Statistics and Its Interface Volume 4 (2011) 197–205

Subset ARMA selection via the adaptive Lasso

Kun Chen* and Kung-Sik Chan*†

Model selection is a critical aspect of subset autoregressive moving-average (ARMA) modelling. This is commonly done by subset selection methods, which may be computationally intensive and even impractical when the true ARMA orders of the underlying model are high. On the other hand, automatic variable selection methods based on regularization do not directly apply to this problem because the innovation process is latent. To solve this problem, we propose to identify the optimal subset ARMA model by fitting an adaptive Lasso regression of the time series on its lags and the lags of the residuals from a long autoregression fitted to the time series data, where the residuals serve as proxies for the innovations. We show that, under some mild regularity conditions, the proposed method enjoys the oracle properties so that the method identifies the correct subset model with probability approaching 1 with increasing sample size, and that the estimators of the nonzero coefficients are asymptotically normal with the limiting distribution the same as the case when the true zero coefficients are known *a priori*. We illustrate the new method with simulations and a real application.

AMS 2000 subject classifications: Primary 62M10; secondary 62J12.

Keywords and phrases: Least squares regression, Oracle properties, Ridge regression, Seasonal ARIMA models, Sparsity.

1. INTRODUCTION

Consider a discrete-time, stationary and ergodic process, $\{y_t\}$, driven by an autoregressive moving-average (ARMA) model:

$$\sum_{j=0}^{p^*} \alpha_j^* y_{t-j} = \sum_{j=0}^{q^*} \beta_j^* \epsilon_{t-j}, \tag{1}$$

where (p^*, q^*) are the AR and MA orders, α_j^*s and β_j^*s are the ARMA parameters with $\alpha_0^* = \beta_0^* = 1$, and the ϵ_ts are the innovations of zero mean, uncorrelated over time and of constant variance $\sigma^2 > 0$. Here for simplicity the data are mean corrected.

In fitting an ARMA model, besides estimating the structural parameters α_j^* $(j = 1, \ldots, p^*)$, β_j^* $(j = 1, \ldots, q^*)$ and the variance σ^2, the AR order p and MA order q must also be determined from the observations y_t $(t = 1, \ldots, T)$. A commonly used approach to determining the ARMA orders is to select a model that minimizes some information criterion, e.g. AIC [1] and BIC [15]. Such methods generally require carrying out maximum likelihood estimation for a large number of ARMA models of different orders. However, maximum likelihood estimation of an ARMA model is prone to numerical problems due to multimodality of the likelihood function and the problem of overfitting when the ARMA orders exceed their true values.

[7] proposed an interesting and practical solution to the order determination problem. Firstly, a long AR(n) model is fitted to the data, with the residuals then serving as proxies for the unobserved innovations ϵ_t [3]:

$$\hat{\epsilon}_t = \sum_{j=0}^{n} \hat{a}_j y_{t-j}, \qquad \hat{a}_0 = 1, \qquad t = n+1, \ldots, T, \tag{2}$$

where the $\hat{a}_j$s $(j = 1, \ldots, n)$ are the autoregressive coefficients estimated by solving the Yule-Walker equations (or by least squares). The AR order n can be determined by minimizing the AIC criterion $\log \hat{\sigma}_n^2 + 2n/T$, where $\hat{\sigma}_n^2$ is the corresponding estimator of the innovation variance. In the second step, the ARMA parameters for various (p, q) orders are estimated by regressing y_t on y_{t-j}, for $j = 1, \ldots, p$ and $\hat{\epsilon}_{t-j}$, for $j = 1, \ldots, q$, where $t = m, \ldots, T$ with $m = n + \max(p, q) + 1$; the innovation variance of the model with ARMA orders (p, q) is then estimated by the residual mean square error, which is denoted by $\tilde{\sigma}_{p,q}^2$. The corresponding BIC values are approximated by $\log \tilde{\sigma}_{p,q}^2 + (p + q) \log T/T$. [7] showed that minimizing the approximate BIC leads to consistent estimation of the ARMA orders, under suitable regularity conditions.

Order determination is related to the more general problem of identifying the nonzero components in a subset ARMA model. A subset ARMA model is an ARMA model with a subset of its coefficients being nonzero, which is a useful and parsimonious way for modeling high-order ARMA processes, e.g. seasonal time series. For ARMA process of high orders, finding a subset ARMA model that adequately approximates the underlying process is more important from a practical standpoint than simply determining the ARMA orders. [2] demonstrated that the method of [7] for estimating the ARMA orders can be extended to solving the problem of finding an optimal subset ARMA model, in which

*The authors gratefully thank the US National Science Foundation (NSF–0934617) for partial financial support.

†Corresponding author.

maximum likelihood estimations are avoided by adopting the aforementioned long AR(n) approximation. Specifically, their method consists of (i) fitting all subset regression models of y_t on its own lags 1 to p and lags 1 to q of the residuals from a long autoregression, where p and q are some known upper bounds of the true ARMA orders; and (ii) selecting an optimal subset model from the pool of all subset regression models, according to some information criterion, e.g. BIC. However, this method still relies on exhaustive subset model selection which requires fitting a large number of subset ARMA(p, q) models (2^{p+q} of them!), which may be computational intensive and even impractical when (p, q) are large.

In recent years, there has been extensive research on automatic variable selection methods via regularization, e.g. Lasso [12, 16] and SCAD [5]. Some main advantages of these methods include computational efficiency and the capability of conducting simultaneous parameter estimation and variable selection. The Lasso method is one of the well-developed automatic model selection approaches for linear regression problems. However, the consistency of Lasso may only hold under some conditions, see [20]. In contrast, as shown by [21] and [10], with appropriate data-driven parameter-specific weighted regularization, the adaptive Lasso approach achieves the oracle properties, i.e. asymptotic normality and model selection consistency. More recently, the regularization approaches have been applied to time series analysis, mainly for the autoregressive models. For example, [17] considered shrinkage estimation of regressive and autoregressive coefficients, and [8] and [14] considered penalized order selection for vector autoregressive models. However, to our knowledge, model selection methods based on regularization have not been applied to the more general ARMA model selection problems, mainly due to the difficulty that the innovations ϵ_ts in the ARMA representation are unobservable.

Motivated by [3, 7] and [2], we propose to find an optimal subset ARMA model by fitting an adaptive Lasso regression of the time series y_t on its own lags and those of the residuals that are obtained from fitting a long autoregression to the y_ts. Besides avoiding troublesome maximum likelihood estimation of ARMA models, the proposed approach also dramatically reduces the computational cost of subset selection to the same order of cost of an ordinary least squares fit. We show that under mild regularity conditions, the proposed method achieves the oracle properties, namely, it identifies the correct subset ARMA model with probability tending to one as the sample size increases to infinity, and that the estimators of the nonzero coefficients are asymptotically normal with the limiting distribution the same as that when the zero coefficients are known *a priori.*

2. ADAPTIVE LASSO PROCEDURE FOR SUBSET ARMA MODEL SELECTION

Throughout this section we assume that $\{y_t\}$ is generated according to model (1), and the underlying true ARMA orders $p^* \leq p$ and $q^* \leq q$, where p, q are known upper bounds of the true orders. Let $\mathbf{y} = (y_m, \ldots, y_T)^T$, $\boldsymbol{\epsilon} = (\epsilon_m, \ldots, \epsilon_T)^T$, $\boldsymbol{\tau}^* = (-\alpha_1^*, \ldots, -\alpha_p^*, \beta_1^*, \ldots, \beta_q^*)^T$ and

$$\begin{aligned}\mathbf{X} &= (\mathbf{x}_1, \ldots, \mathbf{x}_{p+q}) \\ &= \begin{pmatrix} y_{m-1} & \cdots & y_{m-p} & \epsilon_{m-1} & \cdots & \epsilon_{m-q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ y_{T-1} & \cdots & y_{T-p} & \epsilon_{T-1} & \cdots & \epsilon_{T-q} \end{pmatrix}.\end{aligned}$$

Then model (1) can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau}^* + \boldsymbol{\epsilon}.$$

It is assumed that only a subset of the (structural) parameters τ_j^* $(j = 1, \ldots, p+q)$ are nonzero.

Our main goal here is to identify the correct subset of nonzero components in the above subset ARMA model. It has been shown that in linear regression models, the adaptive Lasso method can achieve model selection consistency and produce asymptotically unbiased estimators for the nonzero coefficients. However, the adaptive Lasso method does not directly apply here, due to the difficulty that the design matrix $\mathbf{X}$ involves the latent innovation terms ϵ_t $(t = m-1, \ldots, T-1)$. Motivated by [3, 7] and [2], a long AR(n) process is first fitted to the data to obtain the residuals $\hat{\epsilon}_t$, whose expression is given in (2). Let $\hat{\mathbf{X}}$ denote the approximate design matrix obtained with the entries ϵ_t replaced by $\hat{\epsilon}_t$ $(t = m-1, \ldots, T-1)$. We then propose to select the optimal subset ARMA model by the adaptive Lasso regression model of $\mathbf{y}$ on $\hat{\mathbf{X}}$. The adaptive Lasso estimator of $\boldsymbol{\tau}^*$ is given by

$$(3) \qquad \hat{\boldsymbol{\tau}}^{(T)} = \arg\min_{\boldsymbol{\tau}} \left\{ \|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\tau}\|^2 + \lambda_T \sum_{j=1}^{p+q} \hat{w}_j |\tau_j| \right\},$$

where λ_T is the tuning parameter controlling the degree of penalization, and $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_{p+q})^T$ consists of $p+q$ data-driven weights. (Lasso corresponds to the case of using equal weights, i.e. $w_i \equiv 1$.) Following [21], the weights can be chosen as

$$\hat{\mathbf{w}} = |\tilde{\boldsymbol{\tau}}|^{-\eta},$$

where $\tilde{\boldsymbol{\tau}} = (\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T\mathbf{y}$ is the least squares estimator of $\boldsymbol{\tau}^*$ based on $\hat{\mathbf{X}}$, and η is a prespecified nonnegative parameter; here, the absolute value and the power operators apply component-wise. Based on simulations and as suggested by [21], we use $\eta = 2$ in all numerical studies reported below. Note that the weights can also be constructed based on a ridge regression estimator if sample size is small and multicollinearity is a problem. Yet another alternative approach for deriving the weights is to use the Lasso estimator, which is consistent under some conditions.

The adaptive Lasso is essentially a weighted L_1 regularization method. Its loss function is convex and the entire solution path of various λ_T values can be computed efficiently by a modified LARS algorithm [4], with the same order of computational cost of an ordinary least squares fit. Hence we omit the details of the computational algorithm. An unbiased estimator of the degrees of freedom of a Lasso model is shown to be the number of nonzero coefficients [22], which can be used to construct information criteria for selecting λ_T. After the solution path has been found, we consider both the AIC and BIC criteria for determining the optimal λ_T.

The complete model selection strategy proposed is as follows:

I. Fit an AR(n) autoregressive model to obtain residuals that serve as proxies for the innovations as given in (2). The AR order n can be determined by minimizing the AIC criterion as previously described.

II. Fit an adaptive Lasso regression of the time series $\mathbf{y}$ on $\hat{\mathbf{X}}$ as described in (3).

  (i) Construct adaptive weights $\hat{\mathbf{w}}$ by least squares (alternatively, ridge or Lasso) regression of $\mathbf{y}$ on $\hat{\mathbf{X}}$.

  (ii) Find the solution path of the adaptive Lasso regression.

  (iii) The optimal λ_T is the minimizer of some criterion such as AIC and BIC.

III. (Optional) Do maximum likelihood estimation and model diagnostics for the selected subset ARMA model(s) chosen by some information criterion, e.g. BIC.

3. ASYMPTOTIC PROPERTIES OF THE ADAPTIVE LASSO ESTIMATOR

3.1 Assumptions and preliminary results

In this section, we introduce the main assumptions and some useful results from [7].

A1. For the true system (1), the polynomials

$$\alpha^*(z) = \sum_{j=0}^{p^*} \alpha_j^* z^j, \beta^*(z) = \sum_{j=0}^{q^*} \beta_j^* z^j \text{ with } \alpha_0^* = \beta_0^* = 1$$

are coprime, i.e. have no common factors, and that $\alpha^*(z) \neq 0$, $\beta^*(z) \neq 0$, for $|z| \leq 1$.

A2. Let A_t be the σ-algebra of events determined by ϵ_s ($s \leq t$). We assume

$$E(\epsilon_t | A_{t-1}) = 0, E(\epsilon_t^4) < \infty, \text{ and } E(\epsilon_t^2 | A_{t-1}) = \sigma^2.$$

A3. Assume n increases monotonically to infinity at a rate $c \log T \leq n \leq (\log T)^b$, where $c \geq (2 \log \rho_0)^{-1}$ and ρ_0 is the modulus of a zero of $\beta^*(z)$ nearest to $|z| = 1$, for some $1 < b < \infty$.

Assumption A1 ensures that the true model is stationary and ergodic, and that p^* and q^* are the true model orders. Assumption A2 implies that the innovations are martingale-difference sequence, and hence uncorrelated over time; they also have identical variance. Furthermore, the best linear predictor of y_t is the best in the least squares sense, and it also ensures that $\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 - \sigma^2$ converges to zero at a sufficiently rapid rate [7]. Assumption A3 imposes that the order of the long autoregression increases at a certain rate not slower than $c \log(T)$.

Let $\epsilon_t = \sum_{j=0}^{\infty} a_j y_{t-j}$ be the AR(∞) representation of model (1). Note that a_j decreases at a geometric rate and hence

$$\sum_{j=n}^{\infty} |a_j| = o(T^{-1}). \tag{4}$$

The following lemmas are given either explicitly or implicitly by [7] and [6].

Lemma 3.1. *Under Assumptions A1–A3, almost surely,*

$$\max_{1 \leq j \leq n} \left| \frac{1}{T} \sum_{t=1}^{T} \epsilon_t y_{t-j} \right| = O(Q(T)), \tag{5}$$

$$\max_{1 \leq j \leq n} |\hat{a}_j - a_j| = O(Q(T)), \tag{6}$$

where $Q(T) = (\log \log T / T)^{\frac{1}{2}}$. *Moreover, letting* $c_t = \frac{1}{T}\sum_{s=1}^{T-t} y_s y_{s+t}$, *then*

$$\max_{0 \leq t \leq n} |c_t - \gamma_t| = O(Q(T)), \tag{7}$$

where $\gamma_t = E(y_s y_{s+t})$.

Lemma 3.2. *Under Assumptions A1–A3, almost surely,*

$$\frac{1}{T} \sum_{t=m}^{T} y_{t-j} \hat{\epsilon}_{t-k} = E(y_{t-j} \epsilon_{t-k}) + O(Q(T)), \tag{8}$$

uniformly for $j = 1, \ldots, p$ *and* $k = 1, \ldots, q$.

3.2 Oracle properties

We first prove some results related to the adopted AR(n) approximation. Then we prove our main results, which show that the adaptive Lasso estimator enjoys the oracle properties, i.e. asymptotic normality and model selection consistency. In our proofs, we restrict to the case that the weights are derived from the least squares regression. The proofs can be readily extended to the case for ridge regression based weights, with appropriate conditions on the tuning parameter for the ridge regression. However, the case of Lasso-based weights requires further study.

Lemma 3.3. *Under Assumptions A1–A3, almost surely,*

$$(9)\qquad \frac{1}{\sqrt{T}}\sum_{t=m}^{T}\epsilon_t(\hat{\epsilon}_{t-j}-\epsilon_{t-j}) = O\left(\frac{n\log\log T}{\sqrt{T}}\right),$$

$$(10)\qquad \frac{1}{\sqrt{T}}\sum_{t=m}^{T}(\hat{\epsilon}_t-\epsilon_t)(\hat{\epsilon}_{t-j}-\epsilon_{t-j}) = O\left(\frac{n^2\log\log T}{\sqrt{T}}\right).$$

Proof.

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=m}^{T}\epsilon_t(\hat{\epsilon}_{t-j}-\epsilon_{t-j})\\
&\quad= \frac{1}{\sqrt{T}}\sum_{t=m}^{T}\epsilon_t\left\{\sum_{u=0}^{n}(\hat{a}_u-a_u)y_{t-j-u}\right\}\\
&\qquad - \frac{1}{\sqrt{T}}\sum_{t=m}^{T}\epsilon_t\left(\sum_{u=n+1}^{\infty}a_u y_{t-j-u}\right)\\
&\quad= \sqrt{T}\sum_{u=0}^{n}\left\{(\hat{a}_u-a_u)\left(\frac{1}{T}\sum_{t=m}^{T}\epsilon_t y_{t-j-u}\right)\right\}\\
&\qquad - \sqrt{T}\sum_{u=n+1}^{\infty}\left\{a_u\left(\frac{1}{T}\sum_{t=m}^{T}\epsilon_t y_{t-j-u}\right)\right\}\\
&\quad= \sqrt{T}\cdot n\cdot O(Q(T))\cdot O(Q(T)) + \sqrt{T}\cdot o(T^{-1})\cdot O(Q(T))\\
&\quad= O\left(\frac{n\log\log T}{\sqrt{T}}\right).
\end{aligned}$$

Here we have used (4) and the uniform convergence results in (5) and (6) of Lemma 3.1. Now consider (10),

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=m}^{T}(\hat{\epsilon}_t-\epsilon_t)(\hat{\epsilon}_{t-j}-\epsilon_{t-j})\\
&\quad= \sqrt{T}\sum_{u=0}^{n}\left\{(\hat{a}_u-a_u)\left(\frac{1}{T}\sum_{t=m}^{T}(\hat{\epsilon}_t-\epsilon_t)y_{t-j-u}\right)\right\}\\
&\qquad - \sqrt{T}\sum_{u=n+1}^{\infty}\left\{a_u\left(\frac{1}{T}\sum_{t=m}^{T}(\hat{\epsilon}_t-\epsilon_t)y_{t-j-u}\right)\right\}\\
&\quad= \sqrt{T}\sum_{u=0}^{n}\left\{(\hat{a}_u-a_u)\sum_{v=0}^{n}(\hat{a}_v-a_v)\left(\frac{1}{T}\sum_{t=m}^{T}y_{t-v}y_{t-j-u}\right)\right\}\\
&\qquad - \sqrt{T}\sum_{u=0}^{n}\left\{(\hat{a}_u-a_u)\sum_{v=n+1}^{\infty}a_v\left(\frac{1}{T}\sum_{t=m}^{T}y_{t-v}y_{t-j-u}\right)\right\}\\
&\qquad - \sqrt{T}\sum_{u=n+1}^{\infty}\left\{a_u\left(\frac{1}{T}\sum_{t=m}^{T}(\hat{\epsilon}_t-\epsilon_t)y_{t-j-u}\right)\right\}.
\end{aligned}$$

By (4) and (6), it then suffices to show

$$\begin{aligned}
&\sqrt{T}\sum_{u=0}^{n}\left\{(\hat{a}_u-a_u)\sum_{v=0}^{n}(\hat{a}_v-a_v)\left(\frac{1}{T}\sum_{t=m}^{T}y_{t-v}y_{t-j-u}\right)\right\}\\
&\quad= \sqrt{T}\sum_{u=0}^{n}\left\{(\hat{a}_u-a_u)\sum_{v=0}^{n}(\hat{a}_v-a_v)(\gamma_{v-j-u}+O(Q(T)))\right\}\\
&\quad= \sqrt{T}\cdot n\cdot O(Q(T))\cdot n\cdot O(Q(T))\\
&\quad= O\left(\frac{n^2\log\log T}{\sqrt{T}}\right).
\end{aligned}$$

Here we have used the uniform convergence results in Lemmas 3.1 and 3.2. This completes the proof. □

Lemma 3.4. *Under Assumptions A1–A3,* $\frac{1}{T}\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}} \to \boldsymbol{C}$ *almost surely, where* $\boldsymbol{C}$ *is a nonsingular constant matrix.*

Proof. Write $\hat{\mathbf{X}} = (\mathbf{Y}, \hat{\mathbf{E}})$, where

$$\mathbf{Y} = \begin{pmatrix} y_{m-1} & \cdots & y_{m-p}\\ \vdots & \ddots & \vdots\\ y_{T-1} & \cdots & y_{T-p}\end{pmatrix}, \hat{\mathbf{E}} = \begin{pmatrix} \hat{\epsilon}_{m-1} & \cdots & \hat{\epsilon}_{m-q}\\ \vdots & \ddots & \vdots\\ \hat{\epsilon}_{T-1} & \cdots & \hat{\epsilon}_{T-q}\end{pmatrix}.$$

Then

$$\frac{1}{T}\hat{\mathbf{X}}^T\hat{\mathbf{X}} = \begin{pmatrix} \frac{1}{T}\mathbf{Y}^T\mathbf{Y} & \frac{1}{T}\mathbf{Y}^T\hat{\mathbf{E}}\\ \frac{1}{T}\hat{\mathbf{E}}^T\mathbf{Y} & \frac{1}{T}\hat{\mathbf{E}}^T\hat{\mathbf{E}}\end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_{11} & \hat{\mathbf{X}}_{12}\\ \hat{\mathbf{X}}_{21} & \hat{\mathbf{X}}_{22}\end{pmatrix}.$$

Consider $\hat{\mathbf{X}}_{12}$ and $\hat{\mathbf{X}}_{12}$, with a typical entry given by $\frac{1}{T}\sum_{t=m}^{T}y_{t-j}\hat{\epsilon}_{t-k} = E(y_{t-j}\epsilon_{t-k}) + O(Q(T))$ by Lemma 3.2. Now consider $\hat{\mathbf{X}}_{22}$, a typical entry of which being, for some $j,k = 1,\ldots,q$, equal to

$$\begin{aligned}
&\frac{1}{T}\sum_{t=m}^{T}\hat{\epsilon}_{t-j}\hat{\epsilon}_{t-k}\\
&\quad= \sum_{u=0}^{n}\hat{a}_u\left\{\frac{1}{T}\sum_{t=m}^{T}\hat{\epsilon}_{t-j}y_{t-k-u}\right\}\\
&\quad= \sum_{u=0}^{n}(a_u+O(Q(T)))\{E(y_{t-k-u}\epsilon_{t-j})+O(Q(T))\}\\
&\quad= E\left\{\left(\sum_{u=0}^{n}a_u y_{t-k-u}\right)\epsilon_{t-j}\right\} + O(nQ(T))\\
&\quad= E(\epsilon_{t-k}\epsilon_{t-j}) + O\left(\frac{(\log\log T)^{\frac{1}{2}}(\log T)^b}{\sqrt{T}}\right).
\end{aligned}$$

Here we have used (4) and the uniform convergence results in Lemmas 3.1 and 3.2. Therefore, we have shown that $\frac{1}{T}\hat{\mathbf{X}}^T\hat{\mathbf{X}}$ has the same limit as $\frac{1}{T}\mathbf{X}^T\mathbf{X}$, which converges to a nonsingular constant matrix, almost surely, by ergodicity and the fact that the innovation variance is positive. This completes the proof. □

Lemma 3.5. *Let* $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{X}}\boldsymbol{\tau}^*$. *Under Assumptions A1–A3,* $\frac{\hat{\boldsymbol{X}}^T\tilde{\boldsymbol{\epsilon}}}{\sqrt{T}}$ *has the same limiting distribution as* $\frac{\boldsymbol{X}^T\boldsymbol{\epsilon}}{\sqrt{T}}$, *i.e.* $\frac{\hat{\boldsymbol{X}}^T\tilde{\boldsymbol{\epsilon}}}{\sqrt{T}} \to_d \boldsymbol{W}$, $\frac{\boldsymbol{X}^T\boldsymbol{\epsilon}}{\sqrt{T}} \to_d \boldsymbol{W}$, *where* $\boldsymbol{W} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{C})$.

Proof. We decompose $\frac{\hat{\mathbf{X}}^T\tilde{\boldsymbol{\epsilon}}}{\sqrt{T}}$ into four parts:

$$\frac{\hat{\mathbf{X}}^T\tilde{\boldsymbol{\epsilon}}}{\sqrt{T}} = \frac{\mathbf{X}^T\boldsymbol{\epsilon}}{\sqrt{T}} + \frac{(\hat{\mathbf{X}}-\mathbf{X})^T\boldsymbol{\epsilon}}{\sqrt{T}} + \frac{\mathbf{X}^T(\hat{\mathbf{X}}-\mathbf{X})\boldsymbol{\tau}^*}{\sqrt{T}} + \frac{(\hat{\mathbf{X}}-\mathbf{X})^T(\hat{\mathbf{X}}-\mathbf{X})\boldsymbol{\tau}^*}{\sqrt{T}}. \tag{11}$$

It suffices to show that the second, third and the fourth terms on the right side of (11) are $o(1)$, which follows easily from Lemma 3.3. This completes the proof. □

Lemma 3.6. *Recall $\tilde{\boldsymbol{\tau}} = (\hat{\boldsymbol{X}}^T\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}^T\boldsymbol{y}$ is the least squares estimator of $\boldsymbol{\tau}^*$ based on $\hat{\boldsymbol{X}}$. Under Assumptions A1–A3, $\sqrt{T}(\tilde{\boldsymbol{\tau}}-\boldsymbol{\tau}^*) \to_d N(\boldsymbol{0}, \sigma^2\boldsymbol{C}^{-1})$.*

Proof. We decompose $\sqrt{T}(\tilde{\boldsymbol{\tau}}-\boldsymbol{\tau}^*)$ as follows:

$$\begin{aligned}\sqrt{T}(\tilde{\boldsymbol{\tau}}-\boldsymbol{\tau}^*) = \left(\frac{1}{T}\hat{\mathbf{X}}^T\hat{\mathbf{X}}\right)^{-1}\bigg\{&\frac{1}{\sqrt{T}}\mathbf{X}^T\boldsymbol{\epsilon} + \frac{1}{\sqrt{T}}(\hat{\mathbf{X}}-\mathbf{X})^T\boldsymbol{\epsilon}\\ &-\frac{1}{\sqrt{T}}\mathbf{X}^T(\hat{\mathbf{X}}-\mathbf{X})\boldsymbol{\tau}^*\\ &-\frac{1}{\sqrt{T}}(\hat{\mathbf{X}}-\mathbf{X})^T(\hat{\mathbf{X}}-\mathbf{X})\boldsymbol{\tau}^*\bigg\}.\end{aligned}$$

By Lemma 3.3 and following the same argument used in proving Lemma 3.5, $\sqrt{T}(\tilde{\boldsymbol{\tau}}-\boldsymbol{\tau}^*) = (\frac{1}{T}\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\frac{1}{\sqrt{T}}\mathbf{X}^T\boldsymbol{\epsilon} + o(1)$. The claimed limiting distribution then follows from Lemmas 3.4 and 3.5. □

We first define some notations. Let $\mathcal{A} = \{j : \tau_j^* \neq 0\}$ and $\mathcal{A}^c = \{j : \tau_j^* = 0\}$. Similarly, let $\hat{\mathcal{A}}_T = \{j : \hat{\tau}_j^{(T)} \neq 0\}$ and $\hat{\mathcal{A}}_T^c = \{j : \hat{\tau}_j^{(T)} = 0\}$. Suppose $\mathbf{Z}$ is an $m \times n$ matrix, and $\mathcal{A}$ and $\mathcal{B}$ are subsets of the collection of row and column indices of $\mathbf{Z}$, respectively. We let $\mathbf{Z}_{\mathcal{AB}}$ denote a sub-matrix of $\mathbf{Z}$ whose rows and columns are chosen from $\mathbf{Z}$ according to the index sets $\mathcal{A}$ and $\mathcal{B}$, respectively. For simplicity, we may write $\mathbf{Z}_{\mathcal{AA}} = \mathbf{Z}_{\mathcal{A}}$ when $\mathbf{Z}$ is a square matrix, $\mathbf{Z}_{\mathcal{AB}} = \mathbf{Z}_{\cdot\mathcal{B}}$ ($\mathbf{Z}_{\mathcal{A}\cdot}$) when $\mathcal{A}$ ($\mathcal{B}$) consists of all the row (column) indies, and $\mathbf{Z}_{\mathcal{A}\cdot} = \mathbf{Z}_{\mathcal{A}}$ when $\mathbf{Z}$ is a vector.

Theorem 3.7 (Oracle Properties)**.** *Suppose A1–A3 hold, and assume $\frac{\lambda_T}{\sqrt{T}}T^{\frac{\eta}{2}} \to \infty$ and $\lambda_T/\sqrt{T} \to 0$. Then*

(i) *Asymptotic normality:*
$\sqrt{T}(\hat{\boldsymbol{\tau}}_{\mathcal{A}}^{(T)} - \boldsymbol{\tau}_{\mathcal{A}}^*) \to_d N(\boldsymbol{0}, \sigma^2\boldsymbol{C}_{\mathcal{A}}^{-1})$ *as* $T \to \infty$.

(ii) *Selection consistency:*
$\lim_{T\to\infty} P(\hat{\mathcal{A}}_T = \mathcal{A}) = 1$.

Proof. The proof is similar in structure to the proof of the main result in [21]. Let $\boldsymbol{\tau} = \boldsymbol{\tau}^* + \frac{\mathbf{u}}{\sqrt{T}}$, $\Psi_T(\mathbf{u}) = \|\mathbf{y} - \hat{\mathbf{X}}(\boldsymbol{\tau}^* + \frac{\mathbf{u}}{\sqrt{T}})\|^2 + \lambda_T\sum_{j=1}^{p+q}\hat{w}_j|\tau_j^* + \frac{u_j}{\sqrt{T}}|$, and $\hat{\mathbf{u}}^{(T)} = \arg\min\Psi_T(\mathbf{u})$. Then $\hat{\mathbf{u}}^{(T)} = \sqrt{T}(\hat{\boldsymbol{\tau}}^{(T)} - \boldsymbol{\tau}^*)$. Let $V_T(\mathbf{u}) = \Psi_T(\mathbf{u}) - \Psi_T(\mathbf{0})$. Then we have

$$\begin{aligned}V_T(\mathbf{u}) &= \mathbf{u}^T\left(\frac{1}{T}\hat{\mathbf{X}}^T\hat{\mathbf{X}}\right)\mathbf{u} - 2\frac{\tilde{\boldsymbol{\epsilon}}^T\hat{\mathbf{X}}}{\sqrt{T}}\mathbf{u}\\ &\quad+\frac{\lambda_T}{\sqrt{T}}\sum_{j=1}^{p+q}\hat{w}_j\sqrt{T}\Big(\Big|\tau_j^* + \frac{u_j}{\sqrt{T}}| - |\tau_j^*\Big|\Big).\end{aligned}$$

By Lemmas 3.4–3.6 and following [21], we have $V_T(\mathbf{u}) \to_d V(\mathbf{u})$ for every $\mathbf{u}$, where

$$V(\mathbf{u}) = \begin{cases}\mathbf{u}_{\mathcal{A}}^T\mathbf{C}_{\mathcal{A}}\mathbf{u}_{\mathcal{A}} - 2\mathbf{u}_{\mathcal{A}}^T\mathbf{W}_{\mathcal{A}} & \text{if } u_j = 0\ \forall j \notin \mathcal{A}\\ \infty & \text{otherwise.}\end{cases}$$

$V(\mathbf{u})$ is convex and has a unique minimum. Following [12], we have

$$\hat{\mathbf{u}}_{\mathcal{A}}^{(T)} \to_d \mathbf{C}_{\mathcal{A}}^{-1}\mathbf{W}_{\mathcal{A}} \text{ and } \hat{\mathbf{u}}_{\mathcal{A}^c}^{(T)} \to_d 0. \tag{12}$$

Finally, upon recalling $\mathbf{W}_{\mathcal{A}} \sim N(0, \sigma^2\mathbf{C}_{\mathcal{A}})$, the asymptotic normality result follows.

Next, we show the consistency part. $\forall j \in \mathcal{A}$, the asymptotic normality result indicates that $\hat{\tau}_j^{(T)} \to_p \tau_j^*$; it follows that $P(j \in \hat{\mathcal{A}}_T) \to 1$. It suffices to show that $\forall j \notin \mathcal{A}$, $P(j \in \hat{\mathcal{A}}_T) \to 0$. Consider the event $j \notin \mathcal{A}$ and $j \in \hat{\mathcal{A}}_T$. By the Karush-Kuhn-Tucker (KKT) optimality conditions, we have

$$\frac{2\hat{\mathbf{x}}_j^T(\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\tau}}^{(T)})}{\sqrt{T}} = \frac{\lambda_T\hat{w}_j}{\sqrt{T}}. \tag{13}$$

Note that $\frac{\lambda_T\hat{w}_j}{\sqrt{T}} = \frac{\lambda_T}{\sqrt{T}}T^{\eta/2}|\sqrt{T}\tilde{\tau}_j|^{-\eta} \to \infty$. Consider the left side of (13),

$$\frac{2\hat{\mathbf{x}}_j^T(\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\tau}}^{(T)})}{\sqrt{T}} = \frac{2\hat{\mathbf{x}}_j^T\tilde{\boldsymbol{\epsilon}}}{\sqrt{T}} + \frac{2\hat{\mathbf{x}}_j^T\hat{\mathbf{X}}}{T}\sqrt{T}(\boldsymbol{\tau}^* - \hat{\boldsymbol{\tau}}^{(T)}).$$

By Lemma 3.5, $\frac{2\hat{\mathbf{x}}_j^T\tilde{\boldsymbol{\epsilon}}}{\sqrt{T}} = O_p(1)$. By Lemma 3.4 and (12), $\frac{2\hat{\mathbf{x}}_j^T\hat{\mathbf{X}}}{T}\sqrt{T}(\boldsymbol{\tau}^* - \hat{\boldsymbol{\tau}}^{(T)}) = O_p(1)$. Thus

$$P(j \in \hat{\mathcal{A}}_T) \leq P\left(\frac{2\hat{\mathbf{x}}_j^T(\mathbf{y} - \hat{\mathbf{X}}\hat{\boldsymbol{\tau}}^{(T)})}{\sqrt{T}} = \frac{\lambda_T\hat{w}_j}{\sqrt{T}}\right) \to 0.$$

This completes the proof. □

4. EMPIRICAL PERFORMANCE

We study the empirical performance of the proposed subset model selection method by simulations. Four Gaussian ARMA models are considered:

Model I: $(1 - 0.8B)(1 - 0.7B^6)y_t = \epsilon_t$;

Model II: $(1 - 0.8B)(1 - 0.7B^6)y_t = (1 + 0.8B)(1 + 0.7B^6)\epsilon_t$;

Table 1. Summary statistics of a Monte Carlo study of the empirical performance of the adaptive Lasso subset model selection method for Models I–III. Numbers in the columns with the heading "A" report the relative frequencies of picking all significant variables; those under the heading "T" are the relative frequencies of picking the correct model; those under the heading "−" report the false negative rates; those under the heading "+" report the false positive rates. Numbers under the the heading "N" are the sample sizes. All experiments were replicated 1,000 times

Model I																
	AIC				BIC											
	Lasso				Lasso				Ridge				LS			
N	A	T	−	+	A	T	−	+	A	T	−	+	A	T	−	+
120	0.76	0.19	0.09	0.07	0.75	0.44	0.09	0.04	0.24	0.00	0.26	0.15	0.75	0.19	0.13	0.22
240	0.98	0.30	0.01	0.06	0.98	0.80	0.01	0.01	0.44	0.01	0.19	0.20	0.85	0.40	0.07	0.18
360	1.00	0.32	0.00	0.06	1.00	0.87	0.00	0.01	0.54	0.01	0.15	0.20	0.87	0.49	0.07	0.17

Model II																
	AIC				BIC											
	Lasso				Lasso				Ridge				LS			
N	A	T	−	+	A	T	−	+	A	T	−	+	A	T	−	+
120	0.40	0.01	0.12	0.20	0.32	0.04	0.15	0.14	0.26	0.00	0.18	0.22	0.02	0.00	0.38	0.45
240	0.81	0.03	0.04	0.21	0.76	0.15	0.05	0.13	0.36	0.01	0.18	0.30	0.05	0.00	0.35	0.47
360	0.92	0.04	0.02	0.21	0.92	0.24	0.02	0.10	0.36	0.01	0.20	0.36	0.04	0.00	0.36	0.48

Model III																
	AIC				BIC											
	Lasso				Lasso				Ridge				LS			
N	A	T	−	+	A	T	−	+	A	T	−	+	A	T	−	+
120	0.03	0.00	0.54	0.18	0.03	0.00	0.58	0.13	0.04	0.00	0.47	0.16	0.05	0.00	0.75	0.29
240	0.17	0.00	0.36	0.21	0.14	0.01	0.43	0.14	0.10	0.00	0.37	0.20	0.07	0.00	0.61	0.36
360	0.38	0.00	0.25	0.23	0.34	0.02	0.31	0.15	0.16	0.00	0.33	0.23	0.06	0.00	0.62	0.36

Model III: $y_t = (1 + 0.8B)(1 + 0.7B^6)\epsilon_t$;

Model IV: $y_t = (1 - 0.6B - 0.8B^{12})\epsilon_t$,

where B is the backshift operator so that $B^k y_t = y_{t-k}$, and $\{\epsilon_t\}$ are independent standard normal random variables.

The first three models are multiplicative seasonal models with seasonal period 6, whereas the last model is a non-multiplicative seasonal model with seasonal period 12. For the long autoregressive fits, the AR order was chosen by AIC, with the maximum order set to be $10\log_{10}(T)$. (We have also experimented by fixing the long AR order to the preceding maximum order, but obtained similar results.) As mentioned earlier, for the proposed adaptive Lasso method, there are several ways for determining the weights. (Following [21], the power η in the weights was set to be 2 in all experiments.) The simplest method is ordinary least squares regression (LS), but LS suffers from large variability in the case of low signal to noise ratio. In many applications involving, say monthly data, sample size may range from 120 (10 years) to 360 (30 years). In our simulation experiments for models I to III, we set the sample size to be either 120, 240 or 360, with the maximum AR and MA lags both equal to 14. Consequently, the number of data per parameter is at most slightly higher than 10, for these sample sizes. For such cases, LS is very variable, and the simulation results reported below show that the adaptive Lasso, with weights determined by LS, performed poorly, except for Model I; see Table 1.

The poor performance of the LS-weighted adaptive Lasso may be partly attributed to multicollinearity, which may be somewhat alleviated by using ridge regression, i.e. by minimizing the penalized sum of squares with the penalty equal to the product of a (non-negative) tuning parameter times the L_2-norm of the regression coefficients. The tuning parameter may be determined by minimizing the generalized cross validation (GCV), see [18]. In our simulations, we implemented the ridge regression via the magic function of the mgcv library [18] in the R platform [13]. Yet another method is to use the Lasso to derive the initial weights. The Lasso is known to be consistent under some regularity conditions, see [20]. We have experimented with both AIC and BIC in determining the Lasso tuning parameter. For determining the initial weights, AIC and BIC yielded similar results (unreported); hence, we only report results with the initial weights determined by (i) Lasso with the tuning parameter obtained by minimizing BIC, (ii) ridge regression with the tuning parameter obtained by minimizing GCV and (iii) LS.

Table 1 shows the model selection results of the adaptive Lasso method under the three weighting schemes and different sample sizes. Each experiment was replicated 1,000 times. For each experiment, Table 1 provides 4 statistics: (i) the relative frequency of including all significant variables, (ii) the relative frequency of picking the true model,

(iii) the false negative rate and (iv) the false positive rate. For Models I to III, the maximum AR lag is 14 and so is the maximum MA lag. Recall that the tuning parameter of the adaptive Lasso may be determined by either AIC or BIC. Columns 2–5 of Table 1 report the simulation results for the adaptive Lasso with the weights determined by Lasso, and the tuning parameter of the adaptive Lasso determined by AIC, whereas columns 6–9 report those when the tuning parameter of the adaptive Lasso was determined by BIC. Comparison between the two schemes show that BIC and AIC performed similarly, with BIC having a higher chance of picking the true model, lower false positive rates and slightly higher false negative rates; AIC tends to have a higher chance of including all significant variables at the expense of selecting more complex models than BIC. The proposed method performed very well for the pure AR model, less so for the mixed ARMA model, but performed somewhat poorly for the pure MA model as specified by Model III. Columns 10–13 and columns 14–17 of Table 1 display the results for the proposed adaptive Lasso method with the weighting scheme given by ridge regression and LS, respectively. These results show that the adaptive Lasso with an LS-based weighting scheme generally performed quite poorly except for the AR example, and using weights based on ridge regression alleviated the problem somewhat, but it is still outperformed by the method of adaptive Lasso with Lasso-based weights.

Given that it is difficult to identify an MA model, we also experimented with searching among the subset MA models. In Table 2, we reported results for another three sets of experiments, with models selected by adaptive Lasso with Lasso-based weights. In the first set of experiments, we repeated the experiments with Model III, but with the model selection confined among (subset) MA models, with the maximum MA lag equal to 14. We show the results with the tuning parameter of the adaptive Lasso determined by either AIC or BIC. Again, AIC and BIC performed similarly, and now the proposed method works extremely well if the search is restricted to MA models. This result suggests that, for data analysis, it may be prudent to apply the proposed method with the search among pure AR models, then among mixed ARMA models and finally among the pure MA models. One can then explore further with the optimal models from these three model selection exercises, in order to arrive at an "optimal" model for the data on hand.

Table 2 also reports the results for Model III with larger sample sizes, namely, 480, 600 and 720. The results show that the rates of false negatives and false positives decline steadily with increasing sample size, and the rate of including all significant variables increases steadily with sample size, as well.

Finally, Table 2 reports the results for Model IV, with the maximum AR and MA lags equal to 26, and sample size equal to 240 or 360. While this is also a pure MA model, the performance of the proposed Lasso-weighted adaptive Lasso method is comparable to the case for Model III with sample sizes 600–720; the better performance may be due to the larger gap between the two MA coefficients in Model IV, which induces weaker autocorrelations in the data. The limited Monte Carlo experiments reported here suggest that the proposed method is a promising new tool for identifying sparse stationary ARMA models, especially if the sparsity contains wide gaps in the ARMA coefficients. In practice, real data may be non-stationary, and they must be transformed to stationarity before applying the proposed adaptive Lasso subset model selection method.

Table 2. Summary statistics of a Monte Carlo study of the empirical performance of the adaptive Lasso subset model selection method for Models III–IV. Numbers in the columns with the heading "A" report the relative frequencies of picking all significant variables; those under the heading "T" are the relative frequencies of picking the correct model; those under the heading "−" report the false negative rates; those under the heading "+" report the false positive rates. Numbers under the the heading "N" are the sample sizes. All experiments were replicated 1,000 times

Model III (selection confined among MA models)								
	AIC				BIC			
N	A	T	−	+	A	T	−	+
120	0.90	0.25	0.03	0.17	0.87	0.38	0.05	0.10
240	1.00	0.28	0.00	0.18	1.00	0.51	0.00	0.09
360	1.00	0.30	0.00	0.17	1.00	0.51	0.00	0.08

Model III								
	AIC				BIC			
N	A	T	−	+	A	T	−	+
480	0.54	0.00	0.18	0.24	0.48	0.04	0.24	0.15
600	0.64	0.00	0.14	0.25	0.59	0.06	0.18	0.15
720	0.69	0.01	0.12	0.26	0.66	0.10	0.15	0.14

Model IV								
	AIC				BIC			
N	A	T	−	+	A	T	−	+
240	0.55	0.04	0.26	0.06	0.49	0.08	0.29	0.04
360	0.79	0.03	0.10	0.07	0.74	0.12	0.13	0.05

5. A REAL APPLICATION

As an example, we consider a time series of the monthly CO_2 level from a monitoring site at Alert, Northwest Territories, Canada. (The dataset is contained in the TSA library in R.) The dataset was earlier analyzed by [2, Chapter 12], who fitted a period-12 seasonal $\text{ARIMA}(0,1,1) \times (0,1,1)_{12}$ model. In particular the series is nonstationary. Here, we apply regular differencing and period-12 seasonal differencing to the data before carrying out the Lasso-weighted adaptive Lasso model selection. In practice, it is better to report a few optimal models selected by adaptive Lasso. Figure 1 shows the best 5 models selected by adaptive Lasso with its tuning parameter determined by BIC; each row corresponds to one

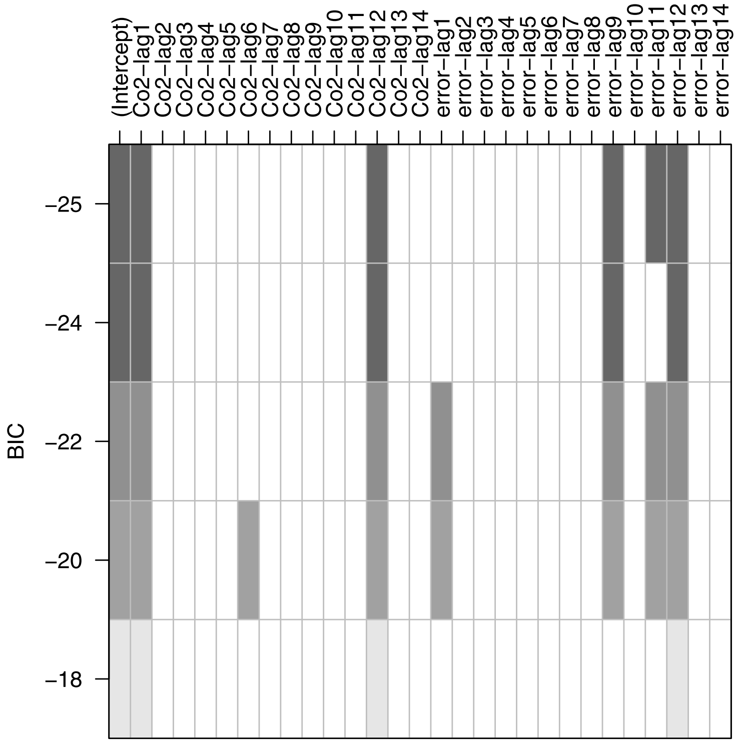

Figure 1. Best five subset models selected for the regularly and seasonally differenced CO_2 series.

selected model with the selected variables shaded in dark, and the models are ordered from top to bottom, according to their BICs. According to Figure 1, the optimal subset model contains lags 1 and 12 of the response variable and lags 9, 11 and 12 of the errors (residuals from the long autoregression, being proxies for the latent innovations). The presence of the error lags 11 and 12 suggests a multiplicative model. Indeed, the lag 1 of the error process appears in some of the selected models shown in Figure 1. The presence of the error lag 9 is harder to interpret. Nevertheless, we fitted a subset $\text{ARIMA}(1,1,9)\times(1,1,1)_{12}$ model with the coefficients of error lags 2 to 8 fixed at zero. The model fit (unreported) suggested that the regular AR(1) and seasonal AR(1) coefficients are non-significant, so these are dropped from a second fitted model:

$$(1-B)(1-B^{12})y_t = (1 - 0.64\ (0.09)\ B - 0.26\ (0.08)\ B^9) \times (1 - 0.81\ (0.1)\ B^{12})\hat{\epsilon}_t,$$

where the numbers in parentheses are the standard errors. This fitted model appears to fit the data well as the residuals were found to be approximately white.

6. CONCLUSION

The numerical studies reported in the preceding two sections illustrate the efficacy of the proposed subset ARMA selection method. There are other time series modeling tasks that require ARMA order section, e.g. VARMA models and transfer-function (dynamic regression) models. It is interesting to extend the proposed method to these settings. While we have derived the oracle properties of the proposed method using LS-based weights, it is worthwhile to investigate the asymptotics for the case of Lasso-based weights, especially in view of the much better empirical properties of the adaptive Lasso selection method using Lasso-based weights. Given the fact that the coefficients in an ARMA model naturally form an AR group and an MA group, it would be interesting to explore bi-level selection penalty forms such as the group bridge penalty [11].

While we focus on subset ARMA models, a similar problem occurs in nonparametric stochastic regression models [19]. A challenging problem consists of lifting some of the automatic model selection methods to the nonparametric setting; see [9] for some recent works in the additive framework.

Received 9 October 2010

REFERENCES

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716–723. MR0423716

[2] Cryer, J. D. and Chan, K.-S. (2008). *Time Series Analysis: With Applications in R*, 2nd ed. Springer, New York.

[3] Durbin, J. (1960). The fitting of time series models. *Int. Statist. Rev.* **28** 233–244.

[4] Efron, B., Hastie, T., Johnstones, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**(2) 407–499. MR2060166

[5] 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. MR1946581

[6] Hannan, E. J. and Kavalieris, L. (1984). A method for autoregressive-moving average estimation. *Biometrika* **71** 273–280. MR0767155

[7] Hannan, E. J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika* **69** 81–94. MR0655673

[8] Hsu, N. J., Hung, H. L. and Chang, Y. M. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics & Data Analysis* **52** 3645–3657. MR2427370

[9] Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics* **38** 2283–2313. MR2676890

[10] Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive Lasso for high-dimensional regression models. *Statistica Sinica* **18** 1603–1618. MR2469326

[11] Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. MR2507147

[12] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics* **28** 1356–1378. MR1805787

[13] R Development Core Team, (2008). R: A Language and Environment for Statistical Computing ISBN 3-900051-07-0.

[14] Ren, Y. and Zhang, X. (2010). Subset selection for vector autoregressive processes via adaptive Lasso. *Statistics & Probability Letters* **80** 1705–1712.

[15] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464. MR0468014

[16] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society (Series B)* **58** 267–288. MR1379242

[17] Wang, H., Li, G. and Tsai, C. (2007). Regression coefficient and autoregressive order shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B.* **69** 63–78. MR2301500

[18] Wood, S. N. (2006). *Generalized Additive Models, An Introduction with R.* Chapman and Hall, London. MR2206355

[19] Yao, Q. and Tong, H. (1994). On subset selection in nonparametric stochastic regression. *Statistica Sinica* **4** 51–70. MR1282865

[20] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[21] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469

[22] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degree of freedom of the Lasso. *The Annals of Statistics* **35** 2173–2192. MR2363967

Kun Chen
Dept. of Statistics and Actuarial Science
University of Iowa
Iowa City, Iowa 52242
USA
E-mail address: kun-chen@uiowa.edu

Kung-Sik Chan
Dept. of Statistics and Actuarial Science
University of Iowa
Iowa City, Iowa 52242
USA
E-mail address: kung-sik-chan@uiowa.edu