Statistics and Probability Letters 80 (2010) 1705–1712

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Subset selection for vector autoregressive processes via adaptive Lasso

Yunwen Ren*, Xinsheng Zhang

Department of Statistics, School of Management, Fudan University, Shanghai, 200433, PR China

ARTICLE INFO

Article history:
Received 1 May 2010
Received in revised form 23 July 2010
Accepted 23 July 2010
Available online 5 August 2010

Keywords:
Adaptive Lasso
Bayesian information criterion
HQ criterion
Oracle property
Vector autoregressive precesses

ABSTRACT

Subset selection is a critical component of vector autoregressive (VAR) modeling. This paper proposes simple and hybrid subset selection procedures for VAR models via the adaptive Lasso. By a proper choice of tuning parameters, one can identify the correct subset and obtain the asymptotic normality of the nonzero parameters with probability tending to one. Simulation results show that for small samples, a particular hybrid procedure has the best performance in terms of prediction mean squared errors, estimation errors and subset selection accuracy under various settings. The proposed method is also applied to modeling the IS-LM data for illustration.

1. Introduction

Vector autoregressive (VAR) models advocated by Sims (1980) have been widely used by applied researchers to capture the evolution of and the interdependence among multiple time series in economics, finance, physical sciences, geophysics, and many other fields. Model selection (i.e., determining the lag order and identifying the nonzero coefficients) is a critical component of VAR modeling. As documented by Hsu et al. (2008), roughly there are three types of existing model selection methods for VAR models: the information-based method, the hypothesis testing-based method and the simulation-based method. Hsu et al. (2008) also proposed several new subset selection procedures based on the Lasso (Tibshirani, 1996), including a simple Lasso procedure and some hybrid Lasso procedures. As pointed out by Zou (2006), Lasso is inconsistent in subset selection and biased in parameter estimation. Zou (2006) proposed the adaptive Lasso to improve upon the Lasso. Similar as the Lasso, the adaptive Lasso selects subset and estimates parameters simultaneously. The computing of the adaptive Lasso can be easily conducted by the Least Angle Regression (LARS) algorithm (Efron et al., 2004). Moreover, the adaptive Lasso is consistent in variable selection, and the nonzero adaptive Lasso estimators have asymptotically normal distribution. The above properties of the adaptive Lasso were established by Zou (2006) for linear regression models. At present, adaptive Lasso model selection procedures for VAR models are not available. The purpose of this paper is to provide subset selection methods for VAR models using the adaptive Lasso and establish their asymptotic properties. Similar as Hsu et al. (2008), we propose two types of methods for VAR model selection: the simple adaptive Lasso method and the hybrid adaptive Lasso method. We prove that under mild conditions, the proposed adaptive Lasso methods can select the correct subset with probability tending to one, and the nonzero estimators are asymptotically normally distributed.

This paper is organized as follows. Section 2 reviews the basics of VAR models. Section 3 presents the adaptive Lasso subset selection procedures for VAR models and the asymptotic property of the resulting estimators. Section 4 discusses the computational issues of the proposed methods. These include the modified LARS algorithm and the BIC criterion for choosing tuning parameters. Section 5 provides simulation studies and a real data analysis. Section 6 concludes the paper.

* Corresponding author.
E-mail addresses: 071025011@fudan.edu.cn (Y. Ren), xszhang@fudan.edu.cn (X. Zhang).

0167-7152/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2010.07.013

2. VAR models

Consider a stationary VAR(p) model:

$$\mathbf{Y}_t = \nu + \mathbf{B}_1\mathbf{Y}_{t-1} + \cdots + \mathbf{B}_p\mathbf{Y}_{t-p} + \epsilon_t = \nu + \mathbf{B}\mathbf{Z}_t + \epsilon_t, \tag{1}$$

where ν is a $k \times 1$ vector of intercept term, $\mathbf{B}_i$'s are $k \times k$ coefficient matrices, $\mathbf{B} = (\mathbf{B}_1, \ldots, \mathbf{B}_p)$, and for $t = 1, \ldots, n$, $\mathbf{Y}_t$ is a $k \times 1$ vector, $\mathbf{Z}_t = (\mathbf{Y}_{t-1}^T, \ldots, \mathbf{Y}_{t-p}^T)^T$, and ϵ_t is a $k \times 1$ white noise with nonsingular covariance matrix Σ. Let $\mathbf{B}^*$ be the true value of $\mathbf{B}$. Once we have an estimator of $\mathbf{B}^*$, say, $\hat{\mathbf{B}}$, we can estimate ν by $\hat{\nu} = \bar{\mathbf{Y}} - \hat{\mathbf{B}}\bar{\mathbf{Z}}$, where $\bar{\mathbf{Y}}$ and $\bar{\mathbf{Z}}$ are the means of $\mathbf{Y}_t$s and $\mathbf{Z}_t$s, respectively. Therefore, in this paper, we consider the centralized VAR(p) model (1) with the same symbols of $\mathbf{Y}_t$ and $\mathbf{Z}_t$ denoting their centralized values. The matrix version of (1) without intercept is

$$\mathbf{Y} = \mathbf{B}\mathbf{Z} + \epsilon, \tag{2}$$

with $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. The vector version of (2) is

$$\mathbf{y} = \mathbf{z}^T\boldsymbol{\beta} + \mathbf{e}, \tag{3}$$

where

$$\mathbf{y} = vec(\mathbf{Y}), \qquad \boldsymbol{\beta} = vec(\mathbf{B}), \qquad \mathbf{z} = \mathbf{Z} \otimes I_k, \qquad \mathbf{e} = vec(\epsilon), \tag{4}$$

and "vec" is the stack operator. The ordinary least square (OLS) estimator of $\boldsymbol{\beta}^* = vec(\mathbf{B}^*)$, the true value of $\boldsymbol{\beta}$, is

$$\hat{\boldsymbol{\beta}} = ([(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}] \otimes I_k)\mathbf{Y} = (\mathbf{z}\mathbf{z}^T)^{-1}\mathbf{z}\mathbf{y}. \tag{5}$$

Notice that the components of $\mathbf{e}$ are not independent, since $Var(\mathbf{e}) = I_n \otimes \Sigma$.

Assumption 1. $\mathbf{Y}_t = \mathbf{B}\mathbf{Z}_t + \epsilon_t$ are stationary processes for all t.

Assumption 2. For $t = 1 \ldots, n$, ϵ_t are independent with mean zeros and covariance $\Sigma > \mathbf{0}$, and for $k_1, k_2, k_3, k_4 = 1, \ldots, k$, $E|\epsilon_{tk_1}\epsilon_{tk_2}\epsilon_{tk_3}\epsilon_{tk_4}| < c$ for some constant c.

Assumptions 1 and 2 are from Fuller (1996) and Lütkepohl (2005). Under Assumptions 1 and 2, we have (1) $\lim_{n\to\infty} \mathbf{Z}\mathbf{Z}^T/n \to \Gamma$, where Γ is nonsingular; and (2) $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}^*$ with asymptotic distribution $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \to N(\mathbf{0}, \Gamma^{-1} \otimes \Sigma)$. These asymptotic properties are useful for proving Theorem 3.1.

3. Subset selection procedures

In this section, we introduce two types of subset selection procedures for VAR models: the simple adaptive Lasso in Section 3.1, and the hybrid adaptive Lasso in Section 3.2. In Section 3.3, we establish the asymptotic properties of the adaptive Lasso estimators. Hereafter, we denote by p^* the true value of the lag order of model (2), and we assume $p \geq p^*$.

3.1. *Simple adaptive Lasso procedure for VAR subset selection*

It is well known that the adaptive Lasso has the Oracle property and produces asymptotically unbiased estimators for the nonzero parameters in linear regression models. The adaptive Lasso can be realized by using the computationally efficient LARS algorithm (Efron et al., 2004). We now consider the model selection of (3) by using the adaptive Lasso. The (simple) adaptive Lasso estimator is given by

$$\hat{\boldsymbol{\beta}}^{(n)} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{z}^T\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{pk^2} \hat{\omega}_j|\beta_j| \right\}, \tag{6}$$

where λ_n is the tuning parameter and $\hat{\omega}_j = |\hat{\beta}_j|^{-\gamma}$ are adaptive weights which impose different shrinkage on different parameters with $\gamma \geq 0$. When $\gamma = 0$, the estimator obtained by (6) is the Lasso estimator. Zou (2006) pointed out that $\hat{\beta}_j$ can be any root-n consistent estimator of $\hat{\beta}_j^*$ as long as $\gamma > 0$. Thus we can use the OLS estimator $\hat{\boldsymbol{\beta}}$ as adaptive weights.

Note that adaptive Lasso is essentially a weighted L_1 penalization method. For the optimization problem in (6), $\lambda_n \oplus \hat{\boldsymbol{\beta}}^{(n)}$ is a critical point if

$$\begin{cases} 2|\mathbf{z}_j(\mathbf{y} - \mathbf{z}^T\hat{\boldsymbol{\beta}}^{(n)})| = \lambda_n\hat{\omega}_j, & \hat{\beta}_j^{(n)} \neq 0, \\ 2|\mathbf{z}_j(\mathbf{y} - \mathbf{z}^T\hat{\boldsymbol{\beta}}^{(n)})| < \lambda_n\hat{\omega}_j, & \hat{\beta}_j^{(n)} = 0. \end{cases} \tag{7}$$

Since the loss function in (6) is convex penalized loss, (7) is actually the Karush–Kuhn–Tucker (KKT) condition for the global minimization of (6). The global minimizer of (6) is unique. The modified LARS algorithm we used to obtain the adaptive estimator $\hat{\boldsymbol{\beta}}^{(n)}$ is based on the KKT condition (7). The computational details are presented in Section 4.

3.2. Hybrid adaptive Lasso procedure for VAR subset selection

We now discuss some hybrid adaptive Lasso procedures for the subset selection of (3). A hybrid adaptive Lasso subset selection procedure for (3) first determines the lag order $\hat{d}$ by some information criteria, and then selects the nonzero components of coefficients of the resulting VAR($\hat{d}$) model by adaptive Lasso. We use the AIC (Akaike, 1973) and the HQ (Hannan and Quinn, 1979) criteria for lag order selection, with

$$\begin{aligned} AIC(d) &= \log(\det(\hat{\Sigma}(d))) + 2mk^2/n, \\ HQ(d) &= \log(\det(\hat{\Sigma}(d))) + 2\log\log(n)mk^2/n, \end{aligned}$$

where d is a selected lag order, $\det(A)$ is the determinant of matrix A and $\hat{\Sigma}(d)$ is the estimator of the covariance matrix Σ under model VAR(d). The definition of the AIC and HQ criteria can be found in Lütkepohl (2005). It is well-known that HQ is a consistent order selection criterion while AIC is not. Let p_{aic}/p_{hq} be the best order selected by the AIC/HQ criterion.

Specifically, the hybrid procedures for VAR subset selection are

(a) *AIC/HQ joint with adaptive Lasso for multiple time series* (AIC + ALasso/HQ + ALasso)
  1. Use AIC/HQ criterion to select the best order for VAR model fitting.
  2. Find the adaptive Lasso estimator for multiple series under the VAR(p_{aic})/VAR(p_{hq}) model.

  For comparison, we also give the hybrid Lasso procedures for VAR subset selection.

(b) *AIC/HQ joint with Lasso for multiple time series* (AIC + Lasso/HQ + Lasso)
  1. Use AIC/HQ criterion to select the best order for VAR model fitting.
  2. Find the Lasso estimator for multiple series under the VAR(p_{aic})/VAR(p_{hq}) model.

3.3. Asymptotic properties of the adaptive Lasso estimator

Let $\mathcal{A} = \{j : \beta_j^* \neq 0\}/\mathcal{A}^c = \{j : \beta_j^* = 0\}$ be the index set of nonzero/zero coefficients. Thus $\mathcal{A}$ is the subset index of interest. Similarly, we denote by $\hat{\mathcal{A}}_n = \{j : \hat{\beta}_j^{(n)} \neq 0\}$ the index set of nonzero estimators selected by using tuning parameter λ_n.

Theorem 3.1 (*Oracle Property*)**.** *Suppose that Assumptions 1 and 2 hold, and assume* $\lambda_n n^{(\gamma-1)/2} \to \infty$ *and* $\lambda_n/\sqrt{n} \to 0$. *Then the adaptive Lasso estimator must satisfy*

(i) *Selection consistency:* $\lim_{n\to\infty} P(\hat{\mathcal{A}}_n = \mathcal{A}) = 1$, *and*

(ii) *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}^{(n)}_{\hat{\mathcal{A}}_n} - \boldsymbol{\beta}^*_{\mathcal{A}}) \to N(\mathbf{0}, (\Gamma^{-1} \otimes \Sigma)_{\mathcal{A}})$.

Proof. The proof of Theorem 3.1 is similar to Zou (2006)'s proof of Theorem 2. We first prove the asymptotic normality part. Similar as Zou (2006) did, let $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \mathbf{u}/\sqrt{n}$, $\Psi_n(\mathbf{u}) = \|\mathbf{y} - \mathbf{z}^T(\boldsymbol{\beta}^* + \mathbf{u}/\sqrt{n})\|^2 + \lambda_n \sum_{j=1}^{pk^2} \hat{\omega}_j |\beta_j^* + u_j/\sqrt{n}|$, and $\hat{\mathbf{u}}^{(n)} = \arg\min \Psi_n(\mathbf{u})$. Then we have $\hat{\boldsymbol{\beta}}^{(n)} = \boldsymbol{\beta}^* + \hat{\mathbf{u}}^{(n)}/\sqrt{n}$.

By Assumptions 1 and 2, we have $n^{-1}\mathbf{z}\mathbf{z}^T \to \Gamma \otimes I_k$ in probability and $\mathbf{e}^T\mathbf{z}^T/\sqrt{n} \to W \sim N(\mathbf{0}, \Gamma \otimes \Sigma)$ in distribution (Fuller, 1996). Since $\hat{\beta}_j$ is the $\sqrt{n}$-consistent estimator of β^*, with probability tending to 1, we have that $\hat{\omega}_j = |\hat{\beta}_j|^{-\gamma} \to |\beta_j^*|^{-\gamma}$ for $j \in \mathcal{A}$, and that $\hat{\omega}_j = |\hat{\beta}_j|^{-\gamma} = O(n^{\gamma/2})$ for $j \in \mathcal{A}^c$. Using the conditions that $\lambda_n n^{(\gamma-1)/2} \to \infty$ and $\lambda_n/\sqrt{n} \to 0$, we have $V_n(\mathbf{u}) = \Psi_n(\mathbf{u}) - \Psi_n(\mathbf{0}) = \mathbf{u}^T(n^{-1}\mathbf{z}^T\mathbf{z})\mathbf{u} - 2\mathbf{e}^T\mathbf{z}^T\mathbf{u}/\sqrt{n} + \lambda_n/\sqrt{n}\sum_{j=1}^{pk^2}\sqrt{n}\hat{\omega}_j(|\beta_j^* + u_j/\sqrt{n}| - |\beta_j^*|) \to V(u)$ in distribution, where

$$V(\mathbf{u}) = \begin{cases} \mathbf{u}_{\mathcal{A}}^T(\Gamma \otimes I_k)_{\mathcal{A}}\mathbf{u}_{\mathcal{A}} - 2\mathbf{u}_{\mathcal{A}}^T W_{\mathcal{A}}, & \text{if } u_j = 0, \forall j \in \mathcal{A}, \\ \infty, & \text{otherwise.} \end{cases}$$

Since $V_n(\mathbf{u})$ is convex, $V(\mathbf{u})$ attains its unique minimum at $((\Gamma \otimes I_k)_{\mathcal{A}}^{-1} W_{\mathcal{A}}, \mathbf{0})$. Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have $\hat{\mathbf{u}}_{\mathcal{A}} \to (\Gamma \otimes I_k)_{\mathcal{A}}^{-1} W_{\mathcal{A}}$ and $\hat{\mathbf{u}}_{\mathcal{A}^c} \to \mathbf{0}$ in distributions. Since $W_{\mathcal{A}} \sim N(\mathbf{0}, (\Gamma \otimes \Sigma)_{\mathcal{A}})$, we have part (ii).

Now we prove the selection consistency part. $\forall j \in \mathcal{A}$, from the proof of the asymptotic normality part, we have $\beta_j^{(n)} \to \beta_j^*$ with probability tending to 1. Thus, $P(j \in \hat{\mathcal{A}}_n) \to 1$. $\forall j' \in \mathcal{A}^c$, we need to show $P(j' \in \hat{\mathcal{A}}_n) \to 0$. If $j' \in \hat{\mathcal{A}}_n$, by the KKT condition (7), we must have $2|\mathbf{z}_{j'}(\mathbf{y} - \mathbf{z}^T\hat{\boldsymbol{\beta}}^{(n)})| = \lambda_n\hat{\omega}_{j'}$. Note again that $\lambda_n\hat{\omega}_{j'} \to \infty$, and that $\mathbf{z}_{j'}(\mathbf{y} - \mathbf{z}^T\hat{\boldsymbol{\beta}}^{(n)})/\sqrt{n} = \mathbf{z}_{j'}\mathbf{z}^T\sqrt{n}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{(n)})/n + \mathbf{z}_{j'}\mathbf{e}/\sqrt{n}$. Note also that $\mathbf{z}_{j'}\mathbf{z}^T\sqrt{n}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{(n)})/n$ converges to some normal distribution by the convergence property of $\hat{\mathbf{u}}_{\mathcal{A}}$, and $\mathbf{z}_{j'}\mathbf{e}/\sqrt{n} \to N(0, \mathbf{z}_{j'}(I_n \otimes \Sigma)\mathbf{z}_{j'}^T)$. Therefore, $P(j' \in \hat{\mathcal{A}}_n) \leq P(2|\mathbf{z}_{j'}(\mathbf{y} - \mathbf{z}^T\hat{\boldsymbol{\beta}}^{(n)})| = \lambda_n\hat{\omega}_{j'}) \to 0$. We thus established part (i) of Theorem 3.1.

Part (i) of Theorem 3.1 demonstrates that we can select the correct subset with probability tending to 1, and part (ii) states that the adaptive Lasso estimator in $\mathcal{A}$ is asymptotic unbiased and asymptotic normally distributed–the nonzero estimators perform as if we knew the zero parameters in advance. Further remarks can be found in Zou (2006). □

Remark 1. The results of Theorem 3.1 are for the simple adaptive Lasso procedure. They also hold for the hybrid adaptive Lasso procedures if $p_{aic}/p_{hq} \geq p^*$. If we adopt the consistent HQ criterion to select the best VAR order, we can select the correct order with probability tending to one. Thus, the important subset can survive after the order selection step with probability tending to one.

4. Computation

The entire solution path of the adaptive Lasso can be computed efficiently by the LARS algorithm, with a computation cost of order $O(npk^3)$, which is the same as the computation cost of an OLS fit.

Algorithm (*The LARS Algorithm for the Adaptive Lasso VAR*)**.**

1. Define $\mathbf{z}_j^* = \mathbf{z}_j\hat{w}_j, j = 1, \ldots, pk^2$;
2. Solve the Lasso problem using the LARS algorithm

$$\hat{\boldsymbol{\beta}}^{*(n)} = \arg\min_{\boldsymbol{\beta}}\left\{\|\mathbf{y}-\mathbf{z}^{*T}\boldsymbol{\beta}\|^2 + \lambda_n\sum\nolimits_{j=1}^{pk^2}|\beta_j|\right\};$$

3. Output $\hat{\beta}_j^{(n)} = \hat{\beta}_j^{*(n)}\hat{w}_j$.

Suppose the OLS estimator $\hat{\boldsymbol{\beta}}$ are adopted to construct the adaptive weights in the adaptive Lasso. For a given γ, we can use BIC criterion (Schwarz, 1978) along with the LARS algorithm to search for the optimal λ_n. Then we can use grid search to find the optimal γ. The BIC score is defined by

$$\mathrm{BIC}_{\lambda_n} = \log(\hat{\sigma}^2) + df_{\lambda_n}\log(nk)/(nk), \tag{8}$$

where $\hat{\sigma}^2 = (\hat{\epsilon}_{\lambda_n}^T\epsilon_{\lambda_n})/(nk)$, $\hat{\epsilon}_{\lambda_n} = (\mathbf{y}-\mathbf{z}^T\hat{\boldsymbol{\beta}}_{\lambda_n})$, with $\hat{\boldsymbol{\beta}}_{\lambda_n}$ the adaptive Lasso estimator given (γ, λ_n), and df_{λ_n} is the size of nonzero components of $\hat{\boldsymbol{\beta}}_{\lambda_n}$.

For the adaptive weights, similar as Zou (2006), we suggest using the OLS estimator when collinearity is not an issue, and using the ridge regression estimator (Hoerl and Kennard, 1988) when collinearity is an issue, since the ridge regression estimator is more stable than OLS estimator in case of collinearity.

5. Numerical examples

In this section, we present simulation studies and a real data analysis to assess the performance of the adaptive Lasso for VAR subset selection. Following Hsu et al. (2008), we use the h-step normalized prediction mean squared error (PMSE) to evaluate the predictive performance of a selected model, RE to measure the averaged relative efficiency of parameter estimates with respect to the full model, and CovRisk to measure the precision for estimating Σ of a selected model. The nearer to 1 the PMSE is, the better the selected model is. RE has the worst value at one if the full model is selected. And the closer to zero CovRisk is, the better the selected model is. The h-step PMSE, RE and CovRisk are defined as

$$\mathrm{PMSE}_h = k^{-1}E[(\hat{\mathbf{Y}}_{n+h}-\mathbf{Y}_{n+h})^T\Sigma_h^{-1}(\hat{\mathbf{Y}}_{n+h}-\mathbf{Y}_{n+h})],$$

$$\mathrm{RE} = (p_{\max}k^2)^{-1}E[(\hat{\boldsymbol{\beta}}^{(n)}-\boldsymbol{\beta}^*)^T[diag(Var(\hat{\boldsymbol{\beta}}^*))]^{-1}(\hat{\boldsymbol{\beta}}^{(n)}-\boldsymbol{\beta}^*)],$$

$$\mathrm{CovRisk} = E[tr(\hat{\Sigma}^{-1}\Sigma) - \log|\hat{\Sigma}^{-1}\Sigma|] - k.$$

where $\Sigma_h = \sum_{j=0}^{h-1}\Phi_j\Sigma\Phi_j^T$, with $\Phi_0 = I_k$ and $\Phi_j = \sum_{i=1}^{j}\Phi_{j-i}\mathbf{A}_i$[(2.1.22) and (2.2.11) in (Lütkepohl, 2005)], $\hat{\mathbf{Y}}_{n+h}$ and $\mathbf{Y}_{n+h}$ are the h-step best linear prediction and the realization, respectively, Σ_h is the h-step prediction variance based on the true model with known parameters, $p_{\max}$ is the largest candidate order, $diag(Var(\hat{\boldsymbol{\beta}}))$ is the diagonal matrix of variance of $\hat{\boldsymbol{\beta}}$-the OLS estimator under the largest candidate model, and $\hat{\Sigma}$ is the estimator of Σ. In the simulation studies, we only consider the one-step PMSE. The one-step PMSE, RE and CovRisk are estimated by the following statistics

$$\mathrm{pmse} = \frac{1}{k}\frac{1}{m}\sum_{i=1}^{m}[(\hat{\mathbf{Y}}_{n+1}^{(i)}-\mathbf{Y}_{n+1}^{(i)})^T\Sigma_1^{-1}(\hat{\mathbf{Y}}_{n+1}^{(i)}-\mathbf{Y}_{n+1}^{(i)})],$$

$$\mathrm{re} = \frac{1}{p_{\max}k^2}\frac{1}{m}\sum_{i=1}^{m}[((\hat{\boldsymbol{\beta}}^{(n)})^{(i)}-\boldsymbol{\beta}^*)^T[diag(\hat{Var}^{(i)}(\hat{\boldsymbol{\beta}}))]^{-1}((\hat{\boldsymbol{\beta}}^{(n)})^{(i)}-\boldsymbol{\beta}^*)],$$

$$\mathrm{covrisk} = \frac{1}{m}\sum_{i=1}^{m}[tr((\hat{\Sigma}^{(i)})^{-1}\Sigma) - \log|(\hat{\Sigma}^{(i)})^{-1}\Sigma|] - k,$$

where the superscript $^{(i)}$ denotes the i-th realization, m is the realization times and $\hat{Var}^{(i)}(\hat{\boldsymbol{\beta}})$ is the estimator of $Var^{(i)}(\hat{\boldsymbol{\beta}})$.

Table 1
Simulation results for Model 1.

		ALasso	Lasso	AIC + ALasso	AIC + Lasso	HQ + ALasso	HQ + Lasso
$n = 100$	PMSE	1.2343	1.7528	1.2126	1.3772	1.2006	1.3201
	RE	0.2322	0.7928	0.2272	0.4141	0.2021	0.3695
	CovRisk	0.0668	0.1912	0.0654	0.1462	0.0614	0.0891
	PROP	0.9077	0.7488	0.9410	0.8113	0.9465	0.8191
	C-fit	0.0720	0	0.2110	0.0010	0.2340	0.0010
$n = 200$	PMSE	1.0919	1.1868	1.0882	1.1455	1.0746	1.1108
	RE	0.2029	0.6664	0.2003	0.3159	0.1831	0.2735
	CovRisk	0.0232	0.0377	0.0347	0.0353	0.0204	0.0222
	PROP	0.9477	0.7114	0.9683	0.8131	0.9712	0.8220
	C-fit	0.2750	0	0.4760	0.0020	0.5010	0.0020

Table 2
Simulation results for Model 2.

		ALasso	Lasso	AIC + ALasso	AIC + Lasso	HQ + ALasso	HQ + Lasso
$n = 100$	PMSE	1.1249	1.2462	1.0533	1.0960	1.0157	1.0618
	RE	0.0083	0.0156	0.0042	0.0089	0.0037	0.0084
	CovRisk	0.0732	0.1074	0.0702	0.0797	0.0673	0.0764
	PROP	0.9903	0.9346	0.9979	0.9620	0.9987	0.9637
	C-fit	0.6080	0.0270	0.8870	0.0450	0.9240	0.0460
$n = 200$	PMSE	1.0378	1.0528	1.0368	1.0579	1.0367	1.0576
	RE	0.0093	0.0292	0.0062	0.0171	0.0060	0.0165
	CovRisk	0.0319	0.0376	0.0317	0.0331	0.0311	0.0332
	PROP	0.9982	0.9419	0.9994	0.9636	0.9996	0.9647
	C-fit	0.8980	0.0260	0.9660	0.0410	0.9750	0.0410

Table 3
Simulation results for Model 3.

		ALasso	Lasso	AIC + ALasso	AIC + Lasso	HQ + ALasso	HQ + Lasso
$n = 100$	PMSE	1.1130	1.2189	1.0574	1.2318	1.0514	1.2241
	RE	0.0159	0.0627	0.0128	0.0492	0.0122	0.0485
	CovRisk	0.0705	0.1852	0.0691	0.1419	0.0681	0.1396
	PROP	0.9877	0.9594	0.9937	0.9631	0.9945	0.9630
	C-fit	0.4610	0.0080	0.6930	0.0410	0.7210	0.0420
$n = 200$	PMSE	1.0238	1.0946	1.0041	1.0628	1.0024	1.0607
	RE	0.0213	0.1367	0.0162	0.0910	0.0157	0.0893
	CovRisk	0.0326	0.0672	0.0314	0.0473	0.0313	0.0467
	PROP	0.9969	0.9566	0.9977	0.9618	0.9981	0.9619
	C-fit	0.8150	0.0430	0.8630	0.0590	0.8830	0.0610

Table 4
Simulation results for IS-LM data.

	ALasso	Lasso	AIC + OLS	AIC + ALasso	AIC + Lasso	HQ + OLS	HQ + ALasso	HQ + Lasso
PMSE$^{(1)}$	0.6263	0.6339	1.2697	0.5947	0.6245	0.6940	0.6358	0.5797
PMSE$^{(2)}$	0.0978	0.0763	0.4091	0.1647	0.0779	0.4475	0.0645	0.1513
PMSE$^{(3)}$	0.5424	0.5424	0.7995	0.5324	0.5324	0.7330	0.5320	0.5267
PMSE$^{(all)}$	1.3327	1.2987	2.3874	1.4188	1.2842	1.9957	1.2605	1.3412

In the simulation studies, "PROP" and "C-fit" are used to evaluate the (adaptive) Lasso's capability of correctly selecting subset. "PROP" denotes the averaged proportion of correct specification among all coefficients in the VAR matrices (Hsu et al. (2008)). "C-fit" is the proportion of correct selection of the subset structure in m time simulations, i.e. the proportion of nonzero coefficients estimated to nonzero value and zero coefficients estimated to zero simultaneously in m realizations. In Tables 1–4, "ALasso"/"Lasso" is the simple adaptive Lasso/Lasso procedure for VAR subset selection, "AIC/HQ + ALasso/Lasso" is the hybrid adaptive Lasso/Lasso procedure for VAR subset selection with AIC/HQ criterion selecting the best order first. In the subset selection step, we use the BIC criterion to choose tuning parameters. The sample sizes are $n = 100$ and $n = 200$. Furthermore, (γ, λ) are chosen by a two-dimension grid search for adaptive Lasso. We repeat the simulation 1000 times for each case.

5.1. Simulation studies

We consider the following three models for data generation:

Model 1 $(I - A_1B)(I - A_2B^4)Y_t = \epsilon_t, \epsilon_t \sim N(\mathbf{0}, \Sigma_{\epsilon 1})$,
Model 2 $(I - A_3B - A_4B^2)Y_t = \epsilon_t, \epsilon_t \sim N(\mathbf{0}, \Sigma_{\epsilon 2})$,
Model 3 $(I - A_5B^4)Y_t = \epsilon_t, \epsilon_t \sim N(\mathbf{0}, \Sigma_{\epsilon 2})$,
where

$$A_1 = \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{13} \end{pmatrix}, \quad A_2 = \begin{pmatrix} a_{14} & 0 \\ 0 & a_{15} \end{pmatrix}, \quad \Sigma_{\epsilon 1} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} a_{21} & 0 & 0 \\ a_{22} & 0 & 0 \\ 0 & a_{23} & a_{24} \end{pmatrix}, \quad A_4 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & a_{25} & 0 \\ a_{26} & 0 & 0 \end{pmatrix}, \quad \Sigma_{\epsilon 2} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

and $vec(A_5) = (0.60, -0.55, 0, 0.63, 0, 0.55, 0, 0.64, 0.56)^T$.

Model 1 and Model 2 are the same models as Hsu et al. (2008) considered. Model 1 is a two-dimensional seasonal VAR(5) with period 4 ($k = 2$), Model 2 is a three-dimensional VAR(2) ($k = 3$), and Model 3 is a three-dimensional seasonal VAR(4) with period 4 ($k = 3$). The largest candidate model considered is VAR(8) (i.e., $p_{\max} = 8$).

In the simulation studies, for each simulation, a_{11}–a_{15} in Model 1 and a_{21}–a_{26} in Model 2 are independent generated from $U(0.5, 1)$, and ρ is set to 0.5 for three models. Results of Models 1–3 are summarized in Tables 1–3, respectively.

We observe from Tables 1–3 that:

(1) In terms of the one-step predictive mean square error (PMSE), the parameter estimation precision (RE), the covariance estimation precision (CovRisk), and the subset selection accuracy (PROP and C-fit), and for both Lasso and adaptive Lasso, the hybrid procedures perform better than the simple procedure for VAR subset selection, and the hybrid procedure with HQ criterion performs better than the hybrid procedure with AIC criterion.
(2) In terms of the performances of prediction, parameter estimation, covariance estimation and subset selection accuracy, and for both simple procedure and hybrid procedures, adaptive Lasso outperforms Lasso.
(3) As sample size increases, all subset selection methods improve in term of PMSE, RE, CovRisk, PROP and C-fit.
(4) The bigger the minimum coefficient of a VAR model is, the better a subset selection method performs in subset selection accuracy.
(5) The hybrid procedure HQ + ALasso performs the best among all these subset selection methods. This coincides with the fact that the HQ criterion is subset selection consistent but the AIC criterion is not, and adaptive Lasso has the Oracle property while Lasso does not.

5.2. Real data analysis (application to the IS-LM data)

We now apply our proposed methods to a trivariate dataset "IS-LM.dat" from the Federal Reserve Economic Data (FRED) database maintained at the Federal Reserve Bank of St. Louis. The data can be downloaded at http://www.jmulti.com/data_atse.html. The dataset consists of quarterly and seasonally adjusted observations of the logarithm of real GDP (q_t), the three-month interbank interest rate (i_t), and the logarithm of the real monetary base (m_t) from Q1, 1970 to Q4, 1998, with a sample size $n = 112$. It has been analyzed by Breitung et al. (2004). The augmented Dickey–Fuller tests for non-stationarity of the three variables and their second differences indicate that q_t, i_t and m_t are non-stationary, and their second differences, $Y_t = (\nabla^2 q_t, \nabla^2 i_t, \nabla^2 m_t)^T$, are stationary. The original and 2nd-differenced series are displayed in Fig. 1. The sample autocorrelations (ACF) and cross-correlations (CCF) of Y_t, which show strong evidence of dependence between and within three series, are given in Fig. 2.

Once the data are made stationary, we apply Lasso and adaptive Lasso to Y_t based on model (1) for subset selecting and forecasting. The first 101 observations are used to select subset and model fitting, and the remaining observations are use to evaluate the selected model using a 9-step PMSE. In this example, we set $p_{\max} = 8$. Besides the six subset selection procedures used in the simulation section, we also use the OLS estimator with AIC/HQ criterion selecting the best order as a benchmark. We denote this benchmark method by "AIC/HQ + OLS". For a selected model, we use the empirical PMSE of the individual series and of all series to evaluate the predictive performance using the validation data. The empirical PMSEs are defined as :

$$\mathrm{PMSE}^{(i)} = \frac{1}{9}\sum_{h=1}^{9}[(\hat{\mathbf{Y}}^{(i)}_{101+h} - \mathbf{Y}^{(i)}_{101+h})^T \hat{\Sigma}_{ii}^{-1}(\hat{\mathbf{Y}}^{(i)}_{101+h} - \mathbf{Y}^{(i)}_{101+h})],$$

$$\mathrm{PMSE}_{all} = \frac{1}{9}\sum_{h=1}^{9}[(\hat{\mathbf{Y}}_{101+h} - \mathbf{Y}_{101+h})^T \hat{\Sigma}^{-1}(\hat{\mathbf{Y}}_{101+h} - \mathbf{Y}_{101+h})],$$

where $\hat{\Sigma}$ is the estimator of Σ, and $\hat{\Sigma}_{ii}$'s are the diagonal elements of $\hat{\Sigma}$.

Based on the first 101 data points and $p_{\max} = 8$, AIC and HQ select the best order 8 and 4, respectively. The results of $\mathrm{PMSE}^{(i)}$, $(i = 1, 2, 3)$ and PMSE_{all} are summarized in Table 4.

As we can see from Table 4, all subset selection methods have smaller predictive mean square errors than the ones without shrinkage (AIC/HQ + OLS). This indicates that a good subset selection method for VAR models can improve the

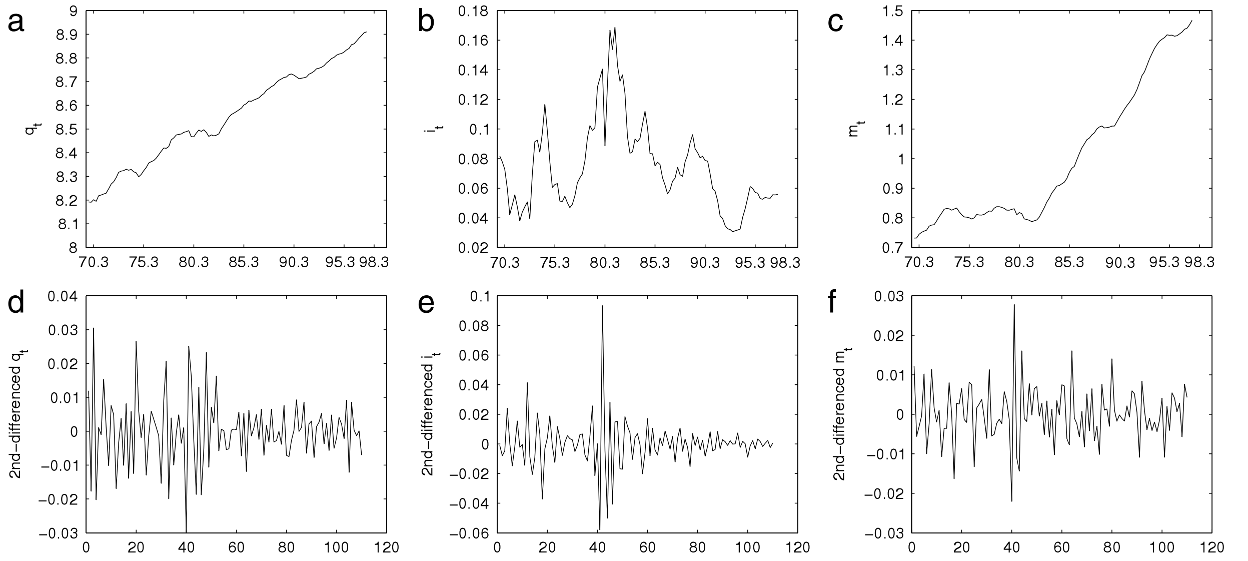
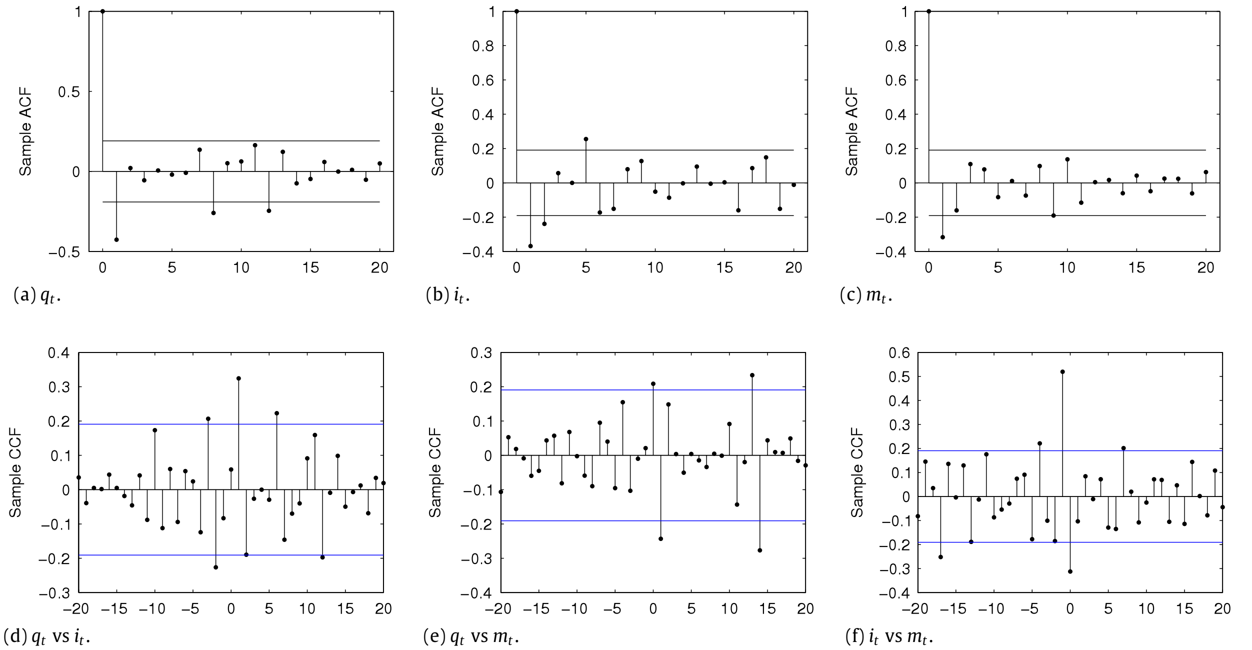


Fig. 1. Quarterly logarithm of real GDP, a three-month interbank interest rate, and the logarithm of the real monetary base from Q1, 1970 to Q4, 1998: (a)–(c) are original series, (d)–(f) are 2nd-differenced series.



(a) q_t.

(b) i_t.

(c) m_t.

(d) q_t vs i_t.

(e) q_t vs m_t.

(f) i_t vs m_t.

Fig. 2. Sample ACF (a)–(c) and Sample CCF (d)–(f) for 2nd-differenced series of quarterly logarithm of real GDP, a three-month interbank interest rate, and the logarithm of the real monetary base from Q1, 1970 to Q4, 1998.

prediction accuracy. Table 4 also shows that among all VAR subset selection methods, the hybrid procedures HQ + ALasso has the best performance in terms of overall prediction accuracy.

6. Conclusion

In this paper, we propose a VAR subset selection procedure based in the adaptive Lasso. Simulation studies and a real data analysis reveal that a hybrid adaptive Lasso procedure is preferable compared with alternatives. For VAR models with collinearity, the adaptive Elastic-net (Zou and Zhang, 2009) might be a better candidate for subset selection than the adaptive Lasso. An investigation into this issue is in order.

Acknowledgements

The authors sincerely thank Dr. Zhiguo Xiao and the referee for their valuable advices and suggestions which greatly improve the manuscript.

The second author's research was partly supported by NSFC Grant 10671037.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Proceedings of the 2nd International Symposium Information Theory. Akademia Kiado, Budapest, pp. 267–281.

Breitung, J., Brüggemann, R., Lütkepohl, H., 2004. Structural vector autoregressive modeling and impulse responses. In: Lütkepohl, H., Krätzig, M. (Eds.), Applied Time Series Econometrics. Cambridge University Press, New York, pp. 159–196.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. The Annals of Statistics 32, 407–499. With discussions, and a rejoinder by the authors.

Fuller, W.A., 1996. Introduction to Statistical Time Series, 2nd ed. John Wiley and Sons, Inc., New York.

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

Hannan, E.J., Quinn, B.C., 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society, Series B 44, 190–195.

Hoerl, A., Kennard, R., 1988. Ridge Regression in Encyclopedia of Statistical Sciences, vol. 8. Wiley, New York, 129–136.

Hsu, N.J., Hung, H.L., Chang, Y.M., 2008. Subset selection for vector autoregressive processes using Lasso. Computational Statistics & Data Analysis 52, 3645–3657.

Knight, K., Fu, W., 2000. Asymptotics for Lasso-type estimators. The Annals of Statistics 28, 1356–1378.

Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer-Verlag, Berlin Heidelberg.

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

Sims, C.A., 1980. Macroeconomics and reality. Econometrica 48, 1–48.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

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

Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters [J]. The Annals of Statistics 37, 1733–1751.