



Sparse Vector Autoregressive Modeling

Richard A. Davis, Pengfei Zang & Tian Zheng


To cite this article: Richard A. Davis, Pengfei Zang & Tian Zheng (2015): Sparse Vector Autoregressive Modeling, Journal of Computational and Graphical Statistics, DOI: [10.1080/10618600.2015.1092978](https://doi.org/10.1080/10618600.2015.1092978)

To link to this article: <http://dx.doi.org/10.1080/10618600.2015.1092978>

 View supplementary material 

 Accepted author version posted online: 29 Sep 2015.

 Submit your article to this journal 

 Article views: 29

 View related articles 

 View Crossmark data 

Sparse Vector Autoregressive Modeling

Richard A. Davis^{*1}, Pengfei Zang^{†1}, and Tian Zheng^{‡1}

¹Department of Statistics, Columbia University
Abstract

The vector autoregressive (VAR) model has been widely used for modeling temporal dependence in a multivariate time series. For large (and even moderate) dimensions, the number of the AR coefficients can be prohibitively large, resulting in noisy estimates, unstable predictions and difficult-to-interpret temporal dependence. To overcome such drawbacks, we propose a 2-stage approach for fitting sparse VAR (sVAR) models in which many of the AR coefficients are zero. The first stage selects non-zero AR coefficients based on an estimate of the partial spectral coherence (PSC) together with the use of BIC. The PSC is useful for quantifying the conditional relationship between marginal series in a multivariate process. A refinement second stage is then applied to further reduce the number of parameters. The performance of this 2-stage approach is illustrated with simulation and real data examples. Supplemental materials for the article are available online.

Keywords: multivariate time series, variable selection, partial spectral coherence (PSC), sparsity.

^{*}rdavis@stat.columbia.edu.

[†]pengfei@stat.columbia.edu. To whom correspondence should be addressed.

[‡]tzheng@stat.columbia.edu.

1 Introduction

The vector autoregressive (VAR) model has been widely used for modeling the temporal dependence of a multivariate time series. Unlike univariate time series, the temporal dependence of a multivariate series consists of not only the serial dependence within each marginal series, but also the interdependence across different components. The VAR model is well suited to describe such temporal dependence structures. In particular, the autoregression (AR) coefficients are crucial in interpreting the temporal dynamics. However, a conventional VAR model is saturatedly-parametrized and the number of its AR coefficients increases quadratically with the dimension of the process. For a large (or even moderate) dimension, this can lead to noisy AR parameter estimates and difficult-to-interpret descriptions of the temporal relationship.

To overcome these drawbacks, we propose a 2-stage approach for fitting sparse VAR (sVAR) models in which many of the AR coefficients are zero. Such sVAR models can enjoy improved efficiency of AR parameter estimates and more interpretable descriptions of AR coefficients. In the literature, a class of popular methods for fitting sVAR models is to re-formulate the VAR model as a penalized regression problem, where the determination of which AR coefficients are zero is equivalent to a variable selection problem in the linear regression setting. One of the most commonly used penalties for the AR coefficients in this context is the Lasso penalty proposed by Tibshirani (1996) and its variants tailored for the VAR modeling purpose, e.g., see Valdés-Sosa (2005); Arnold et al. (2008); Lozano et al. (2009); Shojaie and Michailidis (2010); Song and Bickel (2011). The Lasso-VAR approach has the advantage of performing model selection and parameter estimation simultaneously. However, there are also disadvantages in using this approach. First, Lasso has a tendency to over-select the order of autoregression of VAR models and this phenomenon has been reported in various numerical results, e.g., see Arnold et al. (2008); Lozano et al. (2009); Shojaie and Michailidis (2010). Second, in applying the Lasso-VAR approach, the VAR model is re-formulated as a linear regression model, where current values of the time series

are treated as the response variable and lagged values are treated as the explanatory variables. Such a treatment ignores the temporal dependence in the time series. [Song and Bickel \(2011\)](#) give a theoretical discussion on the consequences of applying Lasso directly to VAR models without taking into account the temporal dependence between the response and the explanatory variables.

In this article, we develop a 2-stage approach of fitting sVAR models. The first stage selects non-zero AR coefficients by screening pairs of distinct marginal series that are conditionally correlated. To compute the conditional correlation between marginal time series, an estimate of the *partial spectral coherence* (PSC) is used in the first stage. PSC is a tool in frequency-domain time series analysis that can be used to quantify direction-free conditional dependence between components of a multivariate time series. An efficient way of computing a non-parametric estimate of PSC is based on results of [Brillinger \(1981\)](#) and [Dahlhaus \(2000\)](#). In conjunction with the PSC, the Bayesian information criterion (BIC) is used in the first stage to determine the number of non-zero off-diagonal pairs of AR coefficients. The VAR model fitted in stage 1 may contain spurious non-zero AR coefficients. To further refine the fitted model, we propose, in stage 2, a screening strategy based on t -ratios of the AR coefficient estimates as well as BIC.

The remainder of this article is organized as follows. In Section 2, we review some results on VAR models for multivariate time series. In Section 3, we describe a 2-stage approach for fitting a sparse VAR model. In Section 4.1, simulation results are presented to compare the performance of the 2-stage approach against the Lasso-VAR approach. In Section 4.2, the 2-stage approach is applied to a real data example: the Google Flu Trends data. Further discussion is contained in Section 5. Supporting materials are provided in the Appendix, which is posted as a supplement to this article.

2 Sparse Vector Autoregressive Models

2.1 Vector Autoregressive Models (VAR)

Suppose $\{Y_t\} = \{(Y_{t,1}, Y_{t,2}, \dots, Y_{t,K})'\}$ is a vector autoregressive process of order p (VAR(p)), which satisfies the recursions,

$$Y_t = \mu + \sum_{k=1}^p A_k Y_{t-k} + Z_t, \quad t = 0, \pm 1, \dots, \quad (2.1)$$

where A_1, \dots, A_p are real-valued $K \times K$ matrices of autoregression (AR) coefficients; $\{Z_t\}$ is K -dimensional iid Gaussian noise with mean $\mathbf{0}$ and non-degenerate covariance matrix Σ_Z . We further assume that the process $\{Y_t\}$ is *causal*, i.e., $\det(I_K - \sum_{k=1}^p A_k z^k) \neq 0$, for $z \in \mathbb{C}, |z| < 1$, e.g., see [Brockwell and Davis \(1991\)](#) and [Reinsel \(1997\)](#), which implies that Z_t is independent of Y_s for $s < t$. Without loss of generality, we also assume that the vector process $\{Y_t\}$ has mean $\mathbf{0}$, i.e., $\mu = \mathbf{0}$ in (2.1).

2.2 Sparse Vector Autoregressive Models (sVAR)

The temporal dependence structure of the VAR model (2.1) is characterized by the AR coefficient matrices A_1, \dots, A_p . Based on T observations Y_1, \dots, Y_T from the VAR model, we want to estimate these AR matrices. However, a VAR(p) model, when fully-parametrized, has $K^2 p$ AR parameters that need to be estimated. For large (and even moderate) dimension K , the number of parameters can be prohibitively large, resulting in noisy estimates, unstable predictions and difficult-to-interpret descriptions of the temporal dependence. It is also generally believed that, for most applications, the true model of the series is sparse, i.e., the number of non-zero coefficients is small. Therefore it is preferable to fit a *sparse* VAR (sVAR) model in which many of its AR parameters are zero. In this article we develop a 2-stage approach of fitting sVAR models. The first stage selects non-zero AR coefficients by screening pairs of distinct marginal series that are conditionally correlated given other marginal series, where the characterization of conditional correlation

between marginal time series is not based on any parametric models. To compute direction-free conditional correlation between components in the time series, we use tools from the frequency-domain, specifically the *partial spectral coherence* (PSC). Below we introduce the basic properties related to PSC.

Let $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$ ($i \neq j$) denote two distinct marginal series of the process $\{Y_t\}$, and $\{Y_{t,-ij}\}$ denote the remaining $(K-2)$ -dimensional process. To compute the conditional correlation between two time series $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$, we need to adjust for the linear effect from the remaining marginal series $\{Y_{t,-ij}\}$. The removal of the linear effect of $\{Y_{t,-ij}\}$ from each of $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$ can be achieved by using results of linear filters, e.g., see [Brillinger \(1981\)](#) and [Dahlhaus \(2000\)](#). Specifically, the optimal linear filter for removing the linear effect of $\{Y_{t,-ij}\}$ from $\{Y_{t,i}\}$ is given by the set of $(K-2)$ -dimensional constant vectors that minimizes the expected squared error of filtering,

$$\{D_{k,i}^{opt} \in \mathbb{R}^{K-2}, k \in \mathbb{Z}\} = \underset{\{D_{k,i}, k \in \mathbb{Z}\}}{\operatorname{argmin}} \mathbf{E}(Y_{t,i} - \sum_{k=-\infty}^{\infty} D_{k,i} Y_{t-k,-ij})^2. \quad (2.2)$$

The *residual series* from the optimal linear filter is defined as,

$$\varepsilon_{t,i} := Y_{t,i} - \sum_{k=-\infty}^{\infty} D_{k,i}^{opt} Y_{t-k,-ij}.$$

Similarly, we use $\{D_{k,j}^{opt} \in \mathbb{R}^{K-2}, k \in \mathbb{Z}\}$ and $\{\varepsilon_{t,j}\}$ to denote the optimal linear filter and the corresponding residual series for another marginal series $\{Y_{t,j}\}$. Then the conditional correlation between $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$ is characterized by the correlation between the two residual series $\{\varepsilon_{t,i}\}$ and $\{\varepsilon_{t,j}\}$. In particular, two distinct marginal series $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$ are *conditionally uncorrelated* after removing the linear effect of $\{Y_{t,-ij}\}$ if and only if their residual series $\{\varepsilon_{t,i}\}$ and $\{\varepsilon_{t,j}\}$ are uncorrelated at all lags, i.e., $\operatorname{cor}(\varepsilon_{t+k,i}, \varepsilon_{t,j}) = 0$, for $k \in \mathbb{Z}$. In the frequency domain, $\{\varepsilon_{t,i}\}$ and $\{\varepsilon_{t,j}\}$ are uncorrelated at all lags is equivalent to the cross-spectral density of the two residual series, denoted by $f_{ij}^{\varepsilon}(\omega)$, is zero at all frequencies ω . Here the residual cross-spectral density is defined by,

$$f_{ij}^{\varepsilon}(\omega) := \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_{ij}^{\varepsilon}(k) e^{-ik\omega}, \quad \omega \in (-\pi, \pi], \quad (2.3)$$

where $\gamma_{ij}^\varepsilon(k) := \text{cov}(\varepsilon_{t+k,i}, \varepsilon_{t,j})$. The cross-spectral density $f_{ij}^\varepsilon(\omega)$ reflects the conditional (or partial) correlation between the two corresponding marginal series $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$, given $\{Y_{t,-ij}\}$. This observation leads to the definition of *partial spectral coherence* (PSC), e.g., see [Brillinger \(1981\)](#); [Brockwell and Davis \(1991\)](#), between two distinct marginal series $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$, which is defined as the scaled cross-spectral density between the two residual series $\{\varepsilon_{t,i}\}$ and $\{\varepsilon_{t,j}\}$, i.e.,

$$\text{PSC}_{ij}(\omega) := \frac{f_{ij}^\varepsilon(\omega)}{\sqrt{f_{ii}^\varepsilon(\omega)f_{jj}^\varepsilon(\omega)}}, \quad \omega \in (-\pi, \pi]. \quad (2.4)$$

[Brillinger \(1981\)](#) showed that the cross-spectral density $f_{ij}^\varepsilon(\omega)$ can be computed from the spectral density $f^Y(\omega)$ of the process $\{Y_t\}$ via,

$$f_{ij}^\varepsilon(\omega) = f_{ii}^Y(\omega) - f_{i,-ij}^Y(\omega)f_{-ij,-ij}^Y(\omega)^{-1}f_{-ij,j}^Y(\omega), \quad (2.5)$$

which involves inverting a $(K-2) \times (K-2)$ dimensional matrix, i.e., $f_{-ij,-ij}^Y(\omega)^{-1}$. Using (2.5) to compute the PSC for all pairs of distinct marginal series of $\{Y_t\}$ requires $\binom{K}{2}$ such matrix inversions, which can be computationally challenging for a large dimension K . [Dahlhaus \(2000\)](#) proposed a more efficient method to simultaneously compute the PSC for all $\binom{K}{2}$ pairs through the inverse of the spectral density matrix, which is defined as $g^Y(\omega) := f^Y(\omega)^{-1}$: Let $g_{ii}^Y(\omega)$, $g_{jj}^Y(\omega)$ and $g_{ij}^Y(\omega)$ denote the i th diagonal, the j th diagonal and the (i, j) th entry of $g^Y(\omega)$, respectively; then the partial spectral coherence between $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$ can be computed as follows,

$$\text{PSC}_{ij}(\omega) = -\frac{g_{ij}^Y(\omega)}{\sqrt{g_{ii}^Y(\omega)g_{jj}^Y(\omega)}}, \quad \omega \in (-\pi, \pi]. \quad (2.6)$$

The computation of all $\binom{K}{2}$ PSC using (2.6) requires only one matrix inversion of the $K \times K$ dimensional matrix $f^Y(\omega)$.

Remark. The regression in (2.2) is illustrative and not operational, i.e., it is not necessary to perform the time domain regression (2.2) in order to produce the PSC. Instead, we use the result in (2.6) to simultaneously compute all pairwise PSCs of a multivariate series.

Following (2.3), (2.4) and (2.6), we can see that,

$$\begin{aligned} \{Y_{t,i}\} \text{ and } \{Y_{t,j}\} \ (i \neq j) \text{ are conditionally uncorrelated} \\ \text{iff } g_{ij}^Y(\omega) = 0, \text{ for all } \omega \in (-\pi, \pi]. \end{aligned} \quad (2.7)$$

In other words, the inverse spectral density matrix $g^Y(\omega)$ encodes the pairwise conditional correlation between the component series of $\{Y_t\}$. The result in (2.7) only requires second-moment stationarity of the multivariate series $\{Y_t\}$ and does not rely on any other assumption of the distribution of $\{Y_t\}$.

3 A 2-stage Approach of Fitting sVAR Models

In this section, we develop a 2-stage approach of fitting sVAR models. The first stage of the approach takes advantage of (2.7) and screens out the pairs of marginal series that are conditionally uncorrelated. For such pairs we set the corresponding AR coefficients to zero for each lag. However, the model fitted in stage 1 may still contain spurious non-zero AR coefficient estimates. To address this possibility, a second stage is used to refine the model further.

3.1 Stage 1: selection

As we have shown in Section 2.2, a zero PSC indicates that the two corresponding marginal series are conditionally uncorrelated. In the first stage of our approach, we use the information of pairwise conditional uncorrelation to reduce the complexity of the VAR model. In particular, we propose to set the estimated AR coefficients between two conditionally uncorrelated marginal series to zero,

i.e.,

if $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$ ($i \neq j$) are conditionally uncorrelated, (3.1)

then set $\hat{A}_k(i, j) = \hat{A}_k(j, i) = 0$ ($k = 1, \dots, \hat{p}$).

where $\hat{A}_k(i, j)$ and \hat{p} are estimated AR coefficients and estimated order of autoregression, respectively. According to the results in (2.6) and (2.7), the conditional uncorrelation between $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$ is equivalent to $\text{PSC}_{ij}(\omega) = 0$ for $\omega \in (-\pi, \pi]$.

Remark. We point out that (3.1) is not an assertion but a strategy we employ to fit sVAR models. This proposed strategy (3.1) of connecting zero PSC with zero AR coefficients may not be exact for some examples. However, numerical results suggest that our 2-stage approach is still able to achieve well-fitted sVAR models in such cases. We will return to this point in Section 5.

From (3.1) we can see that the modeling interest of the first stage is whether or not the AR coefficients belonging to a pair of marginal series at all lags are selected, rather than the selection of an individual AR coefficient. In order to set a group of AR coefficient estimates to zero as in (3.1), we need to find the pairs of marginal series for which their PSC is identically zero. Due to sampling variability, however, the estimated PSC, denoted by $\hat{\text{PSC}}_{ij}(\omega)$ for series $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$, will not be exactly zero even when the two corresponding marginal series are conditionally uncorrelated. Therefore we rank the estimated PSC based on their evidence to be non-zero and decide a cutoff point that separates non-zero PSC from zero PSC. Since the estimate $\hat{\text{PSC}}_{ij}(\omega)$ depends on the frequency ω , we need a quantity to summarize its departure from zero over different frequencies. As in Dahlhaus (2000); Dahlhaus et al. (1997), we use the supremum of the squared modulus of the estimated PSC, i.e.,

$$\hat{S}_{ij} := \sup_{\omega} |\hat{\text{PSC}}_{ij}(\omega)|^2, \quad (3.2)$$

as the summary statistic, where the supremum is taken over all the Fourier frequencies $\{2\pi k/T : k = 1, \dots, T\}$. A large value of \hat{S}_{ij} indicates that the two marginal series are likely to be conditionally correlated. Therefore we can create a sequence \mathbf{Q}_1 of the $\binom{K}{2}$ pairs of distinct marginal series by ranking each pair's summary statistic (3.2) from highest to lowest. This sequence \mathbf{Q}_1 prioritizes the way in which non-zero coefficients are added into the VAR model. Based on the sequence \mathbf{Q}_1 , we need two parameters to fully specify the VAR model: the order of autoregression p and the number of *top* pairs in \mathbf{Q}_1 , denoted by M , that are selected into the VAR model. For the $\frac{(K-1)K}{2} - M$ pairs not selected, their corresponding groups of AR coefficients are set to zero. The two parameters (p, M) control the complexity of the VAR model as the number of non-zero AR coefficients is $(K + 2M)p$. We use the BIC, see Schwarz (1978), to simultaneously choose the values of these two parameters. The BIC is computed as,

$$\text{BIC}(p, M) = -2 \log L(\hat{A}_1, \dots, \hat{A}_p) + \log T \cdot (K + 2M)p, \quad (3.3)$$

where $L(\hat{A}_1, \dots, \hat{A}_p)$ is the maximized likelihood of the VAR model. To compute the maximized likelihood $L(\hat{A}_1, \dots, \hat{A}_p)$, we use results on the constrained maximum likelihood estimation of VAR models as given in Lütkepohl (2007). Details of this estimation procedure can be found Appendix A.1.

Restricting the two parameters p and M to take values in pre-specified ranges \mathbb{P} and \mathbb{M} , respectively, the steps of **stage 1** can be summarized as follows.

Stage 1

1. Estimate the PSC for all $K(K - 1)/2$ pairs of distinct marginal series by inverting a non-parametric estimate of the spectral density matrix and applying equation (2.6).
2. Construct a sequence \mathbf{Q}_1 of the $K(K - 1)/2$ pairs of distinct marginal series by ranking each pair's summary statistic \hat{S}_{ij} (3.2) from highest to lowest.
3. For each $(p, M) \in \mathbb{P} \times \mathbb{M}$, set the order of autoregression to p and select the top M pairs in the sequence \mathbf{Q}_1 into the VAR model, which specifies the parameter constraint on the AR coefficients. Conduct parameter estimation under this constraint using the results in Appendix A.1 and compute the corresponding $\text{BIC}(p, M)$ according to equation (3.3).
4. Choose (\tilde{p}, \tilde{M}) that gives the minimum BIC value over $\mathbb{P} \times \mathbb{M}$.

Remark. In the first step of Stage 1, we use the periodogram smoothed by a modified Daniell kernel, e.g., see Brockwell and Davis (1991), as the non-parametric estimate of the spectral density matrix and obtain estimates of PSC by inverting the smoothed periodogram. If such kernel-smoothed spectral density estimators fail to produce reliable estimates of PSC due to numerical instability, which is indicated by a large condition number, one may consider regularized matrix estimators for the spectral density matrix to improve numerical stability, e.g., the shrinkage spectral density estimators proposed by Böhm and von Sachs (2009) and Fiecas and Ombao (2011).

The model obtained in the first stage contains $(K + 2\tilde{M})\tilde{p}$ non-zero AR coefficients. If only a small proportion of the pairs of marginal series are selected, i.e., $\tilde{M} \ll K(K - 1)/2$, $(K + 2\tilde{M})\tilde{p}$ can be much smaller than $K^2\tilde{p}$, which is the number of AR coefficients in a fully-parametrized $\text{VAR}(\tilde{p})$ model.

3.2 Stage 2: refinement

Stage 1 selects AR parameters related to the most conditionally correlated pairs of marginal series according to BIC. However, it may also have introduced spurious non-zero AR coefficients in the stage 1 model: As PSC can only be evaluated for pairs of series, we cannot select diagonal coefficients in A_1, \dots, A_p , nor can we select within the group of coefficients corresponding to one pair of component series. We therefore apply a second stage to further refine the stage 1 model. To eliminate these possibly spurious coefficients, the $(K + 2\tilde{M})\tilde{p}$ non-zero AR coefficients of the stage 1 model are ranked according to the absolute values of their t -statistic. The t -statistic for a non-zero AR coefficient estimate $\hat{A}_k(i, j)$ ($k = 1, \dots, \tilde{p}$ and $i, j = 1, \dots, K$) is,

$$t_{i,j,k} := \frac{\hat{A}_k(i, j)}{\text{s.e.}(\hat{A}_k(i, j))}. \quad (3.4)$$

Here the standard error of $\hat{A}_k(i, j)$ is computed from the asymptotic distribution of the constrained maximum likelihood estimator of the stage 1 model, which is, e.g., see [Lütkepohl \(2007\)](#),

$$\sqrt{T}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, \tilde{R}[\tilde{R}'(\tilde{\Gamma}_Y(0) \otimes \tilde{\Sigma}_Z^{-1})\tilde{R}]^{-1}\tilde{R}'), \quad (3.5)$$

where $\alpha := \text{vec}(A_1, \dots, A_p)$ is the $K^2p \times 1$ vector obtained by column stacking the AR coefficient matrices A_1, \dots, A_p ; $\hat{\alpha}$, $\tilde{\Gamma}_Y(0)$ and $\tilde{\Sigma}_Z$ are the maximum likelihood estimators of α , $\Gamma_Y(0) := \text{cov}((Y'_t, \dots, Y'_{t-p+1})')$ and Σ_Z , respectively; and \tilde{R} is the *constraint matrix*, defined by equation (A.1) in Appendix A.1, of the stage 1 model. Therefore we can create a sequence \mathbf{Q}_2 of the $(K + 2\tilde{M})\tilde{p}$ triplets (i, j, k) by ranking the absolute values of the t -ratios (3.4) from highest to lowest. The AR coefficients corresponding to the *top* triplets in \mathbf{Q}_2 are more likely to be retained in the model because of their significance. In the second stage, there is only one parameter, denoted by m , controlling the complexity of the model, which is the number of non-zero AR coefficients to be retained. And BIC is used to select the complexity of the final sVAR model. The steps of **stage 2** are as follows.

Stage 2

1. Compute the t -statistic $t_{i,j,k}$ (3.4) for each of the $(K + 2\tilde{M})\tilde{p}$ non-zero AR coefficient estimates of the stage 1 model.
2. Create a sequence \mathbf{Q}_2 of the $(K + 2\tilde{M})\tilde{p}$ triplets (i, j, k) by ranking $|t_{i,j,k}|$ from highest to lowest.
3. For each $m \in \{0, 1, \dots, (K + 2\tilde{M})\tilde{p}\}$, consider the model that selects the m non-zero AR coefficients corresponding to the top m triplets in the sequence \mathbf{Q}_2 . Under this parameter constraint, execute the constrained parameter estimation using results in Appendix A.1 and compute the corresponding BIC according to $\text{BIC}(m) = -2 \log L + \log T \cdot m$.
4. Choose m^* that gives the minimum BIC value.

Our 2-stage approach in the end leads to a sVAR model that contains m^* non-zero AR coefficients corresponding to the top m^* triplets in \mathbf{Q}_2 . We denote this sVAR model by $\text{sVAR}(p^*, m^*)$, where p^* is the order of autoregression and m^* is the number of non-zero AR coefficients.

4 Numerical Results

In this section, we provide numerical results on the performance of our 2-stage approach of fitting sVAR models. In Section 4.1, simulation results are presented to compare the performance of the 2-stage approach against competing Lasso-type methods for fitting an order-1 sVAR model. In Section 4.2, the 2-stage approach is applied to a real data example: the Google Flu Trends data, e.g., see Ginsberg et al. (2009). In Appendix A.3, we provide additional simulation results as well as the application to another real data example: a time series of concentration levels of air pollutants, e.g., see Songsiri et al. (2010).

4.1 Simulation

Simulation results are presented to demonstrate the performance of our 2-stage approach of fitting sVAR models. We compare the 2-stage approach with Lasso-VAR methods. To apply Lasso-VAR methods, the VAR model is re-formulated as a linear regression problem, where current values of the time series are treated as the response variable and lagged values are treated as the explanatory variables. Then Lasso can be applied to select the AR coefficients and fit sVAR models, e.g., see [Valdés-Sosa \(2005\)](#); [Arnold et al. \(2008\)](#); [Lozano et al. \(2009\)](#); [Shojaie and Michailidis \(2010\)](#); [Song and Bickel \(2011\)](#). The Lasso method shrinks the AR coefficients towards zero by minimizing a target function, which is the sum of a loss function and a l_1 penalty on the AR coefficients. Unlike linear regression models, the choice of the loss function between the sum of squared residuals and the minus log likelihood will affect the resulted Lasso-VAR models even if the multivariate time series is Gaussian. This is because the noise covariance matrix Σ_Z is taken into account in the likelihood function of a Gaussian VAR process but not in the sum of squared residuals. In general, this distinction will lead to different VAR models unless the unknown covariance matrix Σ_Z equals to a scalar multiple of the identity matrix, e.g., see [Appendix A.2](#). We notice that this issue of choosing the loss function has not been addressed in the literature of Lasso-VAR models. For example, [Arnold et al. \(2008\)](#); [Lozano et al. \(2009\)](#); [Shojaie and Michailidis \(2010\)](#); [Song and Bickel \(2011\)](#) all used the sum of squared residuals as the loss function and did not consider the possibility of choosing the minus log likelihood as the loss function. The simulation setups in these papers all assume, either explicitly or implicitly, that the covariance matrix Σ_Z is diagonal or simply the identity matrix. Therefore in our simulation we apply Lasso to VAR modeling under both cases: in the first case we choose the sum of squared residuals as the loss function and denote it as the Lasso-SS method; in the second case we use the minus log likelihood as the loss function and denote it as the Lasso-LL method. Details of fitting these two Lasso-VAR models are given in [Appendix A.2](#).

The Lasso-VAR approach simultaneously performs model selection and parameter estimation, which is usually considered as an advantage of the approach. However, our simulation results suggest that simultaneous model selection and parameter estimation can weaken the performance of the Lasso-VAR approach. This is because Lasso-VAR methods, such as Lasso-SS and Lasso-LL, have a tendency to over-select the autoregression order of VAR models, a phenomenon reported by many, see [Arnold et al. \(2008\)](#); [Lozano et al. \(2009\)](#); [Shojaie and Michailidis \(2010\)](#). This over-specified model complexity potentially increases the mean squared error of the AR coefficient estimates of Lasso-VAR models. On the contrary, simulation results show that our 2-stage approach is able to identify the correct set of non-zero AR coefficients more often and it also achieves better parameter estimation efficiency than the two competing Lasso-VAR methods. In addition, simulation results also suggest that the Lasso-SS method, which does not take into account the noise covariance matrix Σ_Z in its model fitting, performs the worst among the three.

Simulation Example: Consider the 6-dimensional VAR(1) process $\{Y_t\} = \{(Y_{t,1}, \dots, Y_{t,6})'\}$ given by,

$$\begin{pmatrix} Y_{t,1} \\ Y_{t,2} \\ Y_{t,3} \\ Y_{t,4} \\ Y_{t,5} \\ Y_{t,6} \end{pmatrix} = \begin{pmatrix} 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.3 & 0 \\ 0.6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.8 \end{pmatrix} \begin{pmatrix} Y_{t-1,1} \\ Y_{t-1,2} \\ Y_{t-1,3} \\ Y_{t-1,4} \\ Y_{t-1,5} \\ Y_{t-1,6} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \\ Z_{t,3} \\ Z_{t,4} \\ Z_{t,5} \\ Z_{t,6} \end{pmatrix}, \quad (4.1)$$

where the $Z_t = (Z_{t,1}, \dots, Z_{t,6})'$ is iid Gaussian noise with mean $\mathbf{0}$ and covariance matrix Σ_Z . The order of autoregression in (4.1) is $p = 1$ and there are 6 non-zero AR coefficients, so (4.1) specifies

a sVAR(1, 6) model. The covariance matrix Σ_Z of the Gaussian noise is,

$$\Sigma_Z = \begin{pmatrix} \delta^2 & \delta/4 & \delta/6 & \delta/8 & \delta/10 & \delta/12 \\ \delta/4 & 1 & 0 & 0 & 0 & 0 \\ \delta/6 & 0 & 1 & 0 & 0 & 0 \\ \delta/8 & 0 & 0 & 1 & 0 & 0 \\ \delta/10 & 0 & 0 & 0 & 1 & 0 \\ \delta/12 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \quad (4.2)$$

We can see that the marginal series $\{Y_{t,1}\}$ is related to all other series via Σ_Z . And we can change the value of δ^2 to compare the impact of the variability of $\{Y_{t,1}\}$ on the performance of the three competing methods. We compare the three methods according to five metrics: (1) the selected order of autoregression \hat{p} ; (2) the number of non-zero AR coefficient estimates \hat{m} ; (3) the squared bias of the AR coefficient estimates,

$$\sum_{k=1}^{p \vee \hat{p}} \sum_{i,j=1}^K [\mathbf{E}[\hat{A}_k(i, j)] - A_k(i, j)]^2;$$

(4) the variance of the AR coefficient estimates,

$$\sum_{k=1}^{p \vee \hat{p}} \sum_{i,j=1}^K \text{var}(\hat{A}_k(i, j));$$

and (5) the mean squared error (MSE) of the AR coefficient estimates,

$$\sum_{k=1}^{p \vee \hat{p}} \sum_{i,j=1}^K \{[\mathbf{E}[\hat{A}_k(i, j)] - A_k(i, j)]^2 + \text{var}(\hat{A}_k(i, j))\},$$

where $p \vee \hat{p} := \max\{p, \hat{p}\}$ and $A_k(i, j) := 0$ for any triplet (k, i, j) such that $k > p$ and $1 \leq i, j \leq K$. The first two metrics show the model selection performance and the latter three metrics reflect the efficiency of parameter estimates of each method. The pre-specified range of the autoregression order p is $\mathbb{P} = \{0, 1, 2, 3\}$. Selection of the tuning parameter for the two Lasso-VAR methods is based on ten-fold cross-validation as described in Appendix A.2. We let δ^2 in Σ_Z take values from $\{1, 4, 25, 100\}$. The sample size T is 100 and the reported results are based on 500 replications.

Remark. For our numerical experiments, our objective is not to solve the justification problem of using cross-validation in time series modeling. Here we simply choose a popular technique of applying cross-validation to select tuning parameters under the Lasso-VAR framework, which is also employed by other authors, e.g., see [Valdés-Sosa \(2005\)](#); [Arnold et al. \(2008\)](#).

The five metrics for comparison are summarized in Table 1. The \hat{p} column shows that the 2-stage approach is able to correctly select the autoregression order $p = 1$ while the two Lasso-VAR methods over-select the autoregression order. Furthermore, the true number of non-zero AR coefficients is $m = 6$. As shown in the \hat{m} column, the average number of non-zero AR coefficient estimates from the 2-stage approach is very close to 6. At the same time, this number from either the Lasso-SS or the Lasso-LL method is much larger than 6, meaning that the two Lasso-VAR methods lead to a lot of spurious non-zero AR coefficients. Second, we compare the efficiency of parameter estimates. The bias^2 column shows that the 2-stage approach has much smaller estimation bias than the two Lasso-VAR methods. This is because the l_1 penalty is known to produce large estimation bias for large non-zero coefficients. In addition, the large number of spurious non-zero AR coefficients also increases the variability of the parameter estimates from the two Lasso-VAR methods. This is reflected in the variance column, showing that the variance of the AR coefficient estimates from the Lasso-SS and the Lasso-LL methods are larger than that from the 2-stage approach. Therefore the 2-stage approach has a much smaller MSE than the two Lasso-VAR methods, and this difference in MSE becomes more notable as the marginal variability δ^2 increases.

A comparison of the AR coefficient estimation performance when $\delta^2 = 1$ is displayed in Figure 1. Panel (a) of Figure 1 displays the true AR coefficient matrix A_1 , where the color of a circle shows the true value of the corresponding AR coefficient. Panels (b) and (c) show the AR coefficient estimates from stages 1 and 2 of the 2-stage approach. The size of each circle is proportional to the percent of times (out of 500 replications) the corresponding AR coefficient is selected and the color of each circle shows the average of the 500 estimates of that AR coefficient. We can

see from panel (b) that the first stage is able to select the AR coefficients belonging to pairs of conditionally correlated marginal series. But the stage 1 model contains spurious non-zero AR coefficients, as indicated by the presence of 6 dominant white circles in panel (b) at 4 diagonal positions, i.e., (2, 2), (3, 3), (4, 4), (5, 5), and 2 off-diagonal positions, i.e., (1, 4), (4, 2). These white circles effectively disappear in panel (c) due to the second stage refinement. This observation demonstrates the effectiveness of the second stage refinement. In addition, the similarity between panel (a) and panel (c) has two implications: first, the presence of 6 dominant color circles in both panels suggests that the 2-stage approach is able to select the true non-zero AR coefficients with high probabilities; second, the other tiny circles in panel (c) indicate that the 2-stage approach leads to only a small number of spurious AR coefficients. These two implications together show that the 2-stage approach is able to correctly select the non-zero AR coefficients for this sVAR model. On the other hand, panels (e) and (f) display the estimated AR coefficients from the Lasso-LL and the Lasso-SS methods, respectively. The most notable aspect in these two panels is the prevalence of medium-sized white circles. The whiteness of these circles indicates that the corresponding AR coefficient estimates are unbiased. However, according to the legend panel, the size of these circles corresponds to an approximate 50% chance that each of these truly zero AR coefficients is selected by the Lasso-VAR methods. As a result, both Lasso-VAR methods lead to a large number of spurious non-zero AR coefficients and their model selection results are highly variable. Consequently, it is more difficult to interpret these Lasso-VAR models. This observed tendency for Lasso-VAR methods to over-select the non-zero AR coefficients is consistent with the numerical findings in [Arnold et al. \(2008\)](#); [Lozano et al. \(2009\)](#); [Shojaie and Michailidis \(2010\)](#).

We also compare the impact of the marginal variability of $\{Y_{1,t}\}$ on the performance of each method. Figure 2 displays the estimated AR coefficients from the 2-stage approach as well as the two Lasso-type methods for $\delta^2 = 4, 25$ and 100, respectively. We can see that the performance of the 2-stage approach remains persistently good against the changing marginal variability δ^2 . This is because the 2-stage approach involves estimating the covariance matrix Σ_Z and therefore

will adjust for the changing variability. On the other hand, both Lasso-VAR methods persistently over-select the AR coefficients as δ^2 varies. But it is interesting to notice that the impact of the changing variability is different for the Lasso-SS and the Lasso-LL methods. The model selection result of the Lasso-SS method is severely impacted by the changing variability. From panels (g), (h) and (i), we can see that as δ^2 increases from 4 to 100, the Lasso-SS method will increasingly over-estimate the temporal influence of the other 5 marginal series into $\{Y_{t,1}\}$ and leads to spurious AR coefficients in the first row of A_1 . On the other hand, panels (d), (e) and (f) show that the model selection result of the Lasso-LL method is not much influenced by the changing variability. Such a difference is due to the fact that the Lasso-LL method takes into account the covariance matrix Σ_Z while the Lasso-SS method does not. The observed distinction between the Lasso-SS and the Lasso-LL methods verifies that the choice of the loss function will affect the resulted Lasso-VAR model, a fact that has not been addressed in the literature of Lasso-VAR modeling. In this simulation example, the Lasso-LL method benefits from modeling the covariance matrix Σ_Z and is superior to the Lasso-SS method.

Finally, it is also interesting to compare between the 2-stage approach and Lasso-VAR methods their behavior of estimating one particular coefficient as the marginal variability δ^2 changes. Such a comparison is deferred to Appendix A.3.1 due to length constraints.

4.2 Google Flu Trends

In this application, we consider the *Google Flu Trends* data, which can be viewed as a measure of the level of influenza activity in the US. It has been noticed by many researchers that the frequencies of certain Internet search terms can be predictive of the influenza activity within a future time period. Based on this fact, a group of researchers at Google applied logistic regression to select the top 45 Google user search terms that are most indicative of the influenza activity. These selected 45 terms were then used to produce the Google Flu Trends data, see [Ginsberg et al. \(2009\)](#). The

Google Flu Trends data consist of weekly predicted numbers of influenza-like-illness (ILI) related visits out of every 100,000 random outpatient visits within a US region. The Google Flu Trends prediction has been shown to be highly consistent with the ILI rate reported by the Centers for Disease Control and Surveillance (CDC), where the ILI rate is the probability that a random outpatient visit is related to an influenza-like-illness. But the Google Flu Trends data have two advantages over the traditional CDC influenza surveillance report: first, the Google Flu Trends predictions are available 1 or 2 weeks before the CDC report is published and therefore provide a possibility to forecast the potential outbreak of influenza epidemics; second, since Google is able to map the I.P. address of each Google user search to a specific geographic area, the Google Flu Trends data enjoy a finer geographic resolution than the CDC report. In particular, the Google Flu Trends data are published not only at the US national level but are also available for the 50 states, the District of Columbia and 122 cities throughout the US. In contrast, the CDC surveillance report is available only at the national level and for 10 major US regions (each region is a group of states).

We apply the 2-stage approach to fit a sVAR model to the weekly Google Flu Trends data from the week of January 1st, 2006 to the week of December 26th, 2010, so the sample size is $T = 260$. Out of the 51 regions (50 states and the District of Columbia), we remove 5 states (Alaska, Hawaii, North Dakota, South Dakota and Wyoming) from our analysis due to incompleteness of the data. So the dimension of the process in this example is $K = 46$ and we refer to these 46 regions as 46 states for simplicity. In applying the 2-stage approach, the pre-specified range of the autoregression order p is $\mathbb{P} = \{0, 1, 2, 3, 4\}$. The first stage selects the autoregression order $\tilde{p} = 2$ and $\tilde{M} = 290$ pairs of distinct marginal series into the model. So the stage 1 model contains $(K + 2\tilde{M})\tilde{p} = (46 + 290 \cdot 2) \cdot 2 = 1252$ non-zero AR coefficients. The second stage follows by further selecting $m^* = 763$ non-zero AR coefficients and leads to a final sVAR(2,763) model, which has only as many as $19.30\% = 763/(46^2 \times 2)$ of the AR coefficients in a fully-parametrized VAR(2) model. For comparison, we also fit an unrestricted VAR(2) model and apply the Lasso-SS method to fit another sVAR model. Based on a ten-fold cross validation, the Lasso-SS

method results in a VAR model with 3123 non-zero AR coefficients, which we denote as Lasso-SS(2,3123).

We compare the temporal dependence structures discovered by the three models, i.e., the VAR(2), the sVAR(2, 763) and the Lasso-SS(2,3123). Figure 3 displays the estimated AR coefficients from the three models at lags 1 and 2, respectively. To illustrate the possible spatial interpretation of the dependence structure, we group the 46 states into 10 regions as suggested in the CDC influenza surveillance report, which is indicated by the solid black lines in Figure 3. From panels (a), (c) and (e), we can see that the AR coefficient estimates on the diagonal of \hat{A}_1 are large and positive in all three models. This observation is reasonable since influenza activity from the previous week should be predictive of influenza activity of the current week within the same region. But panel (a) shows that this diagonal signal is diluted by the noisy off-diagonal AR estimates in the VAR(2) model. And except for this diagonal signal of \hat{A}_1 , the other AR coefficient estimates in the VAR(2) model are noisy and hard to interpret at both lags 1 and 2. In contrast, the diagonal signal of \hat{A}_1 is most dominant in panel (c) of the 2-stage sVAR(2,763) model, in which lots of the off-diagonal AR coefficients are zero. Additionally, the overall interpretability of the sVAR(2,763) and the Lasso-SS(2,3123) models is much better than the VAR(2) model, since both models provide much cleaner descriptions of the temporal dependence structures and reveal some interesting patterns. For example, both the sVAR(2,763) and the Lasso-SS(2,3123) models discover the interdependence among the influenza activity of the 6 states in Region 1, i.e., (CT, MA, ME, NH, RI, VT), as indicated by the first block of states in panels (c), (d), (e) and (f). This within-region dependence is moderately positive at lag 1 and slightly negative at lag 2. In the sVAR(2,763) and the Lasso-SS(2,3123) models, we also observe the cross-region influence from Region 8 of (CO, MT, US) into Region 6 of (AR, LA, NM, OK, TX). In spite of their general resemblance, the Lasso-SS(2,3123) model contains many more non-zero AR coefficients than the sVAR(2,763) model. In fact, the Lasso-SS(2,3123) model has a large number of small (in absolute value) but non-zero AR coefficients, especially those at lag 2 as shown in panel (f).

The fitted sVAR model not only has better interpretability, but also improved forecast performance. To this point, we compare the out-of-sample forecast performance of the sVAR model with competing models. We use the Google Flu Trends data between the week of July 10, 2011 and December 25, 2011 ($T_{\text{test}} = 24$) as the test data. We compute two quantities for the comparison: the first is the h-step-ahead forecast root mean squared error (RMSE), which is defined as,

$$\text{RMSE}(h) = [K^{-1}(T_{\text{test}} - h + 1)^{-1} \sum_{k=1}^K \sum_{t=T}^{T+T_{\text{test}}-h} (\hat{Y}_{t+h,k} - Y_{t+h,k})^2]^{1/2},$$

where $\hat{Y}_{t+h,k}$ is the h-step-ahead forecast of $Y_{t+h,k}$ for $k = 1, \dots, K$; the second is the h-step-ahead logarithmic score (LS), e.g., see [Gneiting and Raftery \(2007\)](#), which is defined as,

$$\text{LS}(h) = (T_{\text{test}} - h + 1)^{-1} \sum_{t=T}^{T+T_{\text{test}}-h} -\log p_{h|t}(\hat{Y}_{t+h}),$$

where $p_{h|t}(\cdot)$ is the conditional probability density function of Y_{t+h} given $\{Y_s, s \leq t\}$. Table 2 and 3, respectively, summarize the $\text{RMSE}(h)$ and the $\text{LS}(h)$ for a forecast horizon $h = 1, 2, 3$ and 4 from four different models: the marginal ar model (where a univariate ar model is fitted to each marginal series of $\{Y_t\}$ and the order of autoregression of each ar model is chosen based on BIC), the sVAR(2,763) model, the LassoSS(2,3123) model and the VAR(2) model. From Table 2 we see that the sVAR(2,763) model fitted by the 2-stage approach has the smallest forecast RMSE among the four models. On the one hand, the sVAR(2,763) model has smaller forecast RMSE than the marginal ar model since the 2-stage approach retains important non-zero off-diagonal AR coefficients that are excluded by the ar model; on the other hand, the sVAR(2,763) model outperforms the most saturated model, i.e., the VAR(2) model, since the 2-stage approach excludes many spurious AR coefficients that are retained in the VAR(2) model. Furthermore, as shown by Table 3, the logarithmic score rule also favors the sVAR(2,763) model among the four models.

5 Discussion and Conclusion

In this article, we propose a 2-stage approach of fitting sVAR models, in which many of the AR coefficients are zero. The first stage of the approach is based on PSC and BIC to select non-zero AR coefficients. The combination of PSC and BIC provides an effective initial selection tool to determine the sparsity constraint on the AR coefficients. The second stage employs t -ratios together with BIC to further refine the stage 1 model. The proposed approach is promising in that the 2-stage fitted sVAR models enjoy improved efficiency of parameter estimates and easier-to-interpret descriptions of temporal dependence, as compared to unrestricted VAR models. In the first stage selection of the 2-stage approach, we use (3.1) to link zero PSCs with zero AR coefficients. For some examples, however, this connection may not be exact. When non-zero AR coefficients correspond to zero PSCs, these AR coefficients are likely to be set to zero in the first stage and thus will not be selected by the 2-stage fitted models. For the cases we have investigated, however, we notice that purely BIC-selected models also tend to discard such AR coefficients. A possible explanation is that if the PSCs are near zero, the corresponding AR coefficients do not increase the likelihood sufficiently to merit their inclusion into the model based on BIC. As a result, the 2-stage approach still leads to sVAR models that perform similarly as the best BIC-selected models. To illustrate this point, we construct a VAR model in which a zero PSC corresponds to non-zero AR coefficients. Consider the following 3-dimensional VAR(1) process $\{Y_t\} = \{(Y_{t,1}, Y_{t,2}, Y_{t,3})'\}$ satisfying the recursions,

$$\begin{pmatrix} Y_{t,1} \\ Y_{t,2} \\ Y_{t,3} \end{pmatrix} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 0.3 \\ 0 & 0.25 & 0.5 \end{pmatrix} \begin{pmatrix} Y_{t-1,1} \\ Y_{t-1,2} \\ Y_{t-1,3} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \\ Z_{t,3} \end{pmatrix}, \quad (5.1)$$

where $\{Z_t = (Z_{t,1}, Z_{t,2}, Z_{t,3})'\}$ are iid Gaussian noise with mean $\mathbf{0}$ and covariance matrix,

$$\Sigma_Z = \begin{pmatrix} 18 & 0 & 6 \\ 0 & 1 & 0 \\ 6 & 0 & 3 \end{pmatrix}.$$

For this example, one can show that $\text{PSC}_{1,2}(\omega) = 0$ for $\omega \in (-\pi, \pi]$ while $A_1(1, 2) = 0.5$. In applying the 2-stage approach to fit sVAR models to (5.1), the first stage estimate of the summary statistic $\sup_{\omega} |\text{PSC}_{1,2}(\omega)|^2$, as defined in (3.2), is likely to be small, so the estimates of $A_1(1, 2)$ and $A_1(2, 1)$ are likely to be automatically set to zero in the first stage.

We compare the performance of the 2-stage approach with a modified 2-stage procedure of fitting sVAR models to (5.1). In the first stage of the modified procedure, we use precise knowledge of which AR coefficients are truly non-zero and conduct constrained maximum likelihood estimation under the corresponding parameter constraint. Then we execute the second stage of the modified procedure in exactly the same way as the original 2-stage approach. In other words, the modified procedure has an “oracle” first stage and uses t -ratios together with BIC for further refinement in its second stage. So the truly non-zero AR coefficients will not be excluded after the first stage of the modified procedure. Such AR coefficients will survive the second stage refinement if the inclusion of them substantially increases the likelihood of the final sVAR model; otherwise they will be discarded after the second stage. For both approaches, the pre-specified range of the autoregression order p is $\mathbb{P} = \{0, 1, 2, 3\}$. The sample size T is 100 and results are based on 500 replications. The comparison of these two approaches using different metrics is shown in Figure 4. In each panel of Figure 4, the x-axis refers to the modified 2-stage procedure and is labeled as “oracle + BIC”; the y-axis refers to the original 2-stage approach and is labeled as “PSC + BIC”. Panel (a) compares the number of non-zero AR coefficients, where these numbers are jittered so that their distributions can be observed; panel (b) compares the out-of-sample one-

step forecast error; panel (c) compares the minus log-likelihood and panel (d) compares the BIC of the fitted models. From panel (a), we can see that the “oracle + BIC” procedure does not lead to more non-zero AR coefficients than the 2-stage approach does. From panels (b), (c) and (d), we can see that the “oracle + BIC” procedure does not provide improvement over the original 2-stage approach with respect to the one-step forecast error, the likelihood, or the BIC of fitted models. So, at least in this example, a non-zero AR coefficient that corresponds to a zero PSC is unlikely to be included in a BIC-selected model. As a result, our 2-stage approach has similar performance as that of the “oracle + BIC” procedure. This phenomenon also raises the connection between the PSC and the likelihood of sVAR processes as an interesting direction for future research.

Supplemental Materials

code_sVAR.zip : A zipped folder including R programs of the 2-stage approach to fitting sparse VAR models and example codes to replicate the simulation study reported in the article. Please read the file **README** in the zipped folder for more details.

Appendix : A file containing supplemental details of model fitting procedures and additional numerical results.

Acknowledgments

The research of Richard A. Davis is supported in part by NSF grant DMS-1107031. The research of Tian Zheng is supported in part by NSF grant SES-1023176 and a 2010 Google faculty research award. We would like to thank Professor Jitkomut Songsiri for providing the air pollutant data. We also want to thank the associate editor and three anonymous referees for their insightful and constructive comments.

References

- Arnold, A., Liu, Y., and Abe, N. (2008), “Temporal causal modeling with graphical Granger methods,” *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Böhm, H. and von Sachs, R. (2009), “Shrinkage estimation in the frequency domain of multivariate time series,” *Journal of Multivariate Analysis*, 100, 913–935.
- Brillinger, D. R. (1981), *Time Series: Data Analysis and Theory*, New York: Holt, Rinehart and Winston.
- Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods*, New York: Springer.
- Dahlhaus, R. (2000), “Graphical interaction models for multivariate time series,” *Metrika*, 51, 157–172.
- Dahlhaus, R., Eichler, M., and Sandkühler, J. (1997), “Identification of synaptic connections in neural ensembles by graphical models,” *Journal of Neuroscience Methods*, 77, 93–107.
- Efron, B., Hastie, T., Johnstone, T., and Tibshirani, R. (2004), “Least angle regression,” *Annals of Statistics*, 32, 408–451.
- Eichler, M. (2006), “Fitting graphical interaction models to multivariate time series,” *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*.
- Fiecas, M. and Ombao, H. (2011), “The generalized shrinkage estimator for the analysis of functional connectivity of brain signals,” *The Annals of Applied Statistics*, 5, 1102–1125.
- Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2009), “Detecting influenza epidemics using search engine query data,” *Nature*, 457, 1012–1014.

- Gneiting, T. and Raftery, A. E. (2007), “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, 102, 359–378.
- Lozano, A. C., Abe, N., Liu, Y., and Rosset, S. (2009), “Grouped graphical Granger modeling for gene expression regulatory networks discovery,” *Bioinformatics*, 25, 110–118.
- Lütkepohl, H. (2007), *New Introduction to Multiple Time Series Analysis*, New York: Springer.
- Reinsel, G. C. (1997), *Elements of Multivariate Time Series Analysis*, New York: Springer.
- Schwarz, G. (1978), “Estimating the dimension of a model,” *Annals of Statistics*, 6, 461–464.
- Shojaie, A. and Michailidis, G. (2010), “Discovering graphical Granger causality using the truncating lasso penalty,” *Bioinformatics*, 26, 517–523.
- Song, S. and Bickel, P. J. (2011), “Large vector auto regressions,” *Arxiv preprint arXiv:1106.3915*.
- Songsiri, J., Dahl, J., and Vandenberghe, L. (2010), “Graphical models of autoregressive processes,” *Convex Optimization in Signal Processing and Communications*, 89–116.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
- Valdés-Sosa, P. A. (2005), “Estimating brain functional connectivity with sparse multivariate autoregression,” *Philosophical Transactions of the Royal Society B*, 360, 969–981.

A Appendix to “Sparse Vector Autoregressive Modeling” published in the Journal of Computational and Graphical Statistics

Richard A. Davis, Pengfei Zang, and Tian Zheng

Department of Statistics, Columbia University

Appendix A.1 gives results on constrained maximum likelihood estimation of sVAR models. Appendix A.2 shows the procedure of implementing the two Lasso-VAR methods, i.e., the Lasso-SS and the Lasso-LL. Appendix A.3 provides additional numerical results of the 2-stage sVAR fitting procedure: in Appendix A.3, continuing the discussion of the simulation example in Section 4.1, the 2-stage approach is compared with Lasso-VAR methods in terms of robustness of their parameter estimation performance against changing variability; in Appendix A.3.2, simulation results of estimating a 12-dimensional order-3 sVAR model are presented; in Appendix A.3.3, application of the 2-stage approach to a 5-dimensional time series of concentration levels of air pollutants is discussed.

A.1 Constrained Maximum Likelihood Estimation of sVAR Models

Continuing with the notation in equation (2.1), the constraint that the AR coefficients of the VAR(p) model are set to zero can be expressed as

$$\alpha := \text{vec}(A_1, \dots, A_p) = R\gamma, \quad (\text{A.1})$$

where $\alpha = \text{vec}(A_1, \dots, A_p)$ is the $K^2p \times 1$ vector obtained by column stacking the AR coefficient matrices A_1, \dots, A_p ; R is a $K^2p \times m$ matrix of known constants with rank m (usually $m \ll K^2p$); γ

is a $m \times 1$ vector of unknown parameters. The matrix R in equation (A.1) is called the *constraint matrix* and it specifies which AR coefficients are set to zero by choosing one entry in each column to be 1 and all the other entries in that column to be 0. The rank m of the constraint matrix R equals the number of non-zero AR coefficients of the VAR model. This formulation is illustrated by the following simple example.

Consider a 2-dimensional zero-mean VAR(2) process $\{Y_t\} = \{(Y_{t,1}, Y_{t,2})'\}$ satisfying the recursions,

$$\begin{pmatrix} Y_{t,1} \\ Y_{t,2} \end{pmatrix} = \begin{pmatrix} A_1(1,1) & 0 \\ A_1(2,1) & A_1(2,2) \end{pmatrix} \times \begin{pmatrix} Y_{t-1,1} \\ Y_{t-1,2} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ A_2(2,1) & 0 \end{pmatrix} \times \begin{pmatrix} Y_{t-2,1} \\ Y_{t-2,2} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \end{pmatrix}, \quad (\text{A.2})$$

where $A_k(i, j)$ is the (i, j) th entry of the AR coefficient matrix A_k ($k = 1, 2$). The VAR(2) model (A.2) contains 4 non-zero AR coefficients, $A_1(1, 1), A_1(2, 1), A_1(2, 2)$ and $A_2(2, 1)$, which can be expressed as

$$\begin{aligned} \alpha &= \text{vec}(A_1, A_2) = R\gamma \\ \Rightarrow \begin{pmatrix} A_1(1,1) \\ A_1(2,1) \\ 0 \\ A_1(2,2) \\ 0 \\ A_2(2,1) \\ 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} A_1(1,1) \\ A_1(2,1) \\ A_1(2,2) \\ A_2(2,1) \end{pmatrix}. \end{aligned} \quad (\text{A.3})$$

The constraint matrix R in (A.3) is of rank $m = 4$, which equals to the number of non-zero AR coefficients.

Lütkepohl (2007) gives results on the constrained maximum likelihood estimation of the AR coefficients. Under the parameter constraint in the form of (A.1), the maximum likelihood estimators of the AR coefficients α and the noise covariance matrix Σ_Z are the solutions to the following equations,

$$\hat{\alpha} = R\{R'(LL' \otimes \hat{\Sigma}_Z^{-1})R\}^{-1}R'(L \otimes \hat{\Sigma}_Z^{-1})y, \quad (\text{A.4})$$

$$\hat{\Sigma}_Z = \frac{1}{T-p} \sum_{t=p+1}^T (Y_t - \hat{Y}_t)(Y_t - \hat{Y}_t)', \quad (\text{A.5})$$

where \otimes is the *Kronecker* product and

$$L_t := (Y_t, Y_{t-1}, \dots, Y_{t-p+1})',$$

$$L := (L_0, L_1, \dots, L_{T-1}),$$

$$y := \text{vec}(Y) = \text{vec}(Y_1, Y_2, \dots, Y_T),$$

$$\hat{Y}_t := \sum_{k=1}^p \hat{A}_k Y_{t-k}.$$

It is known that, e.g., see Lütkepohl (2007); Reinsel (1997), if there is no parameter constraint on the AR coefficients, i.e., $R = I_{K^2p}$ in (A.1), then the maximum likelihood estimator of the AR coefficients does not involve the noise covariance matrix Σ_Z . From equation (A.4), however, we can see that the presence of the parameter constraint (A.1) makes the estimation of the AR coefficients commingled with the estimation of the covariance matrix Σ_Z . Therefore we iteratively update the estimators $\hat{\alpha}$ and $\hat{\Sigma}_Z$ according to equations (A.4) and (A.5), until convergence, to obtain the constrained maximum likelihood estimator of the AR coefficients.

A.2 Implementation of Lasso-VAR Methods

We give details of the two Lasso implementations of fitting VAR models, i.e., the Lasso-SS and Lasso-LL VAR models. Notice that the VAR(p) model (2.1) can be written in the following compact form,

$$y = \text{vec}(Y) = (L' \otimes I_K)\alpha + \text{vec}(Z), \quad (\text{A.6})$$

where vec column stack operator, \otimes is the *Kronecker* product and

$$Y := (Y_1, Y_2, \dots, Y_T),$$

$$y := \text{vec}(Y),$$

$$L_t := (Y_t, Y_{t-1}, \dots, Y_{t-p+1})',$$

$$L := (L_0, L_1, \dots, L_{T-1}),$$

$$Z := (Z_1, Z_2, \dots, Z_T).$$

Since Z_1, \dots, Z_T are iid from the K -dimensional Gaussian $N(0, \Sigma_Z)$, from (A.6) the minus log likelihood of the VAR(p) model (A.6), ignoring an additive constant, is,

$$-2 \log L(\alpha, \Sigma_Z) = T \log |\Sigma_Z| + [y - (L' \otimes I_K)\alpha]' (I_T \otimes \Sigma_Z^{-1}) [y - (L' \otimes I_K)\alpha]. \quad (\text{A.7})$$

For Lasso-penalized VAR models, there are two possible choices of the loss function: one is the sum of squared residuals and the other one is the minus log likelihood. The Lasso-SS method uses the sum of squared residuals as the loss function and the corresponding target function is,

$$Q_\lambda^{SS}(\alpha) := \|y - (L' \otimes I_K)\alpha\|_2^2 + \lambda \|\alpha\|_1; \quad (\text{A.8})$$

while the Lasso-LL method chooses the minus log likelihood as the loss function and its target function is,

$$\begin{aligned} Q_\lambda^{LL}(\alpha, \Sigma_Z) &:= [y - (L' \otimes I_K)\alpha]' (I_T \otimes \Sigma_Z^{-1}) [y - (L' \otimes I_K)\alpha] \\ &\quad + T \log |\Sigma_Z| + \lambda \|\alpha\|_1. \end{aligned} \quad (\text{A.9})$$

In both equations (A.8) and (A.9) the scalar tuning parameter $\lambda \in \mathbb{R}$ controls the amount of penalty. The AR coefficients α of the VAR model are estimated by minimizing the target function $Q_\lambda^{SS}(\alpha)$ (A.8) or $Q_\lambda^{LL}(\alpha, \Sigma_Z)$ (A.9), respectively.

It is worth noting that, unlike the linear regression model, the choice between the sum of squared residuals and minus log likelihood as the loss function will lead to different results of applying the Lasso method to VAR models. This can be seen by taking the first derivative of the Lasso-SS target function (A.8) and the Lasso-LL target function (A.9) with respect to the AR coefficient α ,

$$\frac{\partial Q_\lambda^{SS}(\alpha)}{\partial \alpha} = 2[(LL' \otimes I_K) - (L \otimes I_K)y] + \lambda \cdot \text{sgn}(\alpha), \quad (\text{A.10})$$

$$\frac{\partial Q_\lambda^{LL}(\alpha)}{\partial \alpha} = 2[(LL' \otimes \Sigma_Z^{-1}) - (L \otimes \Sigma_Z^{-1})y] + \lambda \cdot \text{sgn}(\alpha), \quad (\text{A.11})$$

where $\text{sgn}(\cdot)$ is the *signum* function and $\text{sgn}(\alpha)$ is the $K^2p \times 1$ vector in which the k th entry is $\text{sgn}(\alpha_k)$, $k = 1, \dots, K^2p$. We can see that noise covariance matrix Σ_Z is taken into account by the Lasso-LL derivative (A.11) but not by the Lasso-SS derivative (A.10). The two $K^2p \times 1$ vectors of first derivatives (A.10) and (A.11) are in general not equal (up to multiplication by a scalar) unless the covariance matrix Σ_Z is a multiple of the identity matrix I_K . Therefore the Lasso-SS and the Lasso-LL methods will in general result in different VAR models.

Based on (A.8) and (A.9), we describe the estimation procedures of the two Lasso-penalized VAR models. The estimation of Lasso-SS VAR models is straightforward since it can be viewed as standard linear regression problems with the Lasso penalty. Therefore the Lasso-SS VAR model can be fitted efficiently by applying the least angle regression (LARS) algorithm, e.g., see [Efron et al. \(2004\)](#). In this article we use the *R* package *glmnet* for fitting Lasso-SS VAR models. The estimation of Lasso-LL VAR models is more complicated since the target function (A.9) involves the unknown noise covariance matrix Σ_Z . We propose an iterative procedure to fit the Lasso-LL VAR model. The procedure is based on the fact that, for a given covariance matrix Σ_Z , the Lasso-

LL target function (A.9) can be re-cast in a least-squares fashion. In other words, for a $K \times K$ positive-definite matrix Σ_Z , let

$$\Sigma_Z = U \text{diag}\{\kappa_1, \dots, \kappa_K\} U',$$

be its eigenvalue decomposition, where U is an orthonormal matrix and $\kappa_1 \geq \kappa_2 \dots \geq \kappa_K > 0$ are the K positive eigenvalues. Define

$$\Sigma_Z^{-\frac{1}{2}} := U \text{diag}\left\{\frac{1}{\sqrt{\kappa_1}}, \dots, \frac{1}{\sqrt{\kappa_K}}\right\} U' \quad (\text{A.12})$$

to be the *inverse square root* of Σ_Z . Notice that $\Sigma_Z^{-\frac{1}{2}}$ in (A.12) is symmetric and $\Sigma_Z^{-\frac{1}{2}} \Sigma_Z^{-\frac{1}{2}} = \Sigma_Z^{-1}$, then we have

$$\begin{aligned} I_T \otimes \Sigma_Z^{-1} &= (I_T \otimes \Sigma_Z^{-\frac{1}{2}})(I_T \otimes \Sigma_Z^{-\frac{1}{2}}) \\ &= (I_T \otimes \Sigma_Z^{-\frac{1}{2}})'(I_T \otimes \Sigma_Z^{-\frac{1}{2}}), \\ (I_T \otimes \Sigma_Z^{-\frac{1}{2}})[y - (L' \otimes I_K)\alpha] &= (I_T \otimes \Sigma_Z^{-\frac{1}{2}})y - (I_T \otimes \Sigma_Z^{-\frac{1}{2}})(L' \otimes I_K)\alpha \\ &= (I_T \otimes \Sigma_Z^{-\frac{1}{2}})y - (L' \otimes \Sigma_Z^{-\frac{1}{2}})\alpha. \end{aligned}$$

Therefore the Lasso-LL target function (A.9) can be re-written as

$$\begin{aligned} Q_\lambda^{LL}(\alpha, \Sigma_Z) & \quad (\text{A.13}) \\ &= T \log |\Sigma_Z| + [y - (L' \otimes I_K)\alpha]'(I_T \otimes \Sigma_Z^{-1})[y - (L' \otimes I_K)\alpha] + \lambda \|\alpha\|_1 \\ &= T \log |\Sigma_Z| + [y - (L' \otimes I_K)\alpha]'(I_T \otimes \Sigma_Z^{-\frac{1}{2}})'(I_T \otimes \Sigma_Z^{-\frac{1}{2}})[y - (L' \otimes I_K)\alpha] + \lambda \|\alpha\|_1 \\ &= T \log |\Sigma_Z| + [(I_T \otimes \Sigma_Z^{-\frac{1}{2}})y - (L' \otimes \Sigma_Z^{-\frac{1}{2}})\alpha]'[(I_T \otimes \Sigma_Z^{-\frac{1}{2}})y - (L' \otimes \Sigma_Z^{-\frac{1}{2}})\alpha] + \lambda \|\alpha\|_1 \\ &= T \log |\Sigma_Z| + \|(I_T \otimes \Sigma_Z^{-\frac{1}{2}})y - (L' \otimes \Sigma_Z^{-\frac{1}{2}})\alpha\|_2^2 + \lambda \|\alpha\|_1. \end{aligned}$$

The loss function

$$\|(I_T \otimes \Sigma_Z^{-\frac{1}{2}})y - (L' \otimes \Sigma_Z^{-\frac{1}{2}})\alpha\|_2^2,$$

in (A.13) can be viewed as the sum of squared residuals from a linear regression model with the response variable being $(I_T \otimes \Sigma_Z^{-\frac{1}{2}})y$ and the explanatory variables given by $L' \otimes \Sigma_Z^{-\frac{1}{2}}$. Therefore, for a given Σ_Z , minimizing the Lasso-LL target function (A.13) with respect to the AR coefficients α is equivalent to minimizing a Lasso-SS target function corresponding to the response variable $(I_T \otimes \Sigma_Z^{-\frac{1}{2}})y$ and the explanatory variables $L' \otimes \Sigma_Z^{-\frac{1}{2}}$. So we can use the following iterative procedure to fit Lasso-LL VAR models.

An iterative procedure of fitting Lasso-LL VAR models

1. Set an initial value $\Sigma_Z^{(0)}$ for the covariance matrix Σ_Z .
2. Update the AR coefficients α and the covariance matrix Σ_Z at the $(k + 1)$ th iteration, until convergence, as follows,
 - 2.1. $\alpha^{(k+1)} = \underset{\alpha}{\operatorname{argmin}} Q_{\lambda}^{LL}(\alpha, \Sigma_Z^{(k)})$ by applying the coordinate descent algorithm;
 - 2.2. $\Sigma_Z^{(k+1)} = \frac{1}{T}(Y - A^{(k+1)}L)(Y - A^{(k+1)}L)'$,
where $\alpha^{(k+1)} = \operatorname{vec}(A^{(k+1)})$.

Fitting Lasso-penalized VAR models, as all penalized regression methods, also involves choosing the tuning parameter $\lambda \in \mathbb{R}$. The choice of λ is usually based on certain information criterion or cross-validations. In this article we use cross-validations to determine the value of λ . Furthermore, the number of explanatory variables, i.e., the number of lagged values appearing on the right hand side of equation (A.6), also depends on the unknown order of autoregression p . Therefore the values of both p and λ need to be determined in a data-driven manner. Suppose the autoregression order p is restricted to take values in a pre-specified range \mathbb{P} , we use the following steps to fit Lasso-SS as well as Lasso-LL VAR models.

Steps of fitting Lasso-SS and Lasso-LL VAR models

1. For each $p \in \mathbb{P}$, apply the coordinate descent algorithm to minimize the Lasso-SS target function (A.8) and the aforementioned iterative procedure to minimize the Lasso-LL target function (A.9), respectively. For either the Lasso-SS or the Lasso-LL model, the optimal tuning parameter $\lambda^{opt}(p)$, depending on the given autoregression order p , is determined by the minimum average ten-fold cross-validation error, which is denoted by $CV_{min}(p)$.
2. Choose p^* that gives the minimum average cross-validation error over \mathbb{P}

$$p^* = \underset{p \in \mathbb{P}}{\operatorname{argmin}} CV_{min}(p),$$

as the autoregression order for either the Lasso-SS or the Lasso-LL VAR model.

3. Obtain either the Lasso-SS or the Lasso-LL VAR model by setting the autoregression order p equal to p^* and the tuning parameter λ equal to $\lambda^{opt}(p^*)$.

A.3 Numerical Results

A.3.1 Further Remarks on Simulation Example in Section 4.1

Continuing the discussion of the simulation example in Section 4.1, we investigate the estimators of one particular AR coefficient from the three methods in more detail. Figure 5 displays the sampling distributions of the estimator $\hat{A}_1(6, 6)$ from the 2-stage approach as well as the two Lasso-VAR methods for $\delta^2 = 1, 4, 25$ and 100, respectively. Estimation of $A_1(6, 6)$ is of interest because the marginal series $\{Y_{t,6}\}$ is exclusively driven by its own past values. Ideally, due to such “isolation”, the estimation of $A_1(6, 6)$ should not be impacted much by the estimation of the AR coefficients in the 5×5 upper-left sub-matrix of A_1 . Moreover, $A_1(6, 6)$ has a large true value of 0.8 and it is interesting to compare the estimation bias for this large AR coefficient. Figure 5 shows that the

estimators of $A_1(6, 6)$ from the 2-stage approach and the Lasso-LL method are not impacted much by the changing variability of $\{Y_{t,1}\}$. But the Lasso-SS estimator for $A_1(6, 6)$ becomes more biased and volatile as the marginal variability increases from $\delta^2 = 1$ to $\delta^2 = 100$. Although both the 2-stage sVAR and the Lasso-LL estimators of $A_1(6, 6)$ are robust to the changing values of δ^2 , the difference between their bias is significant. The 2-stage approach gives an estimator of $A_1(6, 6)$ that remains nearly unbiased as δ^2 varies. However, there is a systematic bias in the Lasso-LL estimator of $A_1(6, 6)$, which is due to the shrinkage effect of the Lasso penalty on the selected AR coefficients.

A.3.2 Simulation Example 2

Consider the 12-dimensional VAR(3) process $\{Y_t\} = \{(Y_{t,1}, \dots, Y_{t,12})'\}$ given by,

$$Y_t = \begin{pmatrix} B & O \\ O & C \end{pmatrix} Y_{t-1} + \begin{pmatrix} C & O \\ O & B \end{pmatrix} Y_{t-3} + Z_t, \quad (\text{A.14})$$

where O is a 6×6 matrix of zeros; B and C are 6×6 matrices given by

$$B = \begin{pmatrix} 0.32 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.20 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.24 & 0 \\ 0.24 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.24 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.32 \end{pmatrix},$$

$$C = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.40 & 0 & 0 & 0.45 & 0 \\ 0 & 0 & 0.40 & 0 & 0 & 0 \\ 0 & -0.30 & 0 & 0.40 & 0 & 0 \\ 0 & 0 & 0.40 & 0 & 0.40 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.60 \end{pmatrix},$$

and the $Z_t = (Z_{t,1}, \dots, Z_{t,12})'$ is 12-dimensional iid Gaussian noise with mean $\mathbf{0}$ and covariance matrix Σ_Z given by

$$\Sigma_Z = \begin{pmatrix} \tilde{\Sigma}_Z & 0 \\ 0 & \tilde{\Sigma}_Z \end{pmatrix}, \quad (\text{A.15})$$

with $\tilde{\Sigma}_Z$ equal to the 6-dimensional covariance matrix (4.2). Notice that in (A.14), the AR coefficients at lag 2 are all zero. Equation (A.14) specifies a 12-dimensional VAR(3) model with 28 non-zero AR coefficients. As in Example 1, we let δ^2 in Σ_Z (A.15) take values from $\{1, 4, 25, 100\}$ and perform the five-metric comparison between the 2-stage approach and the Lasso-VAR methods. The pre-specified range of the autoregression order p is $\mathbb{P} = \{0, 1, 2, 3, 4\}$. Selection of the tuning parameter for Lasso-VAR methods is based on ten-fold cross-validation. The sample size

T is 200 and the reported results are based on 500 replications.

Table 4 summarizes the comparison of the five metrics between the 2-stage approach and the two Lasso-VAR methods. The \hat{p} column shows that the 2-stage approach is able to correctly select the autoregression order $p = 3$. It is also seen that the two Lasso-VAR methods tend to under-select the autoregression order. Furthermore, as shown by the \hat{m} column, the average number of non-zero AR coefficient estimates from the 2-stage approach is very close to its true value 28 while this number from either the Lasso-SS or the Lasso-LL method is much larger than 28. This observation agrees with the numerical findings in Arnold et al. (2008); Lozano et al. (2009); Shojaie and Michailidis (2010) that Lasso-VAR methods tend to over-select non-zero AR coefficients. Furthermore, the bias², variance and MSE columns altogether suggest that the 2-stage approach is more efficient in parameter estimation than the two Lasso-VAR methods.

Figure 6 displays the true AR coefficient matrices (the 1st row), and a comparison of the AR coefficient estimates from the 2-stage approach (the 2nd row), the Lasso-LL method (the 3rd row) and the Lasso-SS method (the 4th row) when $\delta^2 = 25$. We can see that all of the three methods are able to identify the non-zero AR coefficients at lag 1 with large probabilities. However, the two Lasso-VAR methods also introduce a large number of false non-zero AR coefficient estimates at lag 1, as indicated by the presence of white circles in panels (g) and (j) of Figure 6, while the 2-stage approach is able to eliminate those spurious coefficients from the final model, as shown in panel (d) of Figure 6. Furthermore, as shown by panels (e) and (f) of Figure 6, the 2-stage approach is able to simultaneously detect the zero AR coefficients at lag 2 and the non-zero AR coefficients at lag 3. On the contrary, panels (h), (i), (k) and (l) of Figure 6 suggest that the two Lasso-VAR methods have poor performance in estimating AR coefficients beyond the first lag.

A.3.3 Concentration Levels of Air Pollutants

In this application, we analyze a time series of concentration levels of four air pollutants, CO, NO, NO₂, O₃, as well as the solar radiation intensity R. The data are recorded hourly during the year 2006 at Azusa, California and can be obtained from the Air Quality and Meteorological Informa-

tion System (AQMS). The time series for analysis is of dimension $K = 5$ and with $T = 8370$ observations. The same dataset was previously studied in [Songsiri et al. \(2010\)](#). A similar dataset of the same 5 component series, but recorded at a different location, was analyzed in [Dahlhaus \(2000\)](#); [Eichler \(2006\)](#). The methods employed in [Dahlhaus \(2000\)](#); [Eichler \(2006\)](#); [Songsiri et al. \(2010\)](#) are based on the *partial correlation graph model*, in which VAR models are estimated under sparsity constraints on the inverse spectrum of VAR processes. So the modeling interest of the partial correlation graph approach is sparsity in the frequency domain, i.e., zero constraints on the inverse spectrum, while our 2-stage approach is concerned about sparsity in the time domain, i.e., zero constraints on AR coefficients. For this example, we are interested in comparing the findings from the 2-stage sVAR model and the partial correlation graph model.

We apply the 2-stage approach to fit a sVAR model to the air pollution data. The pre-specified range of the autoregression order p is $\mathbb{P} = \{0, 1, 2, \dots, 8\}$. The same range for p was also used in [Songsiri et al. \(2010\)](#). The first stage does not exclude any pair of marginal series and leads to a stage 1 model with $\tilde{p} = 4$ and $\tilde{M} = 10$, which contains $(5 + 2 \times 10) \times 4 = 100$ non-zero AR coefficients. The second stage further refines the model and leads to a sVAR(4,64) model. The selection of the autoregression order $p^* = 4$ coincides with the result in [Songsiri et al. \(2010\)](#), which also used BIC for VAR order selection. However, the partial correlation graph approach used in [Songsiri et al. \(2010\)](#) is concerned about sparsity in the inverse spectrum rather than in the AR coefficients. So the AR coefficients estimated by the partial correlation graph approach are never exactly zero, and the resulted VAR model will contain spurious non-zeros. The presence of these spurious AR coefficients weakens the interpretability of fitted VAR models and can be viewed as a limitation of the partial correlation graph approach. Another limitation of the partial correlation graph approach is that it only deals with a small dimension, since in the partial correlation graph approach model selection is usually executed based on an exhaustive search of all possible patterns of sparsity constraints on the inverse spectrum, e.g., see [Dahlhaus \(2000\)](#); [Eichler \(2006\)](#); [Songsiri et al. \(2010\)](#). The number of such patterns is $2^{K(K-1)/2}$, which reaches 2×10^6 when $K = 7$.

Therefore the partial correlation graph approach is feasible only for a small dimension. In fact, the largest dimension of all numerical examples considered in [Dahlhaus \(2000\)](#); [Eichler \(2006\)](#); [Songsiri et al. \(2010\)](#) is 6. This is unlike our 2-stage approach, which is able to deal with higher dimensions, such as the 46-dimensional process in the Google Flu Trends example.

Since the 2-stage approach is applied to the same dataset as in [Songsiri et al. \(2010\)](#), it is interesting to compare the findings between the 2-stage sVAR model and the partial correlation graph model. Our comparison is in the frequency domain. Figure 7 displays the estimates of the squared modulus of PSC, i.e., $|\text{PSC}(\omega)|^2$, as computed from the AR coefficient estimates in the 2-stage sVAR(4,64) model as well as the non-parametric estimates of $|\text{PSC}(\omega)|^2$ used in the first stage of the 2-stage approach. We can see the good match-up between the two sets of estimates. So it is implied that it is possible to use the AR coefficient estimates from the 2-stage sVAR model, which are time-domain parameters, to recover the sparsity pattern in the inverse spectrum, which are frequency-domain quantities. We also point out that the estimates of $|\text{PSC}(\omega)|^2$ from the 2-stage sVAR(4,64) model, as displayed in Figure 7, resemble those in Figure 1.9 of [Songsiri et al. \(2010\)](#), which displays the estimates of $|\text{PSC}(\omega)|^2$ from the fitted partial correlation graph model. Furthermore, the findings from Figure 7 agree with the photochemical theory of interactions between the 5 marginal series. For example, the large estimates of $|\text{PSC}(\omega)|^2$ between (CO, NO) comes from the fact that both air pollutants are mainly emitted from cars; the large estimates of $|\text{PSC}(\omega)|^2$ between (O₃, R) reflects the major role of the solar radiation intensity in the generation of ozone, e.g., see [Dahlhaus \(2000\)](#). Additionally, from Figure 7 we observe that the estimates of $|\text{PSC}(\omega)|^2$ between the pairs (CO, O₃), (CO, R), (NO, R) and (NO, O₃) are relatively small as compared to the other pairs. This discovery of weak estimates of $|\text{PSC}(\omega)|^2$ agrees with the findings in [Dahlhaus \(2000\)](#); [Eichler \(2006\)](#); [Songsiri et al. \(2010\)](#), which are summarized in Table 5. For more detailed discussion on the underlying photochemical mechanism of interactions between air pollutants, readers are referred to [Dahlhaus \(2000\)](#).

For certain practical applications, results of VAR models can sometimes be interpreted via *im*-

pulse response analysis, e.g., see Reinsel (1997). In impulse response analysis, the dynamics of a multivariate series are modeled as driven by several uncorrelated noise time series. By computing the *impulse response coefficients*, we can see how a multivariate series reacts over time to exogenous impulses. Below we show some results of applying impulse response analysis to the air pollutants time series data. Details of computing impulse response coefficients of VAR models can be found in Reinsel (1997). In order to compute impulse response coefficients, we use the estimated AR coefficients $\hat{A}_1, \dots, \hat{A}_4$ and the estimated noise covariance matrix $\hat{\Sigma}_Z$ from the 2-stage sVAR(4,64) model, where the Cholesky decomposition of $\hat{\Sigma}_Z$ is used to de-correlate the noise components. Figure 8 plots the impulse response coefficients of the sVAR(4,64) model and it shows how the dynamics of the air pollutants time series are impacted by multiple uncorrelated noise series. For example, the concentration level of CO, which corresponds to the 1st row in Figure 8, is mostly driven by the 1st noise series; while the concentration level of O₃ (the 4th row) and the solar radiation activity R (the 5th row) are largely impacted the 5th noise series.

Figure 8 can also help to explain the conditional correlation between different air pollutants time series. For example, from Figure 8 we see that the concentration levels of CO (the 1st row) and NO₂ (the 3rd row) both react to a strong impulse in the 1st noise series. This agrees with the strong PSC between CO and NO₂ as shown in Figure 7. On the other hand, the concentration level of NO (the 2nd row) and the solar radiation intensity R (the 5th row) are driven by different uncorrelated noise series, i.e., the concentration level of NO is mainly driven by the 1st and the 2nd noise series while the solar radiation intensity R by the 5th noise series, and this helps to explain the weak PSC between NO and R as observed in Figure 7.

Remark. *In the impulse response analysis of the air pollution data, we use the Cholesky decomposition to orthogonalize the noise series. There also exist other decomposition methods and alternative ways of orthogonalization could lead to different results. In addition, interpretations of the Cholesky-orthogonalized noise series are not presented here since such interpretations require specific knowledge of atmospheric chemistry and are beyond the scope of this article.*

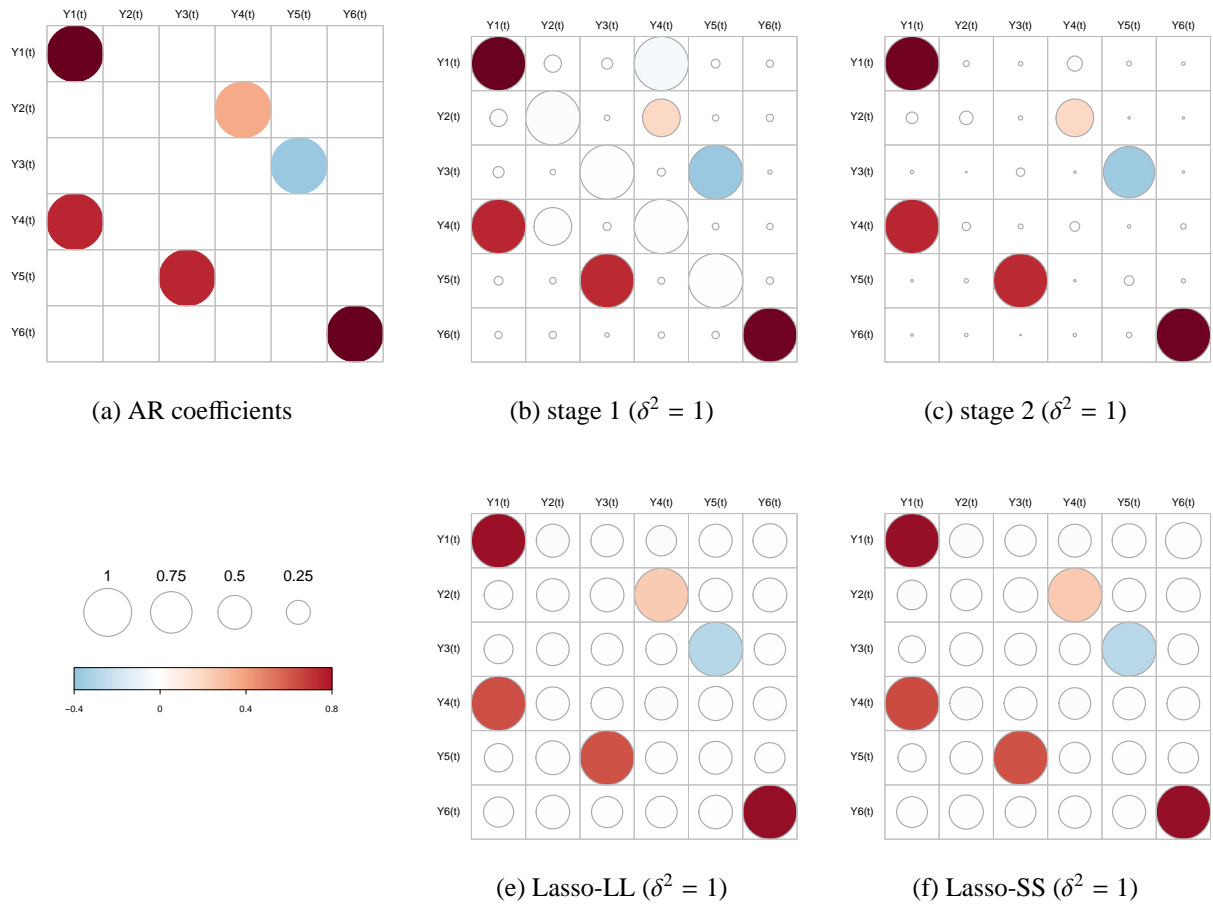


Figure 1: Displays of the AR coefficient estimates from stages 1 and 2 of the 2-stage approach, the Lasso-LL and the Lasso-SS methods when $\delta^2 = 1$.

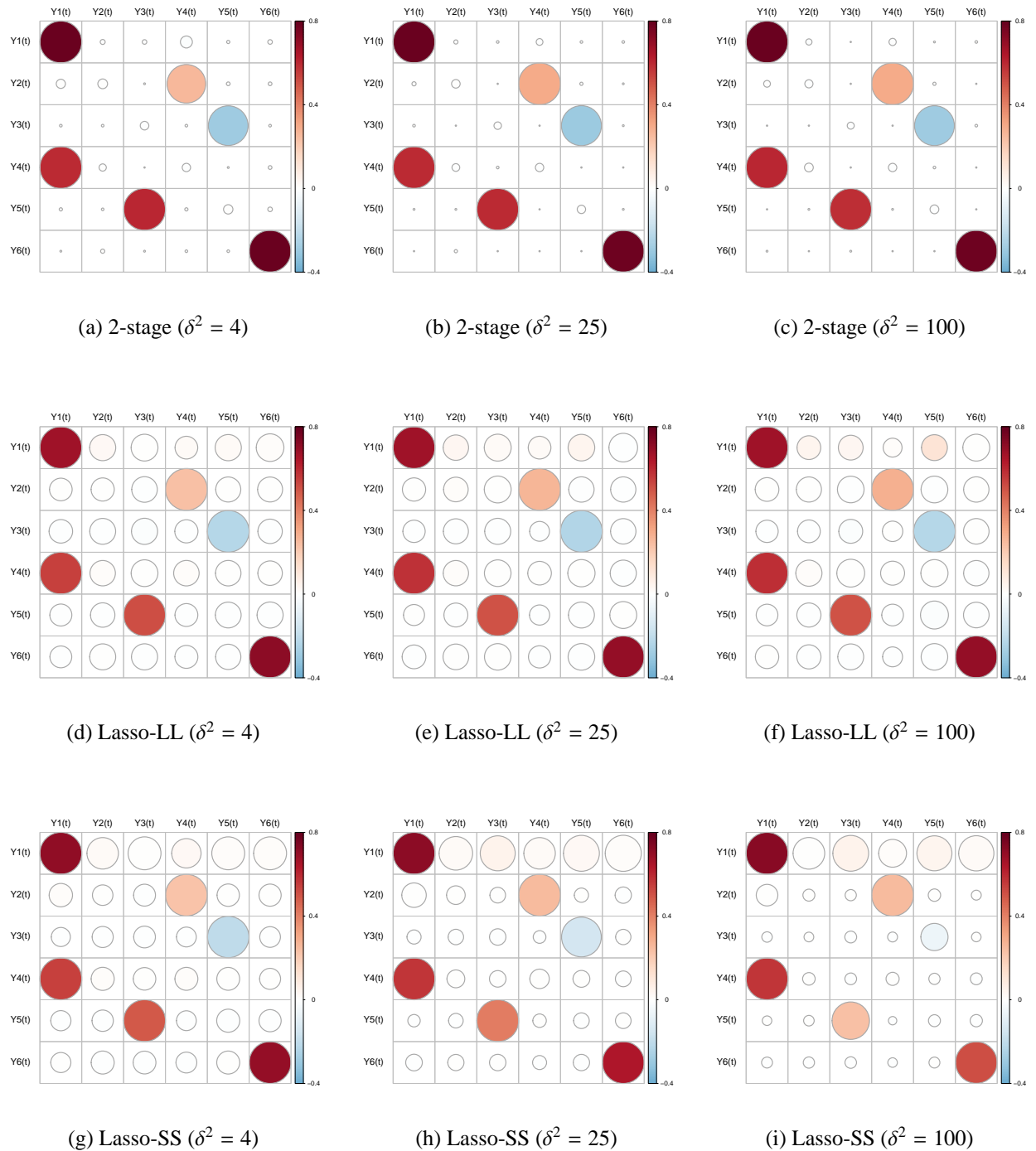


Figure 2: Displays of the AR coefficient estimates from the 2-stage approach, the Lasso-LL and the Lasso-SS methods when $\delta^2 = 4, 25$ and 100 , respectively.

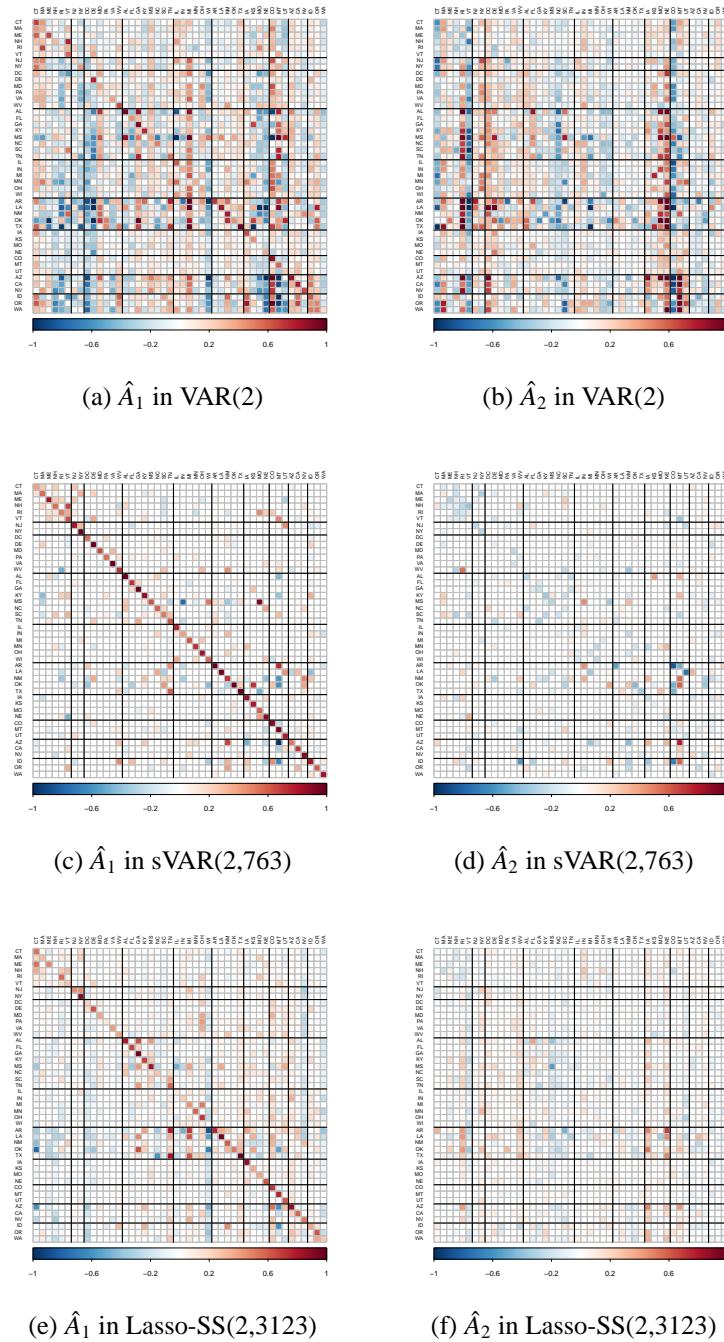


Figure 3: Displays of the AR coefficient estimates from the VAR(2), the sVAR(2,763) and the Lasso-SS(2,3123) models at lags 1 and 2, respectively.

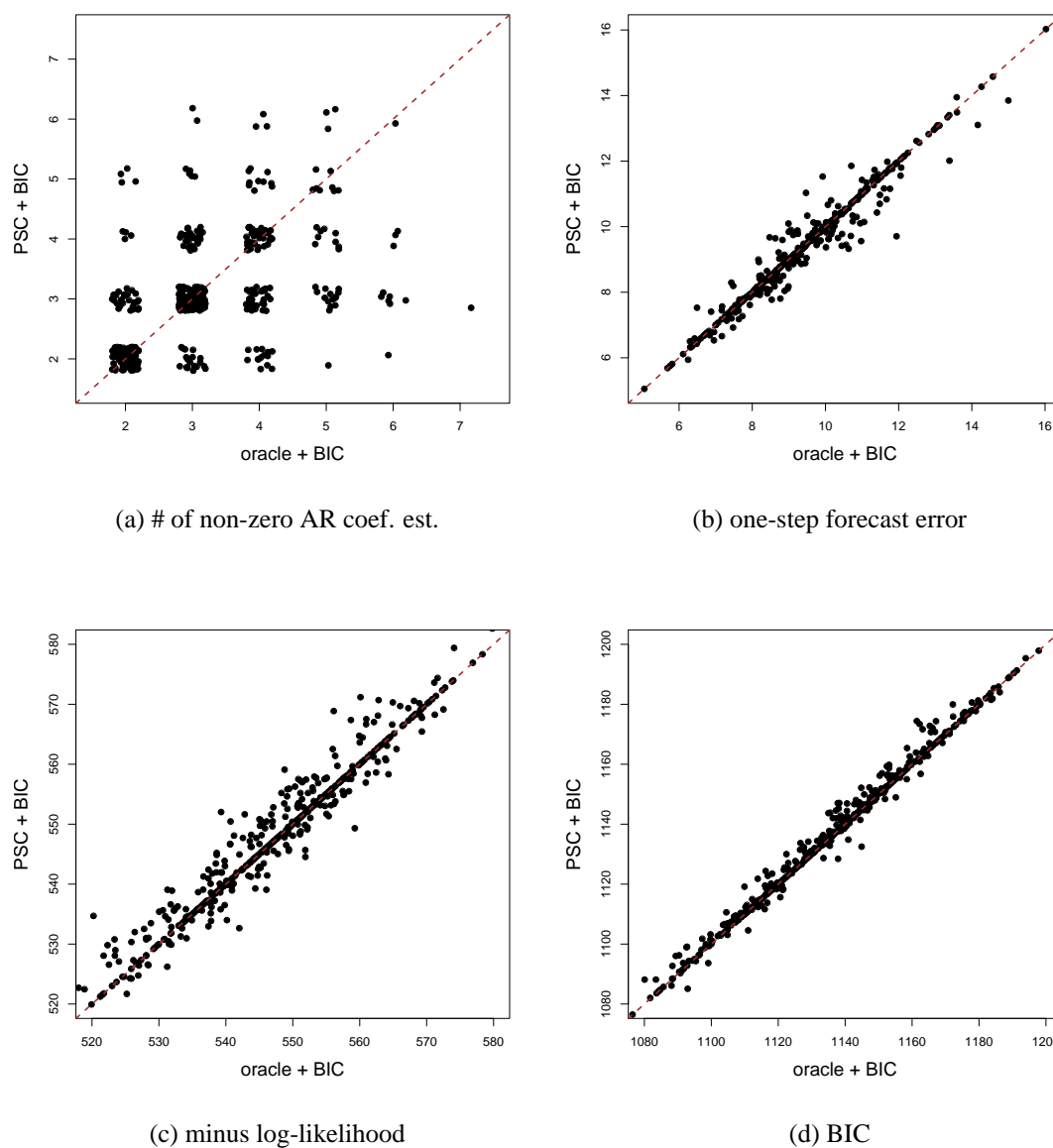


Figure 4: Comparison between the 2-stage approach and the modified 2-stage procedure.

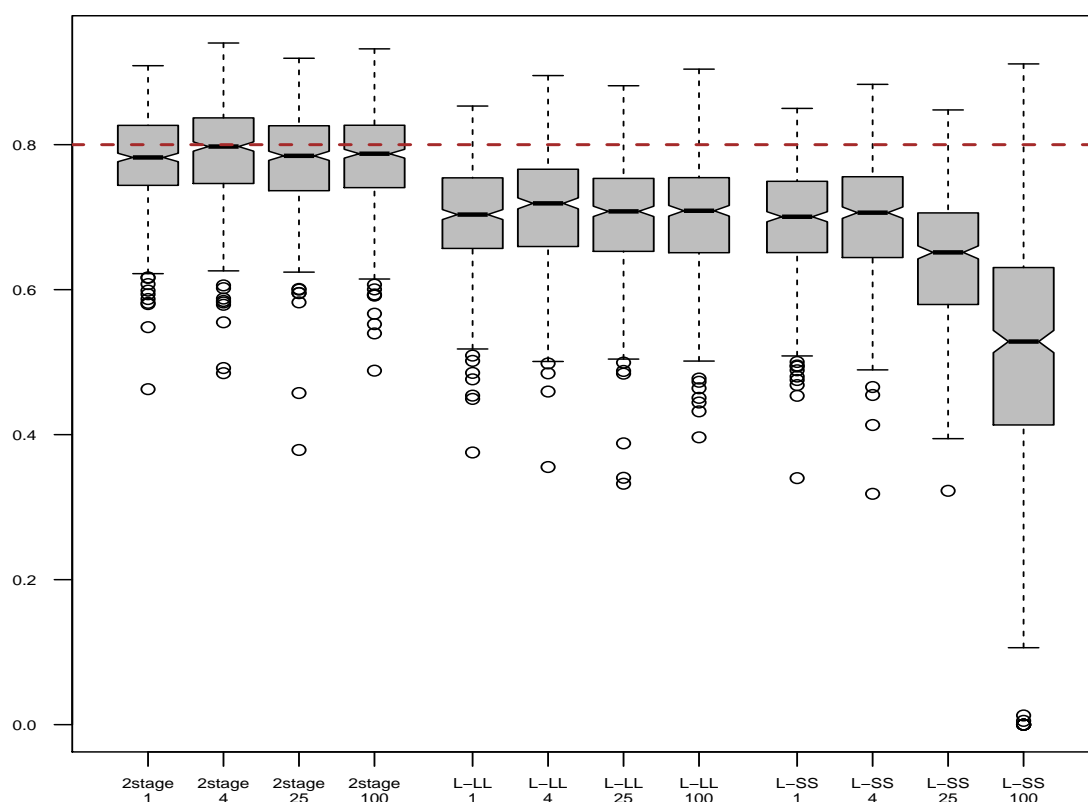


Figure 5: Sampling distributions of the estimators of $A_1(6, 6)$ from the 2-stage approach (the left 4 boxplots), the Lasso-LL method (the middle 4 boxplots) and the Lasso-SS method (the right 4 boxplots) for $\delta^2 = 1, 4, 25$ and 100 , respectively. The dashed horizontal line indicates the true value of $A_1(6, 6) = 0.8$.

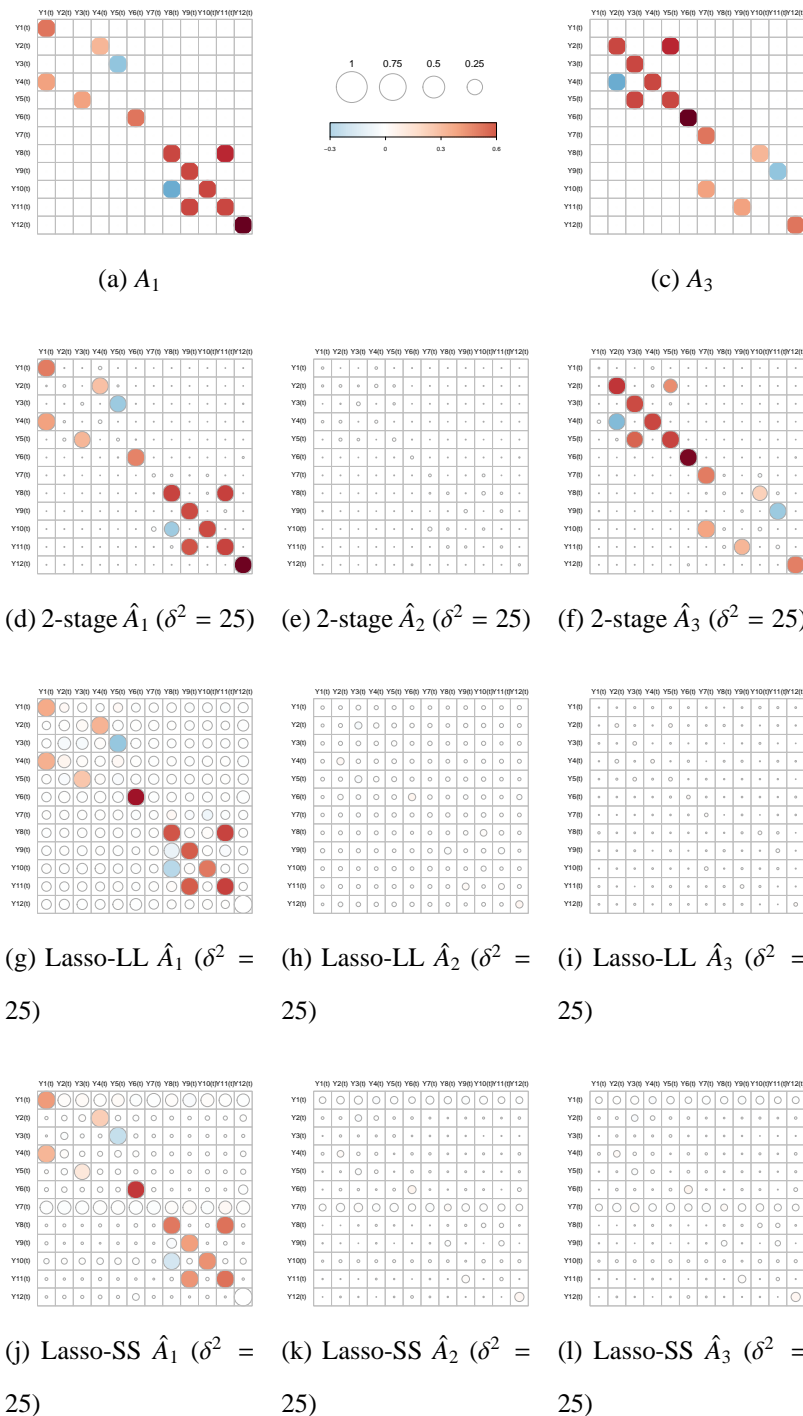


Figure 6: Displays of the true AR coefficient matrices and AR coefficient estimates from the 2-stage approach, the Lasso-LL and the Lasso-SS methods when $\delta^2 = 25$.

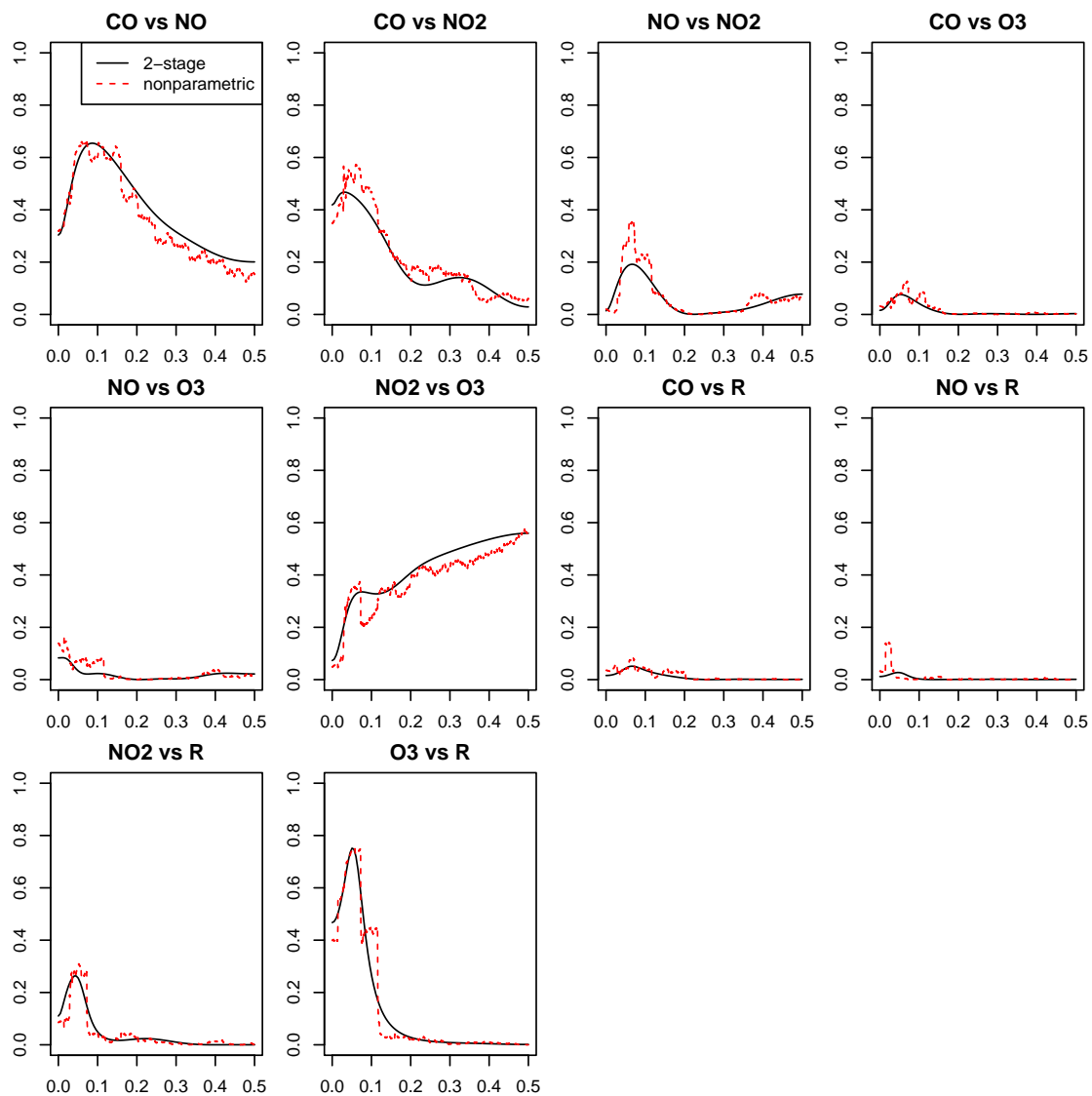


Figure 7: Plots of the parametric estimates of the squared modulus of PSC, i.e., $|\text{PSC}(\omega)|^2$, as computed from the AR coefficient estimates in the 2-stage sVAR(4,64) model (solid lines) and the non-parametric estimates of $|\text{PSC}(\omega)|^2$ used in the first stage selection (dashed lines).

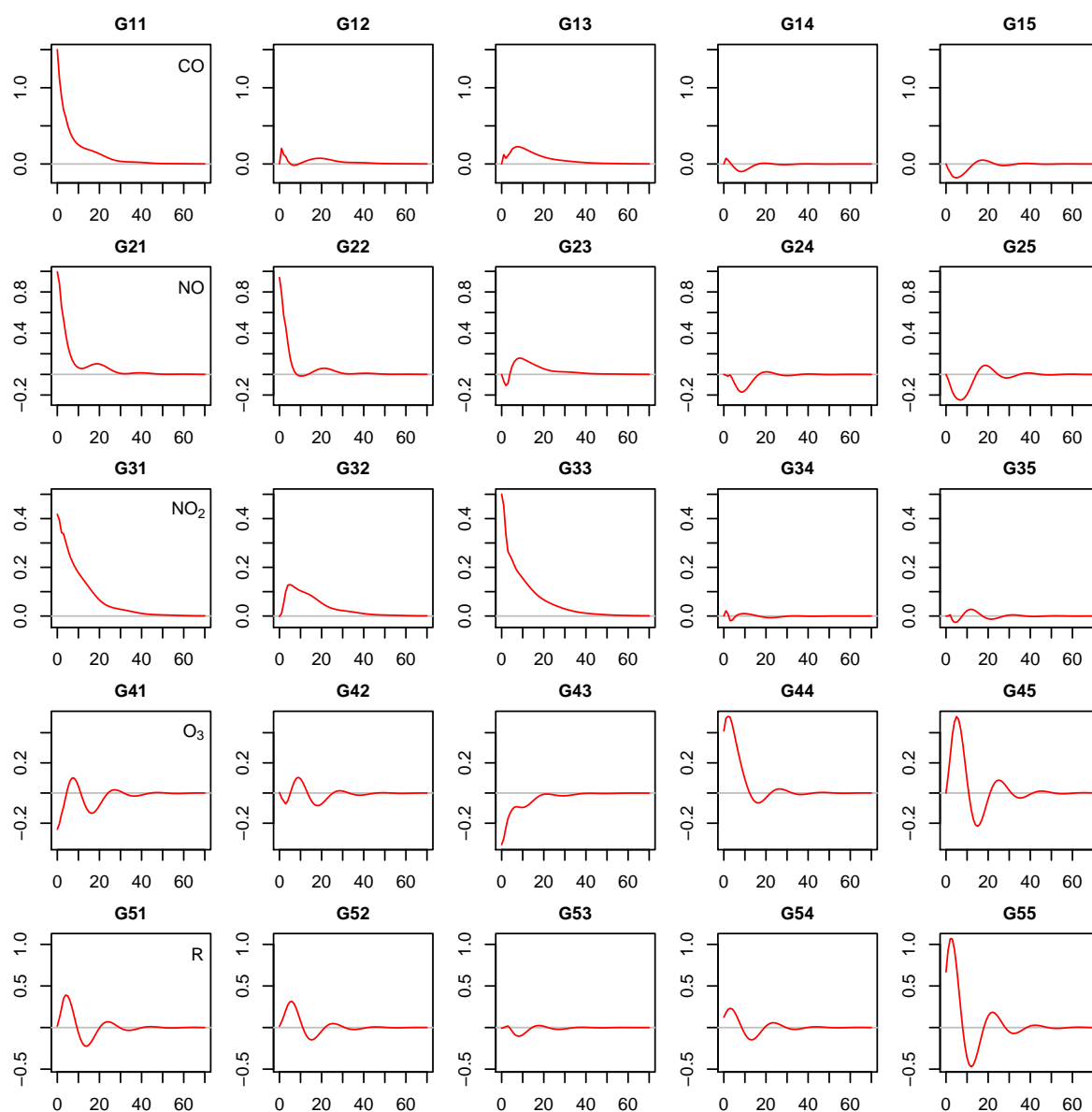


Figure 8: Plots of the impulse response coefficients from the 2-stage sVAR(4,64) model for the air pollutants time series data

Table 1: The five metrics from the 2-stage approach, the Lasso-LL and the Lasso-SS methods.

		\hat{p}	\hat{m}	bias ²	variance	MSE
$\delta^2 = 1$	2-stage	1.000	5.854	0.021	0.092	0.113
	Lasso-LL	1.208	17.852	0.060	0.099	0.159
	Lasso-SS	1.218	17.156	0.054	0.092	0.146
$\delta^2 = 4$	2-stage	1.000	6.198	0.006	0.087	0.093
	Lasso-LL	1.150	17.254	0.046	0.103	0.149
	Lasso-SS	1.246	16.478	0.053	0.136	0.188
$\delta^2 = 25$	2-stage	1.000	6.190	0.002	0.073	0.075
	Lasso-LL	1.179	17.275	0.042	0.274	0.316
	Lasso-SS	1.364	14.836	0.094	0.875	0.969
$\delta^2 = 100$	2-stage	1.000	6.260	0.003	0.175	0.178
	Lasso-LL	1.203	17.464	0.056	0.769	0.825
	Lasso-SS	1.392	11.108	0.298	2.402	2.700

Table 2: The h-step-ahead forecast root mean squared error (RMSE(h)).

Model	$h = 1$	$h = 2$	$h = 3$	$h = 4$
ar	340.5	395.1	477.9	556.2
sVAR(2,763)	321.6	336.0	360.5	397.6
Lasso-SS(2,3123)	331.6	356.5	404.9	439.4
VAR(2)	343.7	401.7	477.7	572.5

Table 3: The h-step-ahead logarithmic score ($LS(h)$).

Model	$h = 1$	$h = 2$	$h = 3$	$h = 4$
ar	319.3	344.1	357.6	365.9
sVAR(2,763)	313.8	322.4	326.6	328.8
Lasso-SS(2,3123)	335.3	348.6	345.9	342.5
VAR(2)	411.8	485.7	428.3	395.7

Table 4: The five metrics from the 2-stage approach, the Lasso-LL and the Lasso-SS methods.

		\hat{p}	\hat{m}	bias^2	variance	MSE
$\delta^2 = 1$	2-stage	3.000	24.962	0.163	0.378	0.541
	Lasso-LL	1.200	60.046	1.874	0.260	2.134
	Lasso-SS	1.576	72.650	1.435	0.466	1.901
$\delta^2 = 4$	2-stage	3.000	26.708	0.070	0.386	0.456
	Lasso-LL	1.256	63.848	1.794	0.314	2.108
	Lasso-SS	1.510	65.858	1.560	0.461	2.021
$\delta^2 = 25$	2-stage	3.000	27.556	0.055	0.408	0.463
	Lasso-LL	1.250	64.302	1.849	0.489	2.338
	Lasso-SS	1.598	44.288	1.633	1.186	2.819
$\delta^2 = 100$	2-stage	3.000	27.950	0.057	0.577	0.634
	Lasso-LL	1.318	66.602	1.867	1.110	2.977
	Lasso-SS	1.672	31.690	2.175	3.213	5.388

Table 5: Pairs with weak estimates of $|\text{PSC}(\omega)|^2$ in the 2-stage sVAR(4,64) model, as well as those found in [Dahlhaus \(2000\)](#), [Eichler \(2006\)](#) and [Songsiri et al. \(2010\)](#). [Songsiri et al. \(2010\)](#) used the same dataset as the sVAR(4,64) model; [Dahlhaus \(2000\)](#) and [Eichler \(2006\)](#) studied a similar dataset with the same 5 component series.

Model	Pairs with small estimates of $ \text{PSC}(\omega) ^2$
2-stage sVAR(4,64)	(CO, O ₃), (CO, R), (NO, R), (NO, O ₃)
Dahlhaus (2000)	(CO, O ₃), (CO, R), (NO, R), (NO, O ₃), (NO, NO ₂)
Eichler (2006)	(CO, O ₃), (CO, R), (NO, R), (NO, O ₃)
Songsiri et al. (2010)	(CO, O ₃), (CO, R), (NO, R)