS estimators for heavy-tailed errors. Third, the MPL works well, whereas the OLS breaks down for the standard Cauchy error. Finally, the efficiency of our MPL estimators is superior to that of the LAD estimators theoretically.

4. Numerical studies

In this section, we conduct several simulation studies to evaluate the finite-sample performance of our proposed estimators (i.e. $\hat{\beta}^{\mathrm{MPL}}$ and $\hat{\beta}^{\mathrm{AMPL}}$) in Section 2 and illustrate the proposed AMPL approach on a real dataset in a health study.

4.1. Study 1: the maximum profile likelihood estimator

In this subsection, we investigate the finite-sample performance of the proposed MPL (i.e. $\hat{\beta}^{\mathrm{MPL}}$) estimator in Section 2. We generate 300 datasets, each consisting of $n = 100$ observations from model (1). Here, $\beta^* = (3, 1.5, 2)$ and (x_1, x_2, x_3) follows a multivariate normal distribution $N(0, \Sigma_X)$, where $(\Sigma_X)_{i,j} = 0.5^{|i-j|}$ for $1 \le i, j \le 3$. In our simulation, we consider the following error distributions: $N(0, 3)$, mixture of normals $0.9N(0, 1) + 0.1N(0, 10^2)$, T-distribution with 3 degrees of freedom, chi-square distribution with 3 degrees of freedom and standard Cauchy. We compare our proposed method with the OLS method, the LAD method, and the CQR approach. Zou & Yuan (2008) used 19 quantiles, and Kai *et al.* (2011) adopted nine quantiles in their respective simulation studies. Following their set-ups, we consider the number of quantiles $K = 9$ and 19 in our simulations. We use Silverman's rule-of-thumb bandwidth for our MPL approach.

Table 1 summarizes the results for the estimates of $\hat{\beta}$ based on 300 simulated datasets. Here, 'Bias' represents the sample average over 300 estimates subtracting β^*, and 'SD' represents the sample standard deviation over 300 estimates. SD can be viewed as the true standard deviation of the resulting estimates and can thus be used to measure the efficiency of the four methods. We observe that our MPL approach is robust against outliers (e.g. the mixed normal error) or the heavy-tailed errors (e.g. the T-distribution error). Our proposed MPL method yields unbiased estimates with smaller SDs when the variance of the error is infinite (e.g. Cauchy distribution). On the contrary, the OLS method produces severely biased estimators with inflated standard deviations. It is obvious that our MPL approach is more efficient than the OLS method for non-normal distributions. It is noted that our MPL estimators tend to have less biases than the OLS estimates.

Our MPL method is more efficient than the LAD for the normal or chi-squared error. It is also important to see that the LAD performs worse (with larger SDs) than the OLS as well as the CQR when the error follows the normal distribution. For the standard Cauchy error, the MPL estimators are slightly more efficient than the LAD estimators in theory. However, we see from our simulation results that the LAD produces slightly more efficient estimators than the MPL. The non-parametric technique used in MPL, which results in loss of efficiency in practice, may cause this phenomenon.

Our MPL approach produces more efficient estimators than the CQR when the error follows the chi-squared or Cauchy distribution. Theoretically, our MPL estimator is slightly more efficient than the CQR estimator for normal and T-distribution errors. In the present simulation studies, the CQR estimate appears to be slightly more efficient than the MPL estimator. This phenomenon is also due to the fact that our MPL approach adopts the non-parametric technique, which may lead to some loss in efficiency of the MPL estimator in practice. We can

Table 1. *Summary of bias and standard deviation over* **300** *simulations for model (1) with different error distributions*

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$	
Methods	Bias	SD	Bias	SD	Bias	SD
$N(0, 3)$						
OLS	0.0134	0.2015	−0.0106	0.2365	0.0034	0.2300
LAD	0.0202	0.2318	−0.0175	0.2626	−0.0060	0.2611
CQR_9	0.0108	0.2058	−0.0060	0.2215	−0.0010	0.2261
CQR_{19}	0.0133	0.2045	−0.0052	0.2247	−0.0038	0.2246
MPL	0.0089	0.2262	−0.0013	0.2329	−0.0027	0.2272
$0.9N(0, 1) + 0.1N(0, 10^2)$						
OLS	0.0143	0.3897	0.0141	0.4248	−0.0017	0.4087
LAD	0.0042	0.1640	−0.0033	0.1628	0.0044	0.1474
CQR_9	0.0067	0.1497	0.0052	0.1574	−0.0008	0.1398
CQR_{19}	0.0071	0.1500	0.0040	0.1583	−0.0008	0.1416
MPL	0.0058	0.1468	−0.0011	0.1658	0.0003	0.1485
T-distribution with $df = 3$						
OLS	0.0138	0.1956	−0.0232	0.2323	0.0141	0.2081
LAD	0.0056	0.1617	−0.0112	0.1628	0.0118	0.1556
CQR_9	0.0089	0.1441	−0.0120	0.1495	0.0066	0.1539
CQR_{19}	0.0081	0.1457	−0.0114	0.1528	0.0064	0.1541
MPL	0.0043	0.1600	−0.0068	0.1860	0.0099	0.1799
Chi-squared with $df = 3$						
OLS	−0.0324	0.4243	−0.0374	0.4925	0.0429	0.4456
LAD	−0.0610	0.5568	0.0247	0.5187	0.0272	0.4993
CQR_9	0.0002	0.2353	−0.0485	0.2451	0.0187	0.2328
CQR_{19}	−0.0002	0.2312	−0.0468	0.2434	0.0175	0.2306
MPL	0.0159	0.1805	−0.0014	0.1918	0.0166	0.1815
Standard Cauchy						
OLS	−1.0815	13.316	−0.5390	10.002	0.9160	12.195
LAD	−0.0087	0.1940	0.0040	0.1864	−0.0014	0.1881
CQR_9	−0.0027	0.2239	−0.0040	0.2271	0.0008	0.2256
CQR_{19}	−0.0054	0.2249	−0.0004	0.2323	−0.0002	0.2268
MPL	0.0004	0.2121	0.0028	0.2260	−0.0008	0.2045

SD, standard deviation; OLS, ordinary least squares; LAD, least absolute deviation; CQR, composite quantile regression; MPL, maximum profile likelihood.

see that the estimator of CQR with $K = 9$ is more efficient than that with $K = 19$ in some cases, whereas CQR with $K = 19$ is more efficient than that with $K = 9$ in other cases. This suggests that the optimal number of quantiles K should be selected carefully in order that the CQR achieves the expected efficiency.

4.2. Study 2: the adaptively penalized maximum profile likelihood estimator

In this subsection, we conduct a simulation study to compare the AMPL method with the adaptive lasso proposed by Zou (2006), the LAD lasso (using the adaptive lasso penalty; Wang *et al.*, 2007), and the adaptive-lasso-penalized CQR (ACQR) approach. Here, 100 datasets, each consisting of $n = 100$ observations, are generated from model (1), where $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and the predictors $(x_1, x_2, \ldots, x_8)$ is distributed as $N(0, \Sigma_X)$ with $(\Sigma_X)_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 8$. This regression model has been considered by Tibshirani (1996), Fan & Li (2001), Zou (2006), and Zou & Yuan (2008). The error distributions considered here are identical to those in study 1.

The simulation results are summarized in Table 2. Here, the model error is computed by $E\{(\hat{\beta} - \beta^*)' \Sigma_X (\hat{\beta} - \beta^*)\}$ (Zou & Yuan, 2008). Column 'C' gives the average number of zero coefficients corrected selected to be zero, and the column 'IC' shows the average number of non-zero coefficients incorrectly estimated to be zero. 'U-fit' gives the probability of datasets excluding any non-zero coefficients in 100 replicates, 'C-fit' presents the proportion of trials selecting the true subset model, and the 'O-fit' tells the probability of trials including the three non-zero coefficients and some zero components (Kai *et al.*, 2011). Table 2 shows that our proposed AMPL approach performs well in variable selection. In particular, the AMPL is robust to outliers (e.g. the mixed normal error) or errors with heavy tails (e.g. the T-distribution error) and works well even for error with infinite variance (e.g. the standard Cauchy error). More precisely, our proposed method yields smaller model errors than the adaptive lasso, which suggests the superiority of our proposed method over the adaptive lasso in variable selection. It is noted that the adaptive lasso breaks down for random error with infinite variance (e.g. Cauchy distribution). Most importantly, for each

Table 2. *Comparisons of variable selection methods for model (1) with different error distributions.*

Methods	Model error	No. of zeros		Proportion of fits		
		C	IC	U-fit	C-fit	O-fit
$N(0, 3)$						
Adaptive lasso	0.1367	4.68	0.00	0.00	0.75	0.25
LAD lasso	0.1563	4.92	0.00	0.00	0.93	0.07
ACQR_9	0.1374	4.71	0.00	0.00	0.72	0.28
ACQR_{19}	0.1679	4.28	0.00	0.00	0.42	0.58
AMPL	0.1241	4.99	0.00	0.00	0.99	0.01
$0.9N(0, 1) + 0.1N(0, 10^2)$						
Adaptive lasso	0.7877	3.92	0.01	0.01	0.36	0.63
LAD lasso	0.0589	4.99	0.00	0.00	0.99	0.01
ACQR_9	0.0639	4.89	0.00	0.00	0.89	0.11
ACQR_{19}	0.0746	4.70	0.00	0.00	0.71	0.29
AMPL	0.0454	5.00	0.00	0.00	1.00	0.00
T-distribution with $df = 3$						
Adaptive lasso	0.1238	4.79	0.00	0.00	0.82	0.18
LAD lasso	0.0705	4.96	0.00	0.00	0.96	0.04
ACQR_9	0.0662	4.91	0.00	0.00	0.92	0.08
ACQR_{19}	0.0736	4.77	0.00	0.00	0.81	0.19
AMPL	0.0731	4.99	0.00	0.00	0.99	0.01
Chi-squared with $df = 3$						
Adaptive lasso	0.6211	4.54	0.00	0.00	0.67	0.33
LAD lasso	0.9716	4.48	0.05	0.05	0.55	0.40
ACQR_9	0.1464	4.72	0.00	0.00	0.78	0.22
ACQR_{19}	0.1759	4.41	0.00	0.00	0.59	0.41
AMPL	0.0718	5.00	0.00	0.00	1.00	0.00
Standard Cauchy						
Adaptive lasso	15.330	3.88	1.04	0.53	0.12	0.35
LAD lasso	0.1105	4.98	0.00	0.00	0.98	0.02
ACQR_9	0.2263	4.59	0.00	0.00	0.68	0.32
ACQR_{19}	0.2772	4.17	0.00	0.00	0.43	0.57
AMPL	0.1125	5.00	0.00	0.00	1.00	0.00

LAD, least absolute deviation; ACQR, adaptive-lasso-penalized composite quantile regression; AMPL, adaptively penalized maximum profile likelihood.

error distribution, the AMPL approach outperforms the adaptive lasso, the LAD lasso, and the ACQR in terms of number of zeros and proportion of fits, which indicates that our AMPL has higher probability of correctly selecting the exact model than the other three methods. That is, the adaptive lasso, the LAD lasso, and the ACQR tend to overselect and estimate a larger model compared with our proposed AMPL approach. Moreover, our proposed AMPL does not overlook significant variables. On the contrary, the 'U-fit' of the adaptive lasso for the mixed normal or Cauchy error is not 0, and the LAD lasso excludes non-zero coefficients for the chi-squared error. Another important contribution of our proposed method is that it demonstrates robustly good performance in selecting the exact model, for example, the 'C-fit' remains no less than 99% regardless of the error distribution. It is interesting to note that our proposed AMPL approach performs no worse than the adaptive lasso method (or even better) in the normal error case.

4.3. Study 3: real data analysis

In this subsection, we illustrate our proposed AMPL method through application to the plasma retinol level dataset collected by a cross-sectional study (Nierenberg *et al.*, 1989). This dataset contains 315 observations. We are interested in the relationships between the plasma retinol level and the following covariates: age, sex (0 = female; 1 = male), smoking status (0 = never; 1 = former; 2 = current smoker), Quetelet index (body mass index), vitamin use (0 = no; 1 = yes, not often; 2 = yes, fairly often), number of calories, grammes of fat, grammes of fibre, number of alcoholic drinks, cholesterol, dietary retinol, dietary beta-carotene, and plasma beta-carotene. We use the linear regression model here. For comparison purposes, we also include the adaptive lasso, the LAD lasso, and the ACQR with the number of quantiles K being 9 and 19, respectively.

A total of 200 observations are randomly chosen as training data to fit the model and to select significant variables, and the remaining 115 observations are used as testing data. Various methods are used to select the best model on the basis of the training dataset. The prediction accuracies of these methods are measured by the root-mean-square prediction error (RMSPE) based on 115 observations of the testing data, which is $\sqrt{\sum_{i=1}^{115}(y_i - \hat{y}_i)^2/115}$. This procedure is repeated 100 times.

The RMSPEs of different methods based on 100 replications are summarized in Fig. 2. From the box plots in Fig. 2, we could see that these methods have similar medians of RMSPEs, and the lower quartiles (25% percentiles) of RMSPEs for various methods are not significantly different. However, the upper quartile (75% percentile) of RMSPEs for our AMPL is lower than that of other approaches, and the range of variation (from the upper extreme to the lower extreme of the box plot; McGill *et al.*, 1978) and the interquartile range of RMSPEs based on our AMPL are both smallest. That is, our AMPL approach tends to produce fewer extreme RMSPEs and smaller variance of RMSPEs than other methods, by which the robustness of the AMPL is verified again. These indicate that the AMPL has better prediction performance than the adaptive lasso, the LAD lasso, and the ACQR under the RMSPE criterion. The empirical selected probabilities for 13 covariates of the model by various methods and the average number of selected variables over 100 replications are given in Table 3. From Table 3, we observe that the model selected by AMPL is sparser than the models selected by the adaptive lasso, the LAD lasso, and the ACQR. It is noteworthy that the adaptive lasso selects 12.02 covariates averagely and the empirical selected probabilities for 13 covariates are all non-zero over 100 replications. This observation may coincide with the simulation results that the adaptive lasso tends to select a larger model than the exact model.

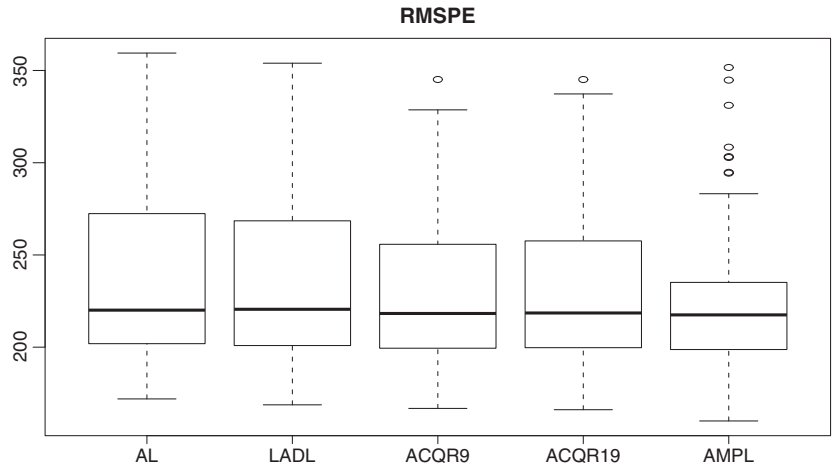

Fig. 2. Box plots of 100 root-mean-square prediction errors (RMSPEs) over 100 replications for the adaptive lasso (AL), the least absolute deviation lasso (LADL), the adaptive-lasso-penalized composite quantile regression (ACQR) with the number of quantiles K being 9 (ACQR9) and 19 (ACQ19), and the adaptively penalized maximum profile likelihood (AMPL) in the real data analysis.

Table 3. *The empirical selected probabilities for* 13 *covariates and the ANSV over* 100 *replications for the plasma retinol level data*

	Adaptive lasso	LAD lasso	ACQR_9	ACQR_{19}	AMPL
Age	1.00	0.99	1.00	1.00	0.48
Sex	1.00	0.96	0.99	0.99	0.56
Smoking	0.99	0.89	0.90	0.95	0.17
Quetelet	0.97	0.81	0.79	0.80	0.00
Vitamin	0.99	0.91	0.91	0.94	0.34
Calories	0.99	0.79	0.79	0.79	0.00
Fat	0.98	0.85	0.92	0.94	0.17
Fibre	0.99	0.84	0.96	0.97	0.15
Alcoholic	1.00	0.94	1.00	1.00	0.51
Cholesterol	0.89	0.41	0.38	0.44	0.00
Dietary retinol	0.68	0.03	0.05	0.05	0.00
Dietary beta	0.59	0.00	0.00	0.00	0.00
Plasma beta	0.95	0.56	0.89	0.89	0.00
ANSV	12.02	8.98	9.58	9.76	2.38

ANSV, average number of selected variables; LAD, least absolute deviation; ACQR, adaptive-lasso-penalized composite quantile regression; AMPL, adaptively penalized maximum profile likelihood.

5. Concluding remarks

For the linear regression model in (1), motivated by the entropy inequality (4), we propose the MPL estimation approach in Section 2. We observe that our proposed MPL is robust against outliers or heavy-tailed errors and behaves nicely even when the error variance is infinite. Moreover, we show by several real examples that the proposed MPL enjoys greater advantages theoretically and practically in terms of ARE when compared with the OLS estimator. Theoretically, our MPL estimator generally possesses smaller asymptotic variance than the LAD and CQR estimators.

We propose the AMPL in Section 2 for robust regression shrinkage and selection in multiple linear regression models. Our proposed AMPL method performs better than the adaptive lasso proposed by Zou (2006) in terms of model error and number of zeros and proportion of fits for all the error distributions considered in Table 2. Also, the AMPL works better than the LAD lasso by Wang *et al.* (2007) and the ACQR by Zou & Yuan (2008) in terms of number of zeros and proportion of fits. For a given likelihood, penalized ML has received much attention. However, one seldom knows the error distribution of the linear regression model in practice. Even worse, if the distribution of the error is incorrectly specified, the existing penalized ML method may perform badly. Our proposed AMPL method works robustly well without any distributional assumption on the error. Hence, our proposed methodology is more flexible and could be widely applied.

It is of great interest to investigate our variable selection theory for the case in which the number of parameters is large and grows with the sample size in model (1). Because of space limitations, we will present the results in another follow-up paper. Our proof using the U-statistics projection theory will play an important part in developing the oracle properties of the AMPL with a diverging number of parameters, and the conditions will be slightly stronger than those in Fan & Peng (2004). We focus on the situation in which the heavy-tailed errors or outliers exist in the responses in this article. We are interested in extending our proposed penalized profile likelihood to the case in which heavy-tailed errors or outliers appear in both responses and predictors (Chi & Scott, to appear).

Acknowledgements

We are grateful to the reviewer, the associate editor, and the editor for their constructive comments that greatly improved the article. This work has been partly supported by the Program for New Century Excellent Talents in University, National Nature Science Foundation of China (no. 11071035), National Nature Science Foundation of China (no. 10931002), National Nature Science Foundation of China (no. 11301245), Specialized Research Fund for the Doctoral Program of Higher Education of China (no. 20130162120086), and China Postdoctoral Science Foundation (no. 2013M531796).

References

Chi, E. C. & Scott, D. W. (to appear). Robust parametric classification and variable selection by a minimum distance criterion. *J. Comput. Graph. Statist.* DOI: 10.1080/10618600.2012.737296.

Fan, J., Huang, T. & Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.* **102**, 632–641.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Fan, J. & Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proc. Madrid Int. Congress of Mathematicians* **3**, 595–622. EMS, Zürich.

Fan, J. & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–961.

Friedman, J., Hastie T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Software* **33**, 1–22.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*, (2nd edn)., Springer, New York.

He, X., Fung, W. K. & Zhu, Z. (2005). Robust estimation in generalized partial linear models for clustered data. *J. Amer. Statist. Assoc.* **100**, 1176–1184.

Kai, B., Li, R. & Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Ann. Statist.* **39**, 305–332.

Lambert-Lacroix, S. & Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron. J. Stat.* **5**, 1015–1053.
Linton, O., Sperlich, S. & Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. *Ann. Statist.* **36**, 686–718.
Lombardia, M.J. & Sperlich, S. (2008). Semiparametric inference in generalized mixed effects models. *J. Roy. Statist. Soc. Ser. B* **70**, 913–930.
Mack, Y. P. & Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **61**, 405–415.
McGill R., Tukey, J. W. & Larsen, W. A. (1978). Variation of box plots. *Amer. Statist.* **32**, 12–16.
Meier, L., van de Geer, S. & Bühlmann, P. (2008). The group lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B* **70**, 53–71.
Newey, W. K. & McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of econometrics*, Vol. 4 (eds R. F. Engle & D. McFadden), Elsevier Science, Amsterdam; 2111–2245.
Nierenberg, D., Stukel, T., Baron, J., Dain, B. & Greenberg, E. (1989). Determinants of plasma levels of beta-carotene and retinol. *Am. J. Epidemiol.* **130**, 511–521.
Pagan, A. & Ullah, A. (1999). *Nonparametric econometrics*, Cambridge University Press, Cambridge.
Rousseeuw, P. J. & Leroy, A. M. (1987). *Robust regression and outlier detection*, Wiley, New York.
Severini, T. A. & Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768–1802.
Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6**, 177–184.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.
Wang, H. (2007). A note on iterative marginal optimization: a simple algorithm for maximum rank correlation estimation. *Comput. Statist. Data Anal.* **51**, 2803–2812.
Wang, H., Li, B. & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. Roy. Statist. Soc. Ser. B* **71**, 671–683.
Wang, H., Li, G. & Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-lasso. *J. Bus. Econom. Statist.* **25**, (3), 347–355.
Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.
Zou, H. & Li, R. (2008). One-step estimation in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509–1533.
Zou, H. & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36**, 1108–1126.

Received January 2013, in final form October 2013

Wei Gao, Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University.
E-mail: gaow@nenu.edu.cn

Appendix

In order to investigate the large-sample properties of the estimators given in this paper, we give the following regular conditions that are mild and can be found in Silverman (1978), Mack & Silverman (1982), Newey & McFadden (1994), Pagan & Ullah (1999), Fan *et al.* (2007), and Linton *et al.* (2008). These conditions may not be the necessary conditions for the theorems presented in this paper but are sufficient conditions to facilitate the proofs.
Conditions:

(a) $(y_i, X_i)_{i=1}^n$ are n independent samples from (1) with $0 < \mathrm{Var}(X_1) < \infty$, X_i is independent of ϵ_i $(i = 1, \ldots, n)$, and $\beta^* \in \mathrm{interior}(\Theta)$ with Θ being a compact subset of R^p.

(b)

(1) $K(\cdot)$ is uniformly continuous with a modulus of continuity W_K, twice continuously differentiable, of bounded variation $V(K)$, and absolutely integrable.
(2) $\int |u\log|u||^{1/2}|\,\mathrm{d}K(u)| < \infty$, $\sup|K(u)| < \infty$, $\sup|K'(u)| < \infty$, and $\int K^2(u)\,\mathrm{d}u < \infty$.
(3) $K(\cdot)$ is symmetric, $\int K(u)\,\mathrm{d}u = 1$, $\int uK(u)\,\mathrm{d}u = 0$, and $\int u^3K(u)\,\mathrm{d}u = 0$.

(c)

(1) $f_{\epsilon(\beta)}(u)$ has a continuous second-order derivative with respect to each u and each $\beta \in \mathcal{N}$, where $\mathcal{N}$ is a neighbourhood of β^*.
(2) $f_{\epsilon(\beta^*)}(\cdot)$ has support R, or $f_{\epsilon(\beta^*)}(\cdot)$ is symmetric about the origin.
(3) $f_{\epsilon(\beta^*)}(\cdot)$ is uniformly continuous.
(4) $E\{\sup_{\beta\in\Theta}|\log f_{\epsilon(\beta)}(\epsilon_1(\beta))|\} < \infty$, $E\{\sup_{\beta\in\mathcal{N}}\|\partial \log f_{\epsilon(\beta)}(\epsilon_1(\beta))/\partial\beta\|\} < \infty$ and $E\{\sup_{\beta\in\mathcal{N}}\|\partial^2 \log f_{\epsilon(\beta)}(\epsilon_1(\beta))/(\partial\beta\partial\beta')\|\} < \infty$.
(5) $\Gamma := [\partial^2 E\{\log f_{\epsilon(\beta)}(\epsilon_1(\beta))\}/(\partial\beta\partial\beta')]\,|_{\beta=\beta^*}$ is finite and negative definite.

(d) The density of X, that is, $f_X(\cdot)$, is bounded away from 0 and ∞ and is Lipschitz continuous on its compact support.
(e) $nh^5/(\log n)^2 \to \infty$, $nh^8 \to 0$, and $n^{1-b}h \to \infty$ for some $b > 0$.

Remarks 4. By Conditions (c, 1) and (c, 4), we have that $E\{\log f_{\epsilon(\beta)}(\epsilon_1(\beta)\}$ is two times continuously differentiable at β^* and $\Gamma = E\{\partial^2 \log f_{\epsilon(\beta^*)}(\epsilon_1(\beta^*))/(\partial\beta\partial\beta')\}$, and then by Lemma 3, we have $\Gamma = [2E\{f_0''(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\} - E\{f_0'(\epsilon_1(\beta^*))/f_0(\epsilon_1(\beta^*))\}^2]\mathrm{Var}X_1$.

Remarks 5. Because of Equation (4), Condition (c, 5) is not strong.

Lemma 1. *Let V and W be independent random variables with $Var(W) \neq 0$. Denote $G := V + W$. Then,*

$$\int f_G(u)\log f_G(u)\,du < \int f_V(u)\log f_V(u)\,du,$$

where $f_G(u)$ and $f_V(u)$ are density functions of the random variables G and V, respectively.

Proof. Because $f_G(u) = \int f_W(x)f_V(u-x)\,\mathrm{d}x$, by Jensen's inequality, we have

$$\begin{aligned}
\int f_G(u)\log f_G(u)\,\mathrm{d}u &= \int\int f_W(x)f_V(u-x)\,\mathrm{d}x\log f_G(u)\,\mathrm{d}u \\
&= \int f_W(x)\int f_V(u-x)\log f_G(u)\,\mathrm{d}u\,\mathrm{d}x \\
&= \int f_W(x)\int f_V(u)\log f_G(u+x)\,\mathrm{d}u\,\mathrm{d}x \\
&< \int f_W(x)\int f_V(u)\log f_V(u)\,\mathrm{d}u\,\mathrm{d}x \\
&= \int f_V(u)\log f_V(u)\,\mathrm{d}u.
\end{aligned}$$

□

Lemma 2. *(Silverman, 1978). Suppose $K(\cdot)$ satisfies condition (b) and $f_{\epsilon(\beta^*)}(\cdot)$ is uniformly continuous. If $h \to 0$ and $n^{1-b}h \to \infty$ for some $b > 0$, then*

$$\sup_u \left| \hat{f}_{\epsilon(\beta^*)}(u) - E\,\hat{f}_{\epsilon(\beta^*)}(u) \right| = O_p\left\{ \left(\frac{\log(1/h)}{nh} \right)^{1/2} \right\}.$$

Detailed proofs of Lemmas 3, 4, and 5 can be found in the Supporting Information.

Lemma 3. *Under conditions (a) and (c, 1), then we have*

$$\dot{f}_{\epsilon(\beta^*)}(u) = f'_{\epsilon(\beta^*)}(u) EX.$$

Lemma 4. *If the conditions in Theorem 2 hold, then we have*

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -\frac{\dot{f}_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))}{f^2_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))} \hat{f}_{\epsilon(\beta^*)(-i)}(\epsilon_i(\beta^*)) + \frac{f'_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))X_i}{f^2_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))} \hat{f}_{\epsilon(\beta^*)(-i)}(\epsilon_i(\beta^*)) \right\} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -\frac{\dot{f}_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))}{f_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))} + \frac{f'_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))X_i}{f_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))} \right\} + o_p(1). \end{aligned}$$

Lemma 5. *If the conditions in Theorem 2 hold, then we have*

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\dot{\hat{f}}_{\epsilon(\beta^*)(-i)}(\epsilon_i(\beta^*))}{f_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\hat{f}'_{\epsilon(\beta^*)(-i)}(\epsilon_i(\beta^*))X_i}{f_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\dot{f}_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))}{f_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{f'_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))X_i}{f_{\epsilon(\beta^*)}(\epsilon_i(\beta^*))} + o_p(1). \end{aligned}$$

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.