Aust. N. Z. J. Stat. 2014 doi: 10.1111/j.1467-842X.XXX

Bridge estimation for linear regression models with mixing properties

TAEWOOK LEE*, CHEOLWOO PARK†, AND YOUNG JOO YOON*‡

Hankuk University of Foreign Studies, University of Georgia, and Daejeon University

Summary

Penalized regression methods have for quite some time been a popular choice for addressing challenges in high dimensional data analysis. Despite their popularity, their application to time series data has been limited. This paper concerns bridge penalized methods in a linear regression time series model. We first prove consistency, sparsity and asymptotic normality of bridge estimators under a general mixing model. Next, as a special case of mixing errors, we consider bridge regression with ARMA error models and develop a computational algorithm that can simultaneously select important predictors and the orders of ARMA models. Simulated and real data examples demonstrate the effective performance of the proposed algorithm and the improvement over ordinary bridge regression.

Key words: ARMA models; asymptotic normality; bridge regression; consistency; mixing processes; variable selection.

1. Introduction

A linear regression model is a conventional technique for modeling the relationship between a response and various explanatory variables. In its applications, practitioners and researchers commonly assume that the errors are independent and identically distributed. Nevertheless serial correlation is frequently present in the observations in which case time series errors are more appropriate. In the literature, Tsay (1984), Glasbey (1988), and Shi & Tsay (2004) study time series regression models of the autoregressive and moving average (ARMA) form. Tsay (1984) verifies the convergence properties of the least squares estimators of the linear regression parameters, and Glasbey (1988) analyzes some real data examples. Later, Shi & Tsay (2004) apply a residual likelihood approach to obtain a residual information criterion (RIC) that can jointly select regression variables and autoregressive orders.

In linear regression models, penalized regression methods have been used increasingly in recent years by both the statistics and machine learning communities due to the prevalence of the high dimensional nature of many of the data sets currently being analyzed. Regression data of high dimensionality often contain noisy and/or correlated variables. Penalized regression methods are applicable to such cases; various types of methods have been developed; these are typically used for the purpose of variable selection. For instance

*Department of Statistics, Hankuk University of Foreign Studies, Yongin, 449-791, Korea. e-mail: twlee@hufs.ac.kr

†Department of Statistics, University of Georgia, Athens, GA 30602, USA. e-mail: cpark@uga.edu

‡Department of Business Information Statistics, Daejeon University, Daejeon, 300-716, Korea. e-mail: yoonyj74@gmail.com

‡*Author to whom correspondence should be addressed.

Acknowledgment. The authors would like to thank an Associate Editor and two referees for their helpful comments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0011222).

Tibshirani (1996) proposes the least absolute shrinkage and selection operator (LASSO) using an L_1 penalty. This operator selects variables and estimates regression parameters simultaneously. The ridge regression method (Hoerl & Kennard 1970) utilizes an L_2 penalty, and although it does not select variables, shows superior empirical performance to the LASSO in the presence of multicollinearity, i.e. when high correlations among predictors exist. The generalized form that includes these two approaches is bridge regression (Frank & Friedman 1993), which utilizes the L_γ ($\gamma > 0$) penalty. The recent explosive growth of penalized regression methods includes smoothly clipped absolute deviation (Fan & Li 2001), the elastic net method (Zou & Hastie 2005), the fused LASSO (Tibshirani, Saunders, Rosset *et al.* 2005), the adaptive LASSO (Zou 2006; Zhang & Lu 2007), the Bayesian LASSO (Park & Casella 2008), and adaptive elastic net methods (Zou & Zhang 2009). While one of the methods is not singled out as being preferable, we focus on bridge regression due to its general form. It fits naturally any situation in which it is necessary to select important predictors or in which there is a multicollinearity issue (Park & Yoon 2011). Bridge regression will be elaborated upon further in Section 2.

As mentioned above, linear regression models with time series errors can serve as a useful tool for analyzing serially correlated data, but there have been few attempts to account for variable selection and multicollinearity in regression analysis of time series. Wang, Li & Tsai (2007) consider linear regression models with autoregressive errors (the REGAR model) and develop a penalized estimation method. They suggest algorithms to obtain LASSO and adaptive LASSO procedures for the REGAR model. Hsu, Hung & Chang (2008) also propose a LASSO-based selection method for vector autoregressive processes. Gelper & Croux (2009) use the least angle regression (LARS) method to identify the most informative predictors in linear regression time series models. In a recent paper Alquier & Doukhan (2011) apply the LASSO and other L_1-penalized methods to dependent observations. Yoon, Park & Lee (2012) propose a computational algorithm with three penalty functions for the REGAR model. This algorithm can select a relevant set of variables and also the order of autoregressive error terms simultaneously.

The main objective of the current paper is to develop bridge regression for a linear model with mixing properties. Mixing processes form a rich family of dependent processes and have been a functional tool for deriving the asymptotic properties of statistical inferences for various dependent processes, including classical stationary ARMA models, models involving nonparametric statistics (Bosq 1998) and Markov processes (Doukhan 1994). We first study the asymptotic properties of bridge estimators such as consistency, sparsity and asymptotic normality under the assumption of mixing errors. One could use the bridge regression method without specifying the underlying serial correlation structure in the model. However, if the underlying serial correlation structure can be determined, it could be included in the bridge estimation model so as to improve the efficiency of the estimators. In this paper, we take ARMA models as a special case of mixing errors and propose an extended bridge regression algorithm that simultaneously determines important predictors and the order of an ARMA model. The algorithm utilizes local quadratic approximation to circumvent the nonconvex optimization problem and adaptively selects tuning parameters in the penalty function to produce flexible solutions in various settings. This algorithm is a generalization of the work in Wang, Li & Tsai (2007) and Yoon, Park & Lee (2012) since ARMA models include AR models.

The remainder of the paper is organized as follows. In Section 2, we present the asymptotic properties of bridge estimators under mixing errors. In Section 3, we first introduce a general computational algorithm for bridge estimators and later improve the algorithm when the errors are modeled by ARMA. A brief discussion of linear regression models with ARMA-GARCH errors is also included. In Sections 4 and 5, we investigate the finite sample performance of bridge estimators using simulated and real data examples, respectively. The Appendix provides proofs of the theorems stated in Section 2.

2. Bridge estimators with mixing error

We consider a linear regression model with p predictors and T observations:

$$Y_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + e_t, \quad t = 1, \dots, T, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top$, $\boldsymbol{x}_t = (x_{t1}, \cdots x_{tp})^\top$ and the error term e_t is a mixing process. Refer to assumption (**A**5) below for the definition of a mixing process. Note that the classical ARMA model is an example of mixing process, and therefore the REGAR model considered in Wang, Li & Tsai (2007) and Yoon, Park & Lee (2012) is a particular example of model (1). We assume that the Y_t are centered and that the covariates $\boldsymbol{x}_t$ are standardized, that is,

$$\sum_{t=1}^{T} Y_t = 0, \quad \sum_{t=1}^{T} x_{tj} = 0 \text{ and } \frac{1}{T}\sum_{t=1}^{T} x_{tj}^2 = 1, \quad j = 1, \dots, p.$$

The coefficient vector $\boldsymbol{\beta}$ can be estimated by minimizing the penalized least squared objective function:

$$L_T(\boldsymbol{\beta}) = \sum_{t=1}^{T} (Y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma, \tag{2}$$

where λ is a penalty parameter. For any given $\gamma > 0$, the value $\hat{\boldsymbol{\beta}}_T$ that minimizes (2) is termed a bridge estimator. It selects relevant predictors when $0 < \gamma \leq 1$, and shrinks the coefficients when $\gamma > 1$.

Since the seminal work of Frank & Friedman (1993), much research has been done on bridge regression when the error terms are assumed to be independent. Fu (1998) carefully examines the structure of bridge estimators and proposes a general algorithm to solve for $\gamma \geq 1$. Knight & Fu (2000) investigate the asymptotic properties of bridge estimators with $\gamma > 0$ when the number of covariates p is fixed. In the case in which p increases along with the sample size, Huang, Horowitz & Ma (2008) study the asymptotic properties of bridge estimators in sparse, high dimensional, linear regression models. Huang, Ma, Xie *et al.* (2009) propose a bridge approach with $0 < \gamma < 1$ that identifies relevant groups of variables, and then selects individual variables within those groups. Recently, Park & Yoon (2011) investigate an adaptive choice of the penalty order γ in bridge regression, and Yoon, Park & Lee (2012) extend this idea to the REGAR model. In the current paper, we further extend it to a broader class of error models and study bridge regression under mixing errors.

In this section, we establish the asymptotic properties of bridge estimators under model (1). The true parameter is denoted by $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^\top, \boldsymbol{\beta}_{20}^\top)^\top$, where $\boldsymbol{\beta}_{10}$ is a $k \times 1$ vector and $\boldsymbol{\beta}_{20}$ is an $m \times 1$ vector. We suppose that $\boldsymbol{\beta}_{10} \neq \mathbf{0}$ and $\boldsymbol{\beta}_{20} = \mathbf{0}$. We write

$\boldsymbol{x}_t = (\boldsymbol{x}_{1,t}^\top, \boldsymbol{x}_{2,t}^\top)^\top$, where $\boldsymbol{x}_{1,t}$ consists of the first k covariates (corresponding to nonzero coefficients), and $\boldsymbol{x}_{2,t}$ consists of the remaining m covariates (those with zero coefficients). Let $\mathbf{X}_T^\top = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ and $\mathbf{X}_{1T}^\top = (\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{1,T})$. Also, let $\Sigma_T = T^{-1}\mathbf{X}_T^\top \mathbf{X}_T$ and $\Sigma_{1T} = T^{-1}\mathbf{X}_{1T}^\top \mathbf{X}_{1T}$. Throughout the paper, the symbol $||\cdot||$ is used to denote the operator norm of matrix and the L_2 norm of any vector $\boldsymbol{u} \in R^p$ is denoted by $\|\boldsymbol{u}\|$; $\|\boldsymbol{u}\| = [\sum_{j=1}^p u_j^2]^{1/2}$.

We introduce the following regularity conditions to build the asymptotic properties of bridge estimators:

(**A**1) The sequence $\{e_t\}$ is ergodic and strictly stationary with mean zero and variance σ_e^2, with $0 < \sigma_e^2 < \infty$. Furthermore,

$$\sum_{k=0}^{\infty} |\gamma_e(k)| < \infty,$$

where $\gamma_e(k) = E(e_i e_{i+k})$.

(**A**2) The sequence $\{\boldsymbol{x}_t\}$ is ergodic and strictly stationary with mean zero and $E\|\boldsymbol{x}_t\|^2 = \sigma_x^2 < \infty$. Furthermore, the matrix $\Sigma = E\left\{\boldsymbol{x}_t \boldsymbol{x}_t^\top\right\}$ is positive definite.

(**A**3) $\{\boldsymbol{x}_t\}$ and $\{e_t\}$ are independent.

(**A**4) (a) $\lambda T^{-1/2} \to 0$; (b) $\lambda T^{-\gamma/2} \to \infty$ for $0 < \gamma < 1$.

Remark 2.1. A stationary sequence is said to be ergodic if it satisfies the strong number of large numbers. Refer to Billingsley (1995) and Francq & Zakoïan (2010) for a more general definition of the ergodicity of stationary and nonstationary sequences.

Remark 2.2. Our main focus is on extending the REGAR model to penalized regression with mixing errors including ARMA errors as a special case. One of the standard models satisfying the condition (**A**3) is the REGAR model considered in Wang, Li & Tsai (2007). This follows from condition (a) on page 66 in their paper. Condition (**A**3) however rules out some important models, for example those in which lagged versions of Y_t are among the covariates or instances in which x_t is Granger-caused by Y_t. Extensions which remove the need for (**A**3) are envisaged for future work but are beyond the scope of the current paper.

Remark 2.3. We note that the assumption of strict stationarity in (**A**1)-(**A**2) is strong and difficult to check. Nevertheless, this assumption is needed here in order to use the results from Peligrad (1986), in which strict stationarity is one of the assumptions to achieve asymptotic normality for a strong mixing process. We refer the reader to Brockwell & Davis (2006) and Francq & Zakoïan (2010) for more details.

We now briefly define mixing coefficients and strong mixing processes in terms of the sequences $\{\boldsymbol{x}_t\}$ and $\{e_t\}$. Let $\mathcal{F}_n^m$ be the $\sigma-$algebra generated by the random variables $\boldsymbol{x}_n, \ldots, \boldsymbol{x}_{n+m}$ for a sequence of random variables $\{\boldsymbol{x}_t\}$ on a probability space $(\Omega, \mathcal{B}, P)$. We define the mixing coefficient of $\{\boldsymbol{x}_t\}$ to be:

$$\alpha_x(m) = \sup\{|P(E \cap F) - P(E)P(F)|; E \in \mathcal{F}_{-\infty}^n, F \in \mathcal{F}_{n+m}^\infty\}.$$

The process $\{\boldsymbol{x}_t\}$ is said to be a strong mixing process if $\alpha_x(m)$ tends to zero as m increases to infinity. See Bradley (1986) and Mokkadem (1988) for more details.

We require one more regularity condition:

(**A**5) The sequences $\{\boldsymbol{x}_t\}$ and $\{e_t\}$ are strong mixing with mixing coefficients $\alpha_x(m)$ and $\alpha_e(m)$ satisfying

$$\sum_{m=0}^{\infty} \alpha_x^{\delta/(2+\delta)}(m) < \infty \quad \text{and} \quad \sum_{m=0}^{\infty} \alpha_e^{\delta/(2+\delta)}(m) < \infty,$$

and

$$E|\boldsymbol{u}^\top \boldsymbol{x}_t|^{2+\delta} < \infty \quad \text{and} \quad E|e_t|^{2+\delta} < \infty,$$

for any unit vector $\boldsymbol{u}$ and some $\delta > 0$.

The following theorems state that bridge estimators are $\sqrt{T}$-consistent and have the oracle property, that is, they can perform as well as if the correct submodel were known. Proofs are given in the Appendix.

Theorem 1. (Consistency) Suppose that $\gamma > 0$ and that the conditions (**A**1)-(**A**3) and (**A**4)(a) hold. Then

$$\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\| = O_P\left(T^{-1/2}\right).$$

Theorem 2. (Oracle property) Let $\hat{\boldsymbol{\beta}}_T = (\hat{\boldsymbol{\beta}}_{1T}^\top, \hat{\boldsymbol{\beta}}_{2T}^\top)^\top$, where $\hat{\boldsymbol{\beta}}_{1T}$ and $\hat{\boldsymbol{\beta}}_{2T}$ are estimators of $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\beta}_{20}$, respectively. Suppose that $0 < \gamma < 1$ and that the conditions (**A**1)-(**A**5) are satisfied. Then the following results hold:

(i) (Sparsity) $\hat{\boldsymbol{\beta}}_{2T} = \mathbf{0}$ with probability converging to 1 as T tends to infinity.

(ii) (Asymptotic normality) Define $\Sigma_1 = E[\boldsymbol{x}_{1,1}\boldsymbol{x}_{1,1}^\top]$. Assume that

$$\boldsymbol{K} = \sigma_e^2 \Sigma_1^{-1} + 2\sum_{j=1}^{\infty} \gamma_e(j) \Sigma_1^{-1} E[\boldsymbol{x}_{1,1}\boldsymbol{x}_{1,1+j}^\top] \Sigma_1^{-1}$$

is a positive definite matrix. Then

$$T^{1/2}(\hat{\boldsymbol{\beta}}_{1T} - \boldsymbol{\beta}_{10}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{K}),$$

where $\xrightarrow{D}$ indicates the convergence in distribution.

3. Computational algorithms

In Section 3.1, we introduce a general bridge regression algorithm when the error time series model is not specified. This general algorithm can be improved in terms of prediction accuracy if the error model is prespecified. In Section 3.2, we develop an alternative algorithm for ARMA error models. A brief discussion of linear regression models with ARMA-GARCH errors is also provided in Section 3.3.

3.1. General algorithm

In this subsection, we introduce an algorithm for bridge regression when e_t in (1) is assumed to be a mixing process. This computational algorithm is introduced in Park & Yoon (2011) with $\gamma > 0$. Here we restrict our attention to $0 < \gamma \leq 1$ because variable selection is our main focus. If $0 < \gamma < 1$, the minimization problem in (2) is not convex, so we apply the local quadratic approximation proposed by Fan & Li (2001) to circumvent this issue. With a nonzero β^*_{0j} that is close to β_j, the penalty term can be locally approximated at $\boldsymbol{\beta}^*_0 = (\beta^*_{01}, \cdots, \beta^*_{0p})^\top$ by a quadratic function:

$$|\beta_j|^\gamma \approx |\beta^*_{0j}|^\gamma + \frac{\gamma}{2}\frac{|\beta^*_{0j}|^{\gamma-1}}{|\beta^*_{0j}|}(\beta_j^2 - \beta^{*2}_{0j}).$$

If β^*_{0j} is close to 0 we set $\beta_j = 0$. The proposed algorithm for bridge estimators can be summarized as follows. For given λ and γ, we obtain $\hat{\boldsymbol{\beta}}^{(l)}$ at the lth iteration from

$$\hat{\boldsymbol{\beta}}^{(l)} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{t=1}^{T} (Y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta})^2 + \frac{\lambda\gamma}{2} \sum_{j=1}^{p} |\hat{\beta}_j^{(l-1)}|^{\gamma-2}\beta_j^2 \right\}.$$

The iteration continues until $||\hat{\boldsymbol{\beta}}^{(l)} - \hat{\boldsymbol{\beta}}^{(l-1)}|| < \eta_0$ where η_0 is a small positive constant, e.g. 10^{-3}. We use the ordinary least squares estimator, with no consideration of the correlation structure, as an initial estimator for the regression coefficients:

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

One can apply a ridge estimator when $(\boldsymbol{X}^\top \boldsymbol{X})$ is close to singular.

We find the optimal combination of λ and γ by grid search using the following criterion (Zou, Hastie & Tibshirani 2007):

$$\mathrm{BIC} = \log(\hat{\sigma}^2) + \hat{\mathrm{df}} \log(T)/T$$

where $\hat{\sigma}^2$ is the sum of squared residuals for the error model divided by T and $\hat{\mathrm{df}}$ is the number of nonzero coefficients. We choose BIC over cross-validation (CV) in accordance with the discussions in Shao (1997) and Wang, Li & Tsai (2007): when the true model is a candidate model and has finite dimension, then BIC is the preferred choice over CV.

Remark 3.1. When the structure of the data is known beforehand, it may be advantageous to use a fixed γ. However, when the data structure is unknown, this approach has disadvantages and it would be better to estimate γ directly from the data. For more information about bridge estimators with an adaptive choice of γ, we refer the reader to Park & Yoon (2011).

Remark 3.2. Since the optimal value of (λ, γ) is found by means of a grid search, the suggested optimization procedure will be time-consuming for large data sets with high dimension. The speed of the procedure might be improved by making use of an efficient algorithm such as the `cleversearch` function in R or pathwise optimization solutions considered in Breheny & Huang (2009) and Hirose, Tateishi & Konishi (2011).

3.2. An Algorithm for ARMA error models

We can further improve the algorithm introduced in Section 3.1 by incorporating the information about the dependence structure of a given time series when this is known in advance. Here we explain this procedure in the case of a linear regression time series model with ARMA errors. The proposed algorithm produces a sparse model and determines the order of the associated ARMA error model. Let $\{e_t\}$ be the sequence of random variables following ARMA(P, Q):

$$e_t = \sum_{i=1}^{P} a_i e_{t-i} + \epsilon_t - \sum_{j=1}^{Q} b_j \epsilon_{t-j}, \tag{3}$$

where $\epsilon_t,\ t = 1, \ldots, T$, is a sequence of independent random variables with zero mean and variance σ^2. The ARMA model parameter vector $(a_1, \ldots, a_P, b_1, \ldots, b_Q)^\top$ is denoted by $\boldsymbol{\theta}$. The true value is denoted by $\boldsymbol{\theta}_0 = (a_{01}, \ldots, a_{0P}, b_{01}, \ldots, b_{0Q})^\top$. The parameter space of $\boldsymbol{\theta}$ is $\Theta \subset R^{P+Q}$. Here P and Q are unknown. We impose the standard constraints for the identifiabilty, invertibility and stationarity of ARMA models as follows. Let $\mathcal{A}_{\boldsymbol{\theta}}(z) = \sum_{i=1}^{P} a_i z^i$ and $\mathcal{B}_{\boldsymbol{\theta}}(z) = 1 - \sum_{j=1}^{Q} b_j z^j$. We assume that for all $\boldsymbol{\theta} \in \Theta$, $\mathcal{A}_{\boldsymbol{\theta}}(z)\mathcal{B}_{\boldsymbol{\theta}}(z) = 0$ implies $|z| > 1$. We also assume that $\mathcal{A}_{\boldsymbol{\theta}_0}(z)$ and $\mathcal{B}_{\boldsymbol{\theta}_0}(z)$ have no common roots, and that neither a_{0P} nor b_{0Q} is equal to 0.

Using the ordinary bridge estimator obtained in Section 3.1 as an initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$, the updated algorithm proceeds from $l = 1$ as follows:

(Step 1) Using $\hat{\boldsymbol{\beta}}^{(l-1)}$, define the residuals

$$\tilde{e}_t^{(l-1)} = Y_t - \boldsymbol{x}_t^\top \hat{\boldsymbol{\beta}}^{(l-1)}, \quad t = 1, \ldots, T.$$

(Step 2) Using the residuals $\{\tilde{e}_t^{(l-1)}\}$ and BIC, obtain the estimated orders $P^{(l-1)}$ and $Q^{(l-1)}$ for P and Q, respectively. We then have the conditional least squares estimator (CLSE) for the parameter θ, denoted by $\hat{\theta}^{(l-1)} = \left(\hat{a}_1^{(l-1)}, \ldots, \hat{a}_{P^{(l-1)}}^{(l-1)}, \hat{b}_1^{(l-1)}, \ldots, \hat{b}_{Q^{(l-1)}}^{(l-1)}\right)^\top$. Using the residuals $\{\tilde{e}_t^{(l-1)}\}$ and the CLSE $\hat{\theta}^{(l-1)}$, the ARMA residuals are defined recursively by

$$\tilde{\epsilon}_t\left(\hat{\theta}^{(l-1)}\right) = \tilde{e}_t^{(l-1)} - \sum_{i=1}^{P^{(l-1)}} \hat{a}_i^{(l-1)} \tilde{e}_{t-i}^{(l-1)} + \sum_{j=1}^{Q^{(l-1)}} \hat{b}_j^{(l-1)} \tilde{\epsilon}_{t-j}\left(\hat{\theta}^{(l-1)}\right), \ t = 1, \ldots, T.$$

(Step 3) Obtain $\hat{\boldsymbol{\beta}}^{(l)}$ in the following way:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(l)} &= \arg\min_{\beta} \Bigg\{ \sum_{t=1}^{T} \Bigg(Y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta} - \sum_{i=1}^{P^{(l-1)}} \hat{a}_i^{(l-1)} (Y_{t-i} - \boldsymbol{x}_{t-i}^\top \boldsymbol{\beta}) \\ &\quad + \sum_{j=1}^{Q^{(l-1)}} \hat{b}_j^{(l-1)} \tilde{\epsilon}_{t-j}\left(\hat{\theta}^{(l-1)}\right) \Bigg)^2 + \frac{\lambda\gamma}{2} \sum_{j=1}^{p} |\hat{\beta}_j^{(l-1)}|^{\gamma-2} \beta_j^2 \Bigg\}.\end{aligned}$$

Set $l = l + 1$ and go to **(Step 1)** until $l < L + 1$ for a predetermined integer $L > 0$ or $||\hat{\boldsymbol{\beta}}^{(l)} - \hat{\boldsymbol{\beta}}^{(l-1)}|| < \eta_0$ where η_0 is a small positive constant, e.g. 10^{-3}.

The proposed method improves the work of Wang, Li & Tsai (2007) in two respects; (i) ARMA models include the AR model as a special case; (ii) The Wang et al. algorithm implicitly suggests that the correct order of error terms could be identified by setting some of the coefficients equal to zero if the initial order is chosen to be larger than the true order. Out proposed computational algorithm includes instead the procedure of explicitly selecting the order as well as important sets of variables. For the sake of simpler computation, we do not include the coefficients of the ARMA model in the penalty term; however, this can easily be done in the proposed algorithm.

Remark 3.3. In **(Step 2)** and **(Step 3)**, the initial values for $\tilde{e}_0, \ldots, \tilde{e}_{1-P^{(l-1)}}$ and $\tilde{\epsilon}_0, \ldots, \tilde{\epsilon}_{1-Q^{(l-1)}}$ are taken to be fixed, and to be neither random nor functions of the parameters.

Remark 3.4. If for all $\theta \in \Theta$, $\mathcal{A}_\theta(z) = 0$ implies $|z| > 1$ and ϵ_t is absolutely continuous with respect to Lebesgue measure, then the ARMA model (4) for $\{e_t\}$ is a strong mixing process satisfying (**A**5). See Mokkadem (1988) for more details.

3.3. Extension to ARMA-GARCH error models

In this subsection, we briefly introduce linear regression models with ARMA-GARCH errors. Let $\{e_t\}$ be a sequence of random variables following an ARMA(P, Q)-GARCH(r, s) model:

$$
\begin{aligned}
e_t &= a_0 + \sum_{i=1}^{P} a_i e_{t-i} + \epsilon_t + \sum_{j=1}^{Q} b_j \epsilon_{t-j}, \\
\epsilon_t &= z_t \sigma_t, \\
\sigma_t^2 &= c_0 + \sum_{i=1}^{r} c_i \epsilon_{t-i}^2 + \sum_{j=1}^{s} d_j \sigma_{t-j}^2. \qquad (4)
\end{aligned}
$$

Here $z_t, t = 1, \ldots, T$, is a sequence of independent random variables with zero mean and unit variance, and P, Q, r and s are unknown. If there is a strong evidence that the errors $\{e_t\}$ follow ARMA-GARCH models in any real applications, one can use the algorithm of Section 3.2, in which, however, **(Step 2)** and **(Step 3)** should be systematically modified to incorporate the ARMA-GARCH error structure.

According to Theorem 8 in Linder (2009), a pure GARCH model for $\{e_t\}$ in (4) is a strong mixing process satisfying (**A**5) provided that $\{\epsilon_t\}$ is strictly stationary, the z_t are absolutely continuous with respect to Lebesgue measure and are strictly positive in a neighborhood of zero, and there exists some $s \in (0, \infty)$ such that $E|z_t|^s < \infty$. Unfortunately it is still unknown whether ARMA-GARCH models are mixing processes or not. Consequently the use of weak dependence may be more suitable for verifying the asymptotic properties of bridge estimators when penalized regression with ARMA-GARCH errors is considered. We intend to investigate this issue in future work.

4. Simulation study

In this section we present a Monte Carlo simulation study to evaluate practical aspects of the proposed methods. We compare four approaches: ordinary least squares (OLS), ordinary bridge regression (ord BRID), the proposed algorithms with 1-step (1-step BRID) and the proposed algorithm with full convergence (full BRID). The 1-step BRID consists in setting the predetermined integer L equal to 1 in **(Step 4)** of the proposed algorithm, and the full BRID consists in continuing the iterations until the estimated parameters converge. We simulated data under the following two settings.

1. Setting I: We generated data from model (1) where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$ and e_t followed the ARMA(1,1) process specified in (5). The covariates $\boldsymbol{x}_t = (x_{t1}, \cdots, x_{t8})^\top$ were independently generated from the multivariate normal distribution with mean $0_{8\times1}$, the variance of each variable 1, and the pairwise correlation between x_{t,j_1} and x_{t,j_2} being $0.5^{|j_1-j_2|}$.
2. Setting II: We generated data from model (1) with $n+8$ covariates. The e_t, the first eight β coefficients, and the first eight covariates were similarly generated as in Setting I. The remaining β coefficients were set to zero and the remaining n noise covariates $(x_{t9}, \cdots, x_{t(n+8)})^\top$ were independently generated from the multivariate normal distribution with mean $0_{n\times1}$, the variance of each variable equal to 1, and the pairwise correlation between x_{t,j_1} and $x_{t,j_2}(9 \leq j_1, j_2 \leq n+8)$ equal to 0. We set the number n of noise covariates equal to 20 and 50.

In these simulation settings, we used the following error model:

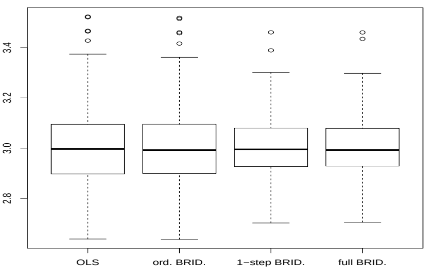
$$e_t = 0.6e_{t-1} + \epsilon_t - 0.4\epsilon_{t-1}. \tag{5}$$

In the simulation signal-to-noise ratios (SNRs) of 5 and 1.25, and sample sizes T equal to 100 and 300 were used. For the tuning parameters, $\lambda = 2^{k-15}$ for $k = 1, 2, \cdots, 20$, and the penalty orders γ = 0.1, 0.4, 0.7, 1 for the bridge estimators were used. For the ARMA orders, $P, Q = 1, 2, 3, 4$ were used. In the fitting process we chose the combination of (λ, γ) and (P, Q) that produced the lowest BIC.

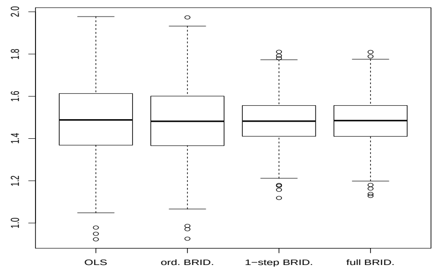
We repeated each setting 100 times and compared the average of the model errors $ME = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\boldsymbol{x}\boldsymbol{x}^\top)^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and the standard error of this average, the average numbers of correct and incorrect zero estimates for $\boldsymbol{\beta}$, the number of the correctly estimated ARMA orders, and the average computational time for a single iteration. We used both 1-step and full BRID in fitting models to the simulated data.

Table 1 reports the simulation results. For the ARMA(1,1) error model, the proposed 1-step and full iteration bridge methods had lower ME versus the OLS and ordinary bridge methods in all cases. The proposed methods correctly select important variables corresponding to the nonzero coefficients, and set most of the truly zero coefficients equal to zero (the average numbers are close to the true value 5). The proposed methods do not perform well for the selection of the ARMA orders when $T = 100$, but their performance significantly improves for $T = 300$ (around 80–90%). The two proposed methods, the 1-step BRID and the full BRID perform similarly in terms of ME, variable selection, and time series model selection. The computation time of the 1-step BRID is slightly longer than the ordinary bridge, but significantly shorter than that of the full BRID. Computation time being taken into account, the 1-step BRID may be preferable to the full BRID.

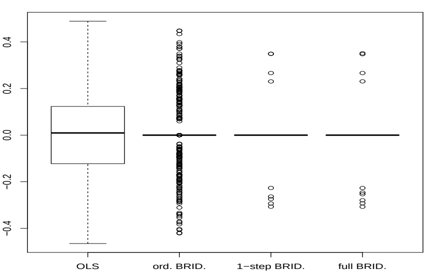
In order to investigate the proposed algorithm more carefully we include boxplots of the coefficients estimated by the four methods when SNR=5 and $T = 300$ in Figure 1. For the nonzero βs (β_1, β_2 and β_5), the centers of the coefficient estimates under all methods



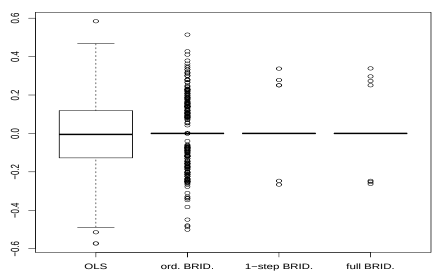
(a) Estimated coefficients for β_1



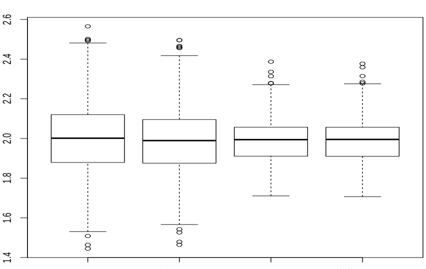
(b) Estimated coefficients for β_2



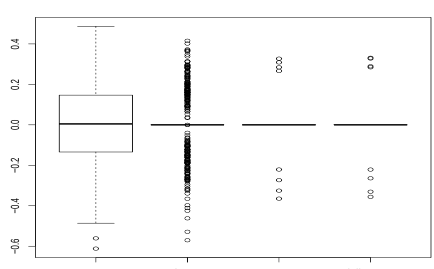
(c) Estimated coefficients for β_3



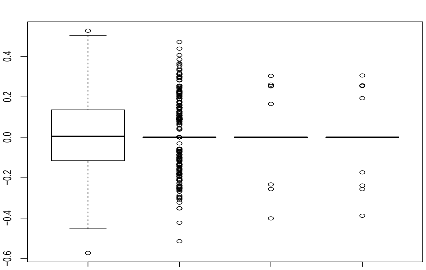
(d) Estimated coefficients for β_4



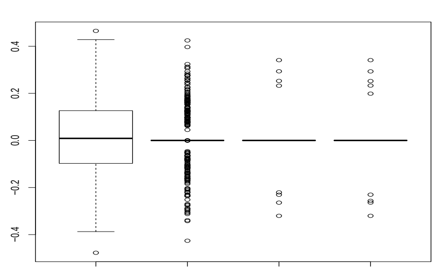
(e) Estimated coefficients for β_5



(f) Estimated coefficients for β_6



(g) Estimated coefficients for β_7



(h) Estimated coefficients for β_8

Figure 1. The distribution of the estimated coefficients for each method

TABLE 1

Results of Setting I. The average of ME and its standard error in the parentheses, average numbers of correct and incorrect zero estimates for $\boldsymbol{\beta}$, the number of the correctly estimated ARMA orders out of 100, and the average computational time for a single iteration.

Method	ME	Reg. coef. Cor	Reg. coef. Incor	error order No. cor	time (sec)	ME	Reg. coef. Cor	Reg. coef. Incor	error order No. cor	time (sec)
$T = 100$		SNR $= 5$					SNR $= 1.25$			
OLS	0.896(0.047)	0	0	-	-	3.584(0.190)	0	0	-	-
ord BRID	0.664(0.041)	2.97	0	-	0.021	2.705(0.168)	2.92	0.11	-	0.021
1-step BRID	0.100(0.009)	4.88	0	44	0.089	0.627(0.070)	4.82	0.14	40	0.091
full BRID	0.093(0.008)	4.84	0	47	0.637	0.604(0.072)	4.82	0.15	35	0.657
$T = 300$		SNR $= 5$					SNR $= 1.25$			
OLS	0.303(0.014)	0	0	-	-	1.212(0.057)	0	0	-	-
ord BRID	0.229(0.013)	2.79	0	-	0.021	0.924(0.053)	2.69	0.02	-	0.026
1-step BRID	0.027(0.003)	4.94	0	89	0.097	0.139(0.024)	4.93	0.02	89	0.098
full BRID	0.030(0.003)	4.93	0	87	0.445	0.143(0.024)	4.92	0.02	86	0.481

are similar to each other and are close to the true values, but the variations produced by the proposed methods are smaller. In cases of the zero βs ($\beta_3, \beta_4, \beta_6, \beta_7$ and β_8), most of the estimated coefficients are zero except for those produced by the OLS method. This demonstrates empirically the sparsity of the bridge method asserted in Theorem 2(i). A more careful look at the figures reveals that the estimates from the ordinary bridge method are more scattered than are those from the proposed algorithm. This implies that the regression parameters can be more accurately estimated by the proposed algorithm than by ordinary bridge regression. Note that the proposed methods with 1-step and full convergence show very similar performances; this suggests that as discussed above one iteration may be sufficient.

TABLE 2

Results of Setting II. The average of ME and its standard error in parentheses, average numbers of correct and incorrect zero estimates for $\boldsymbol{\beta}$, the number of correctly estimated ARMA orders out of 100, and the average computational time for a single iteration.

Method	ME	Reg. coef. Cor	Reg. coef. Incor	error order No. cor	time (sec)	ME	Reg. coef. Cor	Reg. coef. Incor	error order No. cor	time (sec)
					20 noise predictors					
$T = 100$		SNR $= 5$					SNR $= 1.25$			
OLS	4.344(0.182)	0	0	-	-	17.375(0.729)	0	0	-	-
ord BRID	2.768(0.150)	14.39	0.01	-	0.022	10.734(0.614)	14.30	0.17	-	0.021
1-step BRID	0.158(0.023)	24.51	0.01	36	0.095	0.997(0.118)	24.47	0.20	28	0.090
full BRID	0.150(0.023)	24.47	0.01	39	0.849	0.960(0.117)	24.32	0.18	40	0.837
$T = 300$		SNR $= 5$					SNR $= 1.25$			
OLS	1.047(0.031)	0	0	-	-	4.188(0.125)	0	0	-	-
ord BRID	0.649(0.028)	15.04	0	-	0.121	2.586(0.108)	14.51	0	-	0.121
1-step BRID	0.025(0.002)	24.87	0	87	0.279	0.108(0.011)	24.87	0	89	0.260
full BRID	0.025(0.002)	24.84	0	89	0.979	0.101(0.008)	24.86	0	89	0.999
					50 noise predictors					
$T = 100$		SNR $= 5$					SNR $= 1.25$			
OLS	14.786(0.510)	0	0	-	-	59.145(2.039)	0	0	-	-
ord BRID	7.307(0.320)	28.56	0.03	-	0.021	27.023(1.182)	27.92	0.30	-	0.023
1-step BRID	0.492(0.145)	53.56	0.03	39	0.094	2.827(0.620)	52.85	0.36	29	0.091
full BRID	0.470(0.145)	53.29	0.03	42	1.020	2.604(0.621)	52.70	0.33	35	1.184
$T = 300$		SNR $= 5$					SNR $= 1.25$			
OLS	2.423(0.061)	0	0	-	-	9.692(0.243)	0	0	-	-
ord BRID	1.558(0.052)	32.07	0	-	0.037	5.612(0.194)	33.53	0	-	0.034
1-step BRID	0.040(0.004)	54.52	0	86	0.113	0.176(0.016)	54.44	0	87	0.119
full BRID	0.045(0.004)	54.43	0	88	0.445	0.187(0.016)	54.43	0	86	0.445

Table 2 shows the results for Setting II. In this setting, the simulated models contain more noise predictors than in Setting I. As the number of noise predictors increases from 20 to 50, the overall performance of all four methods becomes worse as expected. For both cases $n = 20$ and 50, it is evident that the two proposed methods produce smaller ME values and higher numbers of correct variable selection than do the OLS and ordinary bridge methods. We note that the full BRID tends to yield smaller MEs and higher numbers of correct time series model selection than does the 1-step BRID especially for $T = 100$. However, with the computational time and the relatively small differences in ME kept in mind, we still recommend the use of the 1-step BRID.

5. Real data analysis

In this section, we analyze a data set from Ramanathan (1998) that concerns the consumption of electricity by residential customers served by the San Diego Gas and Electric Company. This data set contains 87 quarterly observations from the second quarter of 1972 through to the fourth quarter of 1993. The response variable is electricity consumption measured by the logarithm of the kwh sales per residential customer. The independent variables are per-capita income (LY), price of electricity (LPRICE), cooling degree days (CDD) and heating degree days (HDD). Wang, Li & Tsai (2007) and Yoon, Park & Lee (2012) analyze a similar data set.

The basic model considered in Ramanathan (1998) is given as:

$$\text{LKWH} = \beta_0 + \beta_1\text{LY} + \beta_2\text{LPRICE} + \beta_3\text{CDD} + \beta_4\text{HDD} + e_t,$$

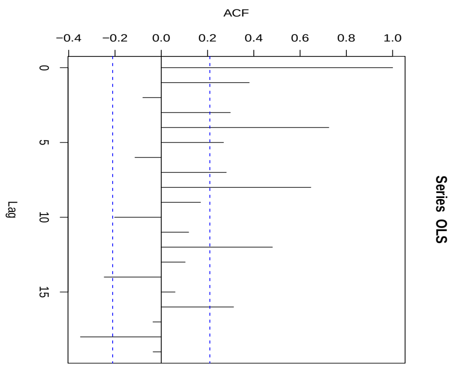
and the expected signs for the β's (except β_0) are (Ramanathan 1998)

$$\beta_1 > 0, \quad \beta_2 < 0, \quad \beta_3 > 0, \quad \beta_4 > 0.$$

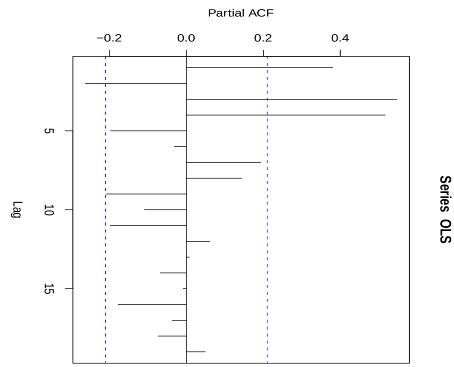
For the ordinary least squares (OLS) method, Table 3 shows that the sign of the estimated LY coefficient is opposite to the expected one. Figures 2 (a) and (b) show the sample autocorrelation function (SACF) and sample partial ACF (SPACF) of the residuals obtained by OLS. From these plots it can be seen that the independence assumption is severely violated i.e. the residuals are serially correlated. Moreover fourth-order serial correlation would appear to be appropriate since the data are quarterly (Ramanathan 1998). For the least squares (LS) with the fourth order AR error model, Table 3 shows that the signs of the coefficients meet the expectation, but the SACF and SPACF shown in Figures 2 (c) and (d) indicate that it may be necessary to seek a more sophisticated time series error model because the residuals still reveal a certain degree of dependence.

We further consider the 1-step bridge with an ARMA(P, Q) error model ($P, Q = 1, 2, 3, 4$) and select the pair of these orders that minimizes BIC. In Table 3, when the 1-step bridge is used, $\hat{\beta}_1$ and $\hat{\beta}_2$ are exactly zero, suggesting that the per-capita income (LY) and the price of electricity (LPRICE) do not contribute to the model. The signs of the estimated coefficients for the other variables are the same as expected. The ARMA orders are estimated as ($P = 4$, $Q = 2$) and the residuals from the 1-step bridge do not have serial correlations based on the SACF and SPACF in Figures 2 (e) and (f), and thus they can be taken to be white noise random variables.

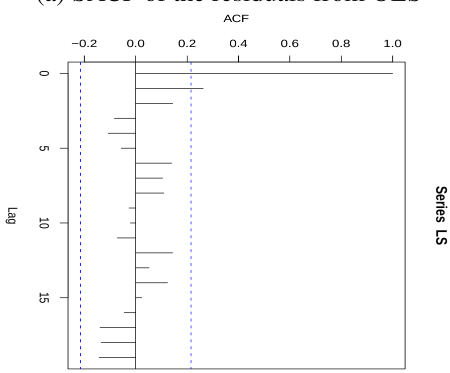
Although electricity demand is well known to be strongly seasonal, it is likely that controlling for heating and cooling days fully eliminates this seasonality, since the SACF



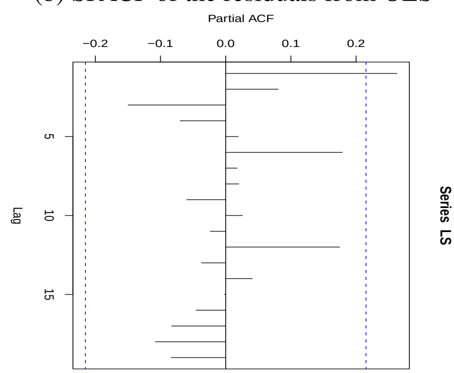
(a) SACF of the residuals from OLS



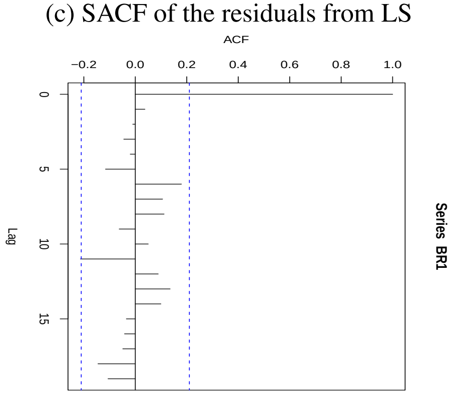
(b) SPACF of the residuals from OLS



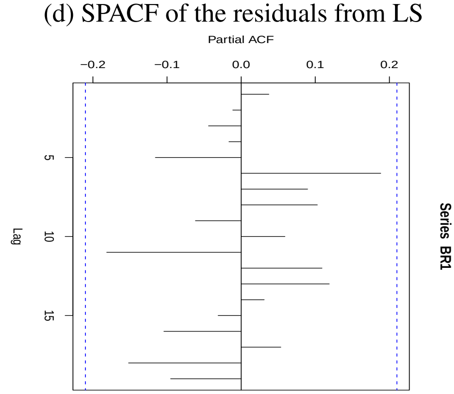
(c) SACF of the residuals from LS



(d) SPACF of the residuals from LS



(e) SACF of the residuals from 1-step bridge



(f) SPACF of the residuals from 1-step bridge

Figure 2. SACF and SPACF of the residuals for the real dataset

and SPACF of the residuals in (e) and (f) of Figure 2 for the one-step bridge method do not suggest a seasonal pattern. However, it would be interesting in future work to consider a seasonal effect in our penalized regression models.

TABLE 3

The estimated coefficients for the real dataset

Variable	OLS	LS	1-step bridge
LY	-0.03627	0.10256	0.00000
LPRICE	-0.09426	-0.09824	0.00000
CDD	0.00027	0.00028	0.00026
HDD	0.00036	0.00023	0.00031
ARMA orders	-	(4,0)	(4,2)

6. Appendix

In this section, we provide the proofs of Theorems 1 and 2. Our proofs are structured similarly to those in Huang, Horowitz & Ma (2008). Throughout this section, C is a generic positive constant that depends on p, taking different values from place to place.

Proof of Theorem 1

Lemma 1. *Let* $\mathbf{u}$ *be a* $p \times 1$ *vector. Under the conditions (**A**1)-(**A**3) in Section 2,*

$$E\left(\sup_{||u||\le\delta}\left|\sum_{i=1}^{T} e_i \boldsymbol{x}_i^\top \mathbf{u}\right|\right) \le C\delta T^{1/2}.$$

Proof: By the Cauchy-Schwarz inequality and conditions (**A**1)-(**A**3), we have

$$\begin{aligned}
E\left(\sup_{||u||\le\delta}\left|\sum_{i=1}^{T} e_i \boldsymbol{x}_i^\top \mathbf{u}\right|^2\right) &\le E\left(\sup_{||u||\le\delta}||\mathbf{u}||^2\left|\left|\sum_{i=1}^{T} e_i\boldsymbol{x}_i\right|\right|^2\right) \le \delta^2 E\left(\left|\left|\sum_{i=1}^{T} e_i\boldsymbol{x}_i\right|\right|^2\right)\\
&= \delta^2 E\left(\sum_{i=1}^{T} e_i^2 \boldsymbol{x}_i^\top\boldsymbol{x}_i\right) + \delta^2 E\left(\sum_{i\neq j}^{T} e_i e_j \boldsymbol{x}_i^\top \boldsymbol{x}_j\right)\\
&\le \delta^2\sigma_e^2 E\left(\sum_{i=1}^{T}\boldsymbol{x}_i^\top\boldsymbol{x}_i\right) + \delta^2\sum_{i\neq j}^{T}|\gamma_e(i-j)|\, E\left(|\boldsymbol{x}_i^\top\boldsymbol{x}_j|\right)\\
&\le \delta^2\sigma_e^2 T E\left(\text{trace}\left(T^{-1}\sum_{i=1}^{T}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\right) + \frac{1}{2}\delta^2\sigma_x^2\sum_{i\neq j}^{T}|\gamma_e(i-j)|\\
&\le \delta^2\sigma_e^2 Tp + \frac{1}{2}\delta^2\sigma_x^2 T\sum_{|m|<T}\left(1-\frac{|m|}{T}\right)|\gamma_e(m)|\\
&\le C\delta^2 T.
\end{aligned}$$

The lemma then follows from Jensen's inequality. $\square$

We now proceed to prove Theorem 1. To begin with, we show the following fact:

$$||\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0|| = O_P\left([(1+\lambda)/T]^{1/2}\right). \tag{6}$$

Using the same arguments as in the proof of Theorem 1 in Huang, Horowitz & Ma (2008), we have the following inequality

$$\|\boldsymbol{\delta}_T\| \le 2\|\boldsymbol{D}_T\boldsymbol{e}_T\| + O(\lambda^{1/2}), \tag{7}$$

where $\boldsymbol{\delta}_T = T^{1/2}\Sigma_T^{1/2}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0)$, $\boldsymbol{D}_T = T^{-1/2}\Sigma_T^{-1/2}\boldsymbol{x}_T^\top$ and $\boldsymbol{e}_T = (e_1, \ldots, e_T)^\top$.

Note that it follows from (**A**2) that

$$\Sigma_T \xrightarrow{a.s.} \Sigma.$$

Therefore it follows from Egorov's theorem that for sufficiently small $\epsilon > 0$ and $\eta > 0$, there exists an event A with $P(A) < \epsilon/2$ and a positive integer $N \ge 1$, such that on A^c (the complement of A) and for all $T \ge N$, there exists a smallest eigenvalue $\rho_{min,T}$ of Σ_T satisfying

$$0 < \rho_{min} - \eta < \rho_{min,T}, \tag{8}$$

and

$$0 < |\Sigma^{1/2}| - \eta < |\Sigma_T^{1/2}| < |\Sigma^{1/2}| + \eta, \;\; 0 < |\Sigma^{-1/2}| - \eta < |\Sigma_T^{-1/2}| < |\Sigma^{-1/2}| + \eta, \tag{9}$$

where ρ_{min} is the smallest eigenvalues of Σ. Note that (8) implies that Σ_T is invertible. Using (7), (9) and (**A**2), we obtain that on A^c

$$T^{1/2}(|\Sigma^{1/2}| - \eta)\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\| \le 2(|\Sigma^{-1/2}| + \eta)\|T^{-1/2}\boldsymbol{X}_T\boldsymbol{e}_T\| + O(\lambda^{1/2}).$$

and thus

$$\begin{aligned} \|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\| &\le CT^{-1/2}\|T^{-1/2}\boldsymbol{X}_T\boldsymbol{e}_T\| + O\left((\lambda/T)^{1/2}\right) \\ &= O_P\left(T^{-1/2}\right) + O\left((\lambda/T)^{-1/2}\right). \end{aligned}$$

The last equality follows from the fact that

$$E\|T^{-1/2}\boldsymbol{X}_T\boldsymbol{e}_T\| \le C,$$

which can be shown as in the proof of Lemma 1. Therefore, given $\epsilon > 0$, there exist $M > 0$ and a positive integer N, such that for any $T \ge N$,

$$\begin{aligned} &P\left([(1+\lambda)/T]^{-1/2}\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\| > M\right) \\ &\qquad \le \;\; P(A) + P\left(([(1+\lambda)/T]^{-1/2}\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\| > M) \cap A^c\right) \; < \; \epsilon. \end{aligned}$$

Hence, (6) is verified.

We next show that

$$\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\| = O_P\left(T^{-1/2}\right). \tag{10}$$

Let $S_{j,T} = \{\boldsymbol{\beta} : 2^{j-1} < T^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < 2^j\}$ with j ranging over the integers. By the definition of $\hat{\boldsymbol{\beta}}_T$, we get, for every $\epsilon > 0$,

$$
\begin{aligned}
&P(T^{1/2}\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\| > 2^M) \\
&\quad = P\left(\bigcup_{j\geq M} (\hat{\boldsymbol{\beta}}_T \in S_{j,T})\right) \\
&\quad = P\left(\bigcup_{j\geq M, 2^j \leq eT^{1/2}} (\hat{\boldsymbol{\beta}}_T \in S_{j,T}) \cap A^c\right) + P(A) + P\left(\bigcup_{j\geq M, 2^j > eT^{1/2}} (\hat{\boldsymbol{\beta}}_T \in S_{j,T})\right) \\
&\quad \leq \sum_{j\geq M, 2^j \leq eT^{1/2}} P\left((\hat{\boldsymbol{\beta}}_T \in S_{j,T}) \cap A^c\right) + P\left(2\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\| \geq e\right) + \frac{\epsilon}{2} \\
&\quad = \sum_{j\geq M, 2^j \leq eT^{1/2}} P\left((\inf_{\boldsymbol{\beta}\in S_{j,T}} (L_T(\boldsymbol{\beta}) - L_T(\boldsymbol{\beta}_0)) \leq 0) \cap A^c\right) \\
&\qquad\qquad + P\left(2\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\| \geq e\right) + \frac{\epsilon}{2}.
\end{aligned}
$$

The second term on the right-hand side converges to zero, because $\hat{\beta}_T$ is consistent by (6) and condition (**A**4)(a). Thus, we only need to show that the first term on the right-hand side converges to zero. This follows from the fact that

$$
\begin{aligned}
L_T(\boldsymbol{\beta}) - L_T(\boldsymbol{\beta}_0) &\geq \sum_{i=1}^T \left[\boldsymbol{x}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right]^2 - 2\sum_{i=1}^T e_i \boldsymbol{x}_i^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \lambda \sum_{j=1}^k \{|\beta_{1j}|^\gamma - |\beta_{01j}|^\gamma\} \\
&\equiv I_{1T} + I_{2T} + I_{3T}.
\end{aligned}
$$

In a manner similar to the proof of Theorem 1 in Huang, Horowitz & Ma (2008), using (9), we can show that for beta in $S_{j,T}$

$$
L_T(\boldsymbol{\beta}) - L_T(\boldsymbol{\beta}_0) \geq -|I_{2T}| + C\left(2^{2(j-1)} - T^{-1/2}\lambda 2^j\right),
$$

and thus

$$
\begin{aligned}
&P\left((\inf_{\boldsymbol{\beta}\in S_{j,T}} (L_T(\boldsymbol{\beta}) - L_T(\boldsymbol{\beta}_0)) \leq 0) \cap A^c\right) \\
&\qquad \leq P\left(\left(\sup_{\boldsymbol{\beta}\in S_{j,T}} |I_{2T}| \geq C\left(2^{2(j-1)} - T^{-1/2}\lambda 2^j\right)\right) \cap A^c\right) \\
&\qquad \leq C\frac{2^j}{2^{2(j-1)} - T^{-1/2}\lambda 2^j} \\
&\qquad = C\frac{1}{2^{j-2} - T^{-1/2}\lambda}
\end{aligned}
$$

where the second inequality follows from Markov's inequality and Lemma 1. Under (**A**4)(a), $\lambda T^{-1/2} \to 0$, and for sufficiently large T, $2^{j-2} - T^{-1/2}\lambda \geq 2^{j-3}$ for all $j \geq 3$. Therefore,

$$
\sum_{j\geq M, 2^j \leq eT^{1/2}} P\left((\inf_{\boldsymbol{\beta}\in S_{j,T}} (L_T(\boldsymbol{\beta}) - L_T(\boldsymbol{\beta}_0)) \leq 0) \cap A^c\right) \leq C\sum_{j\geq M} \frac{1}{2^{j-3}} \leq C2^{-M}
$$

which converges to zero as M tends to infinity. This completes the proof of (10) which leads to the conclusion of Theorem 1. □

Proof of Theorem 2

First we prove the following lemma which is needed for the proof of Theorem 2-(i).

Lemma 2. *Suppose that* $0 < \gamma < 1$. *Let* $\hat{\boldsymbol{\beta}}_T = (\hat{\boldsymbol{\beta}}_{1T}^\top, \hat{\boldsymbol{\beta}}_{2T}^\top)^\top$. *Under conditions (**A**1)-(**A**4),* $\hat{\boldsymbol{\beta}}_{2T} = \mathbf{0}$ *with probability converging to 1 as* T *tends to infinity.*

Proof: By Theorem 1, for any $e > 0$, there exists a positive integer C such that

$$P\left(\hat{\boldsymbol{\beta}}_T \in A_1^c\right) < \epsilon/3 \quad \text{for all} \quad T, \tag{11}$$

where $A_1 = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le T^{-1/2}C\}$. Let $\boldsymbol{\beta}_{1T} = \boldsymbol{\beta}_{10} + T^{-1/2}\boldsymbol{u}_1$, $\boldsymbol{\beta}_{2T} = \boldsymbol{\beta}_{20} + T^{-1/2}\boldsymbol{u}_2 = T^{-1/2}\boldsymbol{u}_2$ with $\|\boldsymbol{u}\|_2^2 = \|\boldsymbol{u}_1\|_2^2 + \|\boldsymbol{u}_2\|_2^2 \le C^2$ and

$$V_T(\boldsymbol{u}_1, \boldsymbol{u}_2) \equiv L_T(\boldsymbol{\beta}_{1T}, \boldsymbol{\beta}_{2T}) - L_T(\boldsymbol{\beta}_{10}, 0) = L_T(\boldsymbol{\beta}_{10} + T^{-1/2}\boldsymbol{u}_1, T^{-1/2}\boldsymbol{u}_2) - L_T(\boldsymbol{\beta}_{10}, 0).$$

Then, $\hat{\boldsymbol{\beta}}_{1T}$ and $\hat{\boldsymbol{\beta}}_{2T}$ can be obtained by minimizing $V_T(\boldsymbol{u}_1, \boldsymbol{u}_2)$ over $\|\boldsymbol{u}\|_2 \le C$, except on an event with probability converging to zero as $T \to \infty$. Now we have

$$V_T(\boldsymbol{u}_1, \boldsymbol{u}_2) - V_T(\boldsymbol{u}_1, 0) \equiv II_{1T} + II_{2T} + II_{3T} + II_{4T},$$

where

$$\begin{aligned} II_{1T} &= T^{-1}\sum_{i=1}^{T}(\boldsymbol{x}_{2,i}^\top \boldsymbol{u}_2)^2, \quad II_{2T} = 2T^{-1}\sum_{i=1}^{T}(\boldsymbol{x}_{1,i}{}^\top \boldsymbol{u}_1)(\boldsymbol{x}_{2,i}^\top \boldsymbol{u}_2), \\ II_{3T} &= -2T^{-1/2}\sum_{i=1}^{T} e_i \boldsymbol{x}_{2,i}^\top \boldsymbol{u}_2, \quad II_{4T} = \lambda T^{-\gamma/2}\sum_{j=1}^{m}|u_{2j}|^\gamma. \end{aligned}$$

For the first two terms, in a manner similar to the proof of Theorem 1, we can show that for sufficiently small $\epsilon > 0$ and $\eta > 0$, there exists an event B_2 with $P(B_2) < \epsilon/3$ and a positive integer $N \ge 1$ such that on B_2^c and for all $T \ge N$, we have

$$\begin{aligned} II_{1T} + II_{2T} &\ge -T^{-1}\sum_{i=1}^{T}(\boldsymbol{x}_{1,i}{}^\top \boldsymbol{u}_1)^2 \\ &\ge -\rho_{max,T}\|\boldsymbol{u}_1\|^2 \\ &\ge -(\rho_{max} - \eta)C, \end{aligned} \tag{12}$$

where, ρ_{max} and $\rho_{max,T}$ are the largest eigenvalues of Σ and Σ_T, respectively. For the third term, in a manner similar to the proofs of Lemma 1 and Theorem 1, we can show that on B_2^c and for all $T \ge N$,

$$\begin{aligned} E\left(\left|\sum_{i=1}^{T} e_i \boldsymbol{x}_{2,i}^\top \boldsymbol{u}_2\right|\right) &\le \left\{E\left(\left(\sum_{i=1}^{T} e_i \boldsymbol{x}_{2,i}^\top \boldsymbol{u}_2\right)^2\right)\right\}^{1/2} \\ &\le \left(\sigma_e^2 T(\rho_{max} - \eta)\|\boldsymbol{u}_1\|^2 + C\sigma_x^2 T\|\boldsymbol{u}_1\|^2\right)^{1/2} \\ &\le CT^{1/2}. \end{aligned}$$

We then have

$$II_{3T} = O_P(1) \tag{13}$$

on B_2^c. For the fourth term, since $[\sum_{j=1}^m |u_{2j}|^\gamma]^{2/\gamma} \ge \|\boldsymbol{u}_2\|^2$, we have

$$II_{4T} \ge \lambda T^{-\gamma/2}\|\boldsymbol{u}_2\|^\gamma \tag{14}$$

for sufficiently large T. Using (11)-(14) and condition (**A**4)(b), we have that, for sufficiently small $\epsilon > 0$, there exists a positive integer $N \ge 1$, such that if $\|\boldsymbol{u}_2\|_2 > 0$, $P\left(V_T(\boldsymbol{u}_1, \boldsymbol{u}_2) - V_T(\boldsymbol{u}_1, 0) < 0\right) < \epsilon$ for any $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ with $\|\boldsymbol{u}\|_2 < C$ and all $T \ge N$. This completes the proof. $\square$

To prove Theorem 2-(ii) we need the following two lemmas.

Lemma 3. *Let $\boldsymbol{\alpha}$ be any $k \times 1$ vector satisfying $\|\boldsymbol{\alpha}\| = 1$. Under the conditions (**A**3) and (**A**5), the sequence $\{e_i\boldsymbol{\alpha}^\top\Sigma_1^{-1}\mathbf{x}_{1,i}\}$ is strong mixing with mixing coefficients $\alpha(i)$ satisfying*

$$\sum_{i=1}^{\infty} \alpha^{\delta/[2+\delta]}(i) < \infty \quad \textit{and} \quad E|e_i\boldsymbol{\alpha}^\top\Sigma_1^{-1}\mathbf{x}_{1,i}|^{2+\delta} < \infty.$$

Proof: The proof follows from Theorem 3.2 in Bradley (1986), page 174. $\square$

Lemma 4. *Let $\boldsymbol{\alpha}$ be any $k \times 1$ vector satisfying $\|\boldsymbol{\alpha}\| = 1$. Under the same conditions of Theorem 2,*

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{T} e_i\boldsymbol{\alpha}^\top\Sigma_1^{-1}\boldsymbol{x}_{1,i} \xrightarrow{D} N(0, \boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha}),$$

where $\boldsymbol{K} = \sigma_e^2\Sigma_1^{-1} + 2\sum_{j=1}^{\infty}\gamma_e(j)\Sigma_1^{-1}E[\boldsymbol{x}_{1,1}\boldsymbol{x}_{1,1+j}^\top]\Sigma_1^{-1}$.

Proof: Note that by Lemma 3, we have

$$\frac{1}{T}Var\left(\sum_{i=1}^{T} e_i\boldsymbol{\alpha}^\top\Sigma_1^{-1}\boldsymbol{x}_{1,i}\right) = \boldsymbol{\alpha}^\top\boldsymbol{K}_T\boldsymbol{\alpha} \to \boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} > 0,$$

where $\boldsymbol{K}_T = \sigma_e^2\Sigma_1^{-1} + \frac{2}{T}\sum_{i\neq j}^{T}\gamma_e(|i-j|)\Sigma_1^{-1}E[\boldsymbol{x}_{1,i}\boldsymbol{x}_{1,j}^\top]\Sigma_1^{-1}$. Thus, the desired result follows from Theorem 1.7 in Peligrad (1986), page 202. $\square$

Proof of Theorem 2: In a manner similar to the proof of Theorem 2 in Huang, Horowitz & Ma (2008), using condition (**A**2) and (8), we can show that

$$\begin{aligned} T^{1/2}\boldsymbol{\alpha}^\top(\hat{\boldsymbol{\beta}}_{1T} - \boldsymbol{\beta}_{10}) &= \frac{1}{\sqrt{T}}\sum_{i=1}^{T} e_i\boldsymbol{\alpha}^\top\Sigma_{1T}^{-1}\boldsymbol{x}_{1,i} + o_P(1) \\ &= \frac{1}{\sqrt{T}}\sum_{i=1}^{T} e_i\boldsymbol{\alpha}^\top\Sigma_1^{-1}\boldsymbol{x}_{1,i} + o_P(1), \end{aligned}$$

for any $k \times 1$ vector $\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\| = 1$. The result then follows from Lemma 4 and Cramér-Wold device. $\square$

References

ALQUIER, P. & DOUKHAN, P. (2011). Sparsity considerations for dependent variables. *Electron. J. Stat.* **5**, 750–774.

BILLINGSLEY, P. (1995). *Probability and measure.* New York: John Wiley & Sons, Inc.

BOSQ, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction.* New York: Springer-verlag.

BRADLEY, R.C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics*, ed. Eberlein, E. and Taqqu, M. S. Birkhäuser, Boston, pp. 165–192.

BREHENY, P. & HUANG, J. (2009). Penalized methods for bi-level variable selection. *Stat. Interface* **2**, 369–380.

BROCKWELL, P.J. & DAVIS, R.A. (2006). *Time Series: Theory and Methods.* New York: Springer-verlag, 2nd edn.

DOUKHAN, P. (1994). *Mixing: properties and examples.* New York: Springer-verlag.

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

FRANCQ, C. & ZAKOÏAN, J.M. (2010). *GARCH moodels: structure, statistical inference and financial applications.* United Kingdom: Wiley.

FRANK, I.E. & FRIEDMAN, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.

FU, W.J. (1998). Penalized regression: the Bridge versus the Lasso. *J. Comput. Graph. Statist.* **7**, 397–416.

GELPER, S. & CROUX, C. (2009). Time series least angle regression for selecting predictive economic sentiment series. Unpublished.

GLASBEY, C.A. (1988). Examples of regression with serially correlated errors. *The Statistician* **37**, 277–291.

HIROSE, K., TATEISHI, S. & KONISHI, S. (2011). Effcient algorithm to select tuning parameters in sparse regression modeling with regularization. Available at http://arxiv.org/pdf/1109.2411.pdf.

HOERL, A.E. & KENNARD, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

HSU, N.J., HUNG, H.L. & CHANG, Y.M. (2008). Subset selection for vector autoregressive processes using LASSO. *Comput. Statist. Data Anal.* **52**, 3645–3647.

HUANG, J., HOROWITZ, J.L. & MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587–613.

HUANG, J., MA, S., XIE, H. & ZHANG, C.H. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339–355.

KNIGHT, K. & FU, W.J. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28**, 1356–1378.

LINDER, A.M. (2009). Stationarity, mixing, distributional properties and moments of GARCH(p,q)-processes. In *Handbook of Financial Time Series*, ed. Andersen, T. G., Davis, R. A., Kreiss, J. P. and Mikosch, T. Springer, New York, pp. 43–69.

MOKKADEM, A. (1988). Mixing properties of ARMA processes. *Stochastic Process. Appl.* **29**, 309–315.

PARK, C. & YOON, Y.J. (2011). Bridge regression: adaptivity and group selection. *J. Statist. Plann. Inference* **141**, 3506–3519.

PARK, T. & CASELLA, G. (2008). The Bayesian Lasso. *J. Amer. Statist. Assoc.* **103**, 681–686.

PELIGRAD, M. (1986). Recent advances in the central limit theorem and its weak invariance princeple for mixing sequences of random variables (a survey). In *Dependence in Probability and Statistics*, ed. Eberlein, E. and Taqqu, M. S. Birkhäuser, Boston, pp. 193–223.

RAMANATHAN, R. (1998). *Introductory Econometrics with Applications.* Fort Worth: Dryden: Harcourt Brace College Publishers.

SHAO, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statist. Sinica* **7**, 221–264.

SHI, P. & TSAY, C.L. (2004). A Joint Regression Variable and Autoregressive Order Selection Criterion. *J. Time Series Anal.* **25**, 923–941.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. Ser. B* **67**, 91–108.

TSAY, R.S. (1984). Regression models with time series errors. *J. Amer. Statist. Assoc.* **79**, 118–124.

WANG, H., LI, G. & TSAI, C. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **69**, 63–78.

YOON, Y.J., PARK, C. & LEE, T. (2012). Penalized regression models with autoregressive error terms. *J. Stat. Comput. Simul.* , DOI:10.1080/00949655.2012.669383.

ZHANG, H.H. & LU, W. (2007). Adaptive-LASSO for Cox's proportional hazard model. *Biometrika* **94**, 691–703.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301–320.

ZOU, H., HASTIE, T. & TIBSHIRANI, R. (2007). On the degrees of freedom of the lasso. *Ann. Statist.* **35**, 2173–2192.

ZOU, H. & ZHANG, H.H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37**, 1733–1751.