Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso

Hansheng WANG
Guanghua School of Management, Peking University, Beijing, P. R. China 100871 (*hansheng@gsm.pku.edu.cn*)

Guodong LI
Department of Statistics and Actuarial Science, University of Hong Kong, P. R. China (*ligd@hkusua.hku.hk*)

Guohua JIANG
Guanghua School of Management, Peking University, Beijing, P. R. China 100871 (*gjiang@gsm.pku.edu.cn*)

The least absolute deviation (LAD) regression is a useful method for robust regression, and the least absolute shrinkage and selection operator (lasso) is a popular choice for shrinkage estimation and variable selection. In this article we combine these two classical ideas together to produce LAD-lasso. Compared with the LAD regression, LAD-lasso can do parameter estimation and variable selection simultaneously. Compared with the traditional lasso, LAD-lasso is resistant to heavy-tailed errors or outliers in the response. Furthermore, with easily estimated tuning parameters, the LAD-lasso estimator enjoys the same asymptotic efficiency as the unpenalized LAD estimator obtained under the true model (i.e., the oracle property). Extensive simulation studies demonstrate satisfactory finite-sample performance of LAD-lasso, and a real example is analyzed for illustration purposes.

KEY WORDS: LAD; LAD-lasso; Lasso; Oracle property.

1. INTRODUCTION

Datasets subject to heavy-tailed errors or outliers are commonly encountered in applications. They may appear in the responses and/or the predictors. In this article we focus on the situation in which the heavy-tailed errors or outliers are found in the responses. In such a situation, it is well known that the traditional ordinary least squares (OLS) may fail to produce a reliable estimator, and the *least absolute deviation* (LAD) estimator can be very useful. Specifically, the $\sqrt{n}$-consistency and asymptotic normality for the LAD estimator can be established without assuming any moment condition of the residual. Due to the fact that the objective function (i.e., the LAD criterion) is a nonsmooth function, the usual Taylor expansion argument cannot be used directly to study the LAD estimator's asymptotic properties. Therefore, over the past decade, much effort has been devoted to establish the $\sqrt{n}$-consistency and asymptotic normality of various LAD estimators (Bassett and Koenker 1978; Pollard 1991; Bloomfield and Steiger 1983; Knight 1998; Peng and Yao 2003; Ling 2005), which left the important problem of robust model selection open for study.

In a regression setting, it is well known that omitting an important explanatory variable may produce severely biased parameter estimates and prediction results. On the other hand, including unnecessary predictors can degrade the efficiency of the resulting estimation and yield less accurate predictions. Hence selecting the best model based on a finite sample is always a problem of interest for both theory and application. Under appropriate moment conditions for the residual, the problem of model selection has been extensively studied in the literature (Shao 1997; Hurvich and Tsai 1989; Shi and Tsai 2002, 2004), and two important selection criteria, the Akaike information criterion (AIC) (Akaike 1973) and the Bayes information criterion (BIC) (Schwarz 1978), have been widely used in practice.

By Shao's (1997) definition, many selection criteria can be classified into two major classes. One class contains all of the *efficient* model selection criteria, among which the most well-known example is the AIC. The efficient criteria have the ability to select the best model by an appropriately defined asymptotic optimality criterion. Therefore, they are particularly useful if the underlying model is too complicated to be well approximated by any finite-dimensional model. However, if an underlying model indeed has a finite dimension, then it is well known that many efficient criteria (e.g., the AIC) suffer from a non-ignorable overfitting effect regardless of sample size. In such a situation, a *consistent* model selection criterion can be useful. Hence the second major class contains all of the consistent model selection criteria, among which the most representative example is the BIC. Compared with the efficient criteria, the consistent criteria have the ability to identify the true model consistently, if such a finite-dimensional true model does in fact exist.

Unfortunately, theoretically there is no general agreement regarding which selection criteria type is preferable (Shi and Tsai 2002). As we demonstrate, the proposed LAD-lasso method belongs to the consistent category. In other words, we make the assumption that the true model is of finite dimension and is contained in a set of candidate models under consideration. Then the proposed LAD-lasso method has the ability to identify the true model consistently. (For a better explanation of model selection efficiency and consistency, see Shao 1997; McQuarrie and Tsai 1998; and Shi and Tsai 2002.)

Because most selection criteria (e.g., AIC, BIC) are developed based on OLS estimates, their finite-sample performance under heavy-tailed errors is poor. Consequently, Hurvich and Tsai (1990) derived a set of useful model selection criteria (e.g., AIC, AICc, BIC) based on the LAD estimates. But despite

Journal of Business & Economic Statistics
???? 0, Vol. 0, No. 00
DOI 10.1198/00

their usefulness, these LAD-based variable selection criteria, have some limitations, the major one being is the computational burden. Note that the number of all possible candidate models increases exponentially as the number of regression variables increases. Hence performing the best subset selection by considering all possible candidate models is difficult if the number of the predictors is relatively large.

To address the deficiencies of traditional model selection methods (i.e., AIC and BIC), Tibshirani (1996) proposed the *least absolute shrinkage and selection operator* (lasso), which can effectively select important explanatory variables and estimate regression parameters simultaneously. Under normal errors, the satisfactory finite-sample performance of lasso has been demonstrated numerically by Tibshirani (1996), and its statistical properties have been studied by Knight and Fu (2000), Fan and Li (2001), and Tibshirani, Saunders, Rosset, Zhu, and Knight (2005). But if the regression error has a very heavy tail or suffers from severe outliers, then the finite-sample performance of the lasso can be poor due to its sensitivity to heavy-tailed errors and outliers.

In this article we attempt to develop a robust regression shrinkage and selection method that can do regression shrinkage and selection (like lasso) and is also resistant to outliers or heavy-tailed errors (like LAD). The basic idea is to combine the usual LAD criterion and the lasso-type penalty together to produce the *LAD-lasso* method. Compared with LAD, LAD-lasso can do parameter estimation and model selection simultaneously. Compared with lasso, LAD-lasso is resistant to heavy-tailed errors and/or outliers. Furthermore, with easily estimated tuning parameters, the LAD-lasso estimator enjoys the same asymptotic efficiency as the oracle estimator.

The rest of the article is organized as follows. Section 2 proposes LAD-lasso and discusses its main theoretical and numerical properties. Section 3 presents extensive simulation results, and Section 4 analyzes a real example. Finally, Section 5 concludes the article with a short discussion. The Appendix provides all of the technical details.

2. ABSOLUTE SHRINKAGE AND SELECTION

2.1 Lasso, Lasso*, and LAD-Lasso

Consider the linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ is the p-dimensional regression covariate, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are the associated regression coefficients, and ϵ_i are iid random errors with median 0. Moreover, assume that $\beta_j \neq 0$ for $j \leq p_0$ and $\beta_j = 0$ for $j > p_0$ for some $p_0 \geq 0$. Thus the correct model has p_0 significant and $(p - p_0)$ insignificant regression variables.

Usually, the unknown parameters of model (1) can be estimated by minimizing the OLS criterion, $\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$. Furthermore, to shrink unnecessary coefficients to 0, Tibshirani (1996) proposed the following lasso criterion:

$$\text{lasso} = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + n\lambda\sum_{j=1}^{p}|\beta_j|,$$

where $\lambda > 0$ is the tuning parameter. Because lasso uses the same tuning parameters for all regression coefficients, the resulting estimators may suffer an appreciable bias (Fan and Li 2001). Hence we further consider the following modified lasso criterion:

$$\text{lasso*} = \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + n\sum_{j=1}^{p}\lambda_j|\beta_j|,$$

which allows for different tuning parameters for different coefficients. As a result, lasso* is able to produce sparse solutions more effectively than lasso.

Nonetheless, it is well known that the OLS criterion used in lasso* is very sensitive to outliers. To obtain a robust lasso-type estimator, we further modify the lasso* objective function into the following LAD-lasso criterion:

$$\text{LAD-lasso} = Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}|y_i - \mathbf{x}_i'\boldsymbol{\beta}| + n\sum_{j=1}^{p}\lambda_j|\beta_j|.$$

As can be seen, the LAD-lasso criterion combines the LAD criterion and the lasso penalty, and hence the resulting estimator is expected to be robust against outliers and also to enjoy a sparse representation.

Computationally, it is very easy to find the LAD-lasso estimator. Specifically, we can consider an augmented dataset $\{(y_i^*, \mathbf{x}_i^*)\}$ with $i = 1, \ldots, n + p$, where $(y_i^*, \mathbf{x}_i^*) = (y_i, \mathbf{x}_i)$ for $1 \leq i \leq n$, $(y_{n+j}^*, \mathbf{x}_{n+j}^*) = (0, n\lambda_j\mathbf{e}_j)$ for $1 \leq j \leq p$, and $\mathbf{e}_j$ is a p-dimensional vector with the jth component equal to 1 and all others equal to 0. It can be easily verified that

$$\text{LAD-lasso} = Q(\boldsymbol{\beta}) = \sum_{i=1}^{n+p}|y_i^* - \mathbf{x}_i^*\boldsymbol{\beta}|.$$

This is just a traditional LAD criterion, obtained by treating $(y_i^*, \mathbf{x}_i^*)$ as if they were the true data. Consequently, any standard unpenalized LAD program (e.g., *l1fit* in S–PLUS, *rq* in the QUANTREG package of R) can be used to find the LAD-lasso estimator without much programming effort.

2.2 Theoretical Properties

For convenience, we decompose the regression coefficient as $\boldsymbol{\beta} = (\boldsymbol{\beta}_a', \boldsymbol{\beta}_b')'$, where $\boldsymbol{\beta}_a = (\beta_1, \ldots, \beta_{p_0})'$ and $\boldsymbol{\beta}_b = (\beta_{p_0+1}, \ldots, \beta_p)'$. Its corresponding LAD-lasso estimator is denoted by $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_a', \hat{\boldsymbol{\beta}}_b')'$, and the LAD-lasso objective function is denoted by $Q(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$. In addition, we also decompose the covariate $\mathbf{x}_i = (\mathbf{x}_{ia}', \mathbf{x}_{ib}')'$ with $\mathbf{x}_{ia} = (x_{i1}, \ldots, x_{ip_0})'$ and $\mathbf{x}_{ib} = (x_{i(p_0+1)}, \ldots, x_{ip})'$. To study the theoretical properties of LAD-lasso, the following technical assumptions are necessarily needed:

Assumption A. The error ϵ_i has continuous and positive density at the origin.

Assumption B. The matrix $\text{cov}(\mathbf{x}_1) = \boldsymbol{\Sigma}$ exists and is positive definite.

Note that Assumptions A and B are both very typical technical assumptions used extensively in the literature (Pollard 1991; Bloomfield and Steiger 1983; Knight 1998). They are needed

for establishing the $\sqrt{n}$-consistency and the asymptotic normality of the unpenalized LAD estimator. Furthermore, define

$$a_n = \max\{\lambda_j, 1 \le j \le p_0\}$$

and

$$b_n = \min\{\lambda_j, p_0 < j \le p\},$$

where λ_j is a function of n. Based on the foregoing notation, the consistency of LAD-lasso estimator can first be established.

Lemma 1 (Root-n consistency). Suppose that $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, are iid and that the linear regression model (1) satisfies Assumptions A and B. If $\sqrt{n}a_n \to 0$, then the LAD-lasso estimator is root-n consistent.

The proof is given in Appendix A. Lemma 1 implies that if the tuning parameters associated with the significant variables converge to 0 at a speed faster than $n^{-1/2}$, then the corresponding LAD-lasso estimator can be $\sqrt{n}$-consistent.

We next show that if the tuning parameters associated with the insignificant variables shrink to 0 slower than $n^{-1/2}$, then those regression coefficients can be estimated exactly as 0 with probability tending to 1.

Lemma 2 (Sparsity). Under the same assumptions as Lemma 1 and the further assumption that $\sqrt{n}b_n \to \infty$, with probability tending to 1, the LAD-lasso estimator $\hat{\boldsymbol{\beta}}' = (\hat{\boldsymbol{\beta}}_a', \hat{\boldsymbol{\beta}}_b')'$ must satisfy $\hat{\boldsymbol{\beta}}_b' = 0$.

The proof is given in Appendix B. Lemma 2 shows that LAD-lasso has the ability to consistently produce sparse solutions for insignificant regression coefficients; hence variable selection and parameter estimation can be accomplished simultaneously.

The foregoing two lemmas imply that the root-n–consistent estimator $\hat{\boldsymbol{\beta}}$ must satisfy $P(\hat{\beta}_2 = 0) \to 1$ when the tuning parameters fulfill appropriate conditions. As a result, the LAD-lasso estimator performs as well as the oracle estimator by assuming that the $\boldsymbol{\beta}_b = \mathbf{0}$ is known in advance. Finally, we obtain the asymptotic distribution of the LAD-lasso estimator.

Theorem (Oracle property). Suppose that $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, are iid and that the linear regression model (1) satisfies Assumptions A and B. Furthermore, if $\sqrt{n}a_n \to 0$ and $\sqrt{n} \times b_n \to \infty$, then the LAD-lasso estimator $\hat{\boldsymbol{\beta}}' = (\hat{\boldsymbol{\beta}}_a', \hat{\boldsymbol{\beta}}_b')'$ must satisfy that $P(\hat{\boldsymbol{\beta}}_b = \mathbf{0}) \to 1$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_a' - \boldsymbol{\beta}_a) \to \mathrm{N}(\mathbf{0}, .25\boldsymbol{\Sigma}_0^{-1}/f^2(0))$, where $\boldsymbol{\Sigma}_0 = \mathrm{cov}(\mathbf{x}_{ia})$ and $f(t)$ is the density of ϵ_i.

The proof is given in Appendix C. Under the conditions $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$, this theorem implies that the LAD-lasso estimator is robust against heavy-tailed errors, because the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\beta}}_a$ is established without making any moment assumptions on the regression error ϵ_i. The theorem also implies that the resulting LAD-lasso estimator has the same asymptotic distribution as the LAD-estimator obtained under the true model. Hence the oracle property of the LAD-lasso estimator is established. Furthermore, due to the convexity and piecewise linearity of the criterion function $Q(\boldsymbol{\beta})$, the LAD-lasso estimator properties discussed in this article are global instead of local, as used by Fan and Li (2001).

2.3 Tuning Parameter Estimation

Intuitively, the commonly used cross-validation, generalized cross-validation (Craven and Wahba 1979; Tibshirani 1996; Fan and Li 2001), AIC, AICc, and BIC (Hurvich and Tsai 1989; Zou, Hastie, and Tibshirani 2004) methods can be used to select the optimal regularization parameter λ_j after appropriate modification, for example, replacing the least squares term by the LAD criterion. However, using them in our situation is difficult for at least two reasons. First, there are a total of p tuning parameters involved in the LAD-lasso criterion, and simultaneously tuning so many regularization parameters is much too expensive computationally. Second, their statistical properties are not clearly understood under heavy-tailed errors. Hence we consider the following option.

Following an idea of Tibshirani (1996), we can view the LAD-lasso estimator as a Bayesian estimator with each regression coefficient following a double-exponential prior with location parameter 0 and scale parameter $n\lambda_j$. Then the optimal λ_j can be selected by minimizing the following negative posterior log-likelihood function:

$$\sum_{i=1}^{n} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + n\sum_{j=1}^{p} \lambda_j|\beta_j| - \log(.5n\lambda_j), \tag{2}$$

which leads to $\lambda_j = 1/(n|\beta_j|)$. For a practical implementation, we do not know the values of β_j; however, they can be easily estimated by the unpenalized LAD estimator $\tilde{\beta}_j$, which produces $\tilde{\lambda}_j = 1/(n|\tilde{\beta}_j|)$. Noting that $\tilde{\beta}_j$ is $\sqrt{n}$-consistent, it follows immediately that $\sqrt{n}\tilde{\lambda}_j \to 0$ for $j \le p_0$. Hence one condition needed in the theorem is satisfied. However, the other condition, $\sqrt{n}\tilde{\lambda}_j \to \infty$ for $j > p_0$, is not necessarily guaranteed. Hence directly using such a tuning parameter estimate tends to produce overfitted results if the underlying true model is indeed of finite dimension. This property is very similar to the AIC. In contrast, note that in (2), the complexity of the final model is actually controlled by $-\log(\lambda)$, which can be viewed as simply the number of significant variables in the AIC. Hence it motivates us to consider the following BIC-type objective function:

$$\sum_{i=1}^{n} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + n\sum_{j=1}^{p} \lambda_j|\beta_j| - \log(.5n\lambda_j)\log(n),$$

which penalizes the model complexity in a similar manner as the BIC with the factor $\log(n)$. This function produces the tuning parameter estimates

$$\hat{\lambda}_j = \frac{\log(n)}{n|\tilde{\beta}_j|}, \tag{3}$$

which satisfies both $\sqrt{n}\hat{\lambda}_j \to 0$ for $j \le p_0$ and $\sqrt{n}\hat{\lambda}_j \to \infty$ for $j > p_0$. Hence the consistent variable selection is guaranteed by the final estimator.

3. SIMULATION RESULTS

In this section we report extensive simulation studies carried out to evaluate the finite-sample performance of LAD-lasso under heavy-tailed errors. For comparison purposes, the finite-sample performance of LAD-based AIC, AICc, and BIC as specified by Hurvich and Tsai (1990), together with the oracle estimator, are evaluated. For the AIC, AICc, and BIC, the best subset selection is used.

Specifically, we set $p = 8$ and $\boldsymbol{\beta} = (.5, 1.0, 1.5, 2.0, 0, 0, 0, 0)$. In other words, the first $p_0 = 4$ regression variables are significant, but the rest are not. For a given i, the covariate $\mathbf{x}_i$ is generated from a standard eight-dimensional multivariate normal distribution. The sample sizes considered are given by $n = 50$, 100, and 200. Furthermore, the response variables are generated according to

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\epsilon_i,$$

where ϵ_i is generated from some heavy-tailed distributions. Specifically, the following three different distributions are considered: the standard double exponential, the standard t-distribution with 5 df (t_5), and the standard t-distribution with 3 df (t_3). Two different values are tested for σ .5 and 1.0, representing strong and weak signal-to-noise ratios.

For each parameter setting, a total of 1,000 simulation iterations are carried out to evaluate the finite-sample performance of LAD-lasso. The simulation results are summarized in Tables 1–3, which include the percentage of correctly (under/over) estimated numbers of regression models, together with the average number of correctly (mistakenly) estimated 0's, in the same manner as done by Tibshirani (1996) and Fan and Li (2001). Also included are the mean and median of the *mean absolute prediction error* (MAPE), evaluated based on another 1,000 independent testing samples for each iteration.

As can be seen from Tables 2–4, all of the methods (AIC, AICc, BIC, and LAD-lasso) demonstrate comparable prediction accuracy in terms of mean and median MAPE. However, the selection results differ significantly. In the case of $\sigma = 1.0$, both the AIC and AICc demonstrate appreciable overfitting effects, whereas the BIC and LAD-lasso have very comparable performance, with both demonstrating a clear pattern of finding the true model consistently. Nevertheless, in the case of $\sigma = .5$, LAD-lasso significantly outperforms the BIC with a margin that can be as large as 30% (e.g., $n = 50$; see Table 3) whereas the overfitting effect of the BIC is quite substantial.

Keep in mind that our simulation results should not be mistakenly interpreted as evidence that the AIC (AICc) is an inferior choice compared with the BIC or LAD-lasso. As pointed out earlier, the AIC (AICc) is a well-known efficient but inconsistent model selection criterion that is asymptotically optimal if the underlying model is of infinite dimension. On the other hand, the BIC and LAD-lasso are useful consistent model selection methods if the underlying model is indeed of finite dimension. Consequently, all of the methods are useful for practical data analysis. Our theory and numerical experience suggest that

Table 1. Simulation results for double-exponential error

						No. of zeros			
σ	n	Method	Underfitted	Correctly fitted	Overfitted	Incorrect	Correct	Average MAPE	Median MAPE
1.0	50	AIC	.079	.347	.574	.079	3.038	1.102	1.011
		AICc	.093	.429	.478	.093	3.252	1.099	.988
		BIC	.160	.591	.249	.161	3.635	1.092	1.110
		LASSO	.184	.633	.183	.185	3.730	1.099	1.045
		ORACLE	0	1.000	0	0	4.000	1.059	1.084
	100	AIC	.003	.385	.612	.003	3.117	1.045	1.013
		AICc	.005	.435	.560	.005	3.232	1.043	1.087
		BIC	.026	.774	.200	.026	3.778	1.035	1.037
		LASSO	.035	.798	.167	.035	3.815	1.040	1.059
		ORACLE	0	1.000	0	0	4.000	1.027	1.073
	200	AIC	0	.475	.525	0	3.295	1.019	.987
		AICc	0	.498	.502	0	3.343	1.018	.946
		BIC	0	.914	.086	0	3.907	1.014	.976
		LASSO	0	.927	.073	0	3.922	1.016	1.020
		ORACLE	0	1.000	0	0	4.000	1.012	1.043
.5	50	AIC	0	.334	.666	0	2.945	.550	.560
		AICc	.001	.449	.550	.001	3.228	.546	.525
		BIC	.002	.684	.314	.002	3.619	.539	.615
		LASSO	.017	.947	.036	.017	3.961	.549	.501
		ORACLE	0	1.000	0	0	4.000	.528	.537
	100	AIC	0	.404	.596	0	3.131	.522	.469
		AICc	0	.461	.539	0	3.249	.521	.542
		BIC	0	.838	.162	0	3.819	.516	.496
		LASSO	0	.987	.013	0	3.987	.520	.496
		ORACLE	0	1.000	0	0	4.000	.513	.513
	200	AIC	0	.429	.571	0	3.224	.510	.526
		AICc	0	.455	.545	0	3.268	.510	.502
		BIC	0	.877	.123	0	3.868	.507	.502
		LASSO	0	.992	.008	0	3.992	.507	.507
		ORACLE	0	1.000	0	0	4.000	.507	.516

Table 2. Simulation results for t_5 error

						No. of zeros			
σ	n	Method	Underfitted	Correctly fitted	Overfitted	Incorrect	Correct	Average MAPE	Median MAPE
1.0	50	AIC	.082	.244	.674	.082	2.825	1.052	1.290
		AICc	.102	.316	.582	.102	3.072	1.048	1.124
		BIC	.177	.492	.331	.179	3.501	1.042	1.164
		LASSO	.225	.566	.209	.228	3.687	1.048	1.073
		ORACLE	0	1.000	0	0	4.000	1.007	.990
	100	AIC	.009	.316	.675	.009	2.995	.994	1.023
		AICc	.011	.364	.625	.011	3.105	.994	.948
		BIC	.027	.721	.252	.027	3.707	.985	.988
		LASSO	.042	.787	.171	.042	3.803	.991	.965
		ORACLE	0	1.000	0	0	4.000	.976	.980
	200	AIC	0	.323	.677	0	3.051	.971	1.161
		AICc	0	.350	.650	0	3.104	.971	.986
		BIC	0	.826	.174	0	3.814	.966	1.020
		LASSO	0	.861	.139	0	3.856	.969	1.059
		ORACLE	0	1.000	0	0	4.000	.963	.907
.5	50	AIC	0	.287	.713	0	2.904	.522	.525
		AICc	.001	.405	.594	.001	3.189	.519	.494
		BIC	.001	.660	.339	.001	3.596	.512	.544
		LASSO	.012	.963	.025	.012	3.975	.524	.489
		ORACLE	0	1.000	0	0	4.000	.501	.522
	100	AIC	0	.301	.699	0	2.957	.499	.546
		AICc	0	.354	.646	0	3.090	.498	.475
		BIC	0	.727	.273	0	3.700	.493	.493
		LASSO	0	.982	.018	0	3.982	.495	.524
		ORACLE	0	1.000	0	0	4.000	.489	.509
	200	AIC	0	.325	.675	0	2.999	.487	.473
		AICc	0	.348	.652	0	3.052	.486	.464
		BIC	0	.800	.200	0	3.781	.483	.496
		LASSO	0	.987	.013	0	3.987	.482	.514
		ORACLE	0	1.000	0	0	4.000	.482	.494

LAD-lasso is a useful consistent model selection method with performance comparable to or even better than that of BIC, but with a much lower computational cost.

4. EARNINGS FORECAST STUDY

4.1 The Chinese Stock Market

Whereas the problem of earnings forecasting has been extensively studied in the literature on the North American and European markets, the same problem regarding one of the world's fastest growing capital markets, the Chinese stock market, remains not well understood. China resumed its stock market in 1991 after a hiatus of 4 decades. Since it reopened, the Chinese stock market has grown rapidly, from a few stocks in 1991 to more than 1,300 stocks today, with a total of more than 450 billion U.S. dollars in market capitalization. Many international institutional investors are moving into China looking for better returns, and earnings forecast research has become extremely important to their success.

In addition, over the past decade, China has enjoyed a high economic growth rate and rapid integration into the world market. Operating in such an environment, Chinese companies tend to experience more turbulence and uncertainties in their operations. As a result, they tend to generate more extreme earnings. For example, in our dataset, the kurtosis of the residuals differentiated from an OLS fit is as large as 90.95, much larger than the value 3 of a normal distribution. Hence the reliability of the usual OLS-based estimation and model selection methods (e.g., AIC, BIC, lasso) is severely challenged, whereas the LAD-based methods (e.g., LAD, LAD-lasso) become more attractive.

4.2 The Dataset

The dataset used here is derived from CCER China Stock Database, which was partially developed by the China Center for Economic Research (CCER) at Peking University. It is considered one of the most authoritative and widely used stock market databases on the Chinese stock market. The dataset contains a total of 2,247 records, with each record corresponding to one yearly observation of one company. Among these, 1,092 come from year 2002 and serve as the training data, whereas the rest come from year 2003 and serve as the testing data.

The response variable is the return on equity (ROE), (i.e., earnings divided by total equity) of the following year

Table 3. Simulation results for t_3 error

σ	n	Method	Underfitted	Correct fitted	Overfitted	No. of zeros: Incorrect	No. of zeros: Correct	Average MAPE	Median MAPE
1.0	50	AIC	.125	.283	.592	.129	2.961	1.217	1.199
		AICc	.149	.373	.478	.153	3.206	1.212	1.258
		BIC	.238	.542	.220	.245	3.640	1.202	1.347
		LASSO	.248	.536	.216	.255	3.671	1.212	1.248
		ORACLE	0	1.000	0	0	4.000	1.164	1.366
	100	AIC	.010	.367	.623	.010	3.108	1.151	1.104
		AICc	.011	.410	.579	.011	3.216	1.151	1.152
		BIC	.043	.742	.215	.043	3.751	1.144	1.339
		LASSO	.046	.748	.206	.046	3.761	1.148	1.218
		ORACLE	0	1.000	0	0	4.000	1.133	1.169
	200	AIC	.001	.385	.614	.001	3.144	1.123	1.148
		AICc	.001	.407	.592	.001	3.191	1.128	1.090
		BIC	.002	.864	.134	.002	3.856	1.120	1.070
		LASSO	.002	.856	.142	.002	3.850	1.123	1.122
		ORACLE	0	1.000	0	0	4.000	1.114	1.110
.5	50	AIC	.002	.309	.689	.002	2.982	.603	.594
		AICc	.003	.412	.585	.003	3.219	.602	.561
		BIC	.004	.652	.344	.004	3.600	.595	.674
		LASSO	.030	.920	.050	.030	3.946	.607	.611
		ORACLE	0	1.000	0	0	4.000	.582	.614
	100	AIC	0	.340	.660	0	3.086	.575	.571
		AICc	0	.391	.609	0	3.197	.574	.581
		BIC	0	.790	.210	0	3.765	.569	.557
		LASSO	0	.978	.022	0	3.977	.572	.573
		ORACLE	0	1.000	0	0	4.000	.566	.594
	200	AIC	0	.355	.645	0	3.085	.564	.560
		AICc	0	.386	.614	0	3.153	.563	.575
		BIC	0	.866	.134	0	3.857	.558	.568
		LASSO	0	.984	.016	0	3.984	.560	.580
		ORACLE	0	1.000	0	0	4.000	.559	.550

(ROE_{t+1}), and the explanatory variables include ROE of the current year (ROE_t), asset turnover ratio (ATO), profit margin (PM), debt-to-asset ratio or leverage (LEV), sales growth rate (GROWTH), price-to-book ratio (PB), account receivables/revenues (ARR), inventory/asset (INV), and the logarithm of total assets (ASSET). These are all measured at year t.

Past studies on earnings forecasting show that these explanatory variables are among the most important accounting variables in predicting future earnings. Asset turnover ratio measures the efficiency of the company using its assets; profit margin measures the profitability of the company's operation; debt-to-asset ratio describes how much of the company is fi-

Table 4. Estimation results of the earning forecast study

Variable	AIC	AICc	BIC	LAD-lasso	LAD	OLS
Int	−.316809	−.316809	−.316809		−.363688	−2.103386
ROE_t	.207533	.207533	.207533	.067567	.190330	−.181815
ATO	.058824	.058824	.058824	.006069	.058553	.165402
PM	.140276	.140276	.140276		.134457	.209346
LEV	−.020943	−.020943	−.020943		−.023592	−.245233
GROWTH	.018806	.018806	.018806		.019068	.033153
PB					.001544	.018385
ARR					−.002520	.009217
INV					.012297	.354148
ASSET	.014523	.014523	.014523	.002186	.016694	.101458
MAPE	.121379	.121379	.121379	.120662	.120382	.233541
STDE	.021839	.021839	.021839	.022410	.021692	.021002

NOTE: "−" represents insignificant variable.

nanced by creditors rather than shareholders; and sales growth rate and price-to-book ratio measure the actual past growth and expected future growth of the company. The account receivable/revenues ratio and the inventory/asset ratio have not been used prominently in studies of North American or European companies, but we use them here because of the observation that Chinese managers have much more discretion over a company's receivable and inventory policies, as allowed by accounting standards. Therefore, it is relatively easy for them to use these items to manage the reported earnings and influence earnings predictability. We include the logarithm of total assets for the same reason, because large companies have more resources under control, which can be used to manage earnings.

4.3 Analysis Results

Various methods are used to select the best model based on the training dataset of 2002. The prediction accuracies of these methods are measured by the MAPE based on the testing data for 2003. For the AIC and BIC, the best subset selection is carried out. For comparison purposes, the results of the full model based on the LAD and OLS estimators are also reported; these are summarized in Table 4.

As can be seen, the MAPE of the OLS is as large as .233541. It is substantially worse than all other LAD-based methods, further justifying the use of the LAD methods. Furthermore, all of the LAD-based selection criteria (i.e., AIC, AICc, and BIC) agree that PB, ARR, and INV are not important, whereas LAD-lasso further identifies the intercept together with PM, LEV, and GROWTH as insignificant variables. Based on a substantially simplified model, the prediction accuracy of the LAD-lasso estimator remains very satisfactory. Specifically, the MAPE of the LAD-lasso is .120662, only slightly larger than the MAPE of the full LAD model (.120382). According to the reported standard error of the MAPE estimate (STDE), we can clearly see that such a difference cannot be statistically significant. Consequently, we conclude that among all of the LAD-based model selection methods, LAD-lasso resulted in the simplest model with a satisfactory prediction accuracy.

According to the model selected by LAD-lasso, a firm with high ATO tends to have higher future profitability (positive sign of ATO). Furthermore, currently more profitable companies continue to earn higher earnings (positive coefficient on ROE_t). Large companies tend to have higher earnings because they tend to have more stable operations, occupy monopolistic industries, and have more resources under control to manage earnings (positive sign of ASSET). Not coincidentally, these three variables are the few variables that capture a company's core earnings ability the most: efficiency (ATO), profitability (ROE_t), and company size (Assets) (Nissim and Penman 2001). It is natural to expect that more accounting variables would also have predictive power for future earnings in developed stock markets such as the North American and European markets. However, in a new emerging stock market with fast economic growth, such as the Chinese stock market, it is not surprising that only the few variables, which capture a company's core earnings ability the most, would survive our model selection and produce accurate predictions.

5. CONCLUDING REMARKS

In this article we have proposed the LAD-lasso method, which combines the ideas of LAD and lasso for robust regression shrinkage and selection. Similar ideas can be further extended to Huber's M-estimation using the following ψ-lasso objective function:

$$Q(\beta) = \sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i'\boldsymbol{\beta}) + n\sum_{j=1}^{p} \lambda_j|\beta_j|,$$

where $\psi(\cdot)$ is the Huber ψ function. Due to the fact that ψ-function contains the LAD function $|\cdot|$ as a special case, ψ-lasso would be expected to be even more efficient than LAD-lasso if an optimal transitional point could be used. However, for real data, how to select such an optimal transitional point automatically within a lasso framework is indeed a challenging and interesting problem. Further study along this line is definitely needed.

ACKNOWLEDGMENTS

The authors are grateful to the editor, the associate editor, the referees, Rong Chen, and W. K. Li for their valuable comments and constructive suggestions, which led to substantial improvement of the manuscript. This research was supported in part by the Natural Science Foundation of China (grant 70532002).

APPENDIX A: PROOF OF LEMMA 1

We want to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P\Big\{\inf_{\|\mathbf{u}\|=C} Q\big(\boldsymbol{\beta} + n^{-1/2}\mathbf{u}\big) > Q(\boldsymbol{\beta})\Big\} \geq 1 - \epsilon, \tag{A.1}$$

where $\mathbf{u} = (u_1, \ldots, u_p)'$ is a p-dimensional vector such that $\|\mathbf{u}\| = C$. From the fact that the function $Q(\boldsymbol{\beta})$ is convex and piecewise linear, the inequality (A.1) implies, with probability at least $1 - \epsilon$, that the LAD-lasso estimator lies in the ball $\{\boldsymbol{\beta} + n^{-1/2}\mathbf{u} : \|\mathbf{u}\| \leq C\}$. For convenience, we define $D_n(\mathbf{u}) \equiv Q(\boldsymbol{\beta} + n^{-1/2}\mathbf{u}) - Q(\boldsymbol{\beta})$. It then follows that

$$\begin{aligned} D_n(\mathbf{u}) &= \sum_{i=1}^{n}\big\{\big|y_i - \mathbf{x}_i'\big(\boldsymbol{\beta} + n^{-1/2}\mathbf{u}\big)\big| - |y_i - \mathbf{x}_i'\boldsymbol{\beta}|\big\} \\ &\quad + n\sum_{j=1}^{p}\lambda_j\big\{\big|\beta_j + n^{-1/2}u_j\big| - |\beta_j|\big\} \\ &\geq \sum_{i=1}^{n}\big\{\big|y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'n^{-1/2}\mathbf{u}\big| - |y_i - \mathbf{x}_i'\boldsymbol{\beta}|\big\} \\ &\quad - \sqrt{n}a_n\sum_{j=1}^{p_0}|u_j|. \end{aligned} \tag{A.2}$$

Denote the first item at the last line of (A.2) by $L_n(\mathbf{u})$. On the other hand, according to Knight (1998), it holds that for $x \neq 0$,

$$\begin{aligned} &|x - y| - |x| \\ &\quad = -y[I(x > 0) - I(x < 0)] + 2\int_0^y [I(x \leq s) - I(x \leq 0)]\,ds. \end{aligned}$$

Applying the foregoing equation, $L_n(\mathbf{u})$ can be expressed as

$$-n^{-1/2}\mathbf{u}'\sum_{i=1}^{n}\mathbf{x}_i[I(\epsilon_i>0)-I(\epsilon_i<0)] + 2\sum_{i=1}^{n}\int_0^{n^{-1/2}\mathbf{u}'\mathbf{x}_i}[I(\epsilon_i\le s)-I(\epsilon_i\le 0)]\,ds. \quad (A.3)$$

By the central limit theorem, the first item on the right side of (A.3) converges in distribution to $\mathbf{u}'\mathbf{W}$, where $\mathbf{W}$ is a p-dimensional normal random vector with mean $\mathbf{0}$ and variance matrix $\mathbf{\Sigma}=\operatorname{cov}(\mathbf{x}_1)$.

Next we show that the second item on the right side of (A.3) converges to a real function of $\mathbf{u}$ in probability. Denote the cumulative distribution function of ϵ_i by F and the item $\int_0^{n^{-1/2}\mathbf{u}'\mathbf{x}_i}[I(\epsilon_i\le s)-I(\epsilon_i\le 0)]\,ds$ by $Z_{ni}(\mathbf{u})$. Hence

$$\begin{aligned} &n\mathrm{E}\big[Z_{ni}^2(\mathbf{u})I\big(n^{-1/2}|\mathbf{u}'\mathbf{x}_i|\ge\eta\big)\big]\\ &\quad\le n\mathrm{E}\left\{\left(\int_0^{n^{-1/2}|\mathbf{u}'\mathbf{x}_i|}2\,ds\right)^2 I\big(n^{-1/2}|\mathbf{u}'\mathbf{x}_i|\ge\eta\big)\right\}\\ &\quad= 4\mathrm{E}\big[|\mathbf{u}'\mathbf{x}_i|^2 I(|\mathbf{u}'\mathbf{x}_i|\ge\sqrt{n}\eta)\big]=o(1).\end{aligned}$$

On the other hand, due to the continuity of f, there exist $\eta>0$ and $0<\kappa<\infty$ such that $\sup_{|x|<\eta}f(x)<f(0)+\kappa$. Then it can be verified that $R=n\mathrm{E}[Z_{ni}^2(\mathbf{u})I(n^{-1/2}|\mathbf{u}'\mathbf{x}_i|<\eta)]$ is dominated by

$$\begin{aligned} R&\le 2n\eta\mathrm{E}\left\{\int_0^{n^{-1/2}|\mathbf{u}'\mathbf{x}_i|}|I(\epsilon_i\le s)-I(\epsilon_i\le 0)|\,ds \times I\big(n^{-1/2}|\mathbf{u}'\mathbf{x}_i|<\eta\big)\right\}\\ &\le 2n\eta\mathrm{E}\left\{\int_0^{n^{-1/2}|\mathbf{u}'\mathbf{x}_i|}[F(s)-F(0)]\,ds\cdot I\big(n^{-1/2}|\mathbf{u}'\mathbf{x}_i|<\eta\big)\right\}\\ &\le 2n\eta\{f(0)+\kappa\}\mathrm{E}\left\{\int_0^{n^{-1/2}|\mathbf{u}'\mathbf{x}_i|}s\,ds\cdot I\big(n^{-1/2}|\mathbf{u}'\mathbf{x}_i|<\eta\big)\right\}\\ &\le \eta\{f(0)+\kappa\}\mathrm{E}|\mathbf{u}'\mathbf{x}_i|^2,\end{aligned}$$

which converges to 0 as $\eta\to 0$. Then it follows that, as $n\to\infty$,

$$\operatorname{var}\left(\sum_{i=1}^{n}Z_{ni}\right)=\sum_{i=1}^{n}\operatorname{var}(Z_{ni})\le n\mathrm{E}Z_{ni}^2(\mathbf{u})\to 0.$$

Hence $\sum_{i=1}^{n}\{Z_{ni}(\mathbf{u})-\mathrm{E}[Z_{ni}(\mathbf{u})]\}=o_p(1)$. Furthermore,

$$\begin{aligned}&\mathrm{E}\left(\sum_{i=1}^{n}Z_{ni}(\mathbf{u})\right)\\ &\quad=n\mathrm{E}[Z_{ni}(\mathbf{u})]=n\mathrm{E}\left\{\int_0^{n^{-1/2}\mathbf{u}'\mathbf{x}_i}[F(s)-F(0)]\,ds\right\}\\ &\quad=\mathrm{E}\left\{\int_0^{n^{-1/2}\mathbf{u}'\mathbf{x}_i}sf(0)\,ds\right\}+o(1)=.5f(0)\mathbf{u}'(\mathbf{x}_i\mathbf{x}_i')\mathbf{u}+o(1),\end{aligned}$$

because

$$\begin{aligned}&P\big\{n^{-1/2}\max(|\mathbf{u}'\mathbf{x}_1|,\ldots,|\mathbf{u}'\mathbf{x}_n|)>\eta^*\big\}\\ &\quad\le nP\big\{|\mathbf{u}'\mathbf{x}_1|>\eta^*n^{1/2}\big\}\\ &\quad\le\frac{1}{(\eta^*)^2}\mathrm{E}\big\{|\mathbf{u}'\mathbf{x}_1|^2I\big(|\mathbf{u}'\mathbf{x}_1|>\eta^*n^{1/2}\big)\big\}\to 0.\end{aligned}$$

It follows from the law of large numbers that

$$\sum_{i=1}^{n}Z_{ni}(\mathbf{u})\to_p\frac{1}{2}f(0)\mathbf{u}'\mathbf{\Sigma}\mathbf{u},$$

where "$\to_p$" represents "convergence in probability." Therefore, the second item on the right side of (A.3) converges to $f(0)\mathbf{u}'\mathbf{\Sigma}\mathbf{u}$ in probability.

By choosing a sufficiently large C, for (A.3), the second item dominates the first item uniformly in $\|\mathbf{u}\|=C$. Furthermore, the second item on the last line of (A.2) converges to 0 in probability and hence is also dominated by the second item on the right side of (A.3). This completes the proof of Lemma 1.

APPENDIX B: PROOF OF LEMMA 2

Using a similar argument as that of Bloomfield and Steiger (1983, p. 4), it can be seen that the LAD-lasso criterion $Q(\boldsymbol{\beta})$ is piecewise linear and reaches the minimum at some breaking point. Taking the first derivative of $Q(\boldsymbol{\beta})$ at any differentiable point $\tilde{\boldsymbol{\beta}}=(\tilde\beta_1,\ldots,\tilde\beta_p)'$ with respect to $\beta_j, j=p_0+1,\ldots,p$, we can obtain that

$$n^{-1/2}\frac{\partial Q(\tilde{\boldsymbol{\beta}})}{\partial\beta_j}=-n^{-1/2}\sum_{i=1}^{n}\operatorname{sgn}(y_i-\mathbf{x}_i'\tilde{\boldsymbol{\beta}})x_{ij}+\sqrt{n}\lambda_j\operatorname{sgn}(\tilde\beta_j),$$

where the function $\operatorname{sgn}(x)$ is equal to 1 for $x>0$, 0 for $x=0$, and -1 for $x<0$. For any $\boldsymbol{\Delta}\in\mathbb{R}^p$, denote

$$\mathbf{V}(\boldsymbol{\Delta})=n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\operatorname{sgn}\big(\epsilon_i-n^{-1/2}\mathbf{x}_i'\boldsymbol{\Delta}\big).$$

By the central limit theorem, it is obvious that

$$\mathbf{V}(\mathbf{0})=n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\operatorname{sgn}(\epsilon_i)\to_d\mathrm{N}(\mathbf{0},\mathbf{\Sigma}),$$

where "$\to_d$" represents "convergence in distribution." Note that $n^{-1/2}\max\{|\mathbf{u}'\mathbf{x}_i|\}_{i=1}^n=o_p(1)$, and lemma A.2 of Koenker and Zhao (1996) can be applied, which leads to

$$\sup_{\|\boldsymbol{\Delta}\|\le M}|\mathbf{V}(\boldsymbol{\Delta})-\mathbf{V}(\mathbf{0})+f(0)\mathbf{\Sigma}\boldsymbol{\Delta}|=o_p(1),$$

where M is any fixed positive number. Then for any $\tilde{\boldsymbol{\beta}}=(\tilde{\boldsymbol{\beta}}_a',\tilde{\boldsymbol{\beta}}_b')'$ satisfying that $\sqrt{n}(\tilde{\boldsymbol{\beta}}_a-\boldsymbol{\beta}_a)=O_p(1)$ and $|\tilde{\boldsymbol{\beta}}_b-\boldsymbol{\beta}_b|\le\epsilon_n=Mn^{-1/2}$,

$$\begin{aligned}&n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\operatorname{sgn}(y_i-\mathbf{x}_i'\tilde{\boldsymbol{\beta}})\\ &\qquad-n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_i\operatorname{sgn}(\epsilon_i)+f(0)\mathbf{\Sigma}\boldsymbol{\Delta}^*=o_p(1),\end{aligned}$$

where $\boldsymbol{\Delta}^* = \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Hence

$$n^{-1/2} \sum_{i=1}^{n} \mathbf{x}_i \operatorname{sgn}(y_i - \mathbf{x}_i' \tilde{\boldsymbol{\beta}}) = O_p(1).$$

On the other hand, as $n \to \infty$, the condition $\sqrt{n} b_n \to \infty$ implies that

$$n^{-1/2} \frac{\partial Q(\tilde{\boldsymbol{\beta}})}{\partial \beta_j} \begin{cases} > 0 & \text{for } 0 < \tilde{\beta}_j < \epsilon_n \\ < 0 & \text{for } -\epsilon_n < \tilde{\beta}_j < 0, \end{cases}$$

where $j = p_0 + 1, \ldots, p$. This completes the proof of Lemma 2.

APPENDIX C: PROOF OF THE THEOREM

For any $\mathbf{v} = (v_1, \ldots, v_{p_0})' \in \mathbb{R}^{p_0}$, define $s_n(\mathbf{v}) = Q(\boldsymbol{\beta}_a + n^{-1/2}\mathbf{v}, 0) - Q(\boldsymbol{\beta}_a, 0)$. Then

$$s_n(\mathbf{v}) = \sum_{i=1}^{n} \left\{ \left| y_i - \mathbf{x}_{ai}' \boldsymbol{\beta}_a - n^{-1/2} \mathbf{v}' \mathbf{x}_{ai} \right| - |y_i - \mathbf{x}_{ai}' \boldsymbol{\beta}_a| \right\} + n \sum_{j=1}^{p_0} \lambda_j \left\{ \left| \beta_j + n^{-1/2} v_j \right| - |\beta_j| \right\}, \tag{A.1}$$

where $\mathbf{x}_{ai} = (x_{i1}, \ldots, x_{ip_0})'$. Similar to the proof of Lemma 1, we know that the first item at the right side of (A.1),

$$\sum_{i=1}^{n} \left\{ \left| y_i - \mathbf{x}_{ai}' \boldsymbol{\beta}_a - n^{-1/2} \mathbf{v}' \mathbf{x}_{ai} \right| - |y_i - \mathbf{x}_{ai}' \boldsymbol{\beta}_a| \right\} \to_d \mathbf{v}' \mathbf{W}_0 + f(0) \mathbf{v}' \boldsymbol{\Sigma} \mathbf{v},$$

where $\mathbf{W}_0$ is a p_0-dimensional normal random vector with mean $\mathbf{0}$ and variance matrix $\boldsymbol{\Sigma}_0$. Furthermore,

$$\left| n \sum_{j=1}^{p_0} \lambda_j \left\{ \left| \beta_j + n^{-1/2} v_j \right| - |\beta_j| \right\} \right| \le \sqrt{n} a_n \sum_{j=1}^{p_0} |v_j| \to 0.$$

Then $s_n(\mathbf{v})$ also converges to $\mathbf{v}' \mathbf{W}_0 + f(0) \mathbf{v}' \boldsymbol{\Sigma} \mathbf{v}$ in distribution. Hence the required central limit theorem follows from lemma 2.2 and remark 1 of Davis et al. (1992). This completes the proof of the theorem.

[Received March 2005. Revised February 2006.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267–281.

Bassett, G., and Koenker, R. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association*, 73, 618–621.

Bloomfield, P., and Steiger, W. L. (1983), *Least Absolute Deviation: Theory, Applications and Algorithms*, Boston: Birkhauser.

Craven, P., and Wahba, G. (1979), "Smoothing Noise Data With Spline Function: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross Validation," *Numerische Mathematik*, 31, 337–403.

Davis, R. A.. Knight, K., and Liu, J. (1992), "M-Estimation for Autoregressions With Infinite Variance," *Stochastic Process and Their Applications*, 40, 145–180.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Hurvich, C. M., and Tsai, C. L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307.

——— (1990), "Model Selection for Least Absolute Deviation Regression in Small Samples," *Statistics and Probability Letters*, 9, 259–265.

Knight, K. (1998), "Limiting Distributions for L_1 Regression Estimators Under General Conditions," *The Annals of Statistics*, 26, 755–770.

Knight, K., and Fu, W. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378.

Koenker, R., and Zhao, Q. (1996), "Conditional Quantile Estimation and Inference for ARCH Models," *Econometric Theory*, 12, 793–813.

Ling, S. (2005), "Self-Weighted Least Absolute Deviation Estimation for Infinite Variance Autoregressive Models," *Journal of the Royal Statistical Society*, Ser. B, 67, 1–13.

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

Nissim, D., and Penman, S. (2001), "Ratio Analysis and Equity Valuation: From Research to Practice," *Review of Accounting Studies*, 6, 109–154.

Peng, L., and Yao, Q. (2003), "Least Absolute Deviation Estimation for ARCH and GARCH Models," *Biometrika*, 90, 967–975.

Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186–199.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, 7, 221–264.

Shi, P., and Tsai, C. L. (2002), "Regression Model Selection: A Residual Likelihood Approach," *Journal of the Royal Statistical Society*, Ser. B, 64, 237–252.

——— (2004), "A Joint Regression Variable and Autoregressive Order Selection Criterion," *Journal of Time Series*, 25, 923–941.

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society*, Ser. B, 67, 91–108.

Zou, H., Hastie, T., and Tibshirani, R. (2004), "On the 'Degrees of Freedom' of Lasso," technical report, Standford University, Statistics Department.