J. R. Statist. Soc. B (2007)
69, *Part* 1, *pp.* 63–78

Regression coefficient and autoregressive order shrinkage and selection via the lasso

Hansheng Wang,

Peking University, Beijing, People's Republic of China

Guodong Li

University of Hong Kong, People's Republic of China

and Chih-Ling Tsai

University of California at Davis, USA

[Received February 2005. Final revision August 2006]

Summary. The *least absolute shrinkage and selection operator* ('lasso') has been widely used in regression shrinkage and selection. We extend its application to the regression model with autoregressive errors. Two types of lasso estimators are carefully studied. The first is similar to the traditional lasso estimator with only two tuning parameters (one for regression coefficients and the other for autoregression coefficients). These tuning parameters can be easily calculated via a data-driven method, but the resulting lasso estimator may not be fully efficient. To overcome this limitation, we propose a second lasso estimator which uses different tuning parameters for each coefficient. We show that this modified lasso can produce the estimator as efficiently as the *oracle*. Moreover, we propose an algorithm for tuning parameter estimates to obtain the modified lasso estimator. Simulation studies demonstrate that the modified estimator is superior to the traditional estimator. One empirical example is also presented to illustrate the usefulness of lasso estimators. The extension of the lasso to the autoregression with exogenous variables model is briefly discussed.

Keywords: Autoregression with exogenous variables; Lasso; Oracle estimator; Regression model with autoregressive errors

1. Introduction

The linear regression model is a commonly used statistical tool for analysis of the relationships between response and explanatory variables. One of its standard assumptions is that different observations are independent. However, significant serial correlation might occur when the data are collected sequentially in time. In this case, the linear regression with autoregressive errors (REGAR) model is often considered, as it takes into account the autocorrelated structure in regression analysis (Shumway and Stoffer, 2000; Tsay, 1984; Harvey, 1981).

In model building, it is known that making the model unnecessarily complex can degrade the efficiency of the resulting parameter estimator and yield less accurate predictions. Hence, two heuristic selection criteria, Akaike's information criterion AIC (Akaike, 1973) and the Bayes information criterion BIC (Schwarz, 1978), are often applied to select regression variables. In the context of time series, both criteria are also employed to choose the order of the autoregressive

Address for correspondence: Hansheng Wang, Guanghua School of Management, Peking University, Beijing 100871, People's Republic of China.
E-mail: hansheng@gsm.pku.edu.cn

 1369–7412/07/69063

process (Brockwell and Davis, 1991; Choi, 1992; McQuarrie and Tsai, 1998; Shumway and Stoffer, 2000). Moreover, Ramanathan (1989) extended the application of AIC and BIC to the linear REGAR model. However, as noted by various researchers, the statistical performance of AIC and BIC can be unstable (Breiman, 1996), and selection bias may cause inference problems (Hurvich and Tsai, 1990).

To amend the deficiencies of classical linear model selections, Tibshirani (1996) developed the *least absolute shrinkage and selection operator* ('lasso'), which selects variables and estimates parameters simultaneously. This motivated us to obtain the shrinkage estimator in the autoregressive process. For this, we employ the lasso-type penalty not only on the regression coefficients but also on the autoregression coefficients. Consequently, a direct extension of the lasso to the REGAR model involves two regularization parameters (i.e. one for regression coefficients and the other for autoregression coefficients), which can be easily tuned via a data-driven method (e.g. cross-validation). We show that the resulting lasso estimator satisfies a Knight–Fu-type asymptotic property (Knight and Fu, 2000). However, it suffers an appreciable bias (Fan and Li, 2001). Hence, the traditional lasso estimator cannot achieve the same efficiency as the *oracle*, i.e. the estimator that is obtained on the basis of the true model (Fan and Li, 2001).

To improve the utility of the traditional lasso approach to the REGAR model, we modify the penalty function so that different tuning parameters can be used for each coefficient. As a result, large amounts of shrinkage can be used for the insignificant variables, whereas small amounts of shrinkage can be used for the significant variables. We show that the resulting modified lasso estimator shares the same asymptotic distribution as the oracle. In practice, however, simultaneously tuning many regularization parameters is not realistic. Therefore, we further propose a tuning parameter algorithm via the unpenalized REGAR estimator. Simulation studies indicate that the resulting lasso estimator outperforms the traditional lasso estimator.

The rest of the paper is organized as follows. Section 2 introduces the REGAR model and the two lasso estimators. Asymptotic theory for the two lasso estimators is established in Section 3. The practical implementations of these two estimators are developed in Section 4, and numerical studies are presented in Section 5. Section 6 concludes the paper with a brief discussion.

2. Least absolute shrinkage and selection operators

Consider the REGAR model

$$y_t = x_t'\beta + e_t \qquad (t = 1, \ldots, n_0), \tag{1}$$

where $x_t = (x_{t1}, \ldots, x_{tp})'$ is the p-dimensional regression covariate and $\beta = (\beta_1, \ldots, \beta_p)'$ is the associated regression coefficient. In addition, the variable e_t follows the autoregressive process

$$e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \ldots + \phi_q e_{t-q} + \varepsilon_t, \tag{2}$$

where $\phi = (\phi_1, \ldots, \phi_q)'$ is the autoregression coefficient and ε_t are independent and identically distributed random variables with mean 0 and variance σ^2. Moreover, we define the regression and autoregressive parameters as $\theta = (\beta', \phi')'$. For practical implementation, it is common practice to standardize the predictor x_{tj} so that it has zero mean and unit variance (Tibshirani, 1996). Analogously, the response y_t is scaled by dividing it by the estimate of $\text{var}(e_t)^{1/2}$.

Suppose that ε_t in model (2) follows a normal distribution and the first q observations are fixed. Then the conditional likelihood function of the remaining $n_0 - q$ observations, $(y_{q+1}, \ldots, y_{n_0})'$, is

$$\left\{\frac{1}{\surd(2\pi)\sigma}\right\}^n \exp\left[-\frac{1}{2\sigma^2}\sum_{t=q+1}^{n_0}\left\{y_t - x_t'\beta - \sum_{j=1}^{q}\phi_j(y_{t-j} - x_{t-j}'\beta)\right\}^2\right],$$

where $n = n_0 - q$ is the effective sample size. Maximizing this likelihood function yields a conditional maximum likelihood estimator of θ. This estimator can also be obtained by minimizing the least-squares-type objective function

$$L_n(\theta) = \sum_{t=q+1}^{n_0}\left\{y_t - x_t'\beta - \sum_{j=1}^{q}\phi_j(y_{t-j} - x_{t-j}'\beta)\right\}^2, \quad (3)$$

where $L_n(\theta)$ is an extension of the method that was proposed by Cochrane and Orcutt (1949) for $q = 1$; see Harvey (1981) and Hamilton (1994).

To shrink unnecessary coefficients to 0, we next adapt Tibshirani's (1996) approach for obtaining the estimator by minimizing the lasso criterion

$$Q_n(\theta) = \sum_{t=q+1}^{n_0}\left\{y_t - x_t'\beta - \sum_{j=1}^{q}\phi_j(y_{t-j} - x_{t-j}'\beta)\right\}^2 + n\sum_{j=1}^{p}\lambda|\beta_j| + n\sum_{j=1}^{q}\gamma|\phi_j|. \quad (4)$$

Because the lasso uses the same tuning parameters λ and γ for the regression and autoregressive coefficients respectively, the resulting estimator $\hat{\theta} = (\hat{\beta}', \hat{\phi}')'$ may suffer an appreciable bias. This is mainly because all the regression (or autoregression) coefficients share the same amount of shrinkage (see Fan and Li (2001)). To overcome this limitation, we propose the modified lasso criterion, which will be denoted by lasso*,

$$Q_n^*(\theta) = \sum_{t=q+1}^{n_0}\left\{y_t - x_t'\beta - \sum_{j=1}^{q}\phi_j(y_{t-j} - x_{t-j}'\beta)\right\}^2 + n\sum_{j=1}^{p}\lambda_j^*|\beta_j| + n\sum_{j=1}^{q}\gamma_j^*|\phi_j|, \quad (5)$$

which allows for different tuning parameters λ_j^* and γ_j^* for different coefficients. As a result, a larger amount of shrinkage can be applied to the insignificant coefficients, whereas a smaller amount of shrinkage can be applied to the significant coefficients. Hence, the resulting estimator $\hat{\theta}^* = (\hat{\beta}^{*\prime}, \hat{\phi}^{*\prime})'$ is expected to have a smaller bias than $\hat{\theta}$. Detailed investigations of these two estimators are given in the next section.

3. Theoretical properties

To study the theoretical properties of the two lasso estimators, we assume that there is a correct model with the regression and autoregression coefficients $\theta^0 = (\beta^{0\prime}, \phi^{0\prime})' = (\beta_1^0, \ldots, \beta_p^0, \phi_1^0, \ldots, \phi_q^0)'$. Furthermore, we assume that there are a total of $p_0 \leqslant p$ non-zero regression coefficients and $q_0 \leqslant q$ non-zero autoregression coefficients. For convenience, we define $\mathcal{S}_1 = \{1 \leqslant j \leqslant p : \beta_j^0 \neq 0\}$ and $\mathcal{S}_2 = \{1 \leqslant j \leqslant q : \phi_j^0 \neq 0\}$. Then, the sets $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively contain the indices of the significant regression and autoregression coefficients, and their complements $\mathcal{S}_1^c$ and $\mathcal{S}_2^c$ respectively contain the indices of the insignificant regression and autoregression coefficients. Next, let $\beta_{\mathcal{S}_1}$ denote the $p_0 \times 1$ significant regression coefficient vector with $\hat{\beta}_{\mathcal{S}_1}$ as its associated lasso estimator. Moreover, other related parameters and their corresponding estimators are analogously defined (e.g. $\beta_{\mathcal{S}_1^c}, \hat{\beta}_{\mathcal{S}_1^c}, \hat{\beta}^*_{\mathcal{S}_1}, \phi_{\mathcal{S}_2}$ and $\hat{\phi}_{\mathcal{S}_2}$). Finally, let $\theta_1^0 = (\beta_{\mathcal{S}_1}^{0\prime}, \phi_{\mathcal{S}_2}^{0\prime})'$ and $\theta_2^0 = (\beta_{\mathcal{S}_1^c}^{0\prime}, \phi_{\mathcal{S}_2^c}^{0\prime})'$. Then, $\hat{\theta}_k$ and $\hat{\theta}_k^*$ $(k = 1, 2)$ are the associated lasso and modified lasso estimators respectively. To investigate the theoretical properties of $\hat{\theta}$ and $\hat{\theta}^*$, we introduce the following conditions.

(a) The sequence $\{x_t\}$ is independent of $\{\varepsilon_t\}$.
(b) All roots of the polynomial $1-\Sigma_{j=1}^{q}\phi_j^0 z^j$ are outside the unit circle.
(c) The error ε_t has a finite fourth-order moment, i.e. $E(\varepsilon_t^4)<\infty$.
(d) The covariate x_t is strictly stationary and ergodic with finite second-order moment (i.e. $E\|x_t\|^2<\infty$). Furthermore, the following matrix is positive definite:

$$B=E\left\{\left(x_t-\sum_{j=1}^{q}\phi_j^0 x_{t-j}\right)\left(x_t-\sum_{j=1}^{q}\phi_j^0 x_{t-j}\right)'\right\}.$$

The technical conditions above are typically used to assure the $\sqrt{n}$-consistency and asymptotic normality of the unpenalized least square estimator.

3.1. The lasso estimator

In this subsection, we study the property of the traditional lasso estimator given below.

Theorem 1. Assume that $\lambda_n\sqrt{n}\to\lambda_0$ and $\gamma_n\sqrt{n}\to\gamma_0$ for some $\lambda_0\geqslant 0$ and $\gamma_0\geqslant 0$. Then, under conditions (a)–(d), we have $(\hat{\theta}-\theta^0)\sqrt{n}\to_{\mathrm{d}}\arg\min\{\kappa(\delta)\}$, where

$$\kappa(\delta)=-2\delta' w+\delta'\Sigma\delta+\lambda_0\sum_{j=1}^{p}\{u_j\,\mathrm{sgn}(\beta_j^0)I(\beta_j^0\neq 0)+|u_j|\,I(\beta_j^0=0)\}$$
$$+\gamma_0\sum_{j=1}^{q}\{v_j\,\mathrm{sgn}(\phi_j^0)\,I(\phi_j^0\neq 0)+|v_j|I(\phi_j^0=0)\},$$

$\delta=(u',v')'$, $w\sim N(0,\sigma^2\Sigma)$, $\Sigma=\mathrm{diag}(B,C)$, $C=(\xi_{|i-j|})$ and $\xi_k=E(e_t e_{t+k})$.

The proof is given in Appendix A. Theorem 1 shows that the lasso estimator has a Knight–Fu-type asymptotic property (Knight and Fu, 2000). This implies that the tuning parameters that are used in the traditional lasso estimator cannot shrink to 0 at a speed faster than $n^{-1/2}$. Otherwise, both λ_0 and γ_0 degenerate to 0 and $\kappa(\delta)$ becomes a standard quadratic function,

$$\kappa(\delta)=\kappa(u,v)=-2(u',v')w+(u',v')\Sigma(u',v')',$$

which cannot produce sparse solutions. Therefore, theorem 1 suggests that $\lambda_0>0$ and $\gamma_0>0$ are needed for obtaining the traditional lasso estimator.

Remark 1. In a standard regression model with independent observations, Fan and Li (2001) noticed that the traditional lasso estimator may suffer an appreciable bias. Therefore, it is of interest to investigate whether the traditional lasso estimator for the REGAR model encounters the same problem. For illustration, we consider a special case with $\beta_j^0>0$ for $1\leqslant j\leqslant p$ and $\phi_j^0=0$ for $1\leqslant j\leqslant q$. If the minimizer of $\kappa(\delta)$ can correctly identify the true model, then $u\neq 0$ but $v=0$. In addition, $\kappa(\delta)$ satisfies the equation

$$\frac{\partial\kappa(u,0)}{\partial u}=-2w_1+2u'B+\lambda_0\mathbf{1}=0,$$

where w_1 consists of the first p components of w and $\mathbf{1}$ is a $p\times 1$ vector with elements 1. As a result,

$$(\hat{\beta}-\beta^0)\sqrt{n} \underset{\mathrm{d}}{\rightarrow} u = B^{-1}(w_1 - 0.5\lambda_0\mathbf{1}),$$

which is distributed as $N(-0.5\lambda_0 B^{-1}\mathbf{1}, B^{-1})$. Because $\lambda_0 > 0$, theorem 1 indicates that the traditional lasso estimator is asymptotically biased. Thus, it is not as efficient as the oracle estimator, whose asymptotic distribution is $N(0, B^{-1})$.

3.2. The modified lasso estimator

In this subsection, we focus on the modified lasso estimator. To facilitate a study of the properties of this estimator, we introduce the notation

$$a_n = \max(\lambda^*_{j_1}, \gamma^*_{j_2}, j_1 \in \mathcal{S}_1, j_2 \in \mathcal{S}_2),$$

$$b_n = \min(\lambda^*_{j_1}, \gamma^*_{j_2}, j_1 \in \mathcal{S}_1^c, j_2 \in \mathcal{S}_2^c),$$

where $\lambda^*_{j_1}$ and $\gamma^*_{j_2}$ are functions of n. We first investigate the consistency of the modified lasso estimator lasso*.

Lemma 1. Assume that $a_n = o(1)$ as $n \to \infty$. Then, under conditions (a)–(d), there is a local minimizer $\hat{\theta}^*$ of $Q^*_n(\theta)$ such that

$$\hat{\theta}^* - \theta^0 = O_p(n^{-1/2} + a_n).$$

The proof is given in Appendix B. Lemma 1 implies that, if the tuning parameters that are associated with the significant regression variables and autoregressive orders converge to 0 at a speed that is faster than $n^{-1/2}$, then there is a local minimizer of $Q^*_n(\theta)$ which is $\sqrt{n}$ consistent.

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

Theorem 2. Assume that $b_n\sqrt{n} \to \infty$ and $\|\hat{\theta}^* - \theta^0\| = O_p(n^{-1/2})$. Then

$$P(\hat{\beta}^*_{\mathcal{S}_1^c} = 0) \to 1 \quad \text{and} \quad P(\hat{\phi}^*_{\mathcal{S}_2^c} = 0) \to 1.$$

The proof is in Appendix C. Theorem 2 shows that the modified lasso can consistently produce a sparse solution for insignificant regression and autoregression coefficients. Furthermore, this theorem, together with lemma 1, indicates that the $\sqrt{n}$-consistent estimator $\hat{\theta}^*$ must satisfy $P(\hat{\theta}^*_2 = 0) \to 1$ when the tuning parameters fulfil the appropriate conditions (for example, λ_j and γ_j are defined as in equations (7) of the next section). Finally, we obtain the asymptotic distribution of the modified lasso estimator.

Theorem 3. Assume that $a_n\sqrt{n} \to 0$ and $b_n\sqrt{n} \to \infty$. Then, under conditions (a)–(d), the component $\hat{\theta}^*_1$ of the local minimizer $\hat{\theta}^*$ that is given in lemma 1 satisfies

$$(\hat{\theta}^*_1 - \theta^0_1)\sqrt{n} \underset{\mathrm{d}}{\rightarrow} N(0, \sigma^2\Sigma_0^{-1}),$$

where Σ_0 is the submatrix of Σ corresponding to θ^0_1.

The proof is given in Appendix D. Theorem 3 implies that, if the tuning parameters satisfy the conditions $a_n\sqrt{n} \to 0$ and $b_n\sqrt{n} \to \infty$, then, asymptotically, the resulting modified lasso estimator can be as efficient as the oracle estimator.

4. Algorithm

After arriving at an understanding of the properties of the two lasso estimators, it is natural to implement them for real applications. For this, we propose the following algorithm to obtain the local minimizers for lasso estimators $\hat{\theta}$ and $\hat{\theta}^*$. In addition, we provide an approach to estimate simultaneously a total of $p+q$ tuning parameters for the modified lasso estimator.

4.1. The iterative process

The objective function $Q_n^*(\theta)$ contains $Q_n(\theta)$ as a special case (i.e. $\lambda_j=\lambda$ and $\gamma_j=\gamma$). Therefore, we focus mainly on the optimization problem of $Q_n^*(\theta)$ in the rest of this section. Because equation (5) contains both regression and autoregression parameters, it is sensible to optimize the objective function $Q_n^*(\theta)$ iteratively by minimizing the following two lasso-type objective functions:

$$\sum_{t=q+1}^{n_0}\left\{y_t-x_t'\beta-\sum_{j=1}^{q}\phi_j(y_{t-j}-x_{t-j}'\beta)\right\}^2+n\sum_{j=1}^{p}\lambda_j|\beta_j| \qquad \text{with a fixed } \phi$$

and

$$\sum_{t=q+1}^{n_0}\left\{y_t-x_t'\beta-\sum_{j=1}^{q}\phi_j(y_{t-j}-x_{t-j}'\beta)\right\}^2+n\sum_{j=1}^{q}\gamma_j|\phi_j| \qquad \text{with a fixed } \beta.$$

As a result, many well-developed procedures can be used to find the solution for the above non-concave penalized functions, e.g. quadratic programming (Tibshirani, 1996), the shooting algorithm (Fu, 1998), the local quadratic approximation (Fan and Li, 2001) and, most recently, the least angle regression method (Efron *et al.*, 2004). For simplicity, we adapt the local quadratic approximation procedure, which was first developed by Fan and Li (2001) and has been used extensively in the literature (for example, see Fan and Li (2002), Fan and Peng (2004) and Cai *et al.* (2005)). Our simulation studies indicate that this procedure converges with a reasonable degree of speed and accuracy.

Remark 2. The solution of the local quadratic approximation does not yield a sparse solution. However, the small parameter estimate that is produced by the local quadratic approximation can be arbitrarily close to 0, as long as a sufficiently small threshold for its tolerance of accuracy is set up. For illustration, ordinary linear regression is considered. In this case, the local quadratic approximation produces the one-step-ahead estimate $\beta^{(m+1)}$ by minimizing

$$\|Y-X\beta^{(m+1)}\|^2+n\sum_{j=1}^{p}\lambda_j\frac{(\beta_j^{(m+1)})^2}{|\beta_j^{(m)}|},$$

where $Y=(y_1,\ldots,y_{n_0})'$ and $X=(x_1,\ldots,x_{n_0})'$. If one of the coefficients (e.g. $\beta_1^{(m)}$) is very small (but not sparse), then the ridge effect that is induced by $\beta_1^{(m)}$, $\lambda_1/|\beta_1^{(m)}|$, can be very large. As a result, the value of $|\beta_1^{(m+1)}|$ is forced to be even smaller. Because this is an iterative process, the value of $|\beta_1^{(m)}|$ can be arbitrarily close to 0 as long as we can have a sufficiently small threshold for accuracy. Therefore, it is possible to set up an arbitrarily small thresholding value to shrink small estimates to be exactly 0. By doing so, the sparse solution is obtained. In simulation studies, we use the thresholding value 10^{-9} so that any coefficient whose absolute value is smaller than 10^{-9} is shrunk to be exactly 0.

4.2. Local convexity

Although the iterative process proposed is easy to implement, we cannot be assured that the resulting estimator converges to the global minimizer. This is because the least squares term, $L_n(\theta)$, in the objective function $Q_n^*(\theta)$ is not a convex function. This motivates us to develop the following theorem, which shows that there is a sufficiently small but fixed local region containing the true parameter in which $L_n(\theta)$ is almost surely guaranteed to be convex.

Theorem 4. There is a probability null set $\mathcal{N}_0$ and a sufficiently small but fixed $\delta > 0$ such that, for any $\omega \notin \mathcal{N}_0$, there is an integer n_ω such that, for any $n > n_\omega$, $L_n(\theta)$ is convex in $\theta \in \mathcal{B}_\delta$, where $\mathcal{B}_\delta = \{\theta : \|\theta - \theta^0\| < \delta\}$ is a ball containing the true value θ^0.

The proof of theorem 4 can be obtained on request from the authors. Theorem 4 indicates that, with probability tending to 1, there will be at most one local minimizer in $\mathcal{B}_\delta$. According to lemma 1, $\hat{\theta}^*$ exists and is consistent in probability. Hence, theorem 4 together with lemma 1 imply that, with probability tending to 1, $\hat{\theta}^*$ is the unique local minimizer in $\mathcal{B}_\delta$. As a result, the desired local minimizer $\hat{\theta}^*$ can be obtained by finding the unique local minimizer in $\mathcal{B}_\delta$.

Remark 3. Theorem 4 is applicable not only for the modified lasso estimator $\hat{\theta}^*$ but also for the traditional lasso estimator $\hat{\theta}$. Specifically, theorem 4 together with theorem 1 imply that $\hat{\theta}$ can be obtained by finding the unique local minimizer in $\mathcal{B}_\delta$. In practice, however, it is not necessary to know $\mathcal{B}_\delta$ exactly. This is because, if the initial estimator is consistent, then it must be within $\mathcal{B}_\delta$ with a probability tending to 1. As a result, the iterative process proposed (with aforementioned initial estimator) leads to the local minimizer (i.e. $\hat{\theta}^*$ or $\hat{\theta}$) in $\mathcal{B}_\delta$ with probability tending to 1.

4.3. Initial estimator

To obtain the consistent estimator for the iterative process, we suggest the following ordinary least squares estimator as an initial estimator for the regression coefficient β^0:

$$\hat{\beta}^{(0)} = (X'X)^{-1}(X'Y).$$

Using the fact that ε_t is independent of x_t (see condition (a) in Section 3), it can be shown that $\hat{\beta}^{(0)}$ is a consistent estimator of β^0 under classical regularity conditions. Then, computing the ordinary residual $\hat{e}_t = y_t - x_t'\hat{\beta}^{(0)}$ and employing the least squares approach by fitting $\hat{e}_t$ *versus* $(\hat{e}_{t-1}, \ldots, \hat{e}_{t-q})$, we obtain the following initial estimator for the autoregressive coefficient ϕ^0:

$$\hat{\phi}^{(0)} = (W'W)^{-1}(W'V),$$

where $V = (\hat{e}_{q+1}, \ldots, \hat{e}_{n_0})'$ and W is an $n \times q$ matrix with tth row given by $(\hat{e}_{t+q-1}, \ldots, \hat{e}_t)$. It can also be shown that $\hat{\phi}^{(0)}$ is a consistent estimator of ϕ^0 under classical regularity conditions.

4.4. Tuning parameters

After obtaining the initial estimator, we need to select the tuning parameters in the iterative process to complete the whole algorithm. The traditional lasso estimator contains only two tuning parameters (i.e. λ and γ). Hence, we can directly apply the commonly used cross-validation (CV) method to select the optimal tuning parameters. Because of the time series structure, we use the first half of the data for model training and the rest for model testing. In the classical linear regression setting, however, Shao (1997) indicated that BIC would perform better than CV if the true model has a finite dimension and is among the candidate models. This motivates us to adapt the BIC-type tuning parameter selector of Zou *et al.* (2004):

$$\mathrm{BIC} = \log(\hat{\sigma}^2) + \widehat{\mathrm{df}}\, \log(n)/n, \tag{6}$$

where

$$\hat{\sigma}^2 = n^{-1} \sum_{t=q+1}^{n_0} \left\{ y_t - x_t'\hat{\beta} - \sum_{j=1}^{q} \hat{\phi}_j (y_{t-j} - x_{t-j}'\hat{\beta}) \right\}^2$$

and $\widehat{\mathrm{df}}$ is the number of non-zero coefficients of $\hat{\theta}$.

As for the modified lasso estimator, it becomes a challenging task since there are $p+q$ regularization parameters that need to be tuned. Following a referee's suggestion, we propose the adaptive estimators

$$\lambda_j^* = \lambda^* \log(n)/n|\tilde{\beta}_j|, \qquad (7)$$
$$\gamma_j^* = \gamma^* \log(n)/n|\tilde{\phi}_j|,$$

where $\tilde{\theta} = (\tilde{\beta}', \tilde{\phi}')'$ is the unpenalized least square estimator by assuming that $\lambda = \gamma = 0$ in equation (4). In addition, both λ^* and γ^* are positive constants and estimated from the data. The advantage of expression (7) is that it converts the original $(p+q)$-dimensional tuning problem for finding λ_j and γ_j into a two-dimensional task for searching λ^* and γ^*, which can be easily determined by using either CV or BIC.

According to theorem 1, $\tilde{\theta}$ is a $\sqrt{n}$-consistent estimator of θ^0. Hence, for any $\beta_j^0 \neq 0$ and $\phi_j^0 \neq 0$, we have $\lambda_j^* = O_p\{\log(n)/n\} = o_p(n^{-1/2})$ and $\gamma_j^* = O_p\{\log(n)/n\} = o_p(n^{-1/2})$. Consequently, both λ_j^* and γ_j^* satisfy the condition $a_n\sqrt{n} \to 0$, where a_n is defined in Section 3.2. In contrast, for any $\beta_j^0 = 0$ and $\phi_j^0 = 0$, theorem 1 implies that $\tilde{\beta}_j = O_p(n^{-1/2})$ and $\tilde{\phi}_j = O_p(n^{-1/2})$. Therefore,

$$\lambda_j^* \sqrt{n} = \lambda^* \log(n)/\tilde{\beta}_j \sqrt{n}$$

and

$$\gamma_j^* \sqrt{n} = \gamma^* \log(n)/\tilde{\phi}_j \sqrt{n},$$

where the denominators of the above equations are $O_p(1)$ and the numerators go to ∞ as $n \to \infty$. As a result, $\lambda_j^* \sqrt{n} \to_{\mathrm{p}} \infty$ and $\gamma_j^* \sqrt{n} \to_{\mathrm{p}} \infty$, which imply that both satisfy the condition $b_n\sqrt{n} \to \infty$, where b_n is defined in Section 3.2. In sum, the proposed tuning parameters λ_j^* and γ_j^* can produce the modified lasso estimator $\hat{\theta}^*$, which is as efficient as the oracle estimator asymptotically.

5. Simulation and example

5.1. Simulation results

We present Monte Carlo simulations to evaluate the finite sample performance of the lasso estimators. They consist of the traditional and modified lasso estimators with the tuning parameters selected by CV and BIC respectively. For the traditional lasso estimator, we adapt Zou and Hastie's (2005) approach to select the optimal tuning parameters $\hat{\lambda}$ and $\hat{\gamma}$ from the grid points $\{0, 0.01, 0.1, 1.0, 10, 100\}$. For the modified lasso estimator, the optimal tuning parameter $\hat{\tau}$ is selected from one of six equally spaced grid points from 0 to 0.5 (i.e. 0, 0.1, $0.2, \ldots, 0.5$). Our simulation experience suggests that such a search region and spacing work satisfactorily. In addition, the estimation algorithm stops if $\Sigma_j |\theta_j^{(m)} - \theta_j^{(m+1)}| < 10^{-12}$, where $\theta^{(m)} = (\theta_1^{(m)}, \ldots, \theta_{p+q}^{(m)})'$ is the estimator of θ at the mth iteration, $\theta_j^{(m)} = \beta_j^{(m)}$ for $j = 1, \ldots, p$ and $\theta_j^{(m)} = \phi_{j-p}^{(m)}$ for $j = p+1, \ldots, p+q$. When convergence is obtained, any parameter estimator whose absolute value is less than 10^{-9} is shrunk to 0. On the basis of our extensive simulation studies, the above proposed stopping and shrinking rules lead to a reasonable speed of convergence.

We generated the data from the REGAR model

$$y_t = 3.0x_{t1} + 1.5x_{t2} + 2.0x_{t5} + e_t, \tag{8}$$

where

$$e_t = 0.5e_{t-1} - 0.70e_{t-3} + \sigma\varepsilon_t, \tag{9}$$

and ε_t were independent and identically standard normal random variables for $t = 1, \ldots, n_0$. The regression and autocorrelated coefficients are $\beta^0 = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ and $\phi^0 = (0.50, 0, -0.70, 0, 0)'$ respectively. In addition, the covariates $x_t = (x_{t1}, \ldots, x_{t8})'$ were independently generated from the multivariate normal distribution with mean $\mathbf{0}_{8\times 1}$, and the pairwise correlation between x_{tj_1} and x_{tj_2} is $\rho^{|j_1 - j_2|}$. Regression model (8) is adapted from Tibshirani (1996) and has been used in other simulation studies (for example, see Fan and Li (2001), Zou *et al.* (2005) and Leng *et al.* (2006)), whereas the autoregression model (9) is modified from Shi and Tsai (2004).

In this study, we consider three sample sizes ($n_0 = 50, 100, 300$) and two standard deviations ($\sigma = 3.0$ and $\sigma = 0.5$). In addition, the correlation coefficients (ρ-values) are 0.75, 0.50 and 0.25, which represent high, moderate and low linear correlations between the covariates. For each setting, 1000 realizations were carried out, and the percentage of correctly estimated, underestimated and overestimated numbers of regression variables, the percentage of correctly estimated, underestimated and overestimated numbers of autoregressive orders, and the percentage of the correct model identified by two lasso estimators were computed.

When $\rho = 0.5$, Table 1 shows that the lasso performs poorly across various sample sizes and noises. This is because its tuning parameter is fixed and therefore cannot effectively shrink non-significant coefficients to 0. As a result, it tends to overfit in both regression and autoregression variable selection. In contrast, the modified lasso with the CV selector (lasso*–CV) demonstrates a considerably improved finite sample performance. Furthermore, the modified lasso with BIC selector (lasso*–BIC) performs the best in correct model identifications across various sample sizes and levels of noise. Moreover, as the sample size increases, the correct model percentage approaches 100% rapidly. In sum, we recommend employing the estimator lasso*–BIC jointly to choose variables and to estimate coefficients.

In addition to the correct model identification, a referee suggested comparing the prediction accuracies of four lasso estimates in terms of their mean-squared prediction error. For this, we generated an additional 10000 independent testing samples within each realization, which are used to evaluate the accuracy of prediction. Analogously to the correct model selection results, Table 1 shows that the lasso–CV estimator performs the worst, and the lasso*–BIC estimator outperforms the rest. Similar patterns (which are not presented here) are also found when $\rho = 0.25$ and $\rho = 0.75$.

5.2. *Electricity demand study*

We consider a data set that was taken from Ramanathan (1989), which studies the consumption of electricity of residential customers served by the San Diego Gas and Electric Company. The data contain a total of 53 quarterly observations, running from the first quarter of 1970 to the first quarter of 1983. The response variable is electricity consumption, which is measured by the log-transformed electricity consumption per residential customer in millions of kilowatt-hours (LKWH). The five explanatory variables are the logarithm of *per capita* real income (LY), the logarithm of real average price of residential electricity in dollars per kilowatt-hour (LELP),

Table 1. Simulation results with $\rho = 0.50$

Estimator	*Tuning method*	*Regression variable*			*Autoregressive order*			*Correctly fitted model*	*Median of mean-squared prediction error*
		Under-fitted	*Correctly fitted*	*Over-fitted*	*Under-fitted*	*Correctly fitted*	*Over-fitted*		
$\sigma = 3.0$, $n = 50$									
lasso	CV	0.019	0.174	0.807	0.043	0.142	0.815	0.026	17.430
	BIC	0.017	0.223	0.760	0.018	0.206	0.776	0.045	16.715
lasso*	CV	0.078	0.412	0.510	0.063	0.585	0.352	0.245	16.228
	BIC	0.101	0.578	0.321	0.074	0.752	0.174	0.455	15.382
$\sigma = 3.0$, $n = 100$									
lasso	CV	0.001	0.235	0.764	0.001	0.126	0.873	0.020	14.713
	BIC	0.001	0.367	0.632	0.000	0.176	0.824	0.054	14.392
lasso*	CV	0.003	0.572	0.425	0.002	0.654	0.344	0.376	13.826
	BIC	0.003	0.852	0.145	0.003	0.932	0.065	0.796	13.504
$\sigma = 3.0$, $n = 300$									
lasso	CV	0.000	0.144	0.856	0.000	0.133	0.867	0.011	13.194
	BIC	0.000	0.167	0.833	0.000	0.233	0.767	0.035	13.111
lasso*	CV	0.000	0.683	0.317	0.000	0.678	0.322	0.449	12.900
	BIC	0.000	0.946	0.054	0.000	0.971	0.029	0.919	12.862
$\sigma = 0.5$, $n = 50$									
lasso	CV	0.000	0.174	0.826	0.047	0.138	0.815	0.026	1.530
	BIC	0.000	0.228	0.772	0.017	0.207	0.776	0.045	1.461
lasso*	CV	0.000	0.566	0.434	0.056	0.579	0.365	0.340	1.320
	BIC	0.000	0.802	0.198	0.071	0.758	0.171	0.636	1.275
$\sigma = 0.5$, $n = 100$									
lasso	CV	0.000	0.234	0.766	0.001	0.126	0.873	0.020	1.289
	BIC	0.000	0.370	0.630	0.000	0.176	0.824	0.056	1.260
lasso*	CV	0.000	0.623	0.377	0.002	0.650	0.348	0.416	1.189
	BIC	0.000	0.941	0.059	0.003	0.930	0.067	0.877	1.165
$\sigma = 0.5$, $n = 300$									
lasso	CV	0.000	0.144	0.856	0.000	0.133	0.867	0.011	1.156
	BIC	0.000	0.168	0.832	0.000	0.233	0.767	0.035	1.148
lasso*	CV	0.000	0.685	0.315	0.000	0.675	0.325	0.452	1.132
	BIC	0.000	0.969	0.031	0.000	0.972	0.028	0.943	1.124

the logarithm of real *ex post* average price of residential gas in dollars per therm (LGSP), the cooling degree days per quarter (CDD) and the heating degree days per quarter (HDD).

We first fit the data with the classical multiple-regression model, and the resulting estimated equation is

$$\text{LKWH} = -8.988 + 0.819\text{LY} + 0.154\text{LELP} - 0.159\text{LGSP} + 0.00012\text{CDD} + 0.00042\text{HDD}.$$

The signs of the parameter estimates of variables LY, CDD and HDD meet our expectations. In other words, an increase in real income (LY), the cooling degree days (CDD) or the heating degree days (HDD) yields more demand for heating. However, the variables LELP and LGSP have unexpected signs since the higher electricity price (LELP) and the higher gas price (LGSP) result in more and less electricity consumption respectively. Because these are time series data, the unexpected signs may occur as a result of ignoring the autocorrelation structure.

Table 2. Three models selected by the lasso and modified lasso for the electricity demand study

Variable	*lasso–CV*	*lasso–BIC*	*lasso*–CV*	*lasso*–BIC*
LY	0.117796	0.196611	—	—
LELP	−0.150010	−0.154907	−0.168853	−0.168853
LGSP	−0.035808	−0.057948	—	—
CDD	0.000237	0.000246	0.000226	0.000226
HDD	0.000216	0.000231	0.000222	0.000222
LAG1	0.608868	0.598635	0.627451	0.627451
LAG2	−0.705199	−0.689985	−0.713343	−0.713343
LAG3	0.590666	0.581069	0.604899	0.604899
LAG4	0.253515	0.271175	0.225005	0.225005
χ^2-test	0.039428	0.005632	0.305896	0.305896

Hence, Ramanathan (1989) naturally recommended the regression model with autoregressive errors.

Following Ramanathan's suggestion, we employ the lasso and modified lasso with CV and BIC to shrink jointly both the regression and the autoregression coefficients. The p and q of the candidate models are 5 and 4 respectively, and the maximum autoregressive order 4 is naturally chosen for the quarterly data. Table 2 indicates that the lasso with CV and BIC yields the most complicated model, which is consistent with the simulation findings. This overfitted model also leads to an unexpected sign on the variable LGSP. In contrast, both the modified lasso with CV and the modified lasso with BIC estimators select the same yet simpler model with variable LELP, CDD, HDD and four lags. It is noteworthy that the two important temperature variables CDD and HDD are successfully identified by the modified lasso. In addition the sign of LELP is corrected as compared with the full model regression estimate. To check the adequacy of the model fitting, the χ^2-test statistics for assessing the autocorrelation of residuals (see Box *et al.* (1994), page 314) are computed (see the last row of Table 2). No statistically significant serial correlation is detected in the residuals. In sum, the modified lasso estimator with either CV or BIC produces the same simple, interpretable, yet adequate model fitting to the electricity demand data.

6. Discussion

In REGAR models, we propose the lasso approach to shrink jointly regression and autoregression coefficients. In contrast with the REGAR model, the autoregression with exogenous variables model (Harvey, 1981; Shumway and Stoffer, 2000) provides an alternative approach to take explicitly into account serial dependence via the lagged variables. Specifically, the autoregression with exogenous variables model is

$$y_t = x_t'\beta + \sum_{j=1}^{q} \phi_{t-j} y_{t-j} + \varepsilon_t.$$

To shrink the regression and lagged coefficients simultaneously, we consider the lasso criterion

$$\sum_{t=q+1}^{n_0} \left(y_t - x_t'\beta - \sum_{j=1}^{q} \phi_j y_{t-j} \right)^2 + n \sum_{j=1}^{p} \lambda_j^* |\beta_j| + n \sum_{j=1}^{q} \gamma_j^* |\phi_j|.$$

Analogously to the REGAR model, it can be shown that the lasso approach produces a sparse solution not only for exogenous variables but also for lagged dependent variables. Moreover, the resulting lasso estimator enjoys the oracle property when the tuning parameters satisfy the proper conditions. Extensive simulation studies (which are not presented here) also indicate their satisfactorily finite sample performance.

Finally, we identify three research areas for further study. The first is extending the application of the lasso to both the dynamic regression model (Greene, 2003) and the regression model with seasonal autoregressive errors. The second is to obtain the lasso estimator for the regression model with autoregressive conditional heteroscedastic errors (Gouriéroux, 1997) and the autoregressive and moving average with exogenous variables model (Shumway and Stoffer, 2000). The third is to investigate autoregressive shrinkage and selection by compressing the partial autocorrelations sequentially. We believe that these efforts would further enhance the usefulness of the lasso estimators in real data analysis.

Acknowledgements

The authors are grateful to the Joint Editor, Associate Editor and two referees for their constructive comments and insightful suggestions, which led to a substantial improvement of the manuscript.

Appendix A: Proof of theorem 1

Let $\delta=(u',v')'$, $u=(u_1,\ldots,u_p)'$ and $v=(v_1,\ldots,v_q)'$, and then define

$$\begin{aligned}\kappa_n(\delta)&=Q_n(\theta^0+n^{-1/2}\delta)-Q_n(\theta^0)\\&=L_n(\theta^0+n^{-1/2}\delta)-L_n(\theta^0)+n\lambda_n\sum_{j=1}^{p}(|\beta_j^0+u_jn^{-1/2}|-|\beta_j^0|)+n\gamma_n\sum_{j=1}^{q}(|\phi_j^0+v_jn^{-1/2}|-|\phi_j^0|).\end{aligned}$$

Adopting Knight and Fu's (2000) approach, we have

$$n\lambda_n\sum_{j=1}^{p}(|\beta_j^0+u_jn^{-1/2}|-|\beta_j^0|)\to\lambda_0\sum_{j=1}^{p}\{u_j\ \mathrm{sgn}(\beta_j^0)\ I(\beta_j^0\neq 0)+|u_j|\ I(\beta_j^0=0)\}$$

$$n\gamma_n\sum_{j=1}^{q}(|\phi_j^0+v_jn^{-1/2}|-|\phi_j^0|)\to\gamma_0\sum_{j=1}^{q}\{v_j\ \mathrm{sgn}(\phi_j^0)\ I(\phi_j^0\neq 0)+|v_j|\ I(\phi_j^0=0)\}.$$

Furthermore,

$$\begin{aligned}L_n(\theta^0+n^{-1/2}\delta)-L_n(\theta^0)&=\sum_t\Big[y_t-x_t'(\beta_0+n^{-1/2}u)-\sum_{j=1}^{q}(\phi_j^0+n^{-1/2}v_j)\,\{y_{t-j}-x_{t-j}'(\beta_0+n^{-1/2}u)\}\Big]^2-\sum_t\varepsilon_t^2\\&=\sum_t\Big[e_t-\sum_{j=1}^{q}(\phi_j^0+n^{-1/2}v_j)e_{t-j}-n^{-1/2}u'\Big\{x_t-\sum_{j=1}^{q}(\phi_j^0+n^{-1/2}v_j)x_{t-j}\Big\}\Big]^2-\sum_t\varepsilon_t^2\\&=\sum_t\Big\{\varepsilon_t-n^{-1/2}\sum_{j=1}^{q}v_je_{t-j}-n^{-1/2}u'\Big(x_t-\sum_{j=1}^{q}\phi_j^0x_{t-j}\Big)+n^{-1}u'\sum_{j=1}^{q}v_jx_{t-j}\Big\}^2-\sum_t\varepsilon_t^2\\&=R_1+R_2+R_3+R_4+R_5,\end{aligned}$$

where

$$R_1=-2n^{-1/2}\sum_t\Big(\varepsilon_t\sum_{j=1}^{q}v_je_{t-j}\Big)-2n^{-1/2}u'\sum_t\varepsilon_t\Big(x_t-\sum_{j=1}^{q}\phi_j^0x_{t-j}\Big),$$

$$R_2=2n^{-1}u'\sum_t\Big\{\Big(\sum_{j=1}^{q}v_je_{t-j}\Big)\Big(x_t-\sum_{j=1}^{q}\phi_j^0x_{t-j}\Big)\Big\}$$

$$R_3=n^{-1}\sum_t\Big(\sum_{j=1}^{q}v_je_{t-j}\Big)^2+n^{-1}u'\sum_t\Big(x_t-\sum_{j=1}^{q}\phi_j^0x_{t-j}\Big)\Big(x_t-\sum_{j=1}^{q}\phi_j^0x_{t-j}\Big)'u,$$

$$R_4 = 2n^{-1} \sum_t \Big(u' \sum_{j=1}^{q} v_j x_{t-j}\Big)\Big\{\varepsilon_t - n^{-1/2} \sum_{j=1}^{q} v_j e_{t-j} - n^{-1/2} u'\Big(x_t - \sum_{j=1}^{q} \phi_j^0 x_{t-j}\Big)\Big\},$$

$$R_5 = n^{-2} u' \sum_t \Big(\sum_{j=1}^{q} v_j x_{t-j}\Big)\Big(\sum_{j=1}^{q} v_j x_{t-j}\Big)' u.$$

Employing the martingale central limit theorem and the ergodic theorem, we can show that $R_1 \rightarrow_{\mathrm{d}} -2\delta' w$, $R_2 = o_p(1)$, $R_3 \rightarrow_{\mathrm{p}} \delta' \Sigma \delta$, $R_4 = o_p(1)$ and $R_5 = o_p(1)$. Consequently,

$$L_n(\theta^0 + n^{-1/2}\delta) - L_n(\theta^0) \underset{\mathrm{d}}{\rightarrow} -2\delta' w + \delta' \Sigma \delta.$$

To show that

$$\arg\min\{\kappa_n(\delta)\} \underset{\mathrm{d}}{\rightarrow} \arg\min\{\kappa(\delta)\},$$

we must prove that $\arg\min\{\kappa_n(\delta)\} = O_p(1)$. Note that

$$\begin{aligned}
\kappa_n(\delta) \geqslant & \sum_t \Big[\Big\{\varepsilon_t - n^{-1/2} \sum_{j=1}^{q} v_j e_{t-j} - n^{-1/2} u'\Big(x_t - \sum_{j=1}^{q} \phi_j^0 x_{t-j}\Big) + n^{-1} u' \sum_{j=1}^{q} v_j x_{t-j}\Big\}^2 - \varepsilon_t^2\Big] \\
& - n\lambda_n \sum_{j=1}^{p} |u_j n^{-1/2}| - n\gamma_n \sum_{j=1}^{q} |v_j n^{-1/2}| \\
\geqslant & \sum_t \Big[\Big\{\varepsilon_t - n^{-1/2} \sum_{j=1}^{q} v_j e_{t-j} - n^{-1/2} u'\Big(x_t - \sum_{j=1}^{q} \phi_j^0 x_{t-j}\Big)\Big\}^2 - \varepsilon_t^2\Big] \\
& - (\lambda_0 + \varepsilon_0) \sum_{j=1}^{p} |u_j| - (\gamma_0 + \varepsilon_0) \sum_{j=1}^{q} |v_j| + \xi_n(\delta) \\
\doteq & \tilde{\kappa}_n(\delta),
\end{aligned}$$

where $\varepsilon_0 > 0$ is some positive constant. In addition, $\kappa_n(0) = \tilde{\kappa}_n(0)$ and $\xi_n(\delta) = o_p(1)$. Moreover, for all δ and sufficiently large n, the quadratic terms in $\tilde{\kappa}_n(\delta)$ grow faster than the $|u_j|$ and $|v_j|$. As a result, $\arg\min\{\tilde{\kappa}_n(\delta)\} = O_p(1)$ and $\arg\min\{\kappa_n(\delta)\} = O_p(1)$. Because $\arg\min\{\kappa(\delta)\}$ is unique with probability 1, the proof is complete.

Appendix B: Proof of lemma 1

Let $\alpha_n = n^{-1/2} + a_n$ and $\{\theta^0 + \alpha_n \delta : \|\delta\| \leqslant d\}$ be the ball around θ^0. Then, for $\|\delta\| = d$, we have

$$\begin{aligned}
D_n(\delta) & \doteq Q_n^*(\theta^0 + \alpha_n\delta) - Q_n^*(\theta^0) \\
& \geqslant L_n(\theta^0 + \alpha_n\delta) - L_n(\theta^0) + n \sum_{j \in \mathcal{S}_1} \lambda_j (|\beta_j^0 + \alpha_n u_j| - |\beta_j^0|) + n \sum_{j \in \mathcal{S}_2} \gamma_j (|\phi_j^0 + \alpha_n v_j| - |\phi_j^0|) \\
& \geqslant L_n(\theta^0 + \alpha_n\delta) - L_n(\theta^0) - n\alpha_n \sum_{j \in \mathcal{S}_1} \lambda_j |u_j| - n\alpha_n \sum_{j \in \mathcal{S}_2} \gamma_j |v_j| \\
& \geqslant L_n(\theta^0 + \alpha_n\delta) - L_n(\theta^0) - n\alpha_n^2 p_0 d - n\alpha_n^2 q_0 d \\
& = L_n(\theta^0 + \alpha_n\delta) - L_n(\theta^0) - n\alpha_n^2 (p_0 + q_0) d. \qquad (10)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
L_n(\theta^0 + \alpha_n\delta) - L_n(\theta^0) & = \sum_t \Big\{\varepsilon_t - \alpha_n \sum_{j=1}^{q} v_j e_{t-j} - \alpha_n u'\Big(x_t - \sum_{j=1}^{q} \phi_j^0 x_{t-j}\Big) + \alpha_n^2 u' \sum_{j=1}^{q} v_j x_{t-j}\Big\}^2 - \sum_t \varepsilon_t^2 \\
& = A_1 + A_2 + A_3 + A_4 + A_5, \qquad (11)
\end{aligned}$$

where

$$A_1 = \alpha_n^2 \sum_t \Big\{\Big(\sum_{j=1}^{q} v_j e_{t-j}\Big)^2 + u'\Big(x_t - \sum_{j=1}^{q} \phi_j^0 x_{t-j}\Big)\Big(x_t - \sum_{j=1}^{q} \phi_j^0 x_{t-j}\Big)' u\Big\},$$

$$A_2 = -2\alpha_n \sum_t \varepsilon_t \Big\{\sum_{j=1}^{q} v_j e_{t-j} + u'\Big(x_t - \sum_{j=1}^{q} \phi_j^0 x_{t-j}\Big)\Big\},$$

$$A_3 = 2\alpha_n^2 \sum_t \Big(\sum_{j=1}^{q} v_j e_{t-j}\Big) u'\Big(x_t - \sum_{j=1}^{q} \phi_j^0 x_{t-j}\Big),$$

$$A_4 = \alpha_n^3 \sum_t \Big(u' \sum_{j=1}^{q} v_j x_{t-j}\Big)\Big\{\alpha_n u' \sum_{j=1}^{q} v_j x_{t-j} - 2u'\Big(x_t - \sum_{j=1}^{q} \phi_j^0 x_{t-j}\Big) - 2\sum_{j=1}^{q} v_j e_{t-j}\Big\},$$

$$A_5 = 2\alpha_n^2 \sum_t \varepsilon_t \Big(u' \sum_{j=1}^{q} v_j x_{t-j}\Big).$$

Moreover, we have

$$A_1 = n\alpha_n^2 \{\delta'\Sigma\delta + o_p(1)\},$$

$$A_2 = \delta' O_p(n\alpha_n^2),$$

$$A_3 = n\alpha_n^2\, o_p(1) = o_p(n\alpha_n^2),$$

$$A_4 = n\alpha_n^3\, O_p(1) = n\alpha_n^2\, o_p(1) = o_p(n\alpha_n^2),$$

$$A_5 = n\alpha_n^2\, o_p(1) = o_p(n\alpha_n^2),$$

because A_1 dominates the rest of the four terms in equation (11) and also $n\alpha_n^2(p_0+q_0)d$ in equation (10). Hence, for any given $\varepsilon > 0$, there is a large constant d such that

$$P[\inf_{\|\delta\|=d}\{Q_n^*(\theta^0 + \alpha_n\delta)\} > Q_n^*(\theta^0)] \geqslant 1-\varepsilon.$$

This implies that, with probability at least $1-\varepsilon$, there is a local minimizer in the ball $\{\theta^0 + \alpha_n\delta : \|\delta\| \leqslant d\}$ (Fan and Li, 2001). Consequently, there is a local minimizer of $Q_n^*(\theta)$ such that $\|\hat{\theta}^* - \theta^0\| = O_p(\alpha_n)$. This completes the proof.

Appendix C: Proof of theorem 2

The proof of theorem 2 follows from the fact that the local minimizer $\hat{\theta}^*$ must satisfy the equation

$$\begin{aligned}\frac{\partial Q_n^*(\hat{\theta}^*)}{\partial \beta_j} &= \frac{\partial L_n(\hat{\theta}^*)}{\partial \beta_j} - n\lambda_j\,\mathrm{sgn}(\hat{\beta}_j^*)\\ &= \frac{\partial L_n(\theta^0)}{\partial \beta_j} + n\Sigma_j(\hat{\theta}^* - \theta^0)\{1+o_p(1)\} - n\lambda_j\,\mathrm{sgn}(\hat{\beta}_j^*),\end{aligned} \tag{12}$$

where Σ_j denotes the jth row of Σ and $j \in \mathcal{S}_1^c$. Employing the central limit theorem, the first term in equation (12) is of order $O_p(n^{1/2})$. Furthermore, the condition in theorem 2 implies that its second term is also of order $O_p(n^{1/2})$. Both are dominated by $n\lambda_j$ since $b_n\sqrt{n} \to \infty$. Therefore, the sign of equation (12) is dominated by the sign of $\hat{\beta}_j^*$. Consequently, we must have $\hat{\beta}_j^* = 0$ in probability. Analogously, we can show that $P(\hat{\phi}^*_{\mathcal{S}_2^c} = 0) \to 1$. This completes the proof.

Appendix D: Proof of theorem 3

Applying lemma 1 and theorem 2, we have $P(\hat{\theta}_2^* = 0) \to 1$. Hence, the minimizer of $Q_n^*(\theta)$ is the same as that of $Q_n^*(\theta_1)$ with probability tending to 1. This implies that the lasso estimator $\hat{\theta}_1^*$ satisfies the equation

$$\frac{\partial Q_n^*(\theta_1)}{\partial \theta_1}\Big|_{\theta_1 = \hat{\theta}_1^*} = 0. \tag{13}$$

According to lemma 1, $\hat{\theta}_1^*$ is a $\sqrt{n}$-consistent estimator. Thus, the Taylor series expansion of equation (13) yields

$$0 = \frac{1}{\sqrt{n}} \frac{\partial L_n(\hat{\theta}_1^*)}{\partial \theta_1} + P(\hat{\theta}_1^*)\sqrt{n}$$
$$= \frac{1}{\sqrt{n}} \frac{\partial L_n(\theta_1^0)}{\partial \theta_1} + P(\theta_1^0)\sqrt{n} + \Sigma_0 \sqrt{n}(\hat{\theta}_1^* - \theta_1^0) + o_p(1),$$

where P is the first-order derivative of the penalty function

$$\sum_{j \in \mathcal{S}_1} \lambda_j |\beta_j| + \sum_{j \in \mathcal{S}_2} \gamma_j |\phi_j|,$$

and $P(\hat{\theta}_1^*) = P(\theta_1^0)$ as n is sufficiently large. Furthermore, it can be easily shown that $P(\theta_1^0)\sqrt{n} = o_p(1)$, which implies that

$$(\hat{\theta}_1^* - \theta_1^0)\sqrt{n} = \frac{\Sigma_0^{-1}}{\sqrt{n}} \frac{\partial L_n(\theta_1^0)}{\partial \theta_1} + o_p(1)$$
$$\xrightarrow{\mathrm{d}} N(0, \sigma^2 \Sigma_0^{-1}).$$

This completes the proof.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.
Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994) *Time Series Analysis—Forecasting and Control*, 3rd edn. New York: Prentice Hall.
Breiman, L. (1996) Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350–2383.
Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*, 2nd edn. New York: Springer.
Cai, J., Fan, J., Li, R. and Zhou, H. (2005) Model selection for multivariate failure time data. *Biometrika*, **92**, 303–316.
Choi, B. S. (1992) *ARMA Model Identification*. New York: Springer.
Cochrane, D. and Orcutt, G. H. (1949) Application of least squares regression to relationships containing auto-correlated error terms. *J. Am. Statist. Ass.*, **44**, 32–61.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Ann. Statist.*, **32**, 407–489.
Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.
Fan, J. and Li, R. (2002) Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, **30**, 74–99.
Fan, J. and Peng, H. (2004) On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.*, **32**, 928–961.
Fu, W. J. (1998) Penalized regression: the bridge versus the lasso. *J. Comput. Graph. Statist.*, **7**, 397–416.
Gouriéroux, C. (1997) *ARCH Models and Financial Applications*. New York: Springer.
Greene, W. H. (2003) *Econometric Analysis*, 5th edn. New York: Prentice Hall.
Hamilton, J. D. (1994) *Time Series Analysis*. Princeton: Princeton University Press.
Harvey, A. C. (1981) *The Econometric Analysis of Time Series*. New York: Wiley.
Hurvich, C. M. and Tsai, C. L. (1990) The impact of model selection on inference in linear regression. *Am. Statistn*, **44**, 214–217.
Knight, K. and Fu, W. (2000) Asymptotics for lasso-type estimators. *Ann. Statist.*, **28**, 1356–1378.
Leng, C., Lin, Y. and Wahba, G. (2006) A note on the lasso and related procedures in model selection. *Statist. Sin.*, **16**, 1273–1284.
McQuarrie, D. R. and Tsai, C. L. (1998) *Regression and Time Series Model Selection*. Singapore: World Scientific Press.
Ramanathan, R. (1989) *Introductory Econometrics with Applications*. Orlando: Harcourt Brace Jovanovich.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Shao, J. (1997) An asymptotic theory for linear model selection (with discussion). *Statist. Sin.*, **7**, 221–264.
Shi, P. and Tsai, C. L. (2004) A joint regression variable and autoregressive order selection criterion. *J. Time Ser.*, **25**, 923–941.
Shumway, R. H. and Stoffer, D. S. (2000) *Time Series Analysis and Its Application*. New York: Springer.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tsay, R. S. (1984) Regression models with time series errors. *J. Am. Statist. Ass.*, **79**, 118–124.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.
Zou, H., Hastie, T. and Tibshirani, R. (2004) On the 'degrees of freedom' of lasso. *Technical Report*. Statistics Department, Stanford University, Stanford.