Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 52 (2008) 3645–3657
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

# Subset selection for vector autoregressive processes using Lasso

Nan-Jung Hsu*, Hung-Lin Hung, Ya-Mei Chang

*Institute of Statistics, National Tsing-Hua University, Taiwan*

Received 6 January 2007; received in revised form 12 August 2007; accepted 7 December 2007
Available online 8 January 2008

**Abstract**

A subset selection method is proposed for vector autoregressive (VAR) processes using the Lasso [Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288] technique. Simply speaking, Lasso is a shrinkage method in a regression setup which selects the model and estimates the parameters simultaneously. Compared to the conventional information-based methods such as AIC and BIC, the Lasso approach avoids computationally intensive and exhaustive search. On the other hand, compared to the existing subset selection methods with parameter constraints such as the top-down and bottom-up strategies, the Lasso method is computationally efficient and its result is robust to the order of series included in the autoregressive model. We derive the asymptotic theorem for the Lasso estimator under VAR processes. Simulation results demonstrate that the Lasso method performs better than several conventional subset selection methods for small samples in terms of prediction mean squared errors and estimation errors under various settings. The methodology is applied to modeling U.S. macroeconomic data for illustration.


## 1. Introduction

Multivariate time series processes are of interest in many fields, such as physical sciences, geophysics, meteorology, social sciences, particularly economics, finance and business. For example, one may be interested in the dynamics between different financial markets and try to understand the impact of changes in one market on the others (see for example Tsay (2002) Chapter 8, and the references therein). As another example, Niu and Tiao (1995) analyzed the satellite ozone data for assessing long-term trends in ozone distributions in which spatial data corrected at many locations over time are treated as a multivariate time series. The vector autoregressive (VAR) models are the most popular and successful models for analyzing multiple time series in the literature, because of their simple specification with easy interpretation for the dynamic relationships between series. In order to find a suitable VAR model and produce good predictions, model selection is the key issue in the statistical analysis. Roughly speaking, there are three types of procedures for determining the best VAR model in previous studies. The first type is the information-based methods such as AIC (Akaike, 1974) and BIC (Schwarz, 1978). Usually, the information-based criteria focus only on selecting the best order instead of exhaustive searching among all possible subset structures due to infeasible computations. For example, suppose the multiple time series are $k$-dimensional and the largest order considered in the

* Corresponding author.
*E-mail address:* njhsu@stat.nthu.edu.tw (N.-J. Hsu).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.12.004

fitting procedure is $p$; then the total number of subset VAR models (without intercept) would be $2^{k^2 p}$ which turns out to be a huge number even for small $k$ and $p$. Under these procedures without further reduction, the selected model often contains unnecessary parameters which leads to less efficient parameter estimates, in particular for modeling seasonal or periodic data.

To overcome this computational difficulty, several searching procedures with parameter constrains such as top-down and bottom-up strategies (Lütkepohl, 1991) were proposed. The basic idea of this type of procedure is first to create a search path with nested structures (by adding or deleting one or more parameters at a time). Then sequentially do the selection between two consecutive models on the basis of some criterion (such as AIC values or BIC values). There are many ways to create the search path of models. The top-down strategy starts from the full model (the largest candidate model) with VAR parameters deleted one at a time, whereas the bottom-up strategy starts from the null model with VAR structure added sequentially one series at a time. Both strategies work for individual series separately or for the full multiple system simultaneously. Other similar approaches can be found in Hsiao (1979), Penm and Terrell (1982, 1984), Krolzig and Hendry (2001) and Brüggemann and Lütkepohl (2001). Although such procedures provide a more time-efficient subset VAR selection compared to the exhaustive search, the result is typically sub-optimal, search path dependent and affected by the order of the variables included into the system which is not a desired property.

The third procedure type is based on hypothesis testing, such as the $t$-test for individual AR coefficient or cross-correlation, the multivariate portmanteau test (Hosking, 1980, 1981; Li and McLeod, 1981) and the likelihood ratio test (see for example Anderson (1971) and Tiao and Box (1981)). These procedures not so much select the best model as eliminate inappropriate models. Moreover, these tests are often applied sequentially or simultaneously which makes it difficult to control the nominal sizes of the tests since the test statistics are usually dependent. In addition to the above-mentioned methods, there are many simulation-based methods such as the bootstrap method (Penm et al., 1992), Bayesian MCMC method (Li and Tsay, 1998), resampling method (Chen et al., 1996). Still, most of the alternatives are computationally intensive.

In this work, we are interested in developing a computationally efficient method for VAR subset selection. We first formulate a VAR model in a regression form in which the past variables for all series are the corresponding explanatory variables. In this setup, the subset selection problem becomes a variable selection problem. Motivated by the success of the Lasso (Tibshirani, 1996) method for analyzing data with large numbers of explanatory variables in a regression framework, we propose using Lasso for VAR subset selection. Simply speaking, Lasso is a shrinkage method in the regression setup which selects variables and estimates the parameters simultaneously. For implementation, the least angle regression (LARS) algorithm (Efron et al., 2004) provides fast computation for solving Lasso which has computational order equivalent to that for solving the ordinary least square (OLS) estimate based on the full model. We further derive the theoretical property and investigate the performance of the proposed method by evaluating the estimation errors and prediction errors compared to the conventional approach.

The rest of the paper is organized as follows. In Section 2, we first review the VAR models, their regression formulation and the least square estimation. In Section 3, we introduce the Lasso estimator and several conventional methods and hybrid methods for VAR subset selection. In Section 4, a simulation study is proposed for investigating the effectiveness of the Lasso method in various settings. The methodology is then applied in Section 5 to U.S. macroeconomic data for modeling the dynamic relationships between the unemployment rate, the gross rate of M1 and the nominal GDP. The asymptotics of the Lasso estimator for a VAR model is established in the Appendix.

## 2. VAR models

Consider a $k$-dimensional time series $\{\boldsymbol{y}_t : t = 1, 2, \ldots, n\}$ where $\boldsymbol{y}_t = (y_{1t}, y_{2t}, \ldots, y_{kt})'$. A vector autoregressive model of order $p$ is defined as

$$\boldsymbol{y}_t = \boldsymbol{v} + \Phi_1 \boldsymbol{y}_{t-1} + \cdots + \Phi_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t, \tag{1}$$

where $\Phi_1, \ldots, \Phi_p$ are $k \times k$ coefficient matrices, $\boldsymbol{v}$ is a $k \times 1$ vector of intercept terms, $\{\boldsymbol{u}_t\}$ is a white noise process with covariance matrix $E\boldsymbol{u}_t\boldsymbol{u}_t' = \Sigma_u$. In order to write (1) in a regression form, we further define the following notation:

$$\boldsymbol{Y}^* = (\boldsymbol{y}_{p+1}, \boldsymbol{y}_{p+2}, \ldots, \boldsymbol{y}_n), \qquad \boldsymbol{Y} = \mathrm{vec}(\boldsymbol{Y}^*),$$
$$X_t = (1, \boldsymbol{y}_t', \ldots, \boldsymbol{y}_{t-p+1}')', \qquad \boldsymbol{X}^* = (X_p, \ldots, X_{n-1}),$$

$$\boldsymbol{B} = (\boldsymbol{v}, \Phi_1, \ldots, \Phi_p), \qquad \boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{B}),$$
$$\boldsymbol{U}^* = (\boldsymbol{u}_{p+1}, \ldots, \boldsymbol{u}_n), \qquad \boldsymbol{U} = \mathrm{vec}(\boldsymbol{U}^*).$$

Then, (1) can be written as

$$\boldsymbol{Y}^* = \boldsymbol{B}\boldsymbol{X}^* + \boldsymbol{U}^*,$$

or equivalently,

$$\boldsymbol{Y} = \left((\boldsymbol{X}^*)' \otimes \boldsymbol{I}_k\right)\boldsymbol{\beta} + \boldsymbol{U} \equiv \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}, \tag{2}$$

where $\boldsymbol{I}_k$ is the $k$-dimensional identity matrix and the covariance matrix of $\boldsymbol{U}$ is $\Sigma_U = \boldsymbol{I}_{n-p} \otimes \Sigma_u$.

Under the regression setup in (2), the least square (LS) estimators of the parameters satisfy

$$\hat{\boldsymbol{\beta}} = \left[\left(\boldsymbol{X}^*(\boldsymbol{X}^*)'\right)^{-1}\boldsymbol{X}^* \otimes \boldsymbol{I}_k\right]\boldsymbol{Y},$$
$$\hat{\Sigma}_u = \frac{1}{n-p}(\boldsymbol{Y}^* - \hat{\boldsymbol{B}}\boldsymbol{X}^*)(\boldsymbol{Y}^* - \hat{\boldsymbol{B}}\boldsymbol{X}^*)',$$

which minimize the weighted least squares:

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\,\Sigma_U^{-1}\,(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\,.$$

The LS estimator $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal under some regularity conditions (Fuller, 1996, Chapter 8).

## 3. Subset selection procedures

We first introduce the Lasso method for VAR subset selection in Section 3.1. Some conventional methods are described in Section 3.2. Several hybrid methods, which combine more than one strategy of Sections 3.1 and 3.2, are suggested in Section 3.3.

### *3.1. Lasso for VAR subset selection*

Tibshirani (1996) proposed the Lasso method (least absolute shrinkage and selection operator) in a linear regression setup which is the least square method with an $L_1$ constraint on the regression parameters. In this work, we adopt the Lasso method for VAR subset selection. Under the regression formulation (2) for VAR models, the Lasso estimator $\tilde{\boldsymbol{\beta}}$ is obtained by minimizing the residual sum of squares:

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\,(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\,,$$

with respect to $\boldsymbol{\beta}$ subject to the constraint $\sum|\beta_j| \le s$, where $\beta_j$ is the $j$th element of $\boldsymbol{\beta}$ and $s$ is a tuning parameter. Equivalently, $\tilde{\boldsymbol{\beta}}$ can be solved by minimizing

$$Q(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\,(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum|\beta_j|, \tag{3}$$

where $\lambda$ is a tuning parameter depending on $s$. If $s$ is large enough, the Lasso estimate is exactly the OLS estimate. If $s$ is small, the Lasso estimate is shrunken toward zero relative to the OLS estimate. In particular, less significant elements in $\tilde{\boldsymbol{\beta}}$ are shrunken to zero so that a sub-model is obtained at the same time. Therefore, Lasso has the advantage of estimating parameters and selecting variables (the past variables of the multiple series) simultaneously. For choosing the tuning parameter $\lambda$, we use cross-validation in this work as suggested in Tibshirani (1996).

The Lasso method has been generalized and improved in many different ways, by such as Fan and Li (2001) and Zou (2006), and successfully applied to applications in many different fields. Recently, Wang et al. (2007) adopted the Lasso method to the regression model with AR errors for the univariate case. Following Knight and Fu (2000), we derive the asymptotics of the Lasso estimator for a VAR model in the Appendix. For implementation, the Lasso estimate in (3) can be solved efficiently using the LARS algorithm (Efron et al., 2004) which has been written in the "lars" package of R and can be downloaded from the Comprehensive R Archive Network (CRAN) website at http://cran.r-project.org.

### 3.2. Conventional procedures for VAR subset selection

In this section, the information-based criteria AIC and BIC for selecting the order and two reduction methods with parameter constraints, top-down and bottom-up strategies, are described.

Under the VAR model in (1), AIC and BIC select the best model with order

$$\hat{p}_{\text{aic}} = \arg\min_{p} \left\{ n \ln \left| \tilde{\Sigma}_u(p) \right| + 2pk^2 \right\}, \tag{4}$$

$$\hat{p}_{\text{bic}} = \arg\min_{p} \left\{ n \ln \left| \tilde{\Sigma}_u(p) \right| + (\ln n) pk^2 \right\}, \tag{5}$$

where $\tilde{\Sigma}_u(p)$ is the maximum likelihood estimate of $\Sigma_u$ under the fitted VAR($p$) model. The difference between AIC and BIC is in the penalized factor for model complexity. BIC tends to select lower order models than AIC since the penalized factor $(\ln n)$ for BIC is larger than the penalized factor 2 for AIC. Usually, the information-based criteria are used for order selection only instead of subset selection for VAR models since the total subset models are numerous and the corresponding computations are infeasible.

To avoid an exhaustive search among all possible subset models, top-down and bottom-up strategies create different search paths for subset selection. Both algorithms can be applied for each individual series one at a time or for the full system simultaneously. In the following, we describe how these strategies work to reduce a VAR($p$) model for each individual series. For further descriptions, we need more notation. Define $\boldsymbol{\beta}_j$ to be the parameter vector in $\boldsymbol{\beta}$ associated with the $j$th series which is a $(kp+1)$-dimensional vector corresponding to the $j$th row in the matrix $\boldsymbol{B}$ defined in Section 2. The $i$th element in $\boldsymbol{\beta}_j$ is denoted by $\boldsymbol{\beta}_{j,i}$. Let $\mathcal{M}(\boldsymbol{\beta}_j)$ denote the full AR($p$) model which consists of all parameters in $\boldsymbol{\beta}_j$ for the $j$th series and $\mathcal{M}(\boldsymbol{\beta}_j \setminus \{\beta_{j,i_1}, \ldots, \beta_{j,i_m}\})$ denote a subset model of $\mathcal{M}(\boldsymbol{\beta}_j)$ with $\beta_{j,i_1} = \cdots = \beta_{j,i_m} = 0$.

*Top-down strategy*

The subset model for the $j$th series is selected via the following algorithm:

1. initial setting: $\mathcal{A} = \boldsymbol{\beta}_j$, $i = kp+1$,
2. if $AIC(\mathcal{M}(\mathcal{A} \setminus \beta_{j,i})) \le AIC(\mathcal{M}(\mathcal{A}))$, update $\mathcal{A}$ to $\mathcal{A} \setminus \beta_{j,i}$,
3. update $i$ to $i-1$,
4. repeat steps 2–3 until $i = 0$.

The final selected model for the $j$th series contains exploratory variables associated with nonzero coefficients in $\mathcal{A}$. Note that, the criterion AIC in step 2 can be replaced by BIC or any other criterion. Basically, this algorithm starts from the full model and then reduces the model by removing one parameter at a time according to some criterion value. The parameter is permanently excluded if its removal does improve the information criterion value compared to the model in the previous step.

*Bottom-up strategy*

The subset model for the $j$th series is selected via the following algorithm:

1. initial setting: $\mathcal{A} = \emptyset$, $i = 1$,
2. keep the explanatory variables in $\mathcal{A}$ and add the past variables of the $i$th series as extra explanatory variables; select the optimal order only for the new added series based on AIC, say $p_{j,i}$,
3. update $\mathcal{A}$ by including the past variables of the $i$th series up to lag $p_{j,i}$,
4. update $i$ to $i+1$,
5. repeat steps 2–4 until $i = k$.

The final selected model for the $j$th series contains variables in $\mathcal{A}$ which includes the past variables of $i$th series up to lag $p_{j,i}$ for $i = 1, 2, \ldots, k$. That is,

$$y_{jt} = \nu_j + \sum_{\ell=1}^{p_{j,1}} \phi_{1,j,\ell} y_{1,t-\ell} + \sum_{\ell=1}^{p_{j,2}} \phi_{2,j,\ell} y_{2,t-\ell} + \cdots + \sum_{\ell=1}^{p_{j,k}} \phi_{k,j,\ell} y_{k,t-\ell} + u_{jt},$$

for some coefficients $\nu_j$ and $\{\phi_{i,j,\ell}\}$. Similarly, AIC in step 2 can be replaced by BIC or any other criterion. Basically, this algorithm starts from the null model and then expends the exploratory variables by sequentially adding a univariate AR structure from the multiple series.

The top-down and the bottom-up procedures provide two different methods for VAR subset selection; their results are typically sub-optimal and affected by the search paths created (which depend on the order of the series included in the system). These two strategies can also work together to produce a more parsimonious subset model. More details about these two strategies can be found in Lütkepohl (1991, Chapter 5) and Brüggemann (2004, Chapter 2).

### 3.3. Hybrid procedures

Beside combining bottom-up and top-down strategies, we suggest several hybrid methods in this section which combine more than one of the above-mentioned strategies, including the information-based criteria, Lasso, top-down and bottom-up ones. Comparisons among them are investigated in Section 4.

*Bottom-up joint with top-down strategy (BU + TD):*

1. use the bottom-up strategy to select the model for each series,
2. use the top-down strategy to reduce the selected model for each series.

*AIC joint with top-down strategy (AIC + TD):*

1. use AIC to select the best order for VAR model fitting in which multiple series are considered simultaneously,
2. use the top-down strategy to reduce the VAR($\hat{p}_{\text{aic}}$) model for each series.

*AIC joint with Lasso for multiple series (AIC + Lasso-f):*

1. use AIC to select the best order for VAR model fitting in which multiple series are considered simultaneously,
2. find Lasso estimates for multiple series (full system) under the VAR($\hat{p}_{\text{aic}}$) model.

*AIC joint with Lasso for individual series (AIC + Lasso-s):*

1. use AIC to select the best order for VAR model fitting in which multiple series are considered simultaneously,
2. find Lasso estimates for each individual series under the VAR($\hat{p}_{\text{aic}}$) model (i.e., solve (3) for each series separately).

## 4. Simulation study

A simulation study is conducted to investigate the performance of eight VAR selection procedures described in Section 3, including AIC, BIC, AIC + TD, BU + TD, Lasso for individual series (Lasso-s), Lasso for multiple series (Lasso-f), AIC+Lasso-s and AIC+Lasso-f. For all Lasso related procedures, we use a tenfold cross-validation method to choose the tuning parameter $\lambda$ in (3). The performance is evaluated using the normalized one-step prediction mean squared error (PMSE) and the estimation precision of the selected model, defined as follows:

$$\begin{aligned}
\text{PMSE} &= \frac{1}{k} E\left[(\boldsymbol{y}_{n+1} - \hat{\boldsymbol{y}}_{n+1})' \Sigma_1^{-1} (\boldsymbol{y}_{n+1} - \hat{\boldsymbol{y}}_{n+1})\right] \\
&= 1 + \frac{1}{k} E\left[\left(\hat{\boldsymbol{y}}^*_{n+1} - \hat{\boldsymbol{y}}_{n+1}\right)' \Sigma_1^{-1} \left(\hat{\boldsymbol{y}}^*_{n+1} - \hat{\boldsymbol{y}}_{n+1}\right)\right], \\
\text{RE} &= \frac{1}{k^2 p_{\max} + k} \sum_{j=1}^{k^2 p_{\max}+k} \frac{E(\hat{\beta}_j - \beta_j)^2}{E_\infty\left(\hat{\beta}^{**}_j - \beta_j\right)^2}, \\
\text{CovRisk} &= E\left[\text{tr}\left(\hat{\Sigma}_u^{-1}\Sigma_u\right) - \log\left|\hat{\Sigma}_u^{-1}\Sigma_u\right| - k\right],
\end{aligned}$$

where $\hat{\boldsymbol{y}}_{n+1}$, $\hat{\beta}_j$ and $\hat{\Sigma}_u$ are the one-step best linear prediction, the estimate of $\beta_j$ (the $j$th element in $\boldsymbol{\beta}$) and the estimate of $\Sigma_u$ for the selected model respectively, $\hat{\boldsymbol{y}}^*_{n+1}$ and $\Sigma_1$ are the one-step best linear prediction and the associated prediction variance based on the true model with known parameters, $p_{\max}$ is the largest order considered in the model fitting procedures, $\hat{\beta}^{**}_j$ is the estimate of $\beta_j$ under the largest candidate model VAR($p_{\max}$) and $E_\infty\left(\hat{\beta}^{**}_j - \beta_j\right)^2$ is the asymptotic variance of $\hat{\beta}^{**}_j$ under VAR($p_{\max}$) which can be computed analytically. The PMSE has the smallest value at one which is attained when $\hat{\boldsymbol{y}}_{n+1}$ is computed under the true model with known parameters (i.e., $\hat{\boldsymbol{y}}_{n+1} = \hat{\boldsymbol{y}}^*_{n+1}$). The RE measures the averaged relative efficiency of parameter estimates with respect to the situation without model

reduction. The RE has the worst value at one when no selection is made (i.e., $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{**}$). The CovRisk measures the precision for estimating $\Sigma_u$ which has the smallest value at zero when the selected model is the true model with known parameters. We are also interested in the performance of correctly identifying the subset structure (i.e., the locations of nonzero coefficients) by each selection procedure, which is evaluated by the averaged proportion (abbreviated as PROP hereafter) of correct specification among all coefficients in the AR matrices. The PROP has the maximal value at one which indicates the selected subset structure is completely correct. Consequently, a procedure with higher value of PROP has better performance in identifying the subset structure.

We consider the following three model structures for data generation:

$$\text{Model 1} : (I - A_1 B)(I - A_2 B^4)\boldsymbol{y}_t = \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim N(\mathbf{0}, \Sigma_{u1}),$$
$$\text{Model 2} : (I - A_3 B - A_4 B^2)\boldsymbol{y}_t = \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim N(\mathbf{0}, \Sigma_{u2}),$$
$$\text{Model 3} : \boldsymbol{y}_t = (I + A_1 B)\boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim N(\mathbf{0}, \Sigma_{u1}),$$

where

$$A_1 = \begin{pmatrix} a_{11} & a_{12} \\ 0 & a_{13} \end{pmatrix}, \qquad A_2 = \begin{pmatrix} a_{21} & 0 \\ 0 & a_{22} \end{pmatrix}, \qquad \Sigma_{u1} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$
$$A_3 = \begin{pmatrix} a_{31} & 0 & 0 \\ a_{32} & 0 & 0 \\ 0 & a_{33} & a_{34} \end{pmatrix}, \qquad A_4 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & a_{41} & 0 \\ a_{42} & 0 & 0 \end{pmatrix}, \qquad \Sigma_{u2} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.$$

Model 1 is a two-dimensional seasonal VAR(5) with period 4 ($k = 2$); Model 2 is a three-dimensional VAR(2) ($k = 3$); Model 3 is a two-dimensional vector moving average (VMA) with order 1 which can be represented as a VAR with infinite order. In order to cover various situations, 100 sets of the parameters $\{a_{ij}\}$ and $\rho$ are randomly generated in which each parameter is generated independently from the uniform distribution on (0, 1). For each set of parameters, 100 realizations with sample size 100 are generated from the model and eight selection procedures are applied in which the largest candidate model considered is VAR(8) (i.e., $p_{\max} = 8$).

The values of PMSE, RE and CovRisk for each procedure under the model with each set of generated parameters are evaluated numerically using the following quantities:

$$\text{pmse} = 1 + \frac{1}{100}\sum_{i=1}^{100}\left(\hat{\boldsymbol{y}}_{n+1}^{*(i)} - \hat{\boldsymbol{y}}_{n+1}^{(i)}\right)' \Sigma_1^{-1}\left(\hat{\boldsymbol{y}}_{n+1}^{*(i)} - \hat{\boldsymbol{y}}_{n+1}^{(i)}\right),$$

$$\text{re} = \frac{1}{100(k^2 p_{\max} + k)}\sum_{i=1}^{100}\sum_{j=1}^{k^2 p_{\max}+k}\frac{\left(\hat{\beta}_j^{(i)} - \beta_j\right)^2}{E_\infty\left(\hat{\beta}_j^{**} - \beta_j\right)^2},$$

$$\text{covrisk} = \frac{1}{100}\sum_{i=1}^{100}\left\{\text{tr}\left[\left(\hat{\Sigma}_u^{(i)}\right)^{-1}\Sigma_u\right] - \log\left|\left(\hat{\Sigma}_u^{(i)}\right)^{-1}\Sigma_u\right| - k\right\},$$

where $\hat{\boldsymbol{y}}_{n+1}^{(i)}$, $\hat{\beta}_j^{(i)}$ and $\hat{\Sigma}_u^{(i)}$ are the one-step best linear prediction, and the estimates of $\beta_j$ and $\Sigma_u$ under the selected model for the $i$th realization, $\hat{\boldsymbol{y}}_{n+1}^{*(i)}$ is the one-step best linear prediction for the $i$th realization based on the true model with known parameters. Finally, the performance of eight selection procedures for each model is evaluated via the values of pmse, re and covrisk averaged over 100 sets of generated parameters. Similarly, the PROP criterion for each procedure is evaluated via the empirical proportion of correctly identifying the subset structure averaged over realizations with different parameters.

The results for VAR Models 1 and 2 are summarized in Table 1 and Table 2 respectively, including the empirical values of PMSE, RE, CovRisk, PROP, their standard errors and the ranks among eight selection methods (rank = 1 indicating the best and rank = 8 indicating the worst). The results for VMA Model 3 are summarized in Table 3 in which only the empirical PMSE and CovRisk are reported since the true model is not included in the candidate models. On the basis of these empirical results, we have the following conclusions.

- For almost all considered cases, AIC + Lasso-f performs best among eight selection procedures. The AIC + TD procedure also works very well, especially in the aspect of correctly identifying the subset structure.

Table 1
Empirical PMSE, RE, CovRisk, PROP, their standard errors and ranks among different selection methods for the two-dimensional VAR(5) models with various parameter values

| | | AIC | BIC | AIC + TD | BU + TD | Lasso-s | Lasso-f | AIC + Lasso-s | AIC + Lasso-f |
|---|---|---|---|---|---|---|---|---|---|
| $\rho = 0$ | PMSE | 1.170 | 1.193 | 1.140 | 1.148 | 1.168 | 1.160 | 1.140 | 1.131 |
| | s.e. | 0.002 | 0.003 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 |
| | Rank | 7 | 8 | 2 | 4 | 6 | 5 | 3 | 1 |
| | RE | 0.744 | 0.821 | 0.608 | 0.639 | 0.778 | 0.761 | 0.607 | 0.582 |
| | s.e. | 0.005 | 0.005 | 0.004 | 0.004 | 0.003 | 0.003 | 0.003 | 0.003 |
| | Rank | 5 | 8 | 3 | 4 | 7 | 6 | 2 | 1 |
| | CovRisk | 0.074 | 0.059 | 0.061 | 0.062 | 0.071 | 0.062 | 0.061 | 0.055 |
| | s.e. | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | Rank | 8 | 2 | 3 | 6 | 7 | 5 | 4 | 1 |
| | PROP | 0.600 | 0.685 | 0.820 | 0.816 | 0.651 | 0.635 | 0.740 | 0.725 |
| | s.e. | 0.005 | 0.005 | 0.004 | 0.004 | 0.005 | 0.005 | 0.004 | 0.004 |
| | Rank | 8 | 5 | 1 | 2 | 6 | 7 | 3 | 4 |
| $\rho \sim U(0, 1)$ | PMSE | 1.177 | 1.209 | 1.151 | 1.162 | 1.218 | 1.187 | 1.179 | 1.156 |
| | s.e. | 0.002 | 0.003 | 0.002 | 0.002 | 0.004 | 0.002 | 0.003 | 0.002 |
| | Rank | 4 | 7 | 1 | 3 | 8 | 6 | 5 | 2 |
| | RE | 0.716 | 0.724 | 0.540 | 0.540 | 0.604 | 0.614 | 0.494 | 0.497 |
| | s.e. | 0.005 | 0.004 | 0.004 | 0.004 | 0.003 | 0.003 | 0.003 | 0.003 |
| | Rank | 7 | 8 | 3 | 4 | 5 | 6 | 1 | 2 |
| | CovRisk | 0.076 | 0.062 | 0.059 | 0.060 | 0.079 | 0.063 | 0.067 | 0.057 |
| | s.e. | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | Rank | 7 | 4 | 2 | 3 | 8 | 5 | 6 | 1 |
| | PROP | 0.584 | 0.675 | 0.803 | 0.795 | 0.659 | 0.643 | 0.741 | 0.728 |
| | s.e. | 0.005 | 0.005 | 0.004 | 0.004 | 0.005 | 0.005 | 0.004 | 0.004 |
| | Rank | 8 | 5 | 1 | 2 | 6 | 7 | 3 | 4 |

- For Lasso related procedures, using AIC to reduce the model first greatly improves the performance. Moreover, applying Lasso to multiple series simultaneously performs better than applying Lasso to individual series. When the errors are dependent between series (i.e., $\Sigma_u$ is not diagonal), the advantage of using the Lasso method among other methods becomes small since $\Sigma_u$ is not incorporated in (3). But, a modified version of (3)

$$Q^*(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' \, \Sigma_U^{-1} \, (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum |\beta_j|$$

  can be considered to further improve the efficiency of the Lasso approach.
- Generally speaking, Lasso procedures are more effective compared to the conventional methods when the model structures are sparse. That is why AIC + Lasso-f is less effective and becomes competitive with other selection methods for the VMA example since the underlying process has a saturated AR structure. Besides, Lasso procedures has fewer advantages for large samples since the estimates are biased due to shrinkage.

We also examine whether the performance of each selection procedure is affected by the sample size. In particular, we look at the PMSE values for different procedures across different sample sizes displayed in Fig. 1 for Models 1 and 3 with various parameter values and $\rho = 0$. For better visualization, the PMSE values in these plots are only displayed for five methods, including AIC, BIC, AIC + TD, Lasso-f and AIC + Lasso-f. Clearly, AIC + Lasso-f outperforms other methods especially for small samples. Similar patterns are found for Models 1 and 3 with $\rho \neq 0$.

We end this simulation study by comparing the computational efficiencies in terms of the CPU time required for implementing each selection procedure. The results are summarized in Table 4 for Models 1–3 which are obtained based on R programs running on a Pentium-4 (3.4 GHz) PC. The AIC and BIC methods are extremely fast since they only determine the order instead of the subset model structures (otherwise, for exhaustive search, the computation order is $O(2^{k^2 p_{\max}})$!). The speed for AIC + TD depends on the model structures but it is surprisingly fast. The speed for BU + TD depends on the data dimension $k$ which is slower when $k$ is larger. For Lasso related procedures, the computational time depends on how the cross-validation is performed for choosing the tuning parameter $\lambda$ in (3). We use tenfold cross-validation in this simulation study; therefore the computation order for the Lasso procedure is about

Table 2
Empirical PMSE, RE, CovRisk, PROP, their standard errors and ranks among different selection methods for the three-dimensional VAR(2) models with various parameter values

| | | AIC | BIC | AIC + TD | BU + TD | Lasso-s | Lasso-f | AIC + Lasso-s | AIC + Lasso-f |
|---|---|---|---|---|---|---|---|---|---|
| $\rho = 0$ | PMSE | 1.097 | 1.101 | 1.074 | 1.135 | 1.115 | 1.111 | 1.074 | 1.064 |
| | s.e. | 0.001 | 0.001 | 0.001 | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 |
| | Rank | 4 | 5 | 3 | 8 | 7 | 6 | 2 | 1 |
| | RE | 0.265 | 0.269 | 0.193 | 0.379 | 0.240 | 0.249 | 0.169 | 0.139 |
| | s.e. | 0.002 | 0.002 | 0.002 | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 |
| | Rank | 6 | 7 | 3 | 8 | 4 | 5 | 2 | 1 |
| | CovRisk | 0.100 | 0.087 | 0.090 | 0.127 | 0.116 | 0.114 | 0.085 | 0.079 |
| | s.e. | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | Rank | 5 | 3 | 4 | 8 | 7 | 6 | 2 | 1 |
| | PROP | 0.820 | 0.857 | 0.945 | 0.899 | 0.820 | 0.745 | 0.912 | 0.913 |
| | s.e. | 0.004 | 0.003 | 0.002 | 0.003 | 0.004 | 0.004 | 0.003 | 0.003 |
| | Rank | 7 | 5 | 1 | 4 | 6 | 8 | 3 | 2 |
| $\rho \sim U(0, 1)$ | PMSE | 1.094 | 1.095 | 1.078 | 1.136 | 1.158 | 1.140 | 1.095 | 1.077 |
| | s.e. | 0.001 | 0.001 | 0.001 | 0.002 | 0.002 | 0.002 | 0.001 | 0.001 |
| | Rank | 3 | 4 | 2 | 6 | 8 | 7 | 5 | 1 |
| | RE | 0.249 | 0.217 | 0.155 | 0.243 | 0.185 | 0.251 | 0.137 | 0.122 |
| | s.e. | 0.002 | 0.001 | 0.002 | 0.002 | 0.001 | 0.002 | 0.001 | 0.001 |
| | Rank | 7 | 5 | 3 | 6 | 4 | 8 | 2 | 1 |
| | CovRisk | 0.103 | 0.089 | 0.089 | 0.119 | 0.118 | 0.126 | 0.084 | 0.079 |
| | s.e. | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | Rank | 5 | 4 | 3 | 7 | 6 | 8 | 2 | 1 |
| | PROP | 0.815 | 0.850 | 0.940 | 0.894 | 0.820 | 0.668 | 0.907 | 0.907 |
| | s.e. | 0.004 | 0.004 | 0.002 | 0.003 | 0.004 | 0.004 | 0.003 | 0.003 |
| | Rank | 7 | 5 | 1 | 4 | 6 | 8 | 3 | 2 |

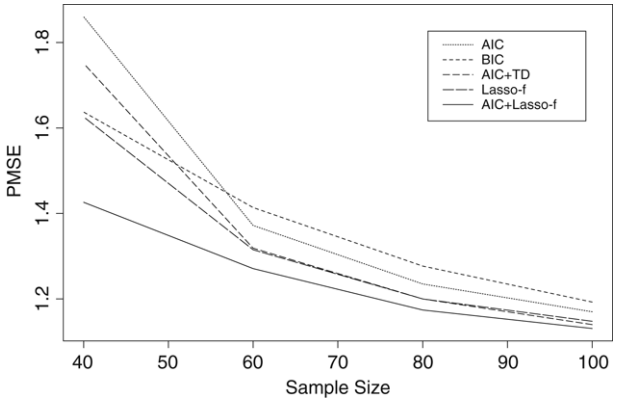
Table 3
Empirical PMSE, CovRisk, their standard errors and ranks among different selection methods for the two-dimensional VMA(1) models with various parameter values

| | | AIC | BIC | AIC + TD | BU + TD | Lasso-s | Lasso-f | AIC + Lasso-s | AIC + Lasso-f |
|---|---|---|---|---|---|---|---|---|---|
| $\rho = 0$ | PMSE | 1.289 | 1.310 | 1.282 | 1.282 | 1.310 | 1.316 | 1.277 | 1.279 |
| | s.e. | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 |
| | Rank | 5 | 7 | 3 | 4 | 6 | 8 | 1 | 2 |
| | CovRisk | 0.057 | 0.049 | 0.051 | 0.051 | 0.055 | 0.054 | 0.047 | 0.048 |
| | s.e. | 0.001 | 0.000 | 0.001 | 0.000 | 0.001 | 0.001 | 0.000 | 0.000 |
| | Rank | 8 | 3 | 5 | 4 | 7 | 6 | 1 | 2 |
| $\rho \sim U(0, 1)$ | PMSE | 1.322 | 1.349 | 1.319 | 1.315 | 1.360 | 1.356 | 1.325 | 1.318 |
| | s.e. | 0.008 | 0.008 | 0.008 | 0.008 | 0.009 | 0.009 | 0.008 | 0.008 |
| | Rank | 4 | 6 | 3 | 1 | 8 | 7 | 5 | 2 |
| | CovRisk | 0.058 | 0.051 | 0.050 | 0.050 | 0.057 | 0.057 | 0.048 | 0.050 |
| | s.e. | 0.001 | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 | 0.000 | 0.000 |
| | Rank | 8 | 5 | 4 | 3 | 7 | 6 | 1 | 2 |

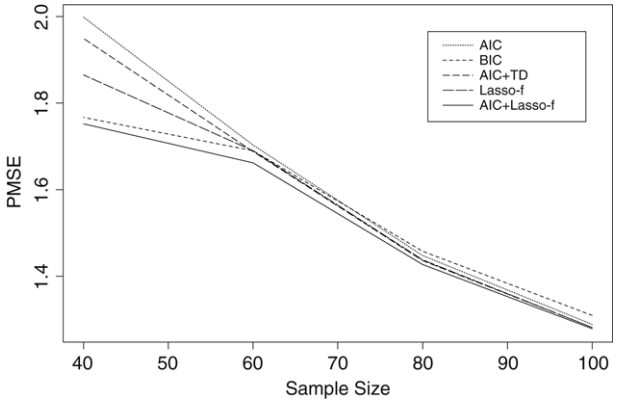
10 times that for the AIC method. Moreover, like for BU + TD, Lasso procedures for individual series (Lasso-s and AIC + Lasso-s) take more time when the data dimension is larger.

## 5. Application

In this section, our methodology is applied to model the relationships between the unemployment rate (UER), the gross rate of M1 and nominal GDP in the U.S. for illustration. The data are three-dimensional multiple series with sample size $n = 185$, consisting of quarterly data from 1960 to 2006. The first 176 data points are used for model fitting and selection; the last 9 data points are left for model evaluation. We first eliminate nonstationarity by




(a) Model 1: VAR(5).




(b) Model 3: VMA(1).

Fig. 1. Empirical PMSE for various selection procedures applied to data with different sample sizes under Models 1 and 3 with various parameter values and $\rho = 0$.

Table 4
Average CPU time (in seconds) required for each selection procedure applied to a realization with sample size 100 under Models 1–3

| Model | $k$ | AIC | BIC | AIC + TD | BU + TD | Lasso-s | Lasso-f | AIC + Lasso-s | AIC + Lasso-f |
|---|---|---|---|---|---|---|---|---|---|
| VAR(5) | 2 | 0.017 | 0.017 | 0.029 | 0.055 | 0.718 | 0.691 | 0.485 | 0.428 |
| VAR(2) | 3 | 0.018 | 0.018 | 0.028 | 0.113 | 1.528 | 1.779 | 0.536 | 0.416 |
| VMA(1) | 2 | 0.017 | 0.017 | 0.022 | 0.047 | 0.717 | 0.690 | 0.388 | 0.330 |

carrying out first-order differencing for each individual series. The original series, differenced series and their sample autocorrelations (ACF), cross-correlations (CCF) are displayed in Fig. 2 and Fig. 3, respectively. The sample ACF and CCF show strong evidence of dependence between and within three series.

We then fit VAR model and apply the above-mentioned model selection procedures ($p_{\max} = 10$) to the differenced series. It turns out that AIC selects VAR(8) and BIC selects VAR(4) as the best models; AIC + TD, AIC + Lasso-s and AIC + Lasso-f further reduce VAR(8) with the proportions of zero coefficients 0.56, 0.293 and 0.413 respectively in the regression matrices. The selected models and the proportions of reduction for all methods are reported in Table 5. According to these selected models, we further evaluate the prediction performance for the validation data in terms of the empirical PMSE for individual series and all series defined as follows:

$$\text{PMSE}_i = \frac{1}{9}\sum_{h=1}^{9}(y_{i,176+h} - \hat{y}_{i,176+h})^2, \quad i = 1, 2, 3,$$

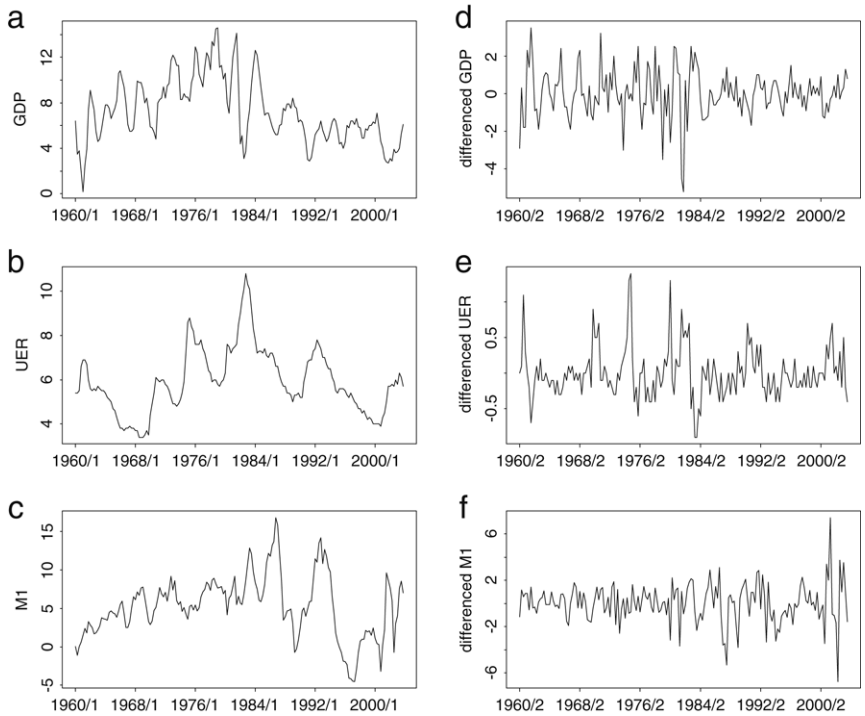



Fig. 2. Quarterly unemployment rates, gross rates of M1 and the nominal GDP in the U.S. from 1960 to 2003: (a)–(c) original series, (d)–(f) differenced series.

Table 5
Selected models, the proportions of zero coefficients, the out-of-sample PMSE and their ranks for the differenced U.S. macroeconomic data under various model selection procedures

| | AIC | BIC | AIC + TD | BU + TD | Lasso-s | Lasso-f | AIC + Lasso-s | AIC + Lasso-f |
|---|---|---|---|---|---|---|---|---|
| Selected VAR order | 8 | 4 | 8 | 9 | 10 | 10 | 8 | 8 |
| proportion of zero | – | – | 0.560 | 0.699 | 0.301 | 0.355 | 0.293 | 0.413 |
| $\text{PMSE}_1$ | 0.673 | 0.219 | 0.344 | 0.208 | 0.239 | 0.290 | 0.191 | 0.224 |
| Rank | 8 | 3 | 7 | 2 | 5 | 6 | 1 | 4 |
| $\text{PMSE}_2$ | 0.090 | 0.040 | 0.051 | 0.035 | 0.046 | 0.028 | 0.039 | 0.024 |
| Rank | 8 | 5 | 7 | 3 | 6 | 2 | 4 | 1 |
| $\text{PMSE}_3$ | 2.498 | 1.775 | 1.986 | 2.059 | 2.344 | 2.376 | 1.891 | 1.495 |
| Rank | 8 | 2 | 4 | 5 | 6 | 7 | 3 | 1 |
| $\text{PMSE}_{\text{all}}$ | 4.026 | 2.400 | 2.676 | 2.175 | 2.781 | 2.493 | 2.345 | 1.745 |
| Rank | 8 | 4 | 6 | 2 | 7 | 5 | 3 | 1 |

$$\text{PMSE}_{\text{all}} = \frac{1}{9}\sum_{h=1}^{9}(\boldsymbol{y}_{176+h} - \hat{\boldsymbol{y}}_{176+h})' \hat{\Sigma}_{\text{aic}}^{-1}(\boldsymbol{y}_{176+h} - \hat{\boldsymbol{y}}_{176+h}),$$

where $\hat{\Sigma}_{\text{aic}}$ is the estimate of $\Sigma_u$ based on AIC which is used to adjust the scales for different series. These out-of-sample PMSEs are reported in Table 5. It turns out that AIC + Lasso-f performs best in terms of overall predictions, which is consistent with the simulation results obtained in Section 4.

## 6. Conclusions

We adopt Lasso method for VAR subset selection, which has good performance in forecasting compared to other conventional methods in finite samples. In addition, this method can be implemented very fast using the LARS

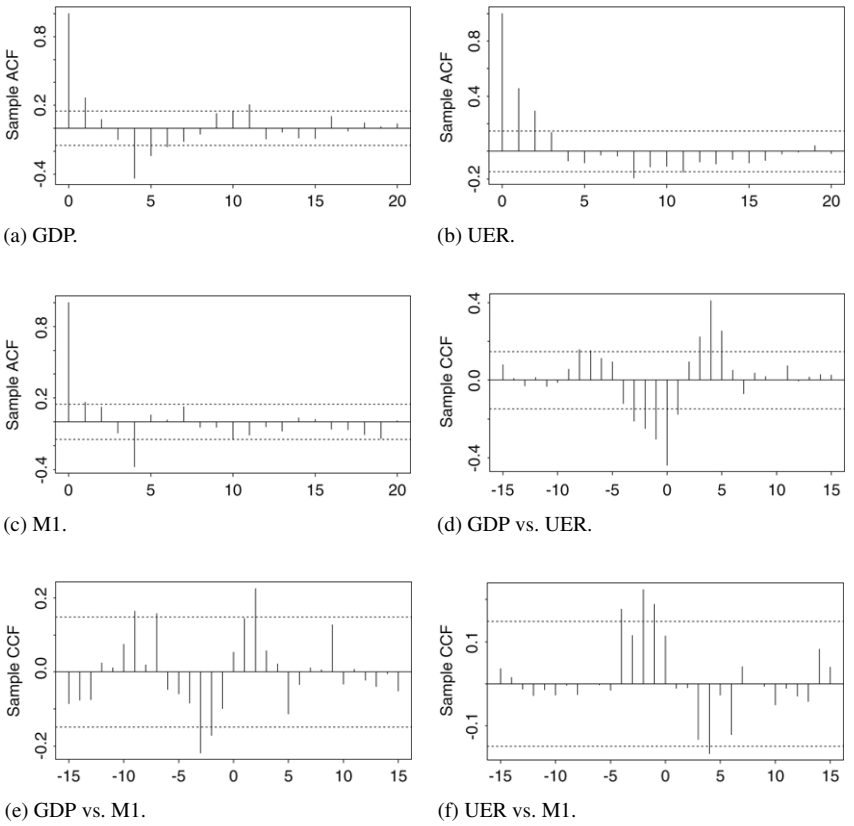



Fig. 3. Sample ACF (a)–(c) and sample CCF (d)–(f) for the differenced series of the unemployment rates, gross rates of M1 and the nominal GDP in the U.S. from 1960 to 2003.

algorithm and, therefore, is computationally efficient even for high dimensional data fitted to high order VAR models. The methodology was applied to modeling the dynamic relationships between the unemployment rates, gross rates of M1 and the nominal GDP in the U.S. As shown for both the simulated and real examples, the proposed AIC + Lasso-f procedure has good performance in forecasting and estimation. For further research, we are currently extending our methodology for vector ARMA subset selection.

## Acknowledgements

This research was supported by the National Science Council (NSC 95-2118-M-007-004-MY2), ROC. The authors wish to thank two anonymous referees for their helpful comments.

## Appendix

In the Appendix, we study the asymptotic property of the Lasso estimator $\tilde{\boldsymbol{\beta}}$ which minimizes (3) for VAR models. We first describe the assumptions required for the theoretical results as follows:

(A1) All roots of $|I - \Phi_1 z - \Phi_2 z^2 - \cdots - \Phi_p z^p| = 0$ are outside the unit circle.
(A2) The noise process $\{\boldsymbol{u}_t\}$ has finite fourth moment.
(A3) There exists an integer $p_0 \leq p$ such that $\Phi_j = \mathbf{0}$ for $j > p_0$.

**Theorem 1.** *Assume that* $\lambda_n/\sqrt{n} \to \lambda_0$ *for some* $\lambda_0 \geq 0$ *and* $\boldsymbol{\beta}_0$ *is the true parameter vector. Under* (A1)–(A3)*, we have* $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to \operatorname{argmin} U(\boldsymbol{\delta})$*, where*

$$U(\boldsymbol{\delta}) = -2\boldsymbol{\delta}' W + \boldsymbol{\delta}'(\Gamma \otimes I_k)\boldsymbol{\delta} + \lambda_0 \sum_j \left\{ \delta_j \operatorname{sign}(\beta_{0j}) I_{\{\beta_{0j} \neq 0\}} + |\delta_j| I_{\{\beta_{0j}=0\}} \right\},$$

*in which $\beta_{0j}$ and $\delta_j$ are the $j$th elements of $\boldsymbol{\beta}_0$ and $\boldsymbol{\delta}$, respectively, $W \sim N(\mathbf{0}, \Gamma \otimes \Sigma_u)$ and $\Gamma$ is a $(1+pk)$-dimensional covariance matrix consisting of 1 and $var((\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_p')')$ as the diagonal blocks.*

**Proof.** Consider the parameterization $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{\delta}$ which builds the sample size into it. Under this parameterization, minimizing $Q(\boldsymbol{\beta})$ in (3) with respect to $\boldsymbol{\beta}$ is equivalent to minimizing

$$\begin{aligned} U_n(\boldsymbol{\delta}) &= Q(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{\delta}) - Q(\boldsymbol{\beta}_0) \\ &= \left(\boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{\delta}\right)\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{\delta}\right)\right) - \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0\right) \\ &\quad + \lambda_n \sum_j (|\beta_{0j} + n^{-1/2}\delta_j| - |\beta_{0j}|) \\ &= -2\boldsymbol{\delta}'\left(n^{-1/2}\boldsymbol{X}'\boldsymbol{U}\right) + \boldsymbol{\delta}'(n^{-1}\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\delta} + (\lambda_n/\sqrt{n})\sqrt{n}\sum_j (|\beta_{0j} + n^{-1/2}\delta_j| - |\beta_{0j}|), \end{aligned}$$

with respective to $\boldsymbol{\delta}$. Following the standard limit theorems for the least square estimator, one can show that

$$n^{-1}\boldsymbol{X}'\boldsymbol{X} \to \Gamma \otimes I_k, \qquad n^{-1/2}\boldsymbol{X}'\boldsymbol{U} \to W.$$

Moreover, according to Knight and Fu (2000),

$$\sqrt{n}\sum_j (|\beta_{0j} + n^{-1/2}\delta_j| - |\beta_{0j}|) \to \sum_j \left\{\delta_j \operatorname{sign}(\beta_{0j}) I_{\{\beta_{0j}\neq 0\}} + |\delta_j| I_{\{\beta_{0j}=0\}}\right\}.$$

Therefore, $U_n(\boldsymbol{\delta}) \to U(\boldsymbol{\delta})$. Since $U_n$ is convex and $U$ has a unique minimum, it follows (Knight and Fu, 2000) that $\operatorname{argmin} U_n(\boldsymbol{\delta}) = \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to \operatorname{argmin} U(\boldsymbol{\delta})$. □

## References


Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC-19, 716–723.

Anderson, T.W., 1971. The Statistical Analysis of Time Series. John Wiley and Sons, Inc., New York.

Brüggemann, R., 2004. Model Reduction Methods for Vector Autoregressive Processes. In: Lecture Notes in Economics and Mathematical Systems, Springer.

Brüggemann, R., Lütkepohl, H., 2001. Lag selection in subset VAR models with an application to a U.S. monetary system. In: Friedmann, R., Knüppel, L., Lütkepohl, H. (Eds.), Econometric Studies—A Festschrift in Honour of Joachim Frohn. LIT, Münster, pp. 107–128.

Chen, C.H., Davis, R.A., Brockwell, P.J., 1996. Order determination for multivariate autoregressive processes using resampling methods. Journal of Multivariate Analysis 57, 175–190.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32, 407–499.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

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

Hosking, J.R.M., 1980. The multivariate Portmanteau statistic. Journal of the American Statistical Association 75, 602–608.

Hosking, J.R.M., 1981. Lagrange-multiplier tests of multivariate time series models. Journal of the Royal Statistical Society, Series B 43, 219–230.

Hsiao, C., 1979. Autoregressive modeling of Canadian money and income data. Journal of the American Statistical Association 74, 553–560.

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

Krolzig, H.-M., Hendry, D.F., 2001. Computer automation of general-to-specific model selection procedures. Journal of Economic Dynamics and Control 25, 831–866.

Li, H., Tsay, R.S., 1998. A unified approach to identifying multivariate time series models. Journal of the American Statistical Association 93, 770–782.

Li, W.K., McLeod, A.I., 1981. Distribution of the residual autocorrelations in multivariate ARMA time series models. Journal of the Royal Statistical Society, Series B 43, 231–239.

Lütkepohl, H., 1991. Introduction to Multiple Time Series Analysis. Springer-Verlag, Berlin.

Niu, X.F., Tiao, G.C., 1995. Modeling satellite ozone data. Journal of the American Statistical Association 90, 969–983.

Penm, J.H.W., Penm, J.H., Terrell, R.D., 1992. Using the bootstrap as an aid in choosing the approximate representation for vector time-series. Journal of Business & Economic Statistics 10, 213–219.

Penm, J.H.W., Terrell, R.D., 1982. On the recursive fitting of subset autoregressions. Journal of Time Series Analysis 3, 43–59.

Penm, J.H.W., Terrell, R.D., 1984. Multivariate subset autoregressive modeling with zero constrains for detecting overall causality. Journal of Econometrics 24, 311–330.

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

Tiao, G.C., Box, G.E.P., 1981. Modeling multiple time series with applications. Journal of the American Statistical Association 76, 802–816.

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

Tsay, R.S., 2002. Analysis of Financial Time Series. John Wiley and Sons, Inc., New York.
Wang, H., Li, G., Tsai, C.-L., 2007. Regression coefficient and autoregressive order shrinkage and selection via Lasso. Journal of the Royal Statistical Society, Series B 69, 63–78.
Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.