Regularization and variable selection for infinite variance autoregressive models

Ganggang Xu[a,*], Yanbiao Xiang[b], Suojin Wang[a], Zhengyan Lin[b]

[a]*Department of Statistics, Texas A&M University, College Station, Texas 77843, USA*
[b]*Department of Mathematics, Zhejiang University, Hangzhou 310027, P. R. China*

Abstract

Autoregressive models with infinite variance are of great importance in modeling heavy-tailed time series and have been well studied. In this paper, we propose a penalized method to conduct model selection for autoregressive models with innovations having Pareto-like distributions with index $\alpha \in (0, 2)$. By combining the least absolute deviation loss function and the adaptive lasso penalty, the proposed method is able to consistently identify the true model and at the same time produce efficient estimators with a convergence rate of $n^{-1/\alpha}$. In addition, our approach provides a unified way to conduct variable selection for autoregressive models with finite or infinite variance. A simulation study and a real data analysis are conducted to illustrate the effectiveness of our method.

Keywords:
Adaptive lasso, Autoregressive model, Infinite variance, Least absolute deviation

1. Introduction

Heavy-tailed time series data is often encountered in a variety of fields, such as hydrology (Castillo, 1988), economics and finance (Koedijk et al., 1990) and teletraffic engineering (Duffy et al., 1994). The infinite variance autoregressive model has been proven to be useful in modeling a wide range of problems; see, for example, Granger and Orr (1972). Because of its practical usefulness, the infinite variance autoregressive model has attracted much attention in statistical research. Extensive studies on its theoretical properties have been conducted in

*Corresponding author
Email address: gang@stat.tamu.edu (Ganggang Xu)

the literature. See Resnick (1997) for a comprehensive review and for further references.

As in the finite variance case, order determination is an important aspect of an autoregressive model with infinite variance. For a time series model with finite variance, traditional model selection criteria such as AIC (Akaike, 1973) and BIC (Schwarz, 1978) can be employed to choose the order of an autoregressive model and their statistical properties have been well studied (McQuarrie and Tsai, 1998). On the contrary, due to the technical difficulties introduced by the infinite variance, few papers have investigated the model selection for autoregressive models with infinite variance. Bhansali (1988) considered the order determination of infinite variance autoregressive processes with innovations in the domain of attraction of a stable law, and proposed a consistent estimator of the order. Knight (1989) studied the same model and showed that the order selection procedure using AIC is weakly consistent. Although order determination can effectively reduce the complexity of the autoregressive model, a selected large order p can still degrade the efficiency of the parameter estimation and lead to less accurate predictions. As a matter of fact, an autoregressive model with a large p is not uncommon in practice, especially considering the fact that a second order stationary time series with symmetric continuous spectral density can be well approximated by an autoregressive model with a sufficiently large order (Brockwell and Davis, 1991). A similar result was obtained by Miamee and Pourahmadi (1988) for infinite variance time series; see Subsection 2.4 for details.

To deal with an autoregressive model with a large order p, one reasonable assumption is that some coefficients are zeros and thus the corresponding variables can be omitted from the model. By correctly identifying those zero coefficients can further help reduce the model complexity and increase the estimation efficiency and the prediction accuracy using the simplified model. However, variable selection for infinite variance autoregressive model is a hard problem mainly due to the fact that most of existing parameter estimators do not even have a closed form limiting distribution (Davis et al., 1992), which made statistical inference of the model extremely difficult. An honorable exception is the self-weighted least absolute deviation estimator proposed in Ling (2005), which is asymptotically normally distributed and thus can be used for statistical inference. Ling (2005) further proposed a variable selection procedure based on a series of hypothesis tests. However, like the other hypothesis-based variable selection procedures such as the subset selection, this method can be unstable and its implementation can be complicated. Another option is to use the shrinkage method for variable selection. Wang et al. (2007a) applied the adaptive lasso method (Zou, 2006) to the

regression model with finite autoregressive errors. But their work is limited to the case where the autoregressive errors have a finite variance. We shall pursue in this direction for the infinite variance autoregressive model.

Furthermore, another difficulty often encountered in real data analysis is that it is generally impossible to know whether a time series of a finite length has infinite variance or not (Granger and Orr, 1972), even though many methods have been proposed to test for infinite variance of a real time series data; see, for example, Hill (1975). The reason why it is important to distinguish these two cases is because that we need to choose a statistical method that suits the data best. For example, Wang et al. (2007a)'s method does not apply to infinite variance autoregressive models and Ling (2005)'s method can cause loss of important information by weighing down large observations. One direct consequence is that Ling (2005)'s estimator has a slower convergence rate ($n^{1/2}$) than the least absolute estimator ($n^{1/\alpha}$, $\alpha \in (0, 2)$) if the autoregressive model truly has infinite variance (Davis et al., 1992). To overcome this, we propose a unified variable selection approach that can efficiently deal with heavy-tailed autoregressive models with either finite or infinite variance. By combining the least absolute deviation as the loss function and the adaptive lasso as the penalty function, we show that under regularity conditions we can identify the true model consistently and obtain a point estimator of the coefficients corresponding to the true model with a convergence rate of $n^{-1/\alpha}$, where $\alpha \in (0, 2)$ is the index of the stable distribution. This convergence rate is faster than that for a finite variance time series.

2. Adaptive lasso for infinite variance autoregressive models

2.1. Notations and Preliminaries

We consider a stationary autoregressive time series $\{y_t\}$ generated by

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t, \tag{1}$$

where $\phi = (\phi_1, \ldots, \phi_p)^T$ is an unknown parameter vector whose true value is $\phi_0 = (\phi_1^0, \ldots, \phi_p^0)^T$ and we assume that there are a total number of $p_0 \leq p$ nonzero coefficients within ϕ_0. In this paper, we shall restrict our attention to the innovation process $\{\epsilon_t\}$ that consists of independent and identically distributed random errors with Pareto-like distributions of index $\alpha \in (0, 2)$, that is,

$$P(|\epsilon_t| > x) = Kx^{-\alpha}(1 + o(1)), \tag{2}$$

and

$$\lim_{x\to\infty} P(\epsilon_t > x)/P(|\epsilon_t| > x) = q, \quad 0 \leq q \leq 1, \tag{3}$$

where K is some constant. It is straightforward to check that $E|Z_t|^\delta = \infty$ if $\delta \geq \alpha$ and $E|Z_t|^\delta < \infty$ if $\delta < \alpha$ (Brockwell and Davis, 1991).

This class of distributions includes the Cauchy and stable distributions, which are popular in modeling infinite variance autoregressive models; see, for example, Brockwell and Davis (1991) and Davis et al. (1992). Granger and Orr (1972) gave a detailed discussion on the appealing properties of the stable distributions in modeling economic time series. To make $\{y_t\}$ strictly stationary and ergodic, we further assume that $\phi(z) = 1 - \phi_1^0 z - \cdots - \phi_p^0 z^p \neq 0$ for all complex z with $|z| \leq 1$ so that Model (1) can be represented as $y_t = \sum_{j=0}^{\infty} \psi_j^0 \epsilon_{t-j}$ for some ψ_j^0's.

2.2. Adaptive lasso with least absolute deviation

In practice, if the true order of the underlying autoregressive model is large, it would be beneficial to further simplify the model by identifying variables with zero coefficients among all the remaining variables, even after the true order has been correctly selected. Another motivation to do so is that a model with a sparse representation may reveal more information about the underlying structure of the observed process. We propose the following procedure to conduct variable selection for an autoregressive model.

Denote $L_{1n}(\phi) = \sum_{t=p+1}^{n} |y_t - X_t^T \phi|$, where $X_t = (y_{t-1}, \ldots, y_{t-p})^T$ and define the least absolute deviation estimator of Model (1) as $\tilde{\phi}_{1n} = \arg\min_\phi \{L_{1n}(\phi)\}$. Then the least absolute deviation adaptive lasso estimator $\hat{\phi}_{1n}$ can be defined as the minimizer of

$$V_n(\phi) = L_{1n}(\phi) + \lambda_n \Sigma_{j=1}^{p} r_j |\phi_j|, \tag{4}$$

where the weight $r_j = |\tilde{\phi}_{1j}|^{-2}$ with $\tilde{\phi}_{1j}$ being the jth element of $\tilde{\phi}_{1n}$. Note that $\tilde{\phi}_{1n}$ can be obtained by setting $\lambda_n = 0$ when minimizing (4). As stated in Davis et al. (1992), although Model (1) has infinite variance and even infinite mean if $\alpha < 1$, the least absolute deviation estimator performs surprisingly well. In fact, $\tilde{\phi}_{1n}$ usually converges in a rate faster than $n^{-1/2}$. Because of this better choice of weights r_j's, for a given sample size n, minimizing (4) would yield better variable selection results than that in the finite variance case.

The choice of weights r_j's can be made more general by incorporating prior information in practice. For example, if previous experience suggests that some variables must be selected, we can simply set $r_j = 0$ for these variables. The choice of penalty term can also be made more general by replacing each $\lambda_n r_j$ term in (4) with $p'_{\lambda_n}(|\tilde{\phi}_{1j}|)$ for some penalty function $p_\lambda(\cdot)$; see Zou and Li (2008).

The asymptotic theory for $\tilde{\phi}_{1n}$ was first established by Davis et al. (1992). Denote

$$W_n(u) = \sum_{t=p+1}^{n} \left(|\epsilon_t - b_n^{-1} X_t^T u| - |\epsilon_t| \right), \tag{5}$$

where $b_n = \inf\{x : P(|\epsilon_t| > x) \leq n^{-1}\}$. As stated in Davis et al. (1992), for Pareto-like distributions we shall take $b_n = n^{1/\alpha}$. Recall that $y_t = \sum_{j=0}^{\infty} \psi_j^0 \epsilon_{t-j}$, and define quantity

$$W(u) = \sum_{i=1}^{\infty} \sum_{k=1}^{\infty} \{ |\epsilon_{k,i} - (\psi_{i-1}^0 u_1 + \cdots + \psi_{i-p}^0 u_p) \varrho_k \Theta_k^{-1/\alpha}| - |\epsilon_{k,i}| \}, \tag{6}$$

where $\{\epsilon_{k,i}\}$, $\{\varrho_k\}$, and $\{\Theta_k\}$ are three independent sequences defined as follows:

1. $\{\epsilon_{k,i}\}$ is independent and identically distributed as ϵ_t;
2. $\{\varrho_k\}$ is independent and identically distributed with $P(\varrho_k = 1) = q$ and $P(\varrho_k = -1) = 1 - q$, with q given in (3);
3. $\Theta_k = \sum_{j=1}^{k} \Gamma_j$, where $\{\Gamma_j\}$ is a sequence of independent and identically distributed unit exponential random variables.

The following lemma is Theorem 4.1 in Davis et al. (1992), which establishes the asymptotic property of $\tilde{\phi}_{1n}$.

Lemma 2.1. *Suppose that $\{\epsilon_t\}$ satisfies (2) and (3) with $\alpha \in (0, 2)$, and has median 0 if $\alpha \geq 1$. If either (a) $\alpha < 1$, or (b) $\alpha > 1$ and $E(|\epsilon_t|^\beta) < \infty$ for some $\beta < 1 - \alpha$, or (c) $\alpha = 1$ and $E(\log |\epsilon_t|) > -\infty$, then $W_n(\cdot) \to W(\cdot)$ in distribution. Moreover, if $W(\cdot)$ has a unique minimum almost surely, then*

$$b_n(\tilde{\phi}_{1n} - \phi_0) \to \xi \quad \textit{in distribution, as } n \to \infty, \tag{7}$$

where $b_n = n^{1/\alpha}$ and ξ is the minimum of $W(\cdot)$.

Remark 2.1. *One sufficient condition for moment conditions in the case of $\alpha \geq 1$ to be met is that the density of ϵ_1 is bounded in a neighborhood of 0. Conditions in Lemma 2.1 guarantee that $W(\cdot)$ is well-defined. To ensure that $W(\cdot)$ has a unique minimum almost surely, Davis et al. (1992) showed that the following condition is sufficient: for all $\varepsilon > 0$ there exists a constant $d > 0$ such that*

$$P(x < \epsilon_t < y) \geq \begin{cases} d(y-x)^{1/\alpha}, & \alpha < 1, \\ d(y-x), & \alpha \geq 1, \end{cases}$$

whenever $-\varepsilon < x < y < \varepsilon$. For the Cauchy distribution and most stable distributions with index $\alpha \in (0, 2)$, this condition is obviously satisfied.

Denote $\mathcal{S} = \{j : \phi_j^0 \neq 0, j = 1, \ldots, p\}$ and $\mathcal{S}^c = \{j : \phi_j^0 = 0, j = 1, \ldots, p\}$. Let $\hat{\phi}_{1n}$ be the minimizer of (4) and $\mathcal{S}^* = \{j : \hat{\phi}_{1j} \neq 0, j = 1, \ldots, p\}$, where $\hat{\phi}_{1j}$ is the jth element of $\hat{\phi}_{1n}$. As our main theoretical result, the next theorem states the variable selection consistency of the adaptive lasso method as well as the weak convergence of coefficient estimators to their true values.

Theorem 2.2. *Suppose that $\{\epsilon_t\}$ satisfies the conditions stated in Lemma 2.1 and Remark 2.1. If $\lambda_n b_n^{-1} \to 0$ and $\lambda_n \to \infty$ with $b_n = n^{1/\alpha}$, then we have $\lim_{n\to\infty} P(\mathcal{S}^* = \mathcal{S}) = 1$ and $b_n(\hat{\phi}_{1\mathcal{S}} - \phi_{\mathcal{S}}^0) = O_p(1)$, where $\phi_{\mathcal{S}}^0$ and $\hat{\phi}_{1\mathcal{S}}$ are the subvectors of ϕ_0 and $\hat{\phi}_{1n}$ corresponding to the non-zero coefficients, respectively.*

The proof of Theorem 2.2 is given in the Appendix. We would like to point out that it is generally not possible to obtain an explicit representation of the limiting distribution of $b_n(\hat{\phi}_{1\mathcal{S}} - \phi_{\mathcal{S}}^0)$.

Remark 2.3. *When $\{\epsilon_t\}$'s in Model (1) have a finite variance, the result in Theorem 2.2 still holds, see Wang et al. (2007b). In this case, we have $b_n = n^{1/2}$ and $\hat{\phi}_{1\mathcal{S}}$ is asymptotically normal.*

Remark 2.4. *The conditions on λ_n seem to be complicated. However, simply taking $\lambda_n = c \log n$ with some constant c would satisfy all conditions for any $\alpha \in (0, 2)$. This choice is appealing in that we do not need to estimate the specific value of the index α, which is generally difficult to do. In addition, this choice of λ_n also works when $\{\epsilon_t\}$ has a finite variance; see Wang et al. (2007b).*

Theorem 2.2 states that the adaptive lasso method is an estimation and variable selection procedure that takes genuine advantage of the least absolute deviation estimators in the infinite variance autoregressive model. Remarks 2.3 and 2.4 point out that the least absolute deviation method combined with the adaptive lasso method works for both finite and infinite variance situations. There is no need to distinguish between these two cases when the primary concern is to obtain a sparse model. As far as the inference of the model concerns, one can simply apply tools such as the self-weighted least absolute deviation method proposed by Ling (2005) to the selected model.

2.3. Comparison with self-weighted least absolute deviation method

To the best of our knowledge, the only existing method that can handle both order determination and variable selection for the infinite variance autoregressive

models is the self-weighted least absolute deviation method proposed by Ling (2005), where the loss function is defined as

$$L_{2n}(\phi) = \sum_{t=p+1}^{n} h_t |y_t - X_t^T \phi|,$$

where $X_t = (y_{t-1}, \ldots, y_{t-p})^T$ and h_t is a given function of $\{y_{t-1}, \ldots, y_{t-p}\}$. The self-weighted least absolute deviation estimator is defined as

$$\tilde{\phi}_{2n} = \arg\min_{\phi} \{L_{2n}(\phi)\}.$$

Ling (2005) showed that, under certain conditions on $\{y_t\}$, h_t and $\{\epsilon_t\}$, one has

$$n^{\frac{1}{2}}(\tilde{\phi}_{2n} - \phi_0) \to N\left\{0, \frac{1}{4f^2(0)}\Sigma^{-1}\Omega\Sigma^{-1}\right\} \quad \text{in distribution,} \tag{8}$$

where $\Sigma = E(h_t X_t X_t^T)$, $\Omega = E(h_t^2 X_t X_t^T)$ and $f(\cdot)$ is the common density of $\{\epsilon_t\}$, which is assumed to be differentiable everywhere in R with $f(0) > 0$ and $sup_{x \in R}|f'(x)| < \infty$.

To implement his method, Ling (2005) suggested the following weight function:

$$h_t = \begin{cases} 1, & a_t = 0, \\ C^3/a_t^3, & a_t \neq 0, \end{cases} \tag{9}$$

where $a_t = \sum_{i=1}^{p} |y_{t-i}| \{I(|y_{t-i}| \geq C)\}$ and $C > 0$ is a constant. As an alternative to the proposed procedure, we also tried combining $L_{2n}(\phi)$ and the adaptive lasso to conduct the model selection. But the finite sample performance of this approach turns out to be less satisfactory in terms of correctly identifying the true model compared to the LAD Adaptive Lasso, which we suspect is largely due to the slower convergence rate of $\tilde{\phi}_{2n}$.

Since $\tilde{\phi}_{2n}$ is asymptotically normal, Ling (2005) proposed a variable selection procedure based on a series of hypothesis tests using Chi-square test statistics. There are a few drawbacks of this approach: (a) when using this method, weighing down large observations would lead to unnecessary loss information; (b) the choice of C is subjective. There is no theoretical justification for a best choice of C. Our simulation results indicate that different choices of C would lead to different model selection results; (c) it is difficult to manage the overall type I error when conducting a series of hypothesis tests; (d) the unknown term $f(0)$ in (8) needs to be estimated, which introduces more uncertainty in the model selection

process since the Chi-square test statistic is proportional to $f^2(0)$. An estimated value might affect the p-values of the tests and thus lead to erroneous conclusions.

The proposed least absolute deviation adaptive lasso method suffers from none of the above problems, although it has a limitation in that there is no closed form limiting distribution for the estimator. Recall that the motivation for our method is to develop a better variable selection strategy and to produce a faster convergent point estimator in the infinite variance case. To make inference, we can apply existing methods such as Ling (2005)'s to the model selected by our method.

2.4. p-Stationary process

The popularity of the autoregressive model in time series studies is largely due to the fact that any second order stationary process with symmetric continuous spectral density can be approximated by an autoregressive process (Brockwell and Davis, 1991). However, when it comes to stationary process with infinite variance, this type of relationship is difficult to establish.

Miamee and Pourahmadi (1988) established such a relationship for the p-stationary process. A discrete time stochastic process $\{y_t\}$ is said to be a p-stationary process if $E|y_t|^p < \infty$, and $E\left|\sum_{k=1}^{n} c_k y_{t_k+h}\right|^p = E\left|\sum_{k=1}^{n} c_k y_{t_k}\right|^p$, $(1 < p \leq 2)$ for all integers $n \geq 1$, $t_1, \ldots, t_n$, h, and scalars $c_1, \ldots, c_n$. Note that, $p = 2$ corresponds to the second order weakly stationary process, which is the only case in this class that has a finite variance. This class of processes includes the harmonizable stable processes of order α with $\alpha \in (1, 2]$ and strictly stationary processes with finite p-th moment. Miamee and Pourahmadi (1988) showed that for a purely nondeterministic p-stationary process $\{y_t\}$ with innovation $\{\epsilon_t\}$, there exists a unique series $\{a_k\}$ such that for all t, one has

$$y_t = \sum_{k=1}^{\infty} a_k y_{t-k} + \epsilon_t,$$

provided that $\sum_{k=1}^{\infty} a_k y_{t-k}$ is convergent in the mean of order p. A sufficient condition for the convergence of $\sum_{k=1}^{\infty} a_k y_{t-k}$ is that $\sum_{k=1}^{\infty} |a_k| < \infty$. For regularity conditions and more recent advances in this area, see Cheng et al. (2000). The autoregressive representation of $\{y_t\}$ above provides justifications for using an autoregressive model for some stationary infinite variance time series in the sense that even though the underlying structure of a time series is not autoregressive, it can be approximated by an autoregressive model under certain conditions.

2.5. Computation and tuning parameter selection

In this subsection we run a simulation study to illustrate the effectiveness of the proposed approach. Minimizing (4) can be written as a standard median regression problem and thus any computer program that can solve median regression problems can be used to find the minimum of (4). In our simulation study, we used the function rq in the R (R Development Core Team, 2009) package QUANTREG.

As pointed out by one referee, the minimization of (4) can have multiple solutions. To see this, let us consider a simple example where n data points $\{x_1, \cdots, x_n\}$ were generated from a continuous random variable X. One way to estimate the median μ_0 of X is to find a μ that minimizes $V(\mu) = \sum_{i=1}^{n} |x_i - \mu|$. If n is an even number, it can be easily shown that any $\mu \in (X_{(n/2)}, X_{(n/2)+1})$ gives the same value of $V(\mu)$, where $x_{(1)} \leq \cdots \leq x_{(n)}$. Although as $n \to \infty$ any solution sequence would converge to μ_0, for a finite sample, one needs to be cautious about this multiplicity issue. When this occurs, proposals such as in Castillo et al. (2008) could be applied to resolve this issue. In our limited empirical studies, the R function rq did not show any alarming problems caused by this issue, which might suggest that its potential impact on the model selection results is manageable when proper functions are employed.

Tuning parameter selection is another key issue in implementing our method. To select the optimal value of λ_n that meets the conditions in Theorem 2.2, we perform a grid search. As stated in Remark 2.4, for any $\alpha \in (0, 2)$, taking $\lambda_n = c \log n$ for a fixed c would always satisfy conditions of Theorem 2.2. Based on this observation, we took $\lambda_n = \lambda^* c \log n$, where λ^* is selected from 20 equally spaced grid points between $(0.1, 5)$ and took c as the median absolute deviation (MAD) of $\{y_1, \cdots, y_n\}$, which is defined as $median_i(|y_i - median_j(y_j)|)$. The reason we chose this value for c is to make the loss function defined in (4) invariant to the scale of $\{y_t\}$. Finally, to select the optimal λ^*, we use the following Schwartz-type information criterion.

$$\text{SIC}_{\lambda^*} = \log\left(\frac{1}{n}\sum_{t=p+1}^{n} |y_t - X_t^T \hat{\phi}_{\lambda^*}|\right) + \hat{df}_{\lambda^*} \times \frac{\log n}{2n}, \tag{10}$$

where $\hat{df}_{\lambda^*}$ is the number of non-zero coefficients in $\hat{\phi}_{\lambda^*}$. This criterion was first suggested by Koenker et al. (1994) and He and Ng (1999) for choosing the regularization parameter in quantile smoothing splines and has been widely used in quantile regression literature. Since least absolute deviation regression is a special case of quantile regression, we can expect SIC to yield reasonably good results. A similar BIC type criterion has also been used in Wang et al. (2007a), where they

showed that such a BIC type criterion performs much better than the cross validation in model selection. We also tried the 5-fold cross validation method to select the best λ^*, but this approach was outperformed by using (10) in almost every case. Therefore, we did not present the cross validation results in this paper.

3. A simulation study

We generated the data from the autoregressive model $y_t = 0.5y_{t-1} - 0.7y_{t-3} + \epsilon_t$, which was also used by Wang et al. (2007b). Three types of innovations, Cauchy, symmetric α-stable distribution with $\alpha = 1.5$ and student t-distribution with degrees of freedom 2, were considered. For the LAD adaptive lasso, we start with a full model of order $p = 5$ and $p = 10$, respectively. All estimates with magnitude smaller than 10^{-8} were numerically set to 0. Using a full order $p = 5$ to identity the true model is easier than using $p = 10$, but our limited experience indicates that as long as p is not too large, the model selection results will not change much.

We compared the variable selection performance of our method and the hypothesis test method proposed by Ling (2005). In each case, we used (9) as the weights for self-weighted least absolute deviation estimator and took C to be the ρ quantile of data $\{y_1, \ldots, y_n\}$. Specifically, we took $\rho = 90\%$ and $\rho = 95\%$. A series of hypothesis tests using the results of Ling (2005) were conducted in the following manner: start from $p = 5$ or 10 and run a Chi-square test for each coefficient at significant level 0.05. If any test statistic is insignificant, we delete the coefficient with the largest p-value and run the procedure again until all remaining coefficients are significant. To run the hypothesis test in Ling (2005), the true value of $f(0)$ was used in each case. One should bear in mind that the performance could be worse if $f(0)$ was estimated in each case.

The sample sizes were chosen as 50, 100, and 200, and the summary statistics were based on 500 replications. To measure variable selection performance, we summarized the average number of correctly identified zero coefficients (CT), the average number of coefficients erroneously set to zero (ICT) and the percent of times when the true model is correctly identified (PCM) using each method. For a fair comparison of the estimation accuracy using our method and Ling (2005)'s method, we calculated the empirical means and standard errors (SE) of estimates obtained from simulation runs where the true model was correctly identified.

Summary results are presented in Tables 1 and 2, where $S(1.5, 0; 1)$ stands for symmetric α-stable distribution with $\alpha = 1.5$ and t_{df} stands for student t-distribution with degrees of freedom df, LAD-alasso stands for the least absolute

deviation adaptive lasso method, Lin-90% and Lin-95% stand for Ling (2005)'s method with $\rho = 90\%$ and 95%, respectively. In all three cases, our method almost always outperforms the method of Ling (2005) on both model selection consistency and the accuracy of coefficient estimators. For the method of Ling (2005), we observe that using $\rho = 90\%$ or $\rho = 95\%$ would yield different model selection results in many cases. In addition, in almost all cases, using $\rho = 95\%$ yields smaller SE. This supports our conjecture that the choice of C in the method of Ling (2005) has an impact on the model selection results as well as on the estimation accuracy.

4. A real data example

In this section, we employ our new method to analyze the Hang Seng Index data, which has been examined by Ling (2005). The data consists of 497 Hang Seng Index daily closing indices from June 3rd, 1996 to May 31st, 1998. Let x_t be the original data and $y_t = \log(x_t/x_{t-1})$. Ling (2005) adopted the Hill estimator to test the tail index of y_t and showed that the data $\{y_t\}$ has infinite variance. Furthermore, to fit the data $\{y_t\}$ with an appropriate infinite variance autoregressive model, Ling (2005) selected the best model by a series of hypothesis tests based on the self-weighted least absolute deviation estimator. The final model used by Ling (2005) is $y_t = \phi_3 y_{t-3} + \epsilon_t$, where the estimator $\tilde{\phi}_3 = 0.123$.

We employed the least absolute deviation adaptive lasso method to fit the data $\{y_t\}$. The criterion defined in (4) was used to select the optimal λ in the same way as described in Subsection 2.5. Three choices of maximum autoregressive order p were taken ($p = 5$, $p = 10$ and $p = 20$) to show the ability of our method to identify the sparse model. The estimation results are presented in Table 3. By using the least absolute deviation adaptive lasso method, all three choices of p lead to the same model with y_{t-3} as the only relevant variable. The coefficients are slightly different because the data need to be formulate differently to a regression problem for different choices of p. This result coincides with the variable selection result of Ling (2005), where the author argued that this is a reasonable model since the residuals from this model passed the independence test and thus may be regarded as white noise.

5. Concluding remarks

In this paper, we have considered the problem of model selection in infinite variance autoregressive models. A unified variable selection approach combining

Table 1: Simulation results with infinite variance innovations when $p = 5$

			Variable Selection			Estimation Accuracy			
						$\hat{\phi}_1$		$\hat{\phi}_3$	
Innovation	n	Method	ICT	CT	PCM (%)	Mean	SE (10^{-2})	Mean	SE (10^{-2})
Cauchy	50	LAD-alasso	0	2.98	98.6	0.496	2.1	−0.695	2.0
		Lin-90%	0	2.36	62.0	0.502	2.5	−0.699	2.3
		Lin-95%	0	2.44	68.0	0.500	2.3	−0.696	2.1
	100	LAD-alasso	0	2.99	99.4	0.499	1.0	−0.698	1.0
		Lin-90%	0	2.46	66.8	0.500	1.4	−0.699	1.5
		Lin-95%	0	2.42	65.0	0.499	1.2	−0.701	1.1
	200	LAD-alasso	0	3.00	100	0.500	0.5	−0.699	0.4
		Lin-90%	0	2.61	73.6	0.500	0.8	−0.700	0.8
		Lin-95%	0	2.43	64.4	0.499	0.8	−0.700	0.7
S(1.5,0;1)	50	LAD-alasso	0.02	2.86	87.3	0.484	6.4	−0.687	5.9
		Lin-90%	0.12	2.64	73.2	0.500	6.4	−0.701	7.1
		Lin-95%	0.05	2.75	81.9	0.497	6.1	−0.696	6.3
	100	LAD-alasso	0	2.97	96.7	0.490	3.5	−0.695	3.2
		Lin-90%	0	2.80	83.9	0.499	4.9	−0.698	4.9
		Lin-95%	0	2.75	80.4	0.497	3.8	−0.697	4.1
	200	LAD-alasso	0	2.99	99.4	0.497	1.9	−0.697	2.1
		Lin-90%	0	2.81	86.6	0.497	3.2	−0.695	3.1
		Lin-95%	0	2.81	84.4	0.499	2.5	−0.698	2.5
t_2	50	LAD-alasso	0.02	2.90	90.2	0.484	5.8	−0.690	5.4
		Lin-90%	0.08	2.65	74.2	0.500	6.8	−0.697	6.6
		Lin-95%	0.04	2.75	80.8	0.500	5.5	−0.696	5.3
	100	LAD-alasso	0	2.98	98.0	0.491	3.1	−0.691	3.2
		Lin-90%	0.01	2.75	80.2	0.502	4.4	−0.697	4.6
		Lin-95%	0	2.78	82.6	0.496	4.0	−0.699	3.7
	200	LAD-alasso	0	3.00	99.6	0.496	1.9	−0.696	2.0
		Lin-90%	0	2.77	81.6	0.502	2.8	−0.700	2.7
		Lin-95%	0	2.82	85.0	0.498	2.3	−0.699	2.4

Table 2: Simulation results with infinite variance innovations when $p = 10$

			Variable Selection			Estimation Accuracy			
						$\hat{\phi}_1$		$\hat{\phi}_3$	
Innovation	n	Method	ICT	CT	PCM (%)	Mean	SE (10^{-2})	Mean	SE (10^{-2})
Cauchy	50	LAD-alasso	0.02	2.94	94.0	0.496	2.1	−0.697	2.0
		Lin-90%	0.11	1.98	43.4	0.502	6.3	−0.692	12.1
		Lin-95%	0.04	2.19	52.2	0.493	4.9	−0.691	6.9
	100	LAD-alasso	0	2.99	99.6	0.498	0.9	−0.699	1.0
		Lin-90%	0	2.35	56.8	0.500	3.1	−0.700	3.8
		Lin-95%	0	2.34	56.2	0.498	2.7	−0.700	3.3
	200	LAD-alasso	0	2.99	99.4	0.499	0.5	−0.699	0.5
		Lin-90%	0	2.52	66.8	0.499	1.6	−0.697	1.9
		Lin-95%	0	2.47	64.0	0.498	1.1	−0.699	1.4
S(1.5,0;1)	50	LAD-alasso	0.06	2.76	77.1	0.476	7.6	−0.683	8.6
		Lin-90%	0.46	2.46	49.8	0.509	11.2	−0.705	13.3
		Lin-95%	0.21	2.66	67.5	0.497	7.3	−0.693	9.7
	100	LAD-alasso	0	2.94	94.1	0.490	3.6	−0.694	3.6
		Lin-90%	0.04	2.79	82.6	0.498	6.7	−0.694	8.5
		Lin-95%	0	2.77	81.4	0.493	5.3	−0.694	6.2
	200	LAD-alasso	0	2.99	99.0	0.496	2.1	−0.696	2.1
		Lin-90%	0	2.83	85.9	0.495	4.3	−0.699	4.9
		Lin-95%	0	2.80	83.1	0.499	3.3	−0.699	4.2
t_2	50	LAD-alasso	0.07	2.76	78.0	0.473	7.5	−0.679	7.7
		Lin-90%	0.45	2.41	49.0	0.497	8.8	−0.710	12.5
		Lin-95%	0.15	2.61	68.8	0.491	7.9	−0.688	10.6
	100	LAD-alasso	0	2.95	95.6	0.491	3.7	−0.691	3.9
		Lin-90%	0.03	2.75	80.2	0.493	6.1	−0.695	7.1
		Lin-95%	0	2.76	81.8	0.494	4.8	−0.696	5.9
	200	LAD-alasso	0	2.99	99.0	0.496	2.1	−0.697	2.0
		Lin-90%	0	2.86	87.2	0.499	3.9	−0.699	5.4
		Lin-95%	0	2.80	83.0	0.499	3.3	−0.698	3.7

Table 3: The final model for the Hang Seng Index data

Method	Selected Model
LAD-alasso($p = 5$)	$y_t = 0.100y_{t-3} + \epsilon_t$
LAD-alasso($p = 10$)	$y_t = 0.108y_{t-3} + \epsilon_t$
LAD-alasso($p = 20$)	$y_t = 0.100y_{t-3} + \epsilon_t$
Lin-95%	$y_t = 0.123y_{t-3} + \epsilon_t$

the least absolute deviation loss function and the adaptive lasso method has been proposed, which can efficiently deal with heavy-tailed autoregressive models with either finite or infinite variance. A simulation study was carried out and a real data example was illustrated. Both theoretical and empirical results show that our method not only works well in selecting appropriate models but also estimates the coefficients efficiently. One remaining issue is whether our method would still work if the order of the autoregressive model $p \to \infty$ as $n \to \infty$, as was considered by some authors in the regression setting; see, for example, Huang et al. (2008). While our empirical results indicate that this might be true as long as p grows at a slower rate, it would be useful to find sound theoretical justifications. This is an interesting topic for future research.

Appendix A. Proof of Theorem 2.2

Proof. Recall the definition of $W_n(u)$ in (5) and denote

$$\tilde{V}_{1n}(u) = W_n(u) + \lambda_n \Sigma_{j=1}^{p} r_j(|\phi_j^0 + b_n^{-1}u_j| - |\phi_j^0|). \tag{A.1}$$

Then we have $b_n(\hat{\phi}_{1n} - \phi_0) = \arg\min\{\tilde{V}_{1n}(u)\}$. By Lemma 2.1, we have, for each u,

$$W_n(u) \to W(u) \quad \text{in distribution}, \tag{A.2}$$

where $W(u)$ is defined in (6). Now consider the second part of (A.1).

If $\phi_j^0 \neq 0$, then by the definition of r_j, we have $r_j \to |\phi_j^0|^{-2} = O_p(1)$ in probability. Furthermore, we have $b_n(|\phi_j^0 + b_n^{-1}u_j| - |\phi_j^0|) \to u_j \text{sgn}(\phi_j^0)$. Thus, by Slutsky's theorem and the condition that $\lambda_n b_n^{-1} \to 0$, we have $\lambda_n r_j(|\phi_j^0 + b_n^{-1}u_j| - |\phi_j^0|) \to 0$ in probability, as $n \to \infty$.

If $\phi_j^0 = 0$, then $b_n(|\phi_j^0 + b_n^{-1}u_j| - |\phi_j^0|) = |u_j|$ and $b_n\tilde{\phi}_{1j} = O_p(1)$ by Lemma 2.1, where $\tilde{\phi}_{1j}$ is the jth element of $\tilde{\phi}_{1n}$. By the condition $\lambda_n \to \infty$, we have

$\lambda_n r_j(|\phi_j^0 + b_n^{-1}u_j| - |\phi_j^0|) = \lambda_n b_n u/(b_n\tilde{\phi}_{1j})^2 \to \infty$ in probability, as $n \to \infty$. To summarize, we have

$$\lambda_n r_j(|\phi_j^0 + b_n^{-1}u_j| - |\phi_j^0|) \to \begin{cases} 0, & \phi_j^0 \neq 0, \\ 0, & \phi_j^0 = 0, u_j = 0, \\ \infty, & \phi_j^0 = 0, u_j \neq 0, \end{cases} \quad \text{in probability as } n \to \infty. \tag{A.3}$$

Combining (A.2) and (A.3) and using Slutsky's theorem, we have $\tilde{V}_{1n}(u) \to \tilde{V}_1(u)$ in distribution, where

$$\tilde{V}_1(u) = \begin{cases} W(u\mid_{u_{\mathcal{S}^c}=0}), & u_j = 0, j \in \mathcal{S}^c, \\ \infty, & \text{otherwise.} \end{cases}$$

Following a discussion similar to that in Davis et al. (1992), it is readily seen that the conditions in Theorem 2.2 guarantee a unique minimum ξ_1 of $W(u\mid_{u_{\mathcal{S}^c}=0})$ almost surely. Therefore, the unique minimum of $\tilde{V}_1(u)$ is $(\xi_1^T, 0^T)^T$. Since $\tilde{V}_{1n}(u)$ is convex, following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$b_n(\hat{\phi}_{1\mathcal{S}} - \phi_{\mathcal{S}}^0) \to \xi_1 \quad \text{in distribution,} \tag{A.4}$$

i.e., $b_n(\hat{\phi}_{1\mathcal{S}} - \phi_{\mathcal{S}}^0) = O_p(1)$.

Next we prove the variable selection consistency. For all $j \in \mathcal{S}$, by (A.4), we have $P(j \in \mathcal{S}^*) \to 1$ immediately. It suffices to show that for all $j \in \mathcal{S}^c$, $P(j \in \mathcal{S}^*) \to 0$. For each $j \in \mathcal{S}^c$, if $j \in \mathcal{S}^*$, we must have $\lambda_n r_j \leq \sum_{t=p+1}^n |X_{tj}|$, where X_{tj} is the jth element of X_t. Thus, it follows that

$$P(j \in \mathcal{S}^*) \leq P\left(\lambda_n|\tilde{\phi}_{1j}|^{-2} \leq \sum_{t=p+1}^n |X_{tj}|\right). \tag{A.5}$$

Using the inequality $|x+y|^\delta \leq |x|^\delta + |y|^\delta$ for $0 < \delta < 1$, we have

$$\frac{1}{n-p}\left(\sum_{t=p+1}^n |X_{tj}|\right)^{\alpha/2} \leq \frac{1}{n-p}\left(\sum_{t=p+1}^n |X_{tj}|^{\alpha/2}\right) \to E(|y_t|^{\frac{\alpha}{2}}) < \infty \tag{A.6}$$

almost surely as $n \to \infty$, since for all $\delta < \alpha$, we have $E|y_t|^\delta < \infty$ by using Lemma 3 of Knight (1989). On the other hand,

$$\frac{1}{n-p}\left(\frac{\lambda_n}{|\tilde{\phi}_{1j}|^2}\right)^{\alpha/2} = \frac{n}{n-p}\left(\frac{\lambda_n b_n^2 n^{-2/\alpha}}{|b_n\tilde{\phi}_{1j}|^2}\right)^{\alpha/2} = \frac{n}{n-p}\frac{\lambda_n^{\alpha/2}}{|b_n\tilde{\phi}_{1j}|^\alpha} \to \infty \tag{A.7}$$

in probability, where (A.7) follows from the condition $\lambda_n \to \infty$ and the facts that $b_n = n^{1/\alpha}$ and $b_n\widetilde{\phi}_{1j} = O_p(1)$ by Lemma 2.1. Combining (A.5)–(A.7), we have $P(j \in \mathcal{S}^*) \to 0$. Thus, we have shown the variable selection consistency, completing the proof of Theorem 2.2. □

References

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

Bhansali, R. J. (1988). Consistent order determination for processes with infinite variance. *Journal of the Royal Statistical Society, Ser. B* **50**, 46–60.

Brockwell, P. J., Davis, R. A. (1991). *Time Series: Theory and Method.* 2nd Ed. Springer, New York.

Castillo, E. (1988). *Extreme Value Theory in Engineering*. Cambridge: Cambridge University Press.

Castillo, E., Minguez, R., Castillo, C., Cofino, A. (2008). Dealing with the multiplicity of solutions of the L_1 and L_∞ regression models. *European Journal of Operational Research* **188**, 460–484.

Cheng,R., Miamee, A. G., Pourahmadi, M. (2000). Regularity and minimality of infinite variance processes. *Journal of Theoretical Probability* **13**, 1115–1122.

Davis, R. A., Knight, K., Liu, J. (1992). M-estimation for autoregressions with infinite variance. *Stochastic Processes and their Applications* **40**, 145–180.

Duffy, D., Mcintosh, A., Rosenstein, M., Willinger, W. (1994). Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks. *IEEE Jornal on Selected Areas in Communications* **12**, 544–551.

Geyer, C. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics* **22**, 1993–2010.

Granger, C., Orr, D. (1972). "Infinite variance" and research strategy in time series analysis. *Journal of the American Statistical Association* **67**, 275–285.

He, X., Ng, P. (1999). COBS: qualitatively constrained smoothing via linear programming. *Computational Statistics* **14**, 315–337.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics* **3**, 1163–1174.

Huang, J., Ma, S., Zhang, C.H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603–1618.

Knight, K. (1989). Consistency of Akaike's information criterion for infinite variance autoregressive processes. *Annals of Statistics* **17**, 824–840.

Knight, K., Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* **28**, 1356–1378.

Koedijk, K., Schafgans, M., De vries, C. (1990). The tail index of exchange rate returns. *Journal of International Economics* **29**, 93–108.

Koenker, R., Ng, P., Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81**, 673-680.

Ling, S. (2005). Self-weighted least absolute deviation estimation for infinite variance autoregressive models. *Journal of the Royal Statistical Society, Ser. B* **67**, 381–393.

Miamee, A. G., Pourahmadi, M. (1988). Wold decomposition, prediction and parameterization of stationary processes with infinite variance. *Probability Theory and Related Fields* **79**, 145–164.

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

R Development Core Team (2009). R: A Language and Enviroment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

Resnick, S. I. (1997). Heavy tail modeling and teletraffic data (with discussion). *Annals of Statistics* **25**, 1805–1869.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Wang, H., Li, G., Tsai, C. L. (2007a). Regression coefficients and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B* **69**, 63–78.

Wang, H., Li, G., Jiang, G. (2007b). Robust regression shrinkage and consistent variable selection via the lad-lasso. *Journal of Business and Economic Statistics* **25**, 347–355.

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

Zou, H., Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36**, 1509–1533.