Modeling Strategies for Large Dimensional Vector Autoregressions

Pengfei Zang

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

ABSTRACT

Modeling Strategies for Large Dimensional Vector Autoregressions

Pengfei Zang

The vector autoregressive (VAR) model has been widely used for describing the dynamic behavior of multivariate time series. However, fitting standard VAR models to large dimensional time series is challenging primarily due to the large number of parameters involved. In this thesis, we propose two strategies for fitting large dimensional VAR models. The first strategy involves reducing the number of non-zero entries in the autoregressive (AR) coefficient matrices and the second is a method to reduce the effective dimension of the white noise covariance matrix.

We propose a 2-stage approach for fitting large dimensional VAR models where many of the AR coefficients are zero. The first stage provides initial selection of non-zero AR coefficients by taking advantage of the properties of partial spectral coherence (PSC) in conjunction with BIC. The second stage, based on t-ratios and BIC, further refines the spurious non-zero AR coefficients post first stage. Our simulation study

suggests that the 2-stage approach outperforms Lasso-type methods in discovering sparsity patterns in AR coefficient matrices of VAR models. The performance of our 2-stage approach is also illustrated with three real data examples.

Our second strategy for reducing the complexity of a large dimensional VAR model is based on a reduced-rank estimator for the white noise covariance matrix. We first derive the reduced-rank covariance estimator under the setting of independent observations and give the analytical form of its maximum likelihood estimate. Then we describe how to integrate the proposed reduced-rank estimator into the fitting of large dimensional VAR models, where we consider two scenarios that require different model fitting procedures. In the VAR modeling context, our reduced-rank covariance estimator not only provides interpretable descriptions of the dependence structure of VAR processes but also leads to improvement in model-fitting and forecasting over unrestricted covariance estimators. Two real data examples are presented to illustrate these fitting procedures.

Contents

List of Tables **iii**

List of Figures **iv**

Acknowledgments **vi**

Chapter 1 Introduction **1**

1.1 Literature review 2

1.2 Overview of the thesis 8

Chapter 2 Sparse Vector Autoregressive Modeling **10**

2.1 Introduction 10

2.2 Sparse vector autoregressive (sVAR) models 13

2.3 A 2-stage approach of fitting sVAR models 17

2.3.1 Stage 1: selection 17

2.3.2 Stage 2: refinement 21

2.4 Numerical results 22

2.4.1 Simulation 23

2.4.2 Real data examples 33

2.5 Discussion 46

2.6 Appendix to Chapter 2 53
2.6.1 Constrained maximum likelihood estimation of sVAR models . 53
2.6.2 Implementation of fitting Lasso-VAR models 55

Chapter 3 Reduced-rank Covariance Estimation in Vector Autoregressive Modeling 61

3.1 Introduction 61
3.2 Reduced-rank covariance estimation 63
3.2.1 For independent observations 64
3.2.2 For VAR series 69
3.3 Numerical results 73
3.3.1 Simulation 73
3.3.2 Real data examples 76
3.4 Appendix to Chapter 3 87
3.4.1 Proof of Proposition 1 in Section 3.2.1 87
3.4.2 Approximation of MSE matrices of VAR forecasting 89

Chapter 4 Conclusions and Future Directions 92

Bibliography 93

List of Tables

2.1 Comparison between the 2-stage approach, the Lasso-LL and the Lasso-SS methods for the simulation example. 26

2.2 The forecast root mean squared error and the logarithmic score of the sVAR(2,763), the Lasso-SS(2,3123) and the VAR(2) models for the Google Flu Trends example. 39

2.3 Pairs with small estimate of $|\text{PSC}|^2$ in the 2-stage sVAR(4,64) model and in the partial correlation models of Dahlhaus (2000), Eichler (2006) and Songsiri et al. (2010) for the air pollutant example. 43

3.1 Comparison of assumptions between the latent variable model and the factor model. 66

3.2 Comparison between the RR, the LW2005 and the SS2005 covariance estimators for the simulation example. 75

List of Figures

2.1 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 the marginal variability is 1. 29

2.2 Displays of the AR coefficient estimates from the 2-stage approach, the Lasso-LL and the Lasso-SS methods when the marginal variability is 4, 25 and 100, respectively. 31

2.3 Sampling distributions of the estimators of $A_1(6,6)$ from the 2-stage approach, the Lasso-LL and the Lasso-SS methods, respectively. . . . 32

2.4 The first 100 observations of the Google Flu Trends example. 34

2.5 BIC curves of stages 1 and 2 of the 2-stage approach for the Google Flu Trends example. 36

2.6 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, for the Google Flu Trends example. 38

2.7 Hourly average series of the air pollutant example. 40

2.8 The parametric and the non-parametric estimates of $|\mathrm{PSC}|^2$ for the air pollutant example. 42

2.9 Displays of the AR coefficient estimates from the VAR(2) and the sVAR(2,42) models for the squared stock returns example. 45

2.10 ACF plots for the squared stock returns example. 47

2.11 Comparison of the estimate $|\hat{\text{PSC}}|^2$ from the 2-stage approach and the modified 2-stage procedure for the discussion example. 51

2.12 Comparison between the 2-stage approach and the modified 2-stage procedure for the discussion example. 52

3.1 The first 60 observations of the stock returns example. 77

3.2 Results of the reduced-rank covariance estimation for the stock returns example. 80

3.3 ACF and CCF plots of the estimated latent variable for the stock returns example. 81

3.4 Comparison of the confidence intervals of the AR coefficient estimates between $d = 54$ and $d = 7$ for the stock returns example. 83

3.5 Display of the difference between the approximate 1-step forecast MSE matrices when $d = 54$ and $d = 7$ for the stock returns example. 83

3.6 The first 36 observations of the temperature example. 84

3.7 Results of the reduced-rank covariance estimation for the temperature example. 86

Acknowledgments

Foremost I offer my sincerest gratitude to my advisors Professor Richard A. Davis and Professor Tian Zheng for their continuous support and encouragement of my PhD study. This thesis would not have been possible without the guidance from them.

I would like to thank Professor Yang Feng, Professor Upmanu Lall and Professor Serena Ng for kindly agreeing to serve on my committee. Their insightful comments and suggestions are valuable to improve my thesis.

I have been blessed with an accompany of numerous friends who gave me so much support, joy and attention. There are too many to thank by name, but I am especially grateful to Li An, Shuoshuo Han, Jie Ren, Tianjiao Yu and Dongzhou Zhang, whose friendship has been so important to me throughout the hard periods of my life.

Finally, I wish to express my profound gratitude to my parents for their everlasting love and encouragement. To them I owe everything that matters to me in this world. To them this thesis is dedicated.

To my parents for their everlasting love.

Chapter 1

Introduction

Large dimensional time series are encountered in many fields, including finance (Tao et al. 2011), environmental science (Lam and Yao 2011), biological study (Holter et al. 2001) and economics (Song and Bickel 2011). Statistical models have proven indispensable for analyzing the temporal evolution of such time series. However, fitting standard time series models to large dimensional series presents many challenges primarily due to the large number of parameters involved. For example, in a vector autoregressive (VAR) model, the number of parameters grows quadratically with the dimension of the series. As a result, standard VAR models are rarely applied to time series with more than 10 dimensions (Fan et al. 2011). Therefore complexity reduction becomes an important aspect in fitting time series models to large dimensional series.

In the literature, there are two major directions for reducing the complexity of time series models: the first direction is to reduce the number of free parameters involved; and the second direction is to reduce the dimension of the original series and model the lower-dimensional process. The focus of this thesis will be on the first

direction under the context of vector autoregressive (VAR) modeling. Specifically, we propose strategies for fitting large dimensional VAR models, where the number of free parameters involved is effectively reduced. In Section 1.1, we provide a review of existing approaches for fitting large dimensional time series models and in Section 1.2 we provide an overview of our VAR model fitting strategies.

1.1 Literature review

Suppose $\{Y_t\} = \{(Y_{t,1}, Y_{t,2}, \ldots, Y_{t,K})'\}$ is a K-dimensional time series and T observations $Y_1, \ldots, Y_T$ from the series are available. When the dimension K is large, fitting time series models to $\{Y_t\}$ is challenging and reducing the complexity of the model becomes an important aspect of the model fitting procedure.

One major direction for complexity reduction of time series models is to reduce the number of parameters via variable selection. We use the vector autoregressive (VAR) model as an example to review different variable selection methods for time series models. The vector autoregressive model of order p (VAR(p)) for the K-dimensional series $\{Y_t\}$ is given by

$$Y_t = \mu + \sum_{k=1}^{p} A_k Y_{t-k} + Z_t, \tag{1.1}$$

where μ is a $K \times 1$ vector of intercepts; $A_1, \ldots, A_p$ are real-valued $K \times K$ matrices of autoregressive (AR) coefficients; and $\{Z_t\}$ is a sequence of iid $K \times 1$ noise with mean $\mathbf{0}$ and covariance matrix Σ_Z. In the VAR model setup, current values of each marginal series of $\{Y_t\}$ are influenced by its own lagged values, as well as lagged values of other marginal series, with additive noise superimposed. Such a setup is well-suited for modeling the joint evolution of multiple series and therefore VAR models have been applied in many fields, such as political science (Freeman et al. 1989), meteorology

(Wilson 2010), macroeconomics (Sims 1980; Stock and Watson 2001), policy analysis (Bernanke et al. 2005), biological science (Holter et al. 2001) and finance (Eun and Shim 1989).

The K-dimensional VAR(p) model (1.1), when fully-parametrized, contains K^2p AR parameters. When the dimension K is small (e.g., $K \leq 5$), the AR coefficient matrices $A_1, \ldots, A_p$ can be efficiently estimated via least squares or maximum likelihood (when $\{Y_t\}$ is assumed to be Gaussian). When K is large (or even moderate), however, the number of AR parameters K^2p becomes comparable to or exceeds the sample size. As a result, least squares or maximum likelihood estimates of the AR parameters are not well behaved since there may exist a large number of spurious AR coefficient estimates, which can weaken the prediction performance and the interpretability of fitted VAR models. This fact limits the applicability of VAR models to large dimensional time series. Therefore different methods have been proposed to reduce the number of AR parameters by setting many entries in the AR coefficient matrices to zero. To determine which AR coefficients are zero, one possibility is based on hypothesis testing, see e.g. Fujita et al. (2007); Opgen-Rhein and Strimmer (2007)). For large dimensional VAR models, a large number of null hypotheses will be involved and the issue of multiple comparison needs to be addressed. Another possibility for determining non-zero AR coefficients is to reformulate the VAR model (1.1) as a linear regression problem

$$\begin{pmatrix} Y_{p+1}' \\ Y_{p+2}' \\ \vdots \\ Y_T' \end{pmatrix} = \begin{pmatrix} \mu' \\ \mu' \\ \vdots \\ \mu' \end{pmatrix} + \begin{pmatrix} Y_p' & Y_{p-1}' & \cdots & Y_1' \\ Y_{p+1}' & Y_p' & \cdots & Y_2' \\ \vdots & \vdots & \vdots & \vdots \\ Y_{T-1}' & Y_{T-2}' & \cdots & Y_{T-p}' \end{pmatrix} \begin{pmatrix} A_1' \\ A_2' \\ \vdots \\ A_p' \end{pmatrix} + \begin{pmatrix} Z_{p+1}' \\ Z_{p+2}' \\ \vdots \\ Z_T' \end{pmatrix},$$

where current values of the time series are treated as the response variable and lagged

values are treated as the explanatory variables. Then penalized regression can be applied to select non-zero AR coefficients. In the literature, 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, see e.g. Valdés-Sosa et al. (2005); Hsu et al. (2008); Arnold et al. (2008); Lozano et al. (2009); Haufe et al. (2010); 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, and it has the advantage of performing model selection and parameter estimation simultaneously. However there are also disadvantages in using Lasso in the context of VAR modeling. First, 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 VAR model even if the multivariate series $\{Y_t\}$ is Gaussian. This is because the noise covariance matrix Σ_Z in (1.1) is taken into account in the likelihood function of a Gaussian VAR series but not in the sum of squared residuals. We notice that this issue of choosing the loss function sometimes is not addressed in fitting Lasso-VAR models. For example, Arnold et al. (2008); Lozano et al. (2009); Haufe et al. (2010); 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. Second, Lasso has a tendency to over-select the number of non-zero AR coefficients and this phenomenon has been reported in various numerical results, see e.g. Arnold et al. (2008); Lozano et al. (2009); Shojaie and Michailidis (2010). The over-selected model complexity can lead to large mean squared error of the AR coefficient estimates and weakened interpretability of the VAR model. Third, in applying Lasso, the VAR model is reformulated as a linear regression problem. Such a formulation ignores the temporal dependence in the time

series. Song and Bickel (2011) give a theoretical discussion on the consequences of applying Lasso directly in VAR models without taking into account the temporal dependence between the response and the explanatory variables.

Fitting the VAR model (1.1) also involves estimating the noise covariance matrix Σ_Z. An estimate of Σ_Z is needed for exploring the dependence structure of the VAR process (Demiralp and Hoover 2003; Moneta 2004) while an estimate of Σ_Z^{-1} is required in constructing confidence intervals for AR coefficient estimates or computing the mean squared error of VAR forecasting (Lütkepohl 1993). A natural estimator for Σ_Z in VAR models is the sample covariance matrix of the residuals from fitting an autoregression (Lütkepohl 1993). To this end, the residuals are treated as independent samples, conditioned on the AR coefficient estimates. Therefore estimating Σ_Z in VAR models can be cast as a covariance estimation problem where independent observations are available. Estimating a $K \times K$ covariance matrix from independent observations is challenging for large K since the number of parameters to be estimated, which is $K(K+1)/2$, grows quadratically in the dimension K. The sample covariance matrix, which serves as a natural estimator, is known to be severely ill-conditioned for large dimension. As a result, various methods are proposed for estimating large dimensional covariance matrices. In the literature, there exist three main approaches for covariance estimation under large dimensionality. The first is the *shrinkage* approach, where the covariance estimator is obtained by shrinking the sample covariance matrix towards a pre-specified covariance structure (Ledoit and Wolf 2004; Schäfer and Strimmer 2005); the second is the *regularization* approach, where the covariance estimator is derived based on regularization methods, such as banding (Bickel and Levina 2008), thresholding (El Karoui 2008) and penalized estimation (Huang et al. 2006); and the third is the *structure* approach, where structural constraints, such as factor structures (Tipping and Bishop 1999) or autoregressive structures (Daniels and

Kass 2001), are imposed to reduce the effective dimension of the covariance estimator. In the context of VAR modeling, these covariance estimators can be applied by viewing the residuals from the fitted autoregression, conditioned on the AR coefficient estimates, as independent samples from a multivariate distribution with covariance matrix Σ_Z.

Another major direction for reducing the complexity of time series models is via dimension-reduction of the original series, where subsequent analysis can be carried out using the dimension-reduced process. One of the most frequently used methods for time series dimension-reduction is via factor models. The factor model for the K-dimensional time series $\{Y_t\}$ is given by

$$Y_t = OX_t + \varepsilon_t, \tag{1.2}$$

where $\{X_t\} = \{(X_{t,1}, \ldots, X_{t,r})'\}$ is a r-dimensional unobserved series with (unknown) $r < K$; O is a $K \times r$ matrix of unknown coefficients; and each $\varepsilon_t = (\varepsilon_{t,1}, \ldots, \varepsilon_{t,K})'$ is a K-dimensional *idiosyncratic* error. In factor analysis, the r components of $\{X_t\}$ are referred to as *factors* and the matrix O is referred to as the *factor loading*. The dimension-reduction of $\{Y_t\}$ is achieved in the sense that the dynamics of $\{Y_t\}$ is explained by the evolution of a lower-dimensional process $\{X_t\}$ and subsequent analysis can be applied to the dimension-reduced series $\{X_t\}$. For example, Pan and Yao (2008) fitted a standard VAR model to the dimension-reduced factor process while such VAR modeling is inappropriate for the original large dimensional series.

The traditional factor models, see e.g. Zellner (1970); Fama and French (1993), are developed under the setting that the sample size T grows large while the dimension of the process K remains bounded. In addition, the following assumptions on the dependence structures of the factors $\{X_t\}$ and the errors $\{\varepsilon_t\}$ are usually made for traditional factor models: first, both the factors $\{X_t\}$ and the errors $\{\varepsilon_t\}$ are

cross-sectionally uncorrelated; second, the factors $\{X_t\}$ are independent with the errors $\{\varepsilon_t\}$. With the recent availability of large dimensional time series data, new results of factor models are developed under the setting where both the number of observations and the dimension of the process are large ("large-T-large-K"). In addition, the aforementioned assumptions on the factors $\{X_t\}$ and the errors $\{\varepsilon_t\}$ in traditional factor models appear to be too restrictive for many applications, especially in economics and finance (Bai and Ng 2008). Therefore efforts have been made to develop "large-T-large-K" factor models where those assumptions are relaxed. For example, Bai and Ng (2002) allow weak serial and cross-sectional dependence for the errors $\{\varepsilon_t\}$; and Pan and Yao (2008) consider including serial dependence between the factors $\{X_t\}$ and the errors $\{\varepsilon_t\}$. Estimation and inference of such "large-T-large-K" factor models with less strict assumptions on the factors $\{X_t\}$ and the errors $\{\varepsilon_t\}$ is an active research area (Bai and Ng 2002; Stock and Watson 2006; Lam and Yao 2011).

In the factor model (1.2), only the series $\{Y_t\}$ is observed while the loading matrix O, the factors $\{X_t\}$ and the errors $\{\varepsilon_t\}$ are all unobserved and need to be estimated from data. For the estimation of factor models, identification of the number of factors r is a central problem. Under certain specific (but unusual) instances, factors can be specified by taking advantage of expert knowledge, see e.g.Engle and Watson (1981), but much more often the number of factors needs to be determined from data. Different methods have been proposed to estimate the number of factors r in large dimensional factor models. For example, Bai and Ng (2002) develop modified versions of AIC- and BIC-type information criteria to determine r; Pan and Yao (2008) identify r by embedding multivariate Portmanteau tests (Li and Mcleod 1981) in a stepwise procedure of expanding the idiosyncratic error space; and Lam et al. (2011) propose a ratio-based estimator for r based on an eigen-analysis of a matrix-

valued function of the autocovariance matrices of $\{Y_t\}$. Once the number of factors r is estimated, the factors $\{X_t\}$ and the loading matrix O can be jointly estimated, under additional identifying conditions, via the eigen-structure of $\{Y_t\}$. For example, Connor and Korajczyk (1986) and Bai and Ng (2002) identify the factors $\{X_t\}$ from a least squares perspective and their methods rely on the eigen-decomposition of the covariance matrix of $\{Y_t\}$; Lam and Yao (2011) and Lam et al. (2011) estimate the factors $\{X_t\}$ by performing an eigen-analysis on a matrix-valued function of the autocovariance matrices of $\{Y_t\}$ at non-zero lags. A survey of recently developed theories and methods for large dimensional factor models is given in Bai and Ng (2008).

1.2 Overview of the thesis

In this thesis, we propose two strategies for fitting large dimensional VAR models. The first strategy involves reducing the number of non-zero AR parameters and the second is a method to reduce the effective dimension of the noise covariance matrix.

In Chapter 2, we propose a 2-stage approach for fitting large dimensional VAR models, where many entries of the AR coefficient matrices $A_1, \ldots, A_p$ are zero. The first stage selects non-zero AR coefficients by screening pairs of distinct marginal series that are conditionally correlated. The conditional correlation between marginal series is quantified by the partial spectral coherence (PSC), a tool in frequency-domain time series analysis. 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 the first stage may contain spurious non-zero coefficients. To further refine the model, in the second stage, we propose a screening strategy based on the t-ratios of the AR coefficient estimates and the BIC.

In the simulation study in Section 2.4.1, our 2-stage approach outperforms Lasso-type methods in discovering non-zero AR coefficients in VAR models. The performance of the 2-stage approach is also illustrated with three real data examples in Section 2.4.2.

In Chapter 3, we consider a method of estimating the noise covariance matrix in a large dimensional VAR model via reducing its effective dimension. We first propose a reduced-rank covariance estimator under the setting of independent observations and give the analytical form of its maximum likelihood estimate. The reduced-rank estimator comes from a latent variable model for the vector observation and it can be viewed as a structure covariance estimator (i.e., the third approach of covariance estimation described in Section 1.1). The effective dimension of the reduced-rank covariance estimator is determined according to information criterion and can be much lower than the dimension of the population covariance matrix. The reduced-rank estimator is attractive since it is not only well-conditioned but also provides an interpretable description of the covariance structure. We demonstrate, in the simulation study in Section 3.3.1, that the reduced-rank covariance estimator outperforms two competing shrinkage estimators in estimating large dimensional covariance structures. Then we describe how to integrate the proposed reduced-rank covariance estimator into the fitting of large dimensional VAR models, for which we consider two scenarios that require different model fitting procedures. The first scenario is that there is no constraint on the AR coefficients, for which the VAR model can be fitted using a 2-step method; while the second scenario is that there exist constraints on the AR coefficients, for which the VAR model needs to be fitted by an iterative procedure. Two real data examples are presented to illustrate these model fitting procedures in Section 3.3.2. Results from the real data example suggest that using the reduced-rank covariance estimator can lead to improvement in model-fitting and forecasting than using unrestricted covariance estimators in large dimensional VAR modeling.

Chapter 2

Sparse Vector Autoregressive Modeling

2.1 Introduction

The vector autoregressive (VAR) model has been widely used for modeling the temporal dependence structure 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 marginal series. The VAR model is well suited to describe such temporal dependence structures. However, the conventional VAR model can be saturatedly-parametrized with the number of AR coefficients prohibitively large for high (and even moderate) dimensional processes. This can result in noisy parameter estimates, unstable predictions and difficult-to-interpret descriptions of the temporal dependence. To overcome these drawbacks, in this chapter we propose a 2-stage approach for fitting sparse VAR (sVAR) models in which many of the autoregressive (AR) coefficients are zero. Such

sVAR models can enjoy improved efficiency of parameter estimates, better prediction accuracy and more interpretable descriptions of the temporal dependence structure. 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 a 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, see e.g. Valdés-Sosa et al. (2005); Hsu et al. (2008); Arnold et al. (2008); Lozano et al. (2009); Haufe et al. (2010); Shojaie and Michailidis (2010); Song and Bickel (2011). The Lasso-VAR modeling approach has the advantage of performing model selection and parameter estimation simultaneously. It can also be applied under the "large-p-small-n" setting. 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, see e.g. 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 the VAR model without taking into account the temporal dependence between the response and the explanatory variables.

In this chapter, 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 component 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 component series 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 selected. The VAR model fitted in the first stage may contain spurious non-zero coefficients. To further refine the fitted model, we propose, in stage 2, a screening strategy based on the t-ratios of the coefficient estimates as well as BIC.

The remainder of this chapter is organized as follows. In Section 2.2, we review some results on the VAR model for multivariate time series and introduce the basic properties related to PSC. In Section 2.3, we describe a 2-stage procedure for fitting sVAR models. Connections between our first stage selection procedure with Granger causal models are give in Section 2.3.1. In Section 2.4.1, simulation results are presented to compare the performance of the 2-stage approach with the Lasso-VAR approach. In Section 2.4.2 the 2-stage approach is applied to fit sVAR models to three real data examples: the first example is the Google Flu Trends data (Ginsberg et al. 2009); the second example is a time series of concentration levels of air pollutants (Songsiri et al. 2010); and the third example is concerned with squared stock returns from S&P 500. Further discussion is contained in Section 2.5. Supplementary material is given in the Appendix.

2.2 Sparse vector autoregressive (sVAR) models

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

$$Y_t = \mu + \sum_{k=1}^{p} A_k Y_{t-k} + Z_t, \ t = 0, \pm 1, \ldots, \tag{2.1}$$

where μ is a $K \times 1$ vector of intercepts; $A_1, \ldots, A_p$ are real-valued $K \times K$ matrices of autoregressive (AR) coefficients; and $\{Z_t\}$ is a sequence of iid $K \times 1$ Gaussian noise with mean $\mathbf{0}$ and non-degenerate covariance matrix Σ_Z. [1] 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 $z \in \mathbb{C}, |z| < 1$, see e.g. 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).

The temporal dependence structure of the VAR model (2.1) is characterized by the AR coefficient matrices $A_1, \ldots, A_p$. Based on T observations $Y_1, \ldots, Y_T$ 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 parameter estimates and difficult-to-interpret descriptions of the temporal dependence structure. It is also generally believed that, for most applications, the true model of the series is sparse, i.e., the number of non-zero AR coefficients is small. Therefore it is preferable to fit a *sparse* VAR (sVAR) model in which many of its AR parameters are zero. For this purpose we develop a 2-stage approach of fitting such sVAR models. The first stage selects non-zero AR coeffi-

[1] In this chapter we assume that the VAR(p) process $\{Y_t\}$ is Gaussian. When $\{Y_t\}$ is non-Gaussian, the 2-stage model fitting approach can still be applied, where now the Gaussian likelihood is interpreted as a quasi-likelihood.

cients by screening pairs of distinct marginal series that are conditionally correlated. To compute direction-free conditional correlation between component series, we use tools from the frequency-domain time series analysis, 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, see e.g. 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

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

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 all $k \in \mathbb{Z}$. 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

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

where $\gamma_{ij}^{\varepsilon}(k) := \mathrm{cov}(\varepsilon_{t+k,i}, \varepsilon_{t,j})$. 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), see e.g. 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.,

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

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

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

which involves inverting a $(K-2) \times (K-2)$ dimensional matrix, i.e., $f_{-ij,-ij}^{Y}(\omega)^{-1}$. Therefore using (2.5) to compute the PSCs 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 PSCs for all $\binom{K}{2}$ pairs through the inverse of the spectral density of the process $\{Y_t\}$, 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 ith diagonal, the jth 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

$$\mathrm{PSC}_{ij}(\omega) = -\frac{g^Y_{ij}(\omega)}{\sqrt{g^Y_{ii}(\omega) g^Y_{jj}(\omega)}},\ \omega \in (-\pi, \pi]. \tag{2.6}$$

Now the computation of all $\binom{K}{2}$ PSCs using (2.6) requires only one matrix inversion of the $K \times K$ dimensional matrix $f^Y(\omega)$. From (2.6), it also follows that

$$\begin{aligned} &\{Y_{t,i}\} \text{ and } \{Y_{t,j}\}\ (i \neq j) \text{ are conditionally uncorrelated} \\ &\text{if and only if } g^Y_{ij}(\omega) = 0,\ \text{for all } \omega \in (-\pi, \pi]. \end{aligned} \tag{2.7}$$

In other words, the inverse spectral density $g^Y(\omega)$ encodes the pairwise conditional correlation between the marginal series of $\{Y_t\}$. This generalizes the problem of *covariance selection* in which independent samples are available, see e.g. Dempster (1972); Friedman et al. (2008). Covariance selection is concerned about studying the conditional relationship between dimensions of a multivariate Gaussian distribution by locating zero entries in the inverse covariance matrix. For example, suppose $X := (X_1, \ldots, X_K)'$ follows a K-dimensional Gaussian distribution $N(0, \Sigma_X)$. It is known that two distinct dimensions of X, say X_i and X_j $(i \neq j)$, are conditionally independent given the other $(K-2)$ dimensions X_{-ij}, if and only if the (i,j)th entry in the inverse covariance matrix Σ_X^{-1} is zero, i.e.,

$$X_i \text{ and } X_j\ (i \neq j) \text{ are conditionally independent iff } \Sigma_X^{-1}(i,j) = 0. \tag{2.8}$$

If the process $\{Y_t\}$ consists of independent replicates of a Gaussian distribution $N(0, \Sigma_Y)$, then its spectral density $f^Y(\omega) = \frac{1}{2\pi}\Sigma_Y$ remains constant over $\omega \in (-\pi, \pi]$ and (2.7) becomes,

$$\{Y_{t,i}\} \text{ and } \{Y_{t,j}\}\ (i \neq j) \text{ are conditionally uncorrelated iff } \Sigma_Y^{-1}(i,j) = 0, \tag{2.9}$$

which coincides with (2.8). Therefore selection of conditionally uncorrelated series using the inverse spectral density contains the covariance selection problem as a special case.

2.3 A 2-stage approach of fitting sVAR models

In this section, we describe 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 the first stage may still contain spurious non-zero AR coefficients. To address this possibility, a second stage is used to refine the model further.

2.3.1 Stage 1: selection

As we have shown, 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 AR coefficients between two conditionally uncorrelated marginal series to zero, i.e.,

$$A_k(i,j) = A_k(j,i) = 0 \text{ for } i \neq j, k = 1, \ldots, p \quad \text{if } \{Y_{t,i}\} \text{ and } \{Y_{t,j}\} \text{ are conditionally uncorrelated,} \tag{2.10}$$

where the latter is equivalent to $\mathrm{PSC}_{ij}(\omega) = 0$ for $\omega \in (-\pi, \pi]$. From (2.10) 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. We point out that our proposed connection from zero PSCs to zero AR coefficients, as described by (2.10), 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 for such examples. We will return to this point in Section 2.5.

In order to set a group of AR coefficients to zero as in (2.10), we need to find the pairs of marginal series for which the PSC is identically zero. Due to sampling variability, however, the estimated PSC, denoted by $\hat{\text{PSC}}_{ij}(\omega)$ for the pair of series $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$, will not be exactly zero even when the two corresponding marginal series are conditionally uncorrelated. To deal with this issue, we rank the estimated PSCs based on their evidence to be non-zero and decide a cutoff point that separates non-zero PSCs from zero PSCs. 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 et al. (1997) and Dahlhaus (2000), we use the supremum of the squared modulus of the estimated PSC, i.e.,

$$\hat{S}_{ij} := \sup_{\omega} |\hat{\text{PSC}}_{ij}(\omega)|^2, \tag{2.11}$$

as the summary statistic, where the supremum is taken over the Fourier frequencies $\{2\pi k/T : k = 1, \ldots, T\}$. A large value of $\hat{S}_{ij}$ indicates that the two marginal series $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$ are likely to be conditionally correlated. Therefore we can create a sequence $\mathbb{Q}_1$ of the $\binom{K}{2}$ pairs of distinct marginal series by ranking each pair's summary statistic (2.11) from highest to lowest. In other words, the sequence $\mathbb{Q}_1$ ranks the $\binom{K}{2}$ pairs of marginal series based on their evidence to be conditionally correlated and thereby, according to (2.10), prioritizes the way in which non-zero coefficients are added into the VAR model. Based on the sequence $\mathbb{Q}_1$, we also need to determine two parameters to fully specify the VAR model: the order of autoregression

p and the number of *top* pairs in $\mathbb{Q}_1$, 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 (Schwarz (1978)) to simultaneously choose the values of these two parameters. The BIC is computed as,

$$\text{BIC}(p, M) = -2 \log L(\hat{A}_1, \ldots, \hat{A}_p, \hat{\Sigma}_Z) + \log T \cdot (K + 2M)p, \tag{2.12}$$

where $L(\hat{A}_1, \ldots, \hat{A}_p, \hat{\Sigma}_Z)$ is the maximized likelihood of the VAR model. To compute $L(\hat{A}_1, \ldots, \hat{A}_p, \hat{\Sigma}_Z)$, we use results on the constrained maximum likelihood estimation of VAR models as given in Lütkepohl (1993). Details of the estimation procedure can be found Appendix 2.6.1.

Restricting the two parameters p and M to take values in pre-specified ranges $\mathbb{P}$ and $\mathbb{M}$, respectively, where $\mathbb{M}$ is usually specified as $\mathbb{M} = \{0, 1, \ldots, K(K-1)/2\}$, the steps of **stage 1** can be summarized as follows.

Step 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 [2] and applying equation (2.6).

Step 2. Construct a sequence $\mathbb{Q}_1$ of the $K(K-1)/2$ pairs of distinct marginal series by ranking each pair's summary statistic $\hat{S}_{ij}$ (2.11) from highest to lowest.

Step 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 $\mathbb{Q}_1$ into the VAR model, which specifies the parameter constraint on the AR coefficients. Conduct parameter estimation under this constraint using the results in Appendix 2.6.1 and compute the corresponding $\text{BIC}(p, M)$ according to (2.12).

Step 4. Choose $(\tilde{p}, \tilde{M})$ that gives the minimum BIC value over $\mathbb{P} \times \mathbb{M}$.

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., $\tilde{M} << K(K-1)/2$, $(K+2\tilde{M})\tilde{p}$ can be much smaller than $K^2\tilde{p}$, which is the number of AR coefficients in a fully-parametrized VAR($\tilde{p}$) model.

In the first stage we execute group selection of AR coefficients by using PSC in conjunction with BIC. This use of the group structure of AR coefficients effectively reduces the number of candidate models to be examined in the first stage. Similar use of the group structure of AR coefficients has also been employed in other settings, one of which is to determine the *Granger causality* between time series. The concept of Granger causality was first introduced by Granger (1969) in econometrics. It is shown that, see e.g. Lütkepohl (1993), a Granger causal relationship can be examined by fitting VAR models to the multivariate time series in question, where non-zero AR coefficients indicate Granger causality between the corresponding series. In the literature, l_1-penalized regression (Lasso) has been widely used to explore sparsity in the pattern of Granger causal relationship by shrinking AR coefficients to zero, see e.g. Arnold et al. (2008); Shojaie and Michailidis (2010). In particular, Lozano et al. (2009) and Haufe et al. (2010) proposed to penalize groups of AR coefficients simultaneously, for which their use of the group structure of AR coefficients is similar to (2.10). In spite of their common purpose of fitting sparse VAR models, simulation results in Section 2.4.1 demonstrate the advantage of using PSC in conjunction with BIC over Lasso in discovering sparsity in AR coefficients. Detailed discussion on using VAR models to determine Granger causality can be found in Granger (1969);

[2]Here we use the periodogram smoothed by a modified Daniell kernel, see e.g. Brockwell and Davis (1991), as the non-parametric estimate of the spectral density. Alternative spectral density estimators, such as the shrinkage estimate proposed by Böhm and von Sachs (2009), can also be applied.

Lütkepohl (1993); Arnold et al. (2008).

2.3.2 Stage 2: refinement

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

$$t_{i,j,k} := \frac{\hat{A}_k(i,j)}{\text{s.e.}(\hat{A}_k(i,j))}. \tag{2.13}$$

Here the standard error of $\hat{A}_k(i,j)$ is computed from the asymptotic distribution of the constrained maximum likelihood estimator of the first stage model, which is, see e.g. Lütkepohl (1993),

$$\sqrt{T}(\hat{\alpha} - \alpha) \overset{d}{\Longrightarrow} N(0,\ \tilde{R}[\tilde{R}'(\tilde{\Gamma}_Y \otimes \tilde{\Sigma}_Z^{-1})\tilde{R}]^{-1}\tilde{R}'), \tag{2.14}$$

where $\alpha := \text{vec}(A_1, \ldots, A_p)$ is the K^2p-dimensional vector obtained by column stacking the AR coefficient matrices $A_1, \ldots, A_p$; $\hat{\alpha}$, $\tilde{\Gamma}_Y$ and $\tilde{\Sigma}_Z$ are the maximum likelihood estimators of α, $\Gamma_Y := \text{cov}((Y_t', \ldots, Y_{t-p+1}')')$ and Σ_Z, respectively; and $\tilde{R}$ is the *constraint matrix*, defined by equation (2.17) in Appendix 2.6.1, of the first stage model. So we can create a sequence $\mathbb{Q}_2$ of the $(K + 2\tilde{M})\tilde{p}$ triplets (i, j, k) by ranking the absolute values of the t-ratios (2.13) from highest to lowest. The AR coefficients

corresponding to the *top* triplets in $\mathbb{Q}_2$ are most likely to be retained in the model because of their significance. In the second stage, there is only one parameter, denoted by m, that controls the complexity of the model, which is the number of non-zero AR coefficients to be retained. And BIC is also used to determine the complexity of the final model. The steps of **stage 2** are as follows.

Step 1. Compute the t-statistic $t_{i,j,k}$ (2.13) for each of the $(K + 2\tilde{M})\tilde{p}$ non-zero AR coefficient estimates of the first stage model.

Step 2. Create a sequence $\mathbb{Q}_2$ of the $(K + 2\tilde{M})\tilde{p}$ triplets (i, j, k) by ranking $|t_{i,j,k}|$ from highest to lowest.

Step 3. For each $m \in \{0, 1, \ldots, (K + 2\tilde{M})\tilde{p}\}$, consider the model that selects the m non-zero AR coefficients corresponding to the top m triplets in the sequence $\mathbb{Q}_2$. Under this parameter constraint, conduct the constrained parameter estimation using the results in Appendix 2.6.1 and compute the corresponding BIC according to $\text{BIC}(m) = -2 \log L + \log T \cdot m$.

Step 4. Choose $\hat{m}$ that gives the minimum BIC value over $\{0, 1, \ldots, (K + 2\tilde{M})\tilde{p}\}$.

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

2.4 Numerical results

In this section we provide numerical results on the performance of our 2-stage approach of fitting sVAR models. In Section 2.4.1, simulation results are presented

to compare the performance of the 2-stage approach against competing Lasso-type methods of fitting sVAR models. In Section 2.4.2, the 2-stage approach is applied to three real data examples. The first example is the Google Flu Trends data; the second example is a time series of concentration levels of air pollutants; and the third example is concerned with squared stock returns from S&P 500.

2.4.1 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, see e.g. Valdés-Sosa et al. (2005); Hsu et al. (2008); Arnold et al. (2008); Lozano et al. (2009); Haufe et al. (2010); 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 series in consideration is Gaussian. This is because the noise covariance matrix Σ_Z in (2.1) 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 (see Appendix 2.6.2). 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); Haufe et al. (2010); 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 noise covariance matrix Σ_Z is diagonal or simply the identity matrix. 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 while 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 2.6.2.

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 order of autoregression 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 much 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 the model fitting procedure, performs the worst among the three.

Here we describe the simulation example. Consider the 6-dimensional VAR(1)

process $\{Y_t\} = \{(Y_{t,1}, \ldots, Y_{t,6})'\}$ given by

$$\begin{pmatrix} Y_{t,1} \\ Y_{t,2} \\ Y_{t,3} \\ Y_{t,4} \\ Y_{t,5} \\ Y_{t,6} \end{pmatrix} = \begin{pmatrix} 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.3 & 0 \\ 0.6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.8 \end{pmatrix} \begin{pmatrix} Y_{t-1,1} \\ Y_{t-1,2} \\ Y_{t-1,3} \\ Y_{t-1,4} \\ Y_{t-1,5} \\ Y_{t-1,6} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \\ Z_{t,3} \\ Z_{t,4} \\ Z_{t,5} \\ Z_{t,6} \end{pmatrix}, \tag{2.15}$$

where $\{Z_t\} = \{(Z_{t,1}, \ldots, Z_{t,6})'\}$ are iid Gaussian noise with mean $\mathbf{0}$ and covariance matrix Σ_Z. The order of autoregression in (2.15) is $p = 1$ and there are 6 non-zero AR coefficients, so (2.15) specifies a sVAR$(1, 6)$ model. The covariance matrix Σ_Z of the Gaussian noise is

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

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 see 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

Table 2.1: 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

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

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

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

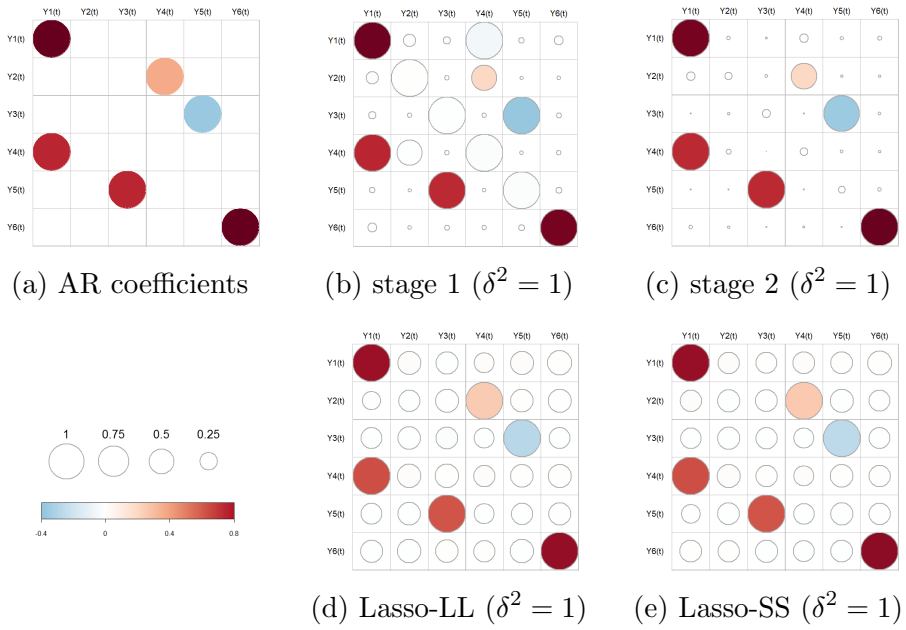
where $K = 6$, $p = 1$, $p \vee \hat{p} := \max\{p, \hat{p}\}$ and $A_k(i, j) := 0$ for any triplet (k, i, j) such that $k > 1$ and $1 \leq i, j \leq K$. The first two metrics show the model selection performance and the latter three metrics reflect the efficiency of parameter estimates of each method. 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 validations, as described in Appendix 2.6.2. We let δ^2 in Σ_Z take values from $\{1, 4, 25, 100\}$. The sample size T is 100 and results are based on 500 replications.

The five metrics for comparison are summarized in Table 2.1. 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 by 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, see Fan and Li (2001). 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 between the three methods when $\delta^2 = 1$ is displayed in Figure 2.1. Panels (b) and (c) of Figure 2.1 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. For comparison, panel (a) displays the true AR coefficient matrix A_1, where the color of a circle shows the true value of the corresponding 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 first stage 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), which demonstrates the effectiveness of the second stage refinement. In addition, the similarity between panels (a) and (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 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, since the true values of these AR coefficients are 0. However, 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 two 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 2.2 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, respec-



(a) AR coefficients (b) stage 1 ($\delta^2 = 1$) (c) stage 2 ($\delta^2 = 1$)

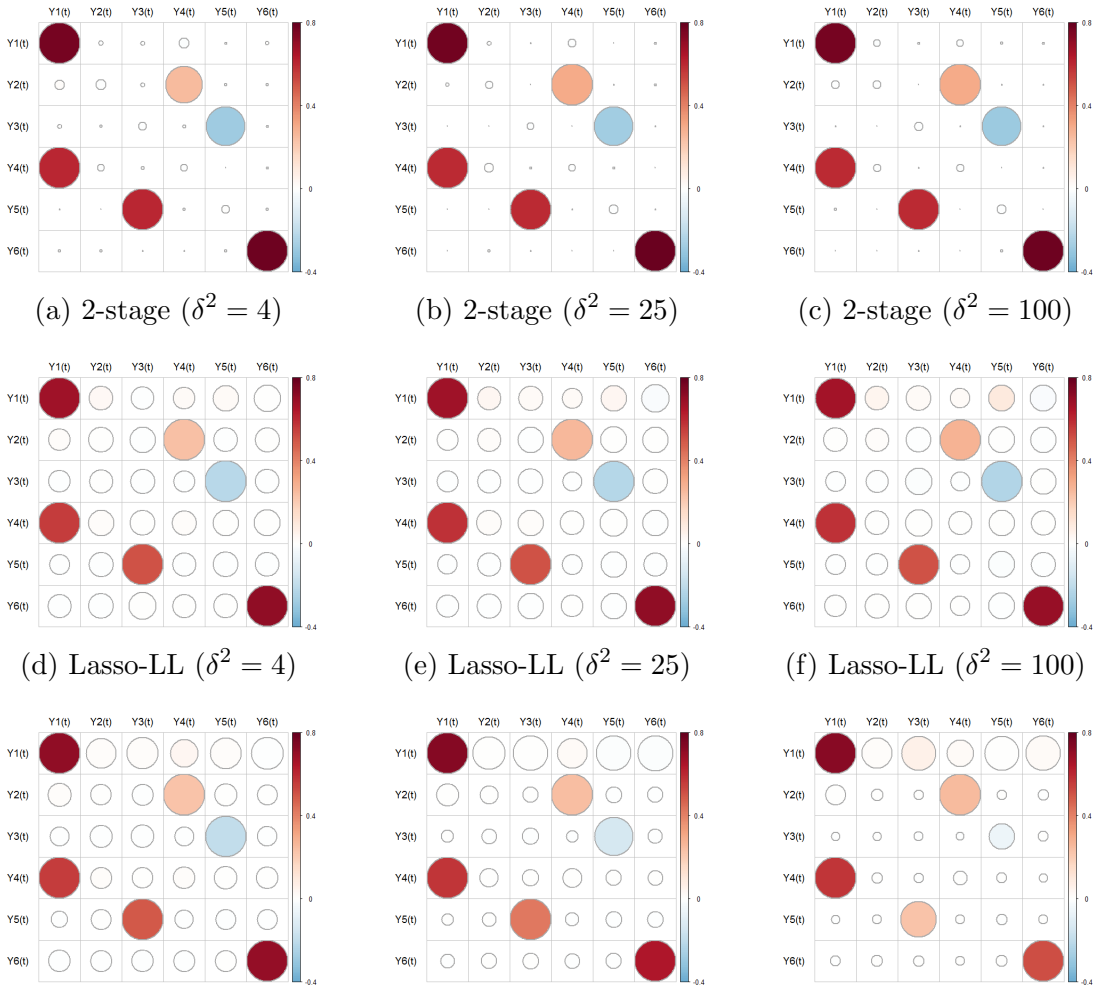
(d) Lasso-LL ($\delta^2 = 1$) (e) Lasso-SS ($\delta^2 = 1$)

Figure 2.1: 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 $\delta^2 = 1$. Panel (a) displays the true AR coefficient matrix A_1, where the color of each circle shows the true value of the corresponding AR coefficient. In panels (b), (c), (e) and (f), the size of each circle is proportional to the percent of times (out of 500 replications) the corresponding AR coefficient is selected; the color of each circle shows the average of the 500 estimates of that AR coefficient.

tively. 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 noise 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 size of the white circles in the first row increases while the size of the white circles in the other five rows decreases. This observation suggests that as the marginal variability of $\{Y_{t,1}\}$ increases, 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 between the Lasso-SS and the Lasso-LL methods is due to the fact that the Lasso-LL method takes into account the noise 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 noise covariance matrix Σ_Z and is superior to the Lasso-SS method.

Finally, we investigate the estimators of one particular AR coefficient from the three methods in more detail. Figure 2.3 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 2.3 shows that the estimators of $A_1(6,6)$ from the 2-stage approach and the Lasso-LL method are not much impacted by the changing variability of $\{Y_{t,1}\}$. But the Lasso-SS estimator for $A_1(6,6)$ becomes more



(a) 2-stage ($\delta^2 = 4$) (b) 2-stage ($\delta^2 = 25$) (c) 2-stage ($\delta^2 = 100$)

(d) Lasso-LL ($\delta^2 = 4$) (e) Lasso-LL ($\delta^2 = 25$) (f) Lasso-LL ($\delta^2 = 100$)

(g) Lasso-SS ($\delta^2 = 4$) (h) Lasso-SS ($\delta^2 = 25$) (i) Lasso-SS ($\delta^2 = 100$)

Figure 2.2: 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. The interpretation of the size and the color of a circle is the same as in Figure 2.1.

biased and volatile as the marginal variability increases from $\delta^2 = 1$ to $\delta^2 = 100$. Although both the 2-stage 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.

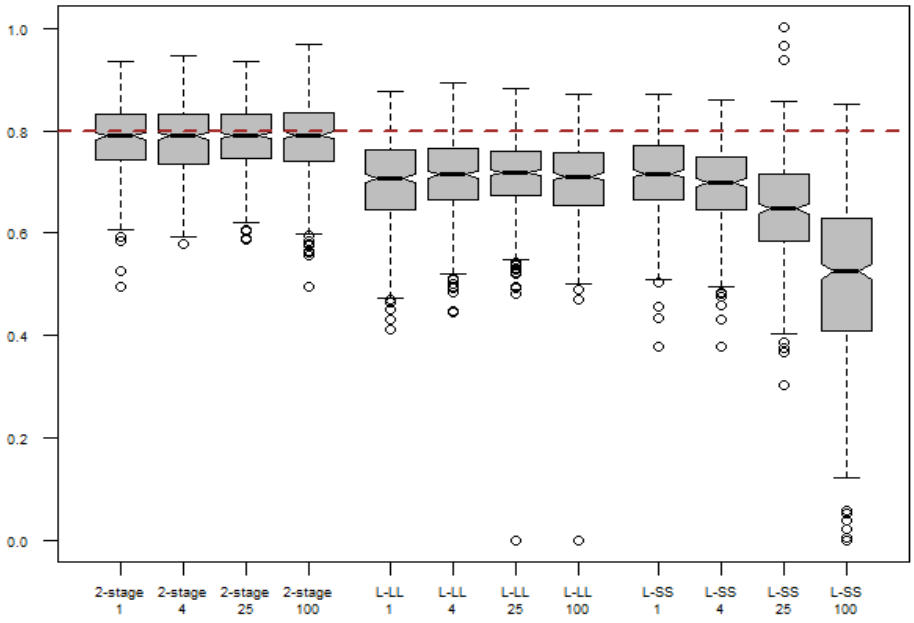


Figure 2.3: Sampling distributions of the estimators of $A_1(6,6)$ from the 2-stage approach (the left 4 boxplots), the Lasso-LL method (the middle 4 boxplots) and the Lasso-SS method (the right 4 boxplots) for $\delta^2 = 1, 4, 25$ and 100, respectively. The dashed horizontal line indicates the true value of $A_1(6,6) = 0.8$.

2.4.2 Real data examples

Google Flu Trends data. In the first example we consider the Google Flu Trends data, which can be viewed as a measure of the level of influenza activity in the U.S.. 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, see e.g. Polgreen et al. (2008); Eysenbach (2009); Hulth et al. (2009). 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) [3] related visits out of every 100,000 random outpatient visits within a U.S. 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 IP 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 national level but are also available for the 50 states, the District of Columbia and 122 cities throughout the U.S.. In contrast,

[3] According to the Centers for Disease Control and Surveillance, an influenza-like-illness is defined as a fever of 100 degrees Fahrenheit (or higher) along with a cough and/or sore throat in the absence of a known cause other than influenza.

the CDC surveillance report is available only at the national level and for 10 major U.S. regions (each region is a group of states). Due to these advantages, there has been increasing interest in modeling the Google Flu Trends data to help monitor the influenza activity in the U.S., see e.g. Dukić et al. (2010); Fox and Dunson (2011).

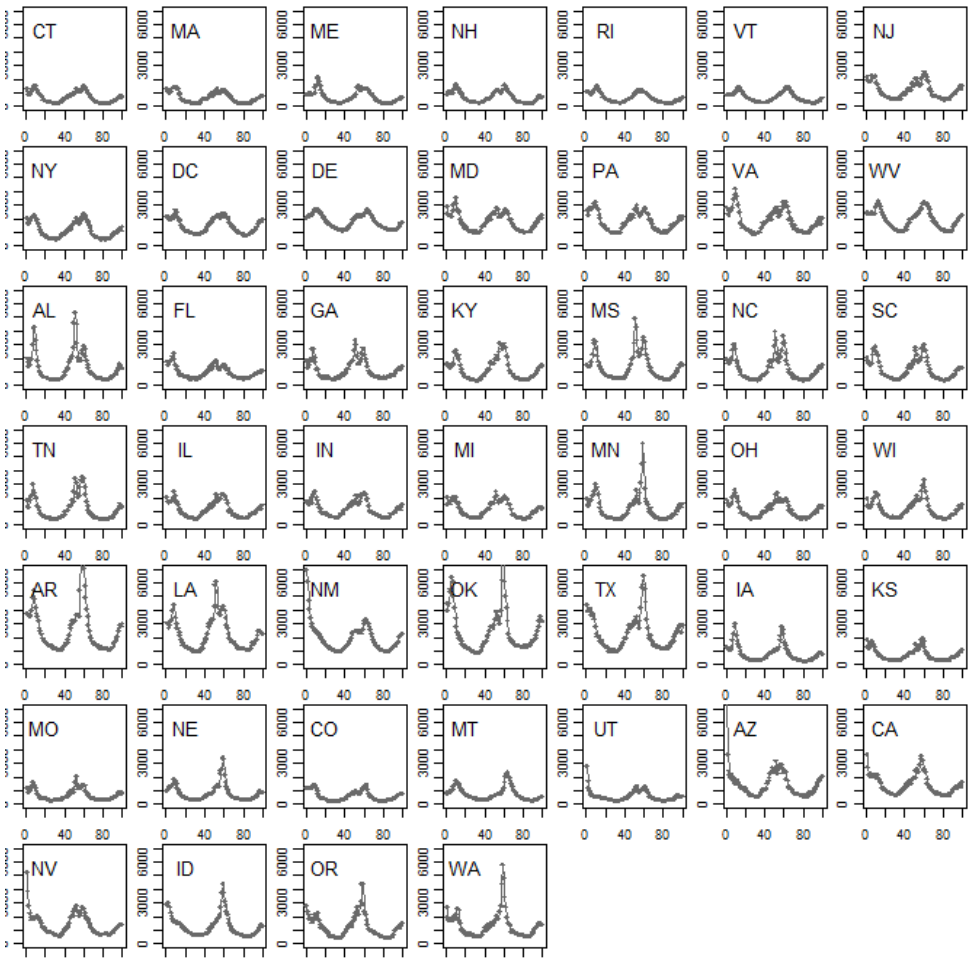


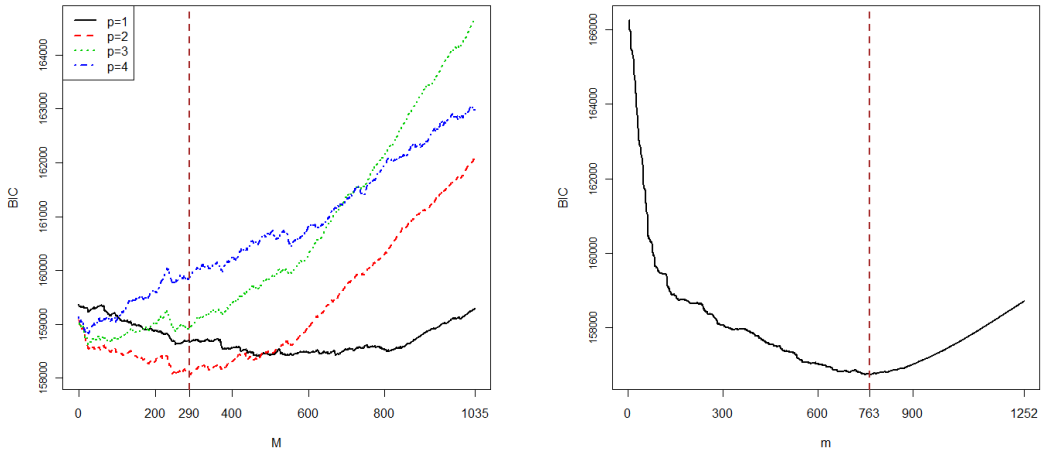
Figure 2.4: The first 100 observations of the Google Flu Trends series.

We apply the 2-stage approach to fit a sVAR model to the weekly Goolge Flu

Trends data from the week of January 1, 2006 to the week of December 26, 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 during the selected time period. So the dimension of the process in this example is $K = 46$ and we refer to these 46 regions as 46 states for simplicity. Figure 2.4 displays the first 100 observations from the 46 states. In applying the 2-stage approach, the pre-specified range of the autoregression order p is $\mathbb{P} = \{0, 1, 2, 3, 4\}$. The 2-stage approach leads to a 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. Figure 2.5 displays the BIC curves from stages 1 and 2 of the 2-stage approach, respectively. From panel (a) of stage 1, we can see that the first stage selects the autoregression order $\tilde{p} = 2$ and $\tilde{M} = 290$ pairs of distinct marginal series into the model. So the first stage model contains $(K + 2\tilde{M})\tilde{p} = (46 + 290 \times 2) \times 2 = 1252$ non-zero AR coefficients. The second stage follows by further selecting $\hat{m} = 763$ non-zero AR coefficients and leads to the final sVAR(2,763) 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 2.6 displays the estimated AR coefficients from the three models at lags 1 and 2, respectively. To illustrate the possible spatial interpretation for the temporal dependence structure, we group the 46 states into 10 regions as suggested in the CDC influenza surveillance report [4], which is indicated by the solid black lines in Figure 2.6. From panels (a), (c)

[4]The CDC 10-region division can be found at http://www.cdc.gov/flu/weekly/



(a) BIC curve of stage 1

(b) BIC curve of stage 2

Figure 2.5: BIC curves of stages 1 and 2 of the 2-stage approach. In panel (a), the x-axis M refers to the number of top pairs selected. Each curve corresponds to one autoregression order $p \in \{1, 2, 3, 4\}$ and shows the BIC values as M varies from 0 to $1035 = \binom{46}{2}$. The BIC value of $p = 0$ is not shown since it is much higher. In panel (b), the x-axis m refers to the number of non-zero AR coefficients retained and the curve shows the BIC values as m varies from 0 to 1252. In both panels, the dashed vertical line indicates where the minimum BIC value occurs.

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 in $\hat{A}_1$, 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

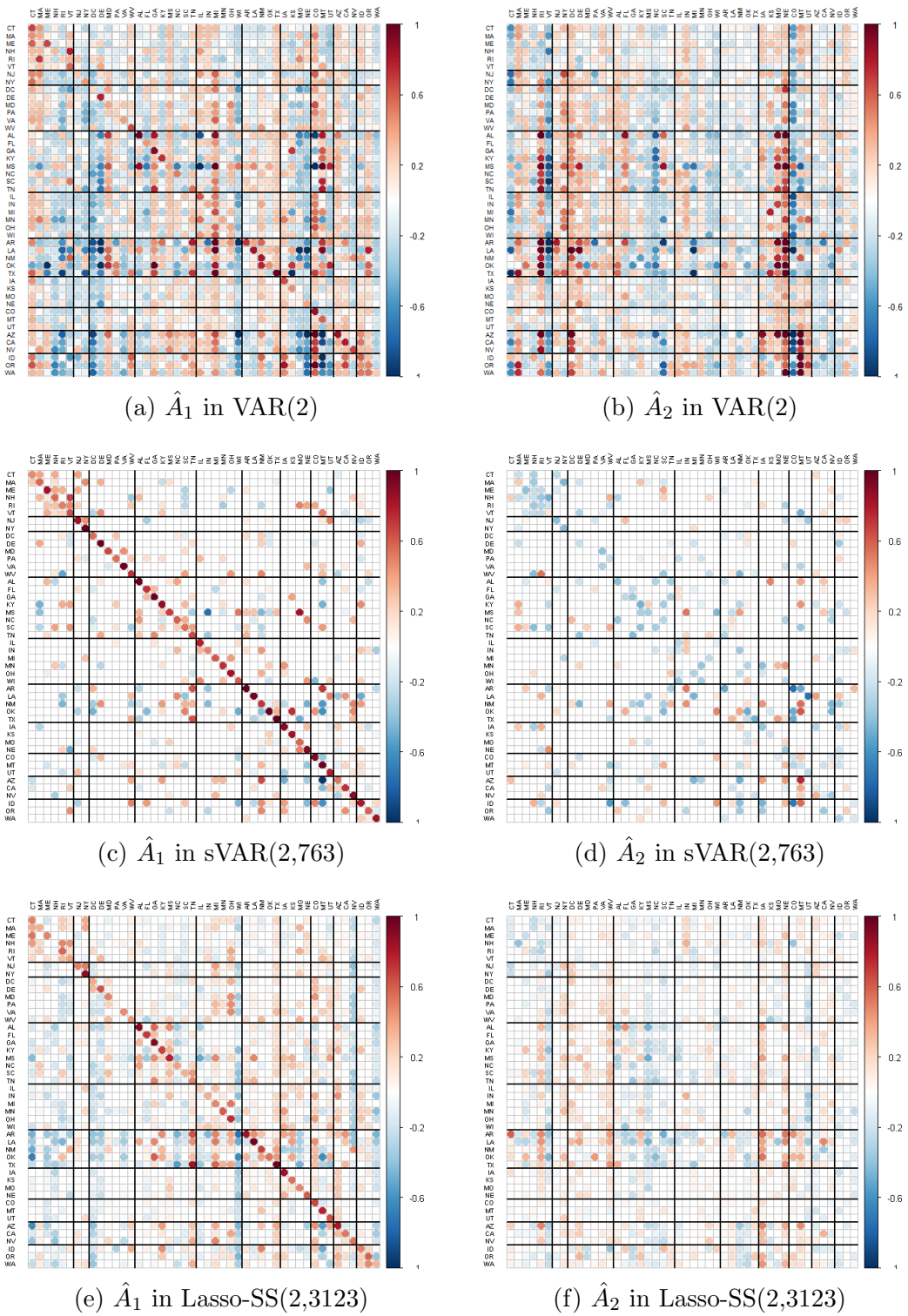
off-diagonal AR coefficients are zero. Additionally, the overall interpretability of the sVAR(2,763) and the Lasso-SS(2,3123) models is much better than the VAR(2) model, since both models provide much cleaner descriptions of the temporal dependence structures and reveal some interesting patterns. For example, both the sVAR(2,763) and the Lasso-SS(2,3123) models discover the interdependence among the influenza activities of the 6 states in Region 1, i.e., (CT, MA, ME, NH, RI, VT), as shown 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 the absolute value) but non-zero AR coefficients, especially those at lag 2 as shown in panel (f).

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

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

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

$$\text{LS} = (T_{\text{test}} - 1)^{-1} \sum_{t=T+1}^{T+T_{\text{test}}-1} - \log p_t(Y_t),$$



(a) $\hat{A}_1$ in VAR(2)

(b) $\hat{A}_2$ in VAR(2)

(c) $\hat{A}_1$ in sVAR(2,763)

(d) $\hat{A}_2$ in sVAR(2,763)

(e) $\hat{A}_1$ in Lasso-SS(2,3123)

(f) $\hat{A}_2$ in Lasso-SS(2,3123)

Figure 2.6: 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.

where $p_t(\cdot)$ is the probability density function of the 1-step forecast distribution. Table 2.2 summarizes the forecast RMSE for each forecast horizon $h = 1, 2, 3$ and 4 as well as the LS of each model. The sVAR(2,763) model fitted by the 2-stage approach has the smallest forecast RMSE among the three models, while the most saturated model, the VAR(2) model, has the worst out-of-sample forecast performance. The 2-stage approach gives the best forecast performance since it excludes many seemingly spurious AR coefficients from the sVAR(2,763) model. But the VAR(2) model contains a large number of spurious AR coefficients and their presence makes the out-of-sample forecast much less reliable. In addition, as seen from the last column of Table 2.2, the LS rule also favors the sVAR(2,763) model among the three.

Table 2.2: The h-step forecast root mean squared error (RMSE) and the logarithmic score (LS) of the sVAR(2,763), the Lasso-SS(2,3123) and the VAR(2) models. The test period is from the week of July 10, 2011 to the week of December 25, 2011 ($T_{\text{test}} = 24$). The forecast horizon is $h = 1, 2, 3$ and 4.

Model	$h = 1$	$h = 2$	$h = 3$	$h = 4$	LS
sVAR(2,763)	315.5	337.8	374.4	420.9	305.2
Lasso-SS(2,3123)	324.7	351.5	400.9	437.2	317.4
VAR(2)	336.4	393.2	468.7	562.3	462.7

Concentration levels of air pollutants. In this example we analyze a time series of concentration levels of four air pollutants, CO, NO, NO_2 and O_3, 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 Information System (AQMIS). The time series under consideration is of dimension $K = 5$ with $T = 8370$ observations. Figure 2.7 displays the hourly averages of the 5 marginal series. 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.

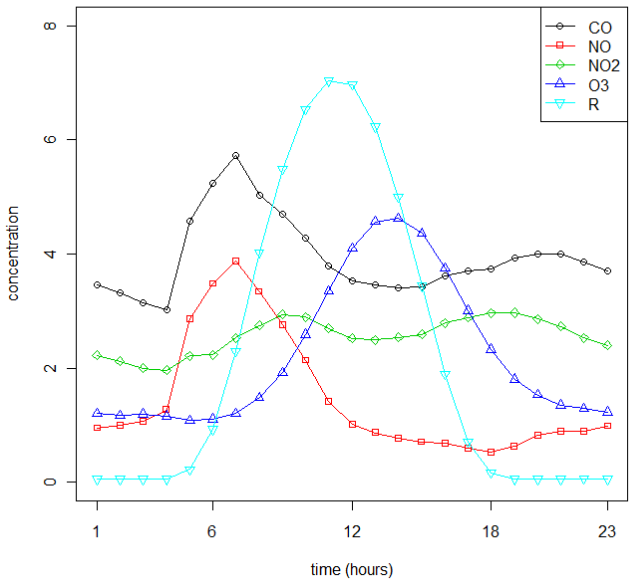


Figure 2.7: Hourly average series of the concentration levels of CO (100 ppb), NO (10 ppb), NO_2 (10 ppb), O_3 (10 ppb) and the solar radiation intensity R ($100W/m^2$) during 2006 at Azusa, California.

We apply the 2-stage approach to fit a sVAR model to the air pollutant data. The pre-specified range of the autoregression order p is $\mathbb{P} = \{0, 1, 2, \ldots, 8\}$. 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 first stage 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 in the end. The selection of the autoregression order $\hat{p} = 4$ coincides with the result in Songsiri et al. (2010), which also used BIC for VAR order selection. However, the BIC value of the 2-stage sVAR(4,64) model is 15301 and it is lower than the best BIC value (15414) reported in Table 1.1 of Songsiri et al. (2010). This is because 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 is one limitation of the partial correlation graph approach: such spurious non-zeros do not substantially increase the likelihood but inflate the BIC, and they also weaken the interpretability of fitted VAR models. 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, see e.g. 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.

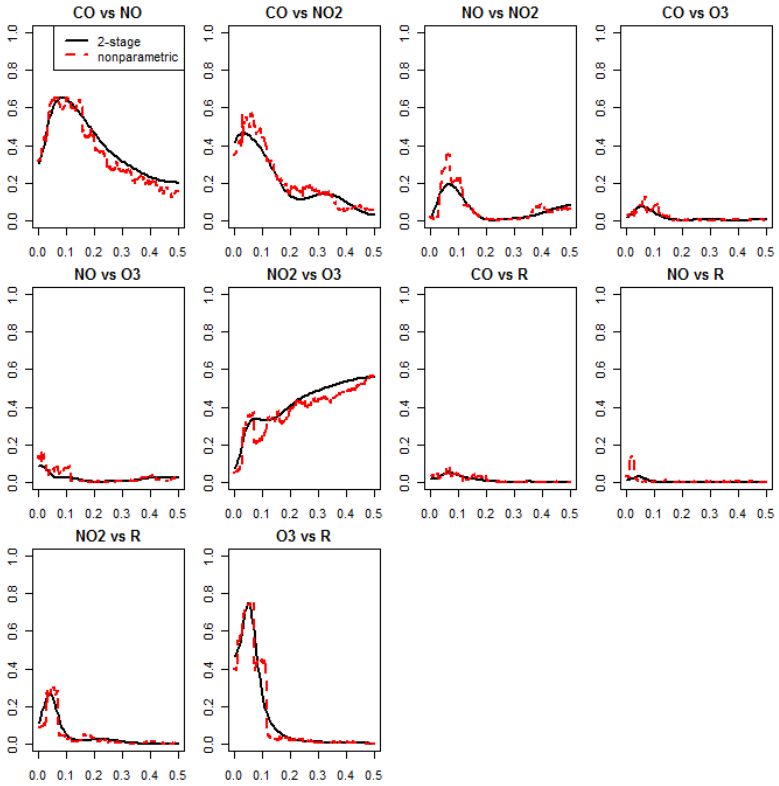


Figure 2.8: Plots of the parametric estimates of the squared modulus of PSC, i.e., $|\text{PSC}|^2$, as computed from the AR coefficient estimates in the 2-stage sVAR(4,64) model (solid curves) and the non-parametric estimates of $|\text{PSC}|^2$ used in the first stage selection (dashed curves).

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 2.8 displays the estimate of the squared modulus of PSC, i.e., $|\text{PSC}|^2$, as computed from the AR coefficient estimate in the 2-stage sVAR(4,64) model as well as the non-parametric estimate of $|\text{PSC}|^2$ 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 estimate 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 estimate of $|\text{PSC}|^2$ from the 2-stage sVAR(4,64) model, as displayed in Figure 2.8, resemble those in Figure 1.9 of Songsiri et al. (2010), which displays the estimate of $|\text{PSC}|^2$ from the fitted partial correlation graph model. Furthermore, the findings from Figure 2.8 agree with the photochemical theory of interactions between the 5 marginal series. For example, as pointed out in Dahlhaus (2000), the large estimate of $|\text{PSC}|^2$ between (CO, NO) comes from the fact that both air pollutants are mainly emitted from cars; and the large estimate of $|\text{PSC}|^2$ between (O_3, R) reflects the major role of the solar radiation intensity in the generation of ozone. Additionally, from Figure 2.8 we observe that the estimates of $|\text{PSC}|^2$ between the pairs (CO, O_3), (CO, R), (NO, R) and (NO, O_3) are relatively small as compared to the other pairs. This discovery of small estimate of $|\text{PSC}|^2$ agrees with the findings in Dahlhaus (2000); Eichler (2006); Songsiri et al. (2010), which are summarized in Table 2.3. For more detailed discussion on the underlying photochemical mechanism of interactions between air pollutants, readers are referred to Dahlhaus (2000).

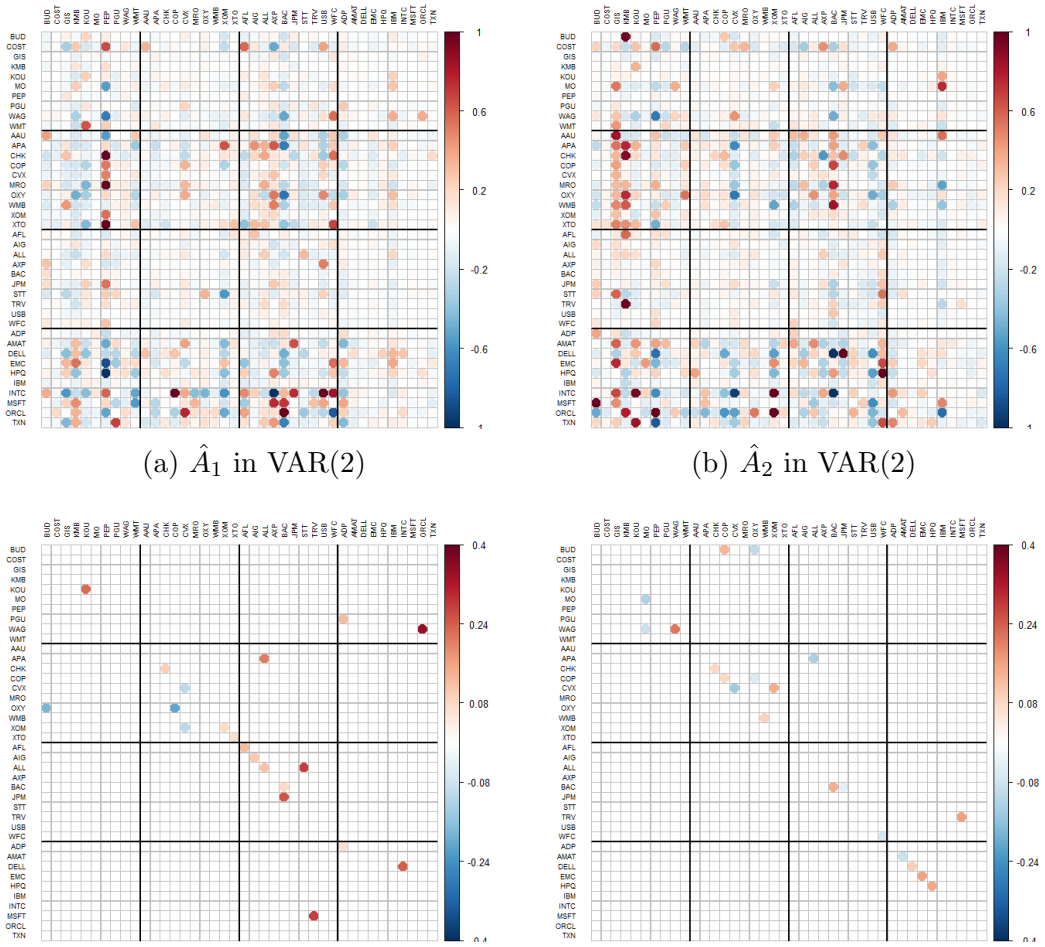
Table 2.3: Pairs with small estimate of $|\text{PSC}|^2$ in the 2-stage sVAR(4,64) model, as well as those found in Dahlhaus (2000), Eichler (2006) and Songsiri et al. (2010). Songsiri et al. (2010) used the same dataset as the sVAR(4,64) model while Dahlhaus (2000) and Eichler (2006) studied a similar 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)

Squared stock returns from S&P 500. In this example we analyze the squared daily returns of $K = 40$ stocks from S&P 500. The 40 stocks come from 4 sectors: *consumer staples*, *energy*, *finance* and *technology*, with 10 stocks from each. The returns are calculated as the logarithm of the ratio between two consecutive daily close prices from the 252 trading days in 2006. We analyze *squared* returns which can serve as proxies for the volatilities of these stocks and we want to investigate the temporal dependence structure between the volatilities.

We apply the 2-stage approach to fit a sVAR model to the squared return series. The pre-specified range of the autoregression order p is $\mathbb{P} = \{0, 1, 2, 3, 4, 5\}$ and the 2-stage approach leads to a sVAR(2,42) model, which has only as many as $2.63\% = 42/(40^2 \times 2)$ of the AR coefficients in a fully-parametrized VAR(2) model. For comparison, we also fit a VAR(2) model to the squared return series. The two estimated AR coefficient matrices from the VAR(2) model are displayed in panels (a) and (b) of Figure 2.9, while the two counterparts from the sVAR(2,42) model are displayed in panels (c) and (d). The solid lines in Figure 2.9 group the 40 stocks according to the 4 sectors.

From Figure 2.9 we have the following observations. First, we compare the temporal dependence structures discovered by the VAR(2) and the sVAR(2,42) models. Panels (a) and (b) in Figure 2.9 show that there exist more non-zero AR coefficients in the first and third block-columns than in the other two block-columns. It suggests that the VAR(2) model detects more occurrences of temporal influence among stock volatilities from the *consumer staples* and the *finance* sectors into the *energy* and the *technology* sectors than in the reverse direction. From panels (c) and (d), however, we can see that most of the AR coefficients in the first and third block-columns are set to zero in the sVAR(2,42) model. And we cannot see the above trend of "direc-



(a) $\hat{A}_1$ in VAR(2)

(b) $\hat{A}_2$ in VAR(2)

(c) $\hat{A}_1$ in sVAR(2,42)

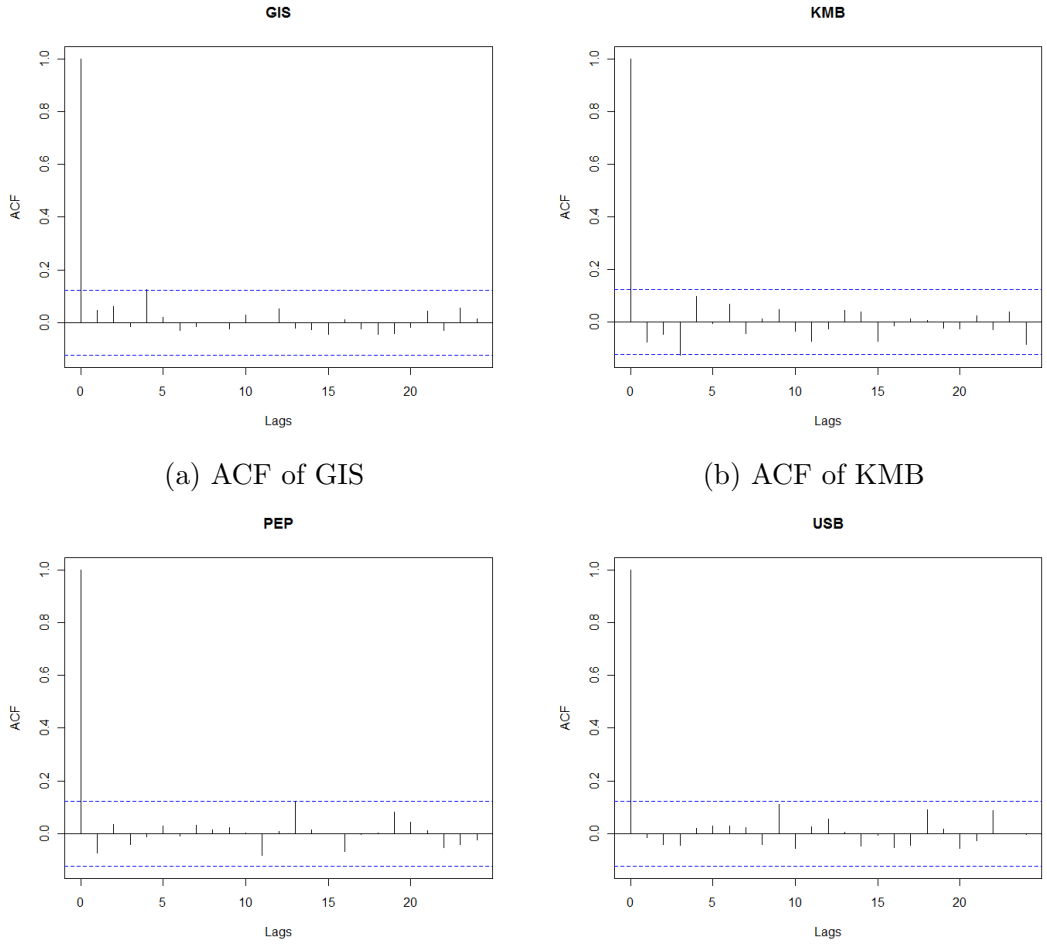
(d) $\hat{A}_2$ in sVAR(2,42)

Figure 2.9: Displays of the AR coefficient estimates from the VAR(2) and the sVAR(2,42) models at lags 1 and 2, respectively. The solid lines indicate the 4 sectors.

tional" temporal influence among stock volatilities between different sectors. This observation indicates that the fully-parametrized VAR(2) model may lead to falsely-detected temporal relationships due to its unstable estimates of AR parameters and the 2-stage sVAR model helps to correct those spurious relationships among the stock volatilities. Second, from panels (c) and (d), we can see that out of the 42 non-zero AR coefficients, only 9 of them occur between stocks from different sectors, which is less than what would happen due to pure randomness ($42 \times 12/16 = 31.5$). Therefore panels (c) and (d) suggest, on a high-level, that stock volatilities are more likely to be temporally related between stocks within the same sector than from different sectors. Third, we notice that the AR coefficients associated with certain stocks, such as GIS, KMB, PEP and USB, are all zero in the final sVAR(2,42) model, since those stocks correspond to both "empty" rows and columns in panels (c) and (d). It means that the squared return series of those stocks simply consist of white noise, which is verified in Figure 2.10. Each panel of Figure 2.10 displays the autocorrelation function (ACF) of the squared return series of the corresponding stock. We can observe little serial dependence in any of the 4 panels. In contrast, those stocks are associated with non-zero AR coefficient estimates in the VAR(2) model, as shown in panels (a) and (b) of Figure 2.9, which does not agree with the fact that their corresponding squared return series consist of white noise.

2.5 Discussion

In this chapter 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 co-



(a) ACF of GIS

(b) ACF of KMB

(c) ACF of PEP

(d) ACF of USB

Figure 2.10: ACF plots of the squared return series of GIS, KMB, PEP and USB.

efficients. The second stage follows using t-ratios together with BIC to further refine the first stage 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. Simulation results show that the 2-stage approach outperforms Lasso-VAR methods in recovering the sparse temporal dependence structure of sVAR models. Applications of the 2-stage approach to two real data examples yield interesting findings about their temporal dynamics.

In the first stage selection of the 2-stage approach, we use (2.10) 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

$$\begin{pmatrix} Y_{t,1} \\ Y_{t,2} \\ Y_{t,3} \end{pmatrix} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 0.3 \\ 0 & 0.25 & 0.5 \end{pmatrix} \begin{pmatrix} Y_{t-1,1} \\ Y_{t-1,2} \\ Y_{t-1,3} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \\ Z_{t,3} \end{pmatrix}, \tag{2.16}$$

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

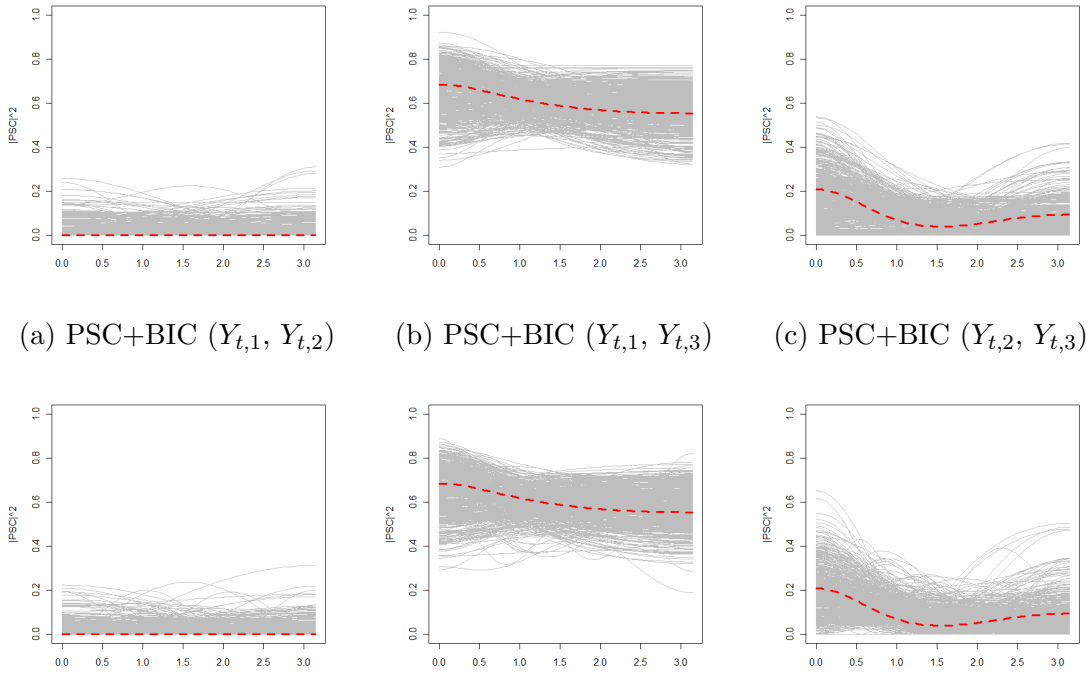
matrix

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

For this example, one can show that $\mathrm{PSC}_{1,2}(\omega) = 0$ for $\omega \in (-\pi, \pi]$ while $A_1(1,2) = 0.5$. In applying the 2-stage approach to fit sVAR models to (2.16), the first stage estimate of the summary statistic $\sup_{\omega} |\mathrm{PSC}_{1,2}(\omega)|^2$, as defined in (2.11), 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 (2.16). 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 in conjunction with BIC for further refinement in its second stage. So the truly non-zero AR coefficients will not be excluded after the first stage in 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. Figure 2.11 displays the comparison between the estimated inverse spectrum from these two approaches. In each panel of Figure 2.11, the dashed curve shows the true $|\mathrm{PSC}|^2$ between one pair of the 3 marginal series of $\{Y_t\}$ (2.16) and each solid curve displays the estimate $|\hat{\mathrm{PSC}}|^2$ from one replication. Panels (a), (b) and (c) correspond

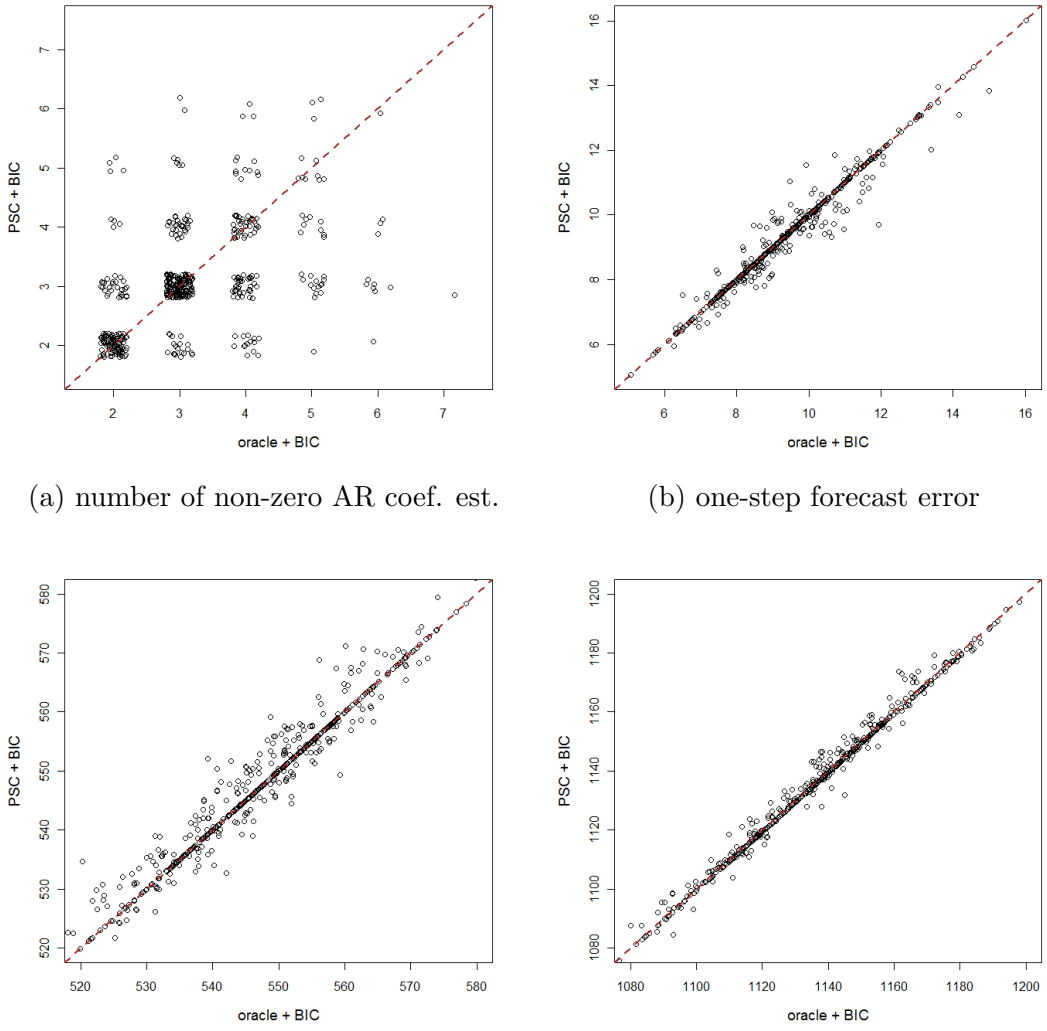
to the original 2-stage approach (labeled as "PSC+BIC") while panels (d), (e) and (f) correspond to the modified 2-stage procedure (labeled as "oracle+BIC"). From Figure 2.11, we can see that the original 2-stage approach leads to very similar estimate of the inverse spectrum as compared to the modified 2-stage procedure, in spite of the difference between their respective first stages. We also compare the two approaches using 4 other metrics, as shown in Figure 2.12. In each panel of Figure 2.12, the x-axis refers to the modified 2-stage procedure while the y-axis refers to the original 2-stage approach. 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 VAR processes as an interesting direction for future research.



(a) PSC+BIC ($Y_{t,1}$, $Y_{t,2}$) (b) PSC+BIC ($Y_{t,1}$, $Y_{t,3}$) (c) PSC+BIC ($Y_{t,2}$, $Y_{t,3}$)

(d) oracle+BIC ($Y_{t,1}$, $Y_{t,2}$) (e) oracle+BIC ($Y_{t,1}$, $Y_{t,3}$) (f) oracle+BIC ($Y_{t,2}$, $Y_{t,3}$)

Figure 2.11: Comparison of the estimate $|\hat{\text{PSC}}|^2$ from the 2-stage approach and the modified 2-stage procedure. The original 2-stage approach is labeled as "PSC+BIC" while the modified 2-stage procedure is labeled as "oracle+BIC". In each panel, the dashed curve shows the true $|\text{PSC}|^2$ and each solid curve corresponds to the estimate $|\hat{\text{PSC}}|^2$ from one replication.



(a) number of non-zero AR coef. est.

(b) one-step forecast error

(c) minus log-likelihood

(d) BIC

Figure 2.12: Comparison between the 2-stage approach and the modified 2-stage procedure using different metrics. Panel (a): number of non-zero AR coefficient estimates. Panel (b): out-of-sample one-step forecast error. Panel (c): minus log-likelihood. Panel (d): BIC. In each panel, the x-axis refers to the modified 2-stage procedure and is labeled as "oracle + BIC" while the y-axis refers to the original 2-stage approach and is labeled as "PSC + BIC".

2.6 Appendix to Chapter 2

2.6.1 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

$$\alpha := \text{vec}(A_1, \ldots, A_p) = R\gamma, \tag{2.17}$$

where $\alpha = \text{vec}(A_1, \ldots, A_p)$ is the $K^2p \times 1$ vector obtained by column stacking the AR coefficient matrices $A_1, \ldots, A_p$; 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 (2.17) 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,

$$\begin{aligned} \begin{pmatrix} Y_{t,1} \\ Y_{t,2} \end{pmatrix} &= \begin{pmatrix} A_1(1,1) & 0 \\ A_1(2,1) & A_1(2,2) \end{pmatrix} \times \begin{pmatrix} Y_{t-1,1} \\ Y_{t-1,2} \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 & 0 \\ A_2(2,1) & 0 \end{pmatrix} \times \begin{pmatrix} Y_{t-2,1} \\ Y_{t-2,2} \end{pmatrix} + \begin{pmatrix} Z_{t,1} \\ Z_{t,2} \end{pmatrix}, \end{aligned} \tag{2.18}$$

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 (2.18) 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

$$\begin{aligned} \alpha &= \text{vec}(A_1, A_2) = R\gamma \\ \Longrightarrow \begin{pmatrix} A_1(1,1) \\ A_1(2,1) \\ 0 \\ A_1(2,2) \\ 0 \\ A_2(2,1) \\ 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} A_1(1,1) \\ A_1(2,1) \\ A_1(2,2) \\ A_2(2,1) \end{pmatrix}. \end{aligned} \tag{2.19}$$

The constraint matrix R in (2.19) is of rank $m = 4$, which equals to the number of non-zero AR coefficients.

Lütkepohl (1993) gives results on the constrained maximum likelihood estimation of the AR coefficients. Under the parameter constraint in the form of (2.17), the maximum likelihood estimators of the AR coefficients α and the noise covariance matrix Σ_Z are the solutions to the following equations,

$$\hat{\alpha} = R\{R'(LL' \otimes \hat{\Sigma}_Z^{-1})R\}^{-1}R'(L \otimes \hat{\Sigma}_Z^{-1})y, \tag{2.20}$$

$$\hat{\Sigma}_Z = \frac{1}{T-p}\sum_{t=p+1}^{T}(Y_t - \hat{Y}_t)(Y_t - \hat{Y}_t)', \tag{2.21}$$

where $\otimes$ is the *Kronecker* product and

$$\begin{aligned} L_t &:= (Y_t, Y_{t-1}, \ldots, Y_{t-p+1})', \\ L &:= (L_0, L_1, \ldots, L_{T-1}), \\ y &:= \mathrm{vec}(Y) = \mathrm{vec}(Y_1, Y_2, \ldots, Y_T), \\ \hat{Y}_t &:= \sum_{k=1}^{p} \hat{A}_k Y_{t-k}. \end{aligned}$$

It is known that, see e.g. Lütkepohl (1993); Reinsel (1997), if there is no parameter constraint on the AR coefficients, i.e., $R = I_{K^2p}$ in (2.17), then the maximum likelihood estimator of the AR coefficients does not involve the noise covariance matrix Σ_Z. From equation (2.20), however, we can see that the presence of the parameter constraint (2.17) 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 (2.20) and (2.21), until convergence, to obtain the constrained maximum likelihood estimator of the AR coefficients.

2.6.2 Implementation of fitting Lasso-VAR models

We give details of the implementation of fitting the two Lasso-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,

$$y = \mathrm{vec}(Y) = (L' \otimes I_K)\alpha + \mathrm{vec}(Z), \tag{2.22}$$

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

$$\begin{aligned} Y &:= (Y_1, Y_2, \ldots, Y_T), \\ y &:= \text{vec}(Y), \\ L_t &:= (Y_t, Y_{t-1}, \ldots, Y_{t-p+1})^{'}, \\ L &:= (L_0, L_1, \ldots, L_{T-1}), \\ Z &:= (Z_1, Z_2, \ldots, Z_T). \end{aligned}$$

Since $Z_1, \ldots, Z_T$ are iid from the K-dimensional Gaussian $N(0, \Sigma_Z)$, from (2.22) the minus log likelihood of the VAR(p) model (2.22), ignoring an additive constant, is,

$$-2\log L(\alpha, \Sigma_Z) = T\log|\Sigma_Z| + [y - (L^{'} \otimes I_K)\alpha]^{'}(I_T \otimes \Sigma_Z^{-1})[y - (L^T \otimes I_K)\alpha]. \quad (2.23)$$

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,

$$Q_\lambda^{SS}(\alpha) := ||y - (L^{'} \otimes I_K)\alpha||_2^2 + \lambda||\alpha||_1; \quad (2.24)$$

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

$$\begin{aligned} Q_\lambda^{LL}(\alpha, \Sigma_Z) := [y - (L^{'} \otimes I_K)\alpha]^{'}(I_T \otimes \Sigma_Z^{-1})[y - (L^{'} \otimes I_K)\alpha] \quad (2.25) \\ + T\log|\Sigma_Z| + \lambda||\alpha||_1. \end{aligned}$$

In both equations (2.24) and (2.25) 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 $Q_\lambda^{SS}(\alpha)$ (2.24) or $Q_\lambda^{LL}(\alpha, \Sigma_Z)$ (2.25), 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 (2.24) and the Lasso-LL target function (2.25) with respect to the AR coefficient α,

$$\frac{\partial Q_\lambda^{SS}(\alpha)}{\partial \alpha} = 2[(LL^{'} \otimes I_K) - (L \otimes I_K)y] + \lambda \cdot \text{sgn}(\alpha), \tag{2.26}$$

$$\frac{\partial Q_\lambda^{LL}(\alpha)}{\partial \alpha} = 2[(LL^{'} \otimes \Sigma_Z^{-1}) - (L \otimes \Sigma_Z^{-1})y] + \lambda \cdot \text{sgn}(\alpha), \tag{2.27}$$

where $\text{sgn}(\cdot)$ is the *signum* function and $\text{sgn}(\alpha)$ is the $K^2p \times 1$ vector in which the kth entry is $\text{sgn}(\alpha_k)$, $k = 1, \ldots, K^2p$. We can see that noise covariance matrix Σ_Z is taken into account by the Lasso-LL derivative (2.27) but not by the Lasso-SS derivative (2.26). The two $K^2p \times 1$ vectors of first derivatives (2.26) and (2.27) 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 (2.24) and (2.25), 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, see e.g. Efron et al. (2004) or the coordinate descent algorithm, see e.g. Friedman et al. (2010). Here we use the coordinate descent algorithm implemented in the *R* package *glmnet* for fitting Lasso-SS VAR models. The estimation of Lasso-LL VAR models is more complicated since the target function (2.25) 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 (2.25) can

be re-cast in a least-squares fashion. In other words, for a $K \times K$ positive-definite matrix Σ_Z, let

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

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

$$\Sigma_Z^{-\frac{1}{2}} := U\text{diag}\{\frac{1}{\sqrt{\kappa_1}}, \ldots, \frac{1}{\sqrt{\kappa_K}}\}U^{'} \tag{2.28}$$

to be the inverse square root of Σ_Z. Notice that $\Sigma_Z^{-\frac{1}{2}}$ in (2.28) 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 (2.25) can be re-written as

$$\begin{aligned} & Q_\lambda^{LL}(\alpha, \Sigma_Z) \qquad (2.29) \\ &= T\log|\Sigma_Z| + [y - (L^{'} \otimes I_K)\alpha]^{'}(I_T \otimes \Sigma_Z^{-1})[y - (L^{'} \otimes I_K)\alpha] + \lambda||\alpha||_1 \\ &= T\log|\Sigma_Z| + [y - (L^{'} \otimes I_K)\alpha]^{'}(I_T \otimes \Sigma_Z^{-\frac{1}{2}})^{'}(I_T \otimes \Sigma_Z^{-\frac{1}{2}})[y - (L^{'} \otimes I_K)\alpha] + \lambda||\alpha||_1 \\ &= T\log|\Sigma_Z| + [(I_T \otimes \Sigma_Z^{-\frac{1}{2}})y - (L^{'} \otimes \Sigma_Z^{-\frac{1}{2}})\alpha]^{'}[(I_T \otimes \Sigma_Z^{-\frac{1}{2}})y - (L^{'} \otimes \Sigma_Z^{-\frac{1}{2}})\alpha] + \lambda||\alpha||_1 \\ &= T\log|\Sigma_Z| + ||(I_T \otimes \Sigma_Z^{-\frac{1}{2}})y - (L^{'} \otimes \Sigma_Z^{-\frac{1}{2}})\alpha||_2^2 + \lambda||\alpha||_1. \end{aligned}$$

The loss function

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

in (2.29) 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 (2.29) 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.

Step 1. Set an initial value $\Sigma_Z^{(0)}$ for the covariance matrix Σ_Z.

Step 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. $\Sigma_Z^{(k+1)} = \frac{1}{T-p}(Y - A^{(k+1)}L)(Y - A^{(k+1)}L)'$,
where $\alpha^{(k+1)} = \operatorname{vec}(A^{(k+1)})$.

Fitting Lasso-penalized VAR models, as other penalized regression methods, also involves choosing the tuning parameter $\lambda \in \mathbb{R}$. Furthermore, the number of explanatory variables, i.e., the number of lagged values appearing on the right hand side of equation (2.22), depends on the unknown order of autoregression p. Therefore values of both p and λ need to be determined in a data-driven manner. Here we use a ten-fold cross-validation to determine their values. Restricting the order of autoregression p to take values in a pre-specified range $\mathbb{P}$, the following steps are used to fit 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 (2.24) or the aforementioned iterative procedure to minimize the Lasso-LL target function (2.25). For either the Lasso-SS or Lasso-LL method,

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.

2. Obtain either the Lasso-SS or Lasso-LL VAR model by setting the autoregression order to $\hat{p}$, which gives the minimum average cross-validation error over $\mathbb{P}$, and the tuning parameter equal to $\lambda^{opt}(\hat{p})$.

Chapter 3

Reduced-rank Covariance Estimation in Vector Autoregressive Modeling

3.1 Introduction

In this chapter, we propose a strategy to estimate the covariance matrix Σ_Z of the vector noise Z_t in a large dimensional VAR model as given by (1.1). Covariance estimation is important for VAR modeling: an estimate of the noise covariance matrix Σ_Z is needed for exploring the dependence structure of the VAR process (Demiralp and Hoover 2003; Moneta 2004) while an estimate of the inverse of the noise covariance matrix Σ_Z^{-1} is required in constructing the confidence intervals for AR coefficient estimates or computing the mean squared error of VAR forecasting (Lütkepohl 1993). A natural estimator for Σ_Z in a VAR model is the sample covariance matrix of the residuals from fitting an autoregression (Lütkepohl 1993). To this end, the residuals

are viewed as independent samples, conditioned on the AR coefficient estimates, from an underlying distribution with covariance matrix Σ_Z. Therefore estimating the noise covariance matrix in a VAR model can be cast as a covariance estimation problem where independent observations are available.

Covariance estimation from independent observations is a fundamental problem in many areas, such as portfolio selection (Ledoit and Wolf 2004), functional genomics (Schäfer and Strimmer 2005), fMRI study (Daniels and Kass 2001) and graphical models (Lauritzen and Wermuth 1989). Estimating a $K \times K$ covariance matrix posits many challenges for large K since the number of parameters to be estimated, which is $K(K+1)/2$, grows quadratically in the dimension K. The sample covariance matrix of the observations serves as a natural estimator when the dimension K is much smaller than the sample size. But it is also well known that the sample covariance matrix can be severely ill-conditioned in small- to medium- samples. As a result, various methods are proposed to estimate large dimensional covariance matrices. In the literature, there exist three main approaches for covariance estimation under large dimensionality. The first is the *shrinkage* approach, where the covariance estimator is obtained by shrinking the sample covariance matrix towards a pre-specified covariance structure (Ledoit and Wolf 2004; Schäfer and Strimmer 2005); the second is the *regularization* approach, where the covariance estimator is derived based on regularization methods, such as banding (Bickel and Levina 2008), thresholding (El Karoui 2008) and penalized estimation (Huang et al. 2006); and the third is the *structure* approach, where structural constraints, such as factor structures (Tipping and Bishop 1999) or autoregressive structures (Daniels and Kass 2001), are imposed to reduce the effective dimension of the covariance estimator.

In this chapter, we propose a reduced-rank estimator for the noise covariance matrix in a large dimensional VAR model. In Section 3.2 we first derive the reduced-

rank estimator under the setting where independent observations are available. The reduced-rank estimator is based on a latent variable model for the vector observation and its effective dimension can be much lower than the dimension of the population covariance matrix. So the reduced-rank estimator can be viewed as a structure covariance estimator. The reduced-rank estimator is attractive since it is not only well-conditioned but also provides an interpretable description of the covariance structure. Our simulation study shows that the reduced-rank covariance estimator outperforms two competing shrinkage estimators for estimating large dimensional covariance matrices from independent observations. In Section 3.2.2, we proceed to the context of VAR modeling. We describe how to integrate the proposed reduced-rank estimator into the fitting of large dimensional VAR models, for which we consider two scenarios that require different model fitting procedures. The first scenario is that there is no constraint on the AR coefficients, for which the VAR model can be fitted using a 2-step method; while the second scenario is that there exist constraints on the AR coefficients, where the VAR model needs to be fitted by an iterative procedure. In Section 3.3.2, the reduced-rank covariance estimator is applied to the VAR modeling of two real data examples. The first example is concerned with stock returns from S&P 500 and the second example is a time series of temperatures in southeast China.

3.2 Reduced-rank covariance estimation

In this section, we first derive the reduced-rank covariance estimator based on independent observations. Then we proceed to VAR modeling and describe how to integrate the reduced-rank estimator into the fitting of large dimensional VAR models.

3.2.1 For independent observations

We assume that $Z_1, \ldots, Z_T$ are T independent replicates from a K-dimensional Gaussian with mean $\mathbf{0}$ and covariance matrix Σ_Z. [1] The problem of interest is to estimate Σ_Z, which can be large dimensional. To derive our covariance estimator, we further assume that each vector observation Z_t follows the latent variable model

$$Z_t = U\delta_t + \varepsilon_t, \text{ for } t = 1, \ldots, T, \tag{3.1}$$

where the latent variables δ_t $(t = 1, \ldots, T)$ are independent replicates from a d-dimensional $(0 \leq d \leq K-1)$ Gaussian with mean $\mathbf{0}$ and diagonal covariance matrix $\Lambda := \text{diag}\{\lambda_1, \ldots, \lambda_d\}$ (the diagonal entries are positive and in decreasing order); U is a $K \times d$ column-orthonormal matrix, i.e., $U'U = I_d$; and the errors ε_t $(t = 1, \ldots, T)$ are independent replicates from a K-dimensional Gaussian with mean $\mathbf{0}$ and *isotropic* covariance matrix $\text{cov}(\varepsilon_t) = \sigma^2 I_K$.

Under the latent variable model (3.1), the covariance matrix Σ_Z is seen to be

$$\Sigma_Z = U\Lambda U' + \sigma^2 I_K. \tag{3.2}$$

The first component $U\Lambda U'$ in the decomposition (3.2) has reduced-rank d $(d < K)$ and contains the core information about the dependence structure between the K marginals of Z_t. The second component $\sigma^2 I_K$ has sparse structure and accounts for unexplained variability in individual marginals. The decomposition (3.2) approximates the K-dimensional dependence structure encoded by Σ_Z with a rank-d matrix $U\Lambda U'$. Such an approximation is useful for separating important dependence patterns from large dimensional noisy observations.

[1] Here we make the assumption of Gaussian distribution. As in Chapter 2, when Z_t is non-Gaussian, the reduced-rank covariance estimation method can still be applied, where the Gaussian likelihood is interpreted as a quasi-likelihood.

Connection and distinction with factor models.

The motivation of the latent variable model (3.1) is that the K-dimensional vector Z_t can be related to a d-dimensional vector δ_t of latent (unobserved) variables through a column-orthonormal matrix U. With $d < K$, the latent variable δ_t provides a more parsimonious description of the dependence structure of Z_t. This motivation is similar to that of factor models, see e.g. Anderson (2003). In the factor model setup, the relation (3.1) is also used to link the observation with the latent variable and the matrix U is called the *factor loading*; but it is usually assumed that the latent variable δ_t has an isotropic covariance matrix while the error ε_t has a non-isotropic covariance matrix. It is known that factor models have non-identifiability issues. Specifically, for any $d \times d$ orthogonal matrix C, the pairs (U, δ_t) and $(UC', C\delta_t)$ will lead to two factor models that are observationally equivalent. In contrast, our latent variable model (3.1) avoids such non-identifiability issues. This is because in the latent variable model we make different assumptions on the covariance structures of the latent variable δ_t and the error ε_t, as summarized in Table 3.1. In the latent variable model, the covariance matrix of the latent vector post an orthogonal rotation C is $\mathrm{cov}(C\delta_t) = C\mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}C'$, which in general is not equal to the original covariance matrix $\mathrm{cov}(\delta_t) = \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$. So the two latent variable models corresponding to the pairs (U, δ_t) and $(UC', C\delta_t)$ are not observationally equivalent. In other words, the assumption of the non-isotropic covariance matrix for the latent vector δ_t leads to the identifiability of the latent variable model (3.1). Due to the identifiability, interpretation of the matrix parameter U becomes meaningful.

Table 3.1: Comparison of assumptions between the latent variable model and the factor model.

Model	$\mathrm{cov}(\delta_t)$	$\mathrm{cov}(\varepsilon_t)$
latent variable model (3.1)	$\mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$	$\sigma^2 I_K$
factor model	$\sigma^2 I_d$	$\mathrm{diag}\{\lambda_1, \ldots, \lambda_K\}$

Maximum likelihood estimation.

We derive the maximum likelihood estimator of the reduced-rank covariance matrix $\Sigma_Z = U\Lambda U' + \sigma^2 I_K$ (3.2). Based on observations $Z_1, \ldots, Z_T$, $-\frac{2}{T}$log-likelihood, ignoring an additive constant, is given by

$$-\frac{2}{T}\log L(U, \Lambda, \sigma^2) = \log|\Sigma_Z| + \mathrm{tr}(\Sigma_Z^{-1} S), \tag{3.3}$$

where $S := \frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'$ is the sample covariance matrix. The following proposition shows that there exists an analytical form for the maximum likelihood estimator of the reduced-rank covariance matrix Σ_Z.

Proposition 1 *Let $c_1 \geq c_2 \ldots \geq c_K \geq 0$ be the eigenvalues of the sample covariance matrix S and assume that the reduced-rank d is known. The maximum likelihood estimator of the reduced-rank covariance matrix Σ_Z is given by*

$$\hat{\Sigma}_Z = \hat{U}\hat{\Lambda}\hat{U}' + \hat{\sigma}^2 I_K, \tag{3.4}$$

where

$$\hat{U} = (\hat{U}_1, \ldots, \hat{U}_d), \text{ and } \hat{U}_i \text{ is the eigenvector of } S \text{ corresponding to } c_i; \tag{3.5}$$

$$\hat{\sigma}^2 = \frac{1}{K-d}\sum_{i=d+1}^{K} c_i; \tag{3.6}$$

$$\hat{\Lambda} = \mathrm{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_d\}, \text{ where } \hat{\lambda}_i = c_i - \hat{\sigma}^2, \ i = 1, \ldots, d. \tag{3.7}$$

We defer the proof to Appendix 3.4.1.

Properties of the reduced-rank covariance estimator.

From (3.5) we can see that there exist links between the latent variable model (3.1) and *principal component analysis* (PCA), which is perhaps the most widely used statistical tool for dimension reduction. The common setup of PCA is based on a series of mutually-orthogonal projections of vector observations that maximize the retained variance, where the directions of these projections are called *principal axes*, see e.g. Jolliffe (2002). This setup is *not* based on a probabilistic model but comes from a projection perspective. In contrast, the latent variable model (3.1) provides a model-based formulation of PCA, in which the principal axes coincide with the columns of the maximum likelihood estimator $\hat{U}$ as given by (3.5). In the literature, such a probabilistic formulation of PCA is first investigated by Lawley (1953) within the context of factor analysis and is then reiterated by Tipping and Bishop (1999) under the name *probabilistic principal component analysis* (PPCA). A discussion on the advantages of this probabilistic formulation of PCA over the traditional projection-based setup is given in Tipping and Bishop (1999).

We also investigate the conditioning property of the reduced-rank estimator $\hat{\Sigma}_Z$ (3.4). It can be shown that the eigenvalues, denoted by b_i $(i = 1, \ldots, K)$, of the reduced-rank estimator $\hat{\Sigma}_Z$ are

$$b_i = \begin{cases} \hat{\lambda}_i + \hat{\sigma}^2 & = c_i, \qquad\qquad\qquad \text{for } i = 1, \ldots, d; \\ \hat{\sigma}^2 & = \frac{1}{K-d} \sum_{i=d+1}^{K} c_i, \quad \text{for } i = d+1, \ldots, K, \end{cases}$$

which means that the reduced-rank estimator $\hat{\Sigma}_Z$ retains the d largest eigenvalues but shrinks towards their average the remaining $(K-d)$ eigenvalues of the sample covariance matrix S. Therefore, the *condition number*, i.e., the ratio between the largest

and smallest eigenvalues of the covariance estimator, of the reduced-rank estimator can be much smaller than that of the sample covariance matrix. In other words, the reduced-rank estimator can be much more well-conditioned than the sample covariance matrix. In addition, as long as the reduced-rank d is smaller than the sample size T, the reduced-rank estimator will be invertible even if the dimension K exceeds the sample size T.

Next we discuss how to control the complexity of a reduced-rank covariance estimator through the choice of its reduced-rank d. From (3.4) we can see that there exist two extremes for $\hat{\Sigma}_Z$ as the reduced-rank d varies: when $d = K - 1$, i.e., there is no dimension reduction, $\hat{\Sigma}_Z = S$ becomes the full covariance model; and when $d = 0$, i.e., there is no structured component $\hat{U}\hat{\Lambda}\hat{U}'$, $\hat{\Sigma}_Z = \bar{c}I_K$ becomes the isotropic covariance model. In other words, the reduced-rank covariance estimator is obtained by balancing between the unbiased but highly variable sample covariance matrix and the biased but well-conditioned isotropic covariance matrix, where the balance is controlled by the reduced-rank d. In practice, the reduced-rank d is unknown and needs to be estimated from data. Here we use the *Bayesian information criterion* (BIC) (Schwarz 1978) to determine the reduced-rank d. The BIC is computed as

$$\text{BIC}(d) \quad = \quad -2\log L(\hat{U}, \hat{\Lambda}, \hat{\sigma}^2) + \log(T) \times (Kd - \frac{d(d-1)}{2} + 1), \tag{3.8}$$

where $L(\hat{U}, \hat{\Lambda}, \hat{\sigma}^2)$ is the maximized likelihood and $Kd - d(d-1)/2 + 1$ is the number of free parameters in the reduced-rank covariance estimator. We select the reduced-rank d from $\{0, 1, \ldots, K-1\}$ according to minimum BIC. Tipping and Bishop (1999) give similar results on controlling the complexity of PPCA.

Finally we describe a diagnostic tool for the reduced-rank covariance model. The latent variable δ_t in (3.1) can be estimated as

$$\hat{\delta}_t \quad = \quad \hat{U}' Z_t, \text{ for } t = 1, \ldots, T, \tag{3.9}$$

where $\hat{U}$ is given by (3.5). According to model assumptions, $\hat{\delta}_1, \ldots, \hat{\delta}_T$ should behave like independent replicates from a d-dimension Gaussian with *diagonal* covariance matrix. So the correlation function of the estimated latent variable $\hat{\delta}_t$ (3.9) can be used for model diagnostics.

3.2.2 For VAR series

In this section, we proceed from the setting of independent observations to VAR processes and apply the reduced-rank covariance estimator to the noise covariance matrix Σ_Z in a VAR model (1.1).

As described in Section 3.1, the reduced-rank estimator for Σ_Z in a VAR model is computed based on the residuals from fitting the autoregression. Therefore, in order to apply the reduced-rank covariance estimator, we need to estimate the AR coefficient matrices $A_1, \ldots, A_p$ in (1.1) as well, for which we consider two scenarios. The first scenario is that there is no constraint on the AR coefficient matrices $A_1, \ldots, A_p$; while the second scenario is that there exist constraints on the AR coefficients. The second scenario occurs, for example, when some of the AR coefficients are constrained be to zero. Such zero constraints on AR coefficients arise when we model the *Granger causality* of $\{Y_t\}$, see e.g. Granger (1969); Lutkepohl (1993), or when we fit *sparse vector autoregressive* models to $\{Y_t\}$, see e.g. Davis et al. (2012) and Chapter 2. Here we use zero constraints on AR coefficients as the example of the second scenario. Zero constraints on the AR coefficient matrices $A_1, \ldots, A_p$ can be expressed as

$$\alpha := \text{vec}(A_1, \ldots, A_p) = R\gamma, \tag{3.10}$$

where $\alpha := \text{vec}(A_1, \ldots, A_p)$ is the K^2p-dimensional vector obtained by stacking the columns of the AR coefficient matrices $A_1, \ldots, A_p$; R is a $K^2p \times m$ matrix of known constants with rank m; and γ is a m-dimensional vector of unknown parameters. The

matrix R is referred to as the *constraint matrix* (Davis et al. 2012) and it specifies which AR coefficients are 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 is equal to the number of non-zero AR coefficients. Using results on the constrained VAR estimation in Lütkepohl (1993) and on the reduced-rank covariance estimation in Section 3.2.1, we can show that, under the constraint (3.10) and the reduced-rank covariance model (3.2), the maximum likelihood estimator of the AR coefficients α is given by

$$\hat{\alpha} = R[R'(LL' \otimes \hat{\Sigma}_Z^{-1})R]^{-1}R'(L \otimes \hat{\Sigma}_Z^{-1})y, \tag{3.11}$$

where

$$\begin{aligned} L_t &:= (Y_t, Y_{t-1}, \ldots, Y_{t-p+1})', \\ L &:= (L_0, L_1, \ldots, L_{T-1}), \\ y &:= \mathrm{vec}(Y) = \mathrm{vec}(Y_1, Y_2, \ldots, Y_T). \end{aligned}$$

And $\hat{\Sigma}_Z$ in (3.11) is the reduced-rank maximum likelihood estimator for the noise covariance matrix Σ_Z based on the residuals $\hat{Z}_t := Y_t - \sum_{k=1}^{p} \hat{A}_k Y_{t-k}$ $(t = p+1, \ldots, T)$ from the fitted autoregression.

The model fitting procedure for the first scenario.

When there is no constraint on the AR coefficients (scenario 1), we have $R = I_{K^2p}$ in (3.10) and (3.11) becomes

$$\begin{aligned} \hat{\alpha} &= I_{K^2p}[I'_{K^2p}(LL' \otimes \hat{\Sigma}_Z^{-1})I_{K^2p}]^{-1}I'_{K^2p}(L \otimes \hat{\Sigma}_Z^{-1})y \\ &= [(LL')^{-1} \otimes \hat{\Sigma}_Z](L \otimes \hat{\Sigma}_Z^{-1})y \\ &= [(LL')^{-1}L \otimes I_K]y. \end{aligned} \tag{3.12}$$

So for the first scenario, (3.12) shows that the estimation of the AR coefficients α does not involve the reduced-rank estimation of the noise covariance matrix Σ_Z. Therefore the reduced-rank covariance estimator can be applied to a VAR model using the following 2-step method.

Step 1. Fit an unconstrained VAR model to $\{Y_t\}$ and obtain the AR coefficient estimates $\hat{\alpha}$ according to (3.12).

Step 2. Compute the reduced-rank covariance estimator $\hat{\Sigma}_Z$ using the results in Proposition 1 based on the residuals from the autoregression conditioned on the AR coefficient estimates $\hat{\alpha}$.

The model fitting procedure for the second scenario.

Where there exist zero constraints on the AR coefficients (scenario 2), (3.11) shows that the estimation of the AR coefficients α is commingled with the reduced-rank estimation of the noise covariance matrix Σ_Z. Therefore the reduced-rank covariance estimator is applied to a VAR model using the following iterative procedure.

- Start with initial estimators $\hat{\alpha}^{(0)}$ and $\hat{\Sigma}_Z^{(0)}$.
- Assume that at the rth iteration, the current estimators are $\hat{\alpha}^{(r)}$ and $\hat{\Sigma}_Z^{(r)}$, respectively. Repeat the following steps 1 and 2 until convergence.

  Step 1. Compute $\hat{\alpha}^{(r+1)}$ according to (3.11) by replacing $\hat{\Sigma}_Z$ with the current reduced-rank covariance estimator $\hat{\Sigma}_Z^{(r)}$.

  Step 2. Compute $\hat{\Sigma}_Z^{(r+1)}$ by applying the results of Proposition 1 based on the residuals from the autoregression conditioned on the current constrained AR coefficient estimates $\hat{\alpha}^{(r+1)}$.

A latent space interpretation.

We conclude this section by introducing a *latent space* setup that helps to interpret results from a reduced-rank covariance estimator when it is applied in VAR modeling. In particular, this latent space setup is useful in exploring the *contemporaneous* dependence structure of the VAR process $\{Y_t\}$, which describes how synchronous values of different marginal series of $\{Y_t\}$ impact each other, see e.g Reale and Wilson (2001); Demiralp and Hoover (2003); Moneta (2004). For $i = 1, \ldots, K$, let $u_i := (U_{i,1}, \ldots, U_{i,d})'$ be the ith row of the $K \times d$ matrix U in (3.2). Then for two different marginal series of $\{Y_t\}$, say $\{Y_{t,i}\}$ and $\{Y_{t,j}\}$ $(i \neq j)$, we have

$$\operatorname{cov}(Y_{t,i}, Y_{t,j} |\, Y_s, s \leq t-1) = u_i' \Lambda u_j. \tag{3.13}$$

The relation (3.13) shows that the conditional contemporaneous covariance between two different marginal series of $\{Y_t\}$ is represented by a weighted inner-product of the corresponding rows of U. To help interpret (3.13), we postulate the existence of a d-dimensional Euclidean space of unobserved (latent) characteristics. The latent characteristics determine the contemporaneous dependence between the marginal series of $\{Y_t\}$. We further assume that each marginal series of $\{Y_t\}$ is associated with a position in this latent space and the pattern of contemporaneous dependence among the K marginal series of $\{Y_t\}$ can be characterized by their latent positions. Such a setup is also used in latent space network models, see e.g. Hoff et al. (2002); Hoff (2005). From (3.13) we can see that, when the above latent space setup is adopted to the reduced-rank covariance model (3.2), the d dimensions of the latent space are represented by the columns of U while the K latent positions are given by the rows of U. Therefore the matrix U provides a tool to represent the K-dimensional contemporaneous dependence structure in a lower-dimensional space. In addition, if we are able

to find interpretations for different columns of U by taking advantage of exogenous information, such interpretations will help to identify the unobserved characteristics that are important in forming the contemporaneous dependence relationship. The heuristics behind such a latent space setup is similar to that of multidimensional scaling (MDS), see e.g. Borg and Groenen (1997), in that both methods are concerned with "spatial" representations of observed patterns of dependence among a group of subjects, such as the K marginal series of $\{Y_t\}$ in our case. However, the MDS method is not model-based and it constructs spatial representations in an ad-hoc manner; in contrast, the above latent space setup leads to model-based graphical representations of the contemporaneous dependence structure via inference of the reduced-rank covariance model. In Section 3.3.2, we illustrate via real data examples the use of this latent space setup in interpreting results from the reduced-rank covariance estimator in a VAR model.

3.3 Numerical results

3.3.1 Simulation

As pointed out in Section 3.1, there exist three major classes of covariance estimators under large dimensionality: the shrinkage, the regularization and the structure covariance estimators. The reduced-rank (RR) estimator can be viewed as a structure covariance estimator, as discussed in Section 3.2.1. One difference between the three classes of covariance estimators is that, under finite samples, invertibility holds for the shrinkage and the structure estimators, but not guaranteed for the regularization estimator. Due to this difference, in the simulation study we compare the reduced-rank covariance estimator with shrinkage estimators for their performance of estimating

large dimensional covariance matrices from independent observations. The earliest attempt of shrinkage covariance estimation is given in Stein (1975) and since then many shrinkage estimators have been proposed, see e.g. Dey and Srinivasan (1985); Daniels and Kass (2001); Ledoit and Wolf (2003; 2004); Schäfer and Strimmer (2005). A shrinkage covariance estimator is obtained by shrinking the sample covariance matrix towards a target covariance structure. The balance between these two extremes is controlled by the *shrinkage intensity*, a tuning parameter that needs to be estimated from data. A review of commonly-used target covariance structures is given in Schäfer and Strimmer (2005).

We consider two shrinkage covariance estimators: one is proposed in Ledoit and Wolf (2004) (LW2004) and the other one is given by Schäfer and Strimmer (2005) (SS2005). The two shrinkage estimators differ in their choices of the target covariance structure. We generate independent replicates from a K-dimensional Gaussian $N(0, \Sigma_Z)$ under three cases: (I) $\Sigma_Z = I_K$; (II) Σ_Z has all variances set to 1 and all covariances set to 0.1; (III) Σ_Z has variances set to $\{1, \ldots, 1, 0.8, \ldots, 0.8\}$ (the first five entries are 1 and the others are 0.8) and all covariances set to 0.1. Cases (I) and (II) are also used for the simulation study in Daniels and Kass (2001). We let the dimension $K = 20$ and the sample size $T = 20, 40, 100, 200, 400$, respectively. In applying the RR covariance estimator, the reduced-rank d is selected from $\{0, 1, \ldots, 19\}$ according to minimum BIC, which is computed as (3.8); while in applying the two shrinkage estimators LW2004 and SS2005, their shrinkage intensities are determined analytically as described in Ledoit and Wolf (2004) and Schäfer and Strimmer (2005), respectively. We use two metrics to compare the performance of different covariance estimators: the first metric is the *Stein's loss* (SL) (James and Stein 1961), which is defined by $\mathrm{SL}(\hat{\Sigma}_Z) := \mathrm{tr}(\hat{\Sigma}_Z \Sigma_Z^{-1}) - \log|\hat{\Sigma}_Z \Sigma_Z^{-1}| - K$. It can be shown that the Stein's loss $\mathrm{SL}(\hat{\Sigma}_Z)$ is equal to (up to a 1/2 multiplier) the *Kullback-Leibler diver-*

gence (Kullback and Leibler 1951) between two K-dimensional Gaussians $N(0, \hat{\Sigma}_Z)$ and $N(0, \Sigma_Z)$; and the second metric is the mean squared error (MSE), which is defined by $\text{MSE}(\hat{\Sigma}_Z) := \mathbb{E}||\hat{\Sigma}_Z - \Sigma_Z||_2^2$.

Table 3.2: Percentage reduction in SL and MSE of the RR, the LW2004 and the SS2005 covariance estimators as compared to the sample covariance matrix. For each setting, the largest reduction among the three is marked in bold. The results are based on 1000 replications.

		percentage reduction in SL			percentage reduction in MSE		
Σ_Z	T	RR	LW2004	SS2005	RR	LW2004	SS2005
I	20	**99.9**	99.6	99.6	**99.3**	98.0	98.2
	40	**99.5**	98.8	98.7	**99.4**	98.6	98.4
	100	**99.5**	99.0	98.6	**99.5**	98.9	98.5
	200	**99.6**	99.0	98.6	**99.5**	99.0	98.6
	400	**99.5**	99.1	98.7	**99.5**	99.1	98.6
II	20	**98.8**	98.5	98.6	78.2	83.3	**83.6**
	40	**91.8**	90.0	90.3	62.6	72.3	**72.8**
	100	**87.4**	75.4	75.8	**56.6**	50.5	52.6
	200	**90.2**	57.8	58.9	**71.7**	33.2	37.3
	400	**90.2**	38.4	41.3	**71.8**	19.2	25.6
III	20	**98.3**	98.0	98.1	68.8	77.8	**78.2**
	40	**89.0**	86.5	86.5	52.0	**64.5**	63.8
	100	**84.9**	67.7	67.0	**58.2**	41.1	41.9
	200	**81.0**	48.3	47.2	**60.1**	25.2	26.1
	400	**70.9**	30.1	29.3	**51.3**	13.7	14.7

For small- to medium- sample sizes, the three covariance estimators under consideration all provide improvement over the sample covariance matrix. Table 3.2 summarizes the percentage reduction in SL and MSE of each covariance estimator as compared to the sample covariance matrix, where the largest reduction among the three is marked in bold. We can see that for case (I), where $\Sigma_Z = I_K$ has very simple structure, all three covariance estimators achieve similar improvement over the sample covariance matrix. When the structure of Σ_Z becomes more complicated,

such as in cases (II) and (III), the RR estimator still leads to considerable improvement over the sample covariance matrix. Specifically, for small sample sizes, such as $T = 20, 40$ and 100, the improvement of the RR estimator is comparable to that of the two shrinkage estimators; while for medium sample sizes, such as $T = 200$ and 400, the RR estimator achieves significantly better performance than the two competing shrinkage estimators with respect to both SL and MSE. In addition, the advantage of the RR estimator in cases (II) and (III) is seen to increase with the sample size T. For example, when T is 400, the percentage reduction of the RR estimator is twice as large as that of the two shrinkage estimators with respect to SL, and almost four times as large with respect to MSE.

3.3.2 Real data examples

We apply the reduced-rank covariance estimator in the VAR modeling of two real data examples. The first example is concerned with stock returns from S&P 500 and corresponds to the first scenario in Section 3.2.2, i.e., there is no constraint on the AR coefficients of the VAR model. The second example is a time series of temperatures in southeast China and corresponds to the second scenario, i.e., there exist zero constraints on the AR coefficients. For both examples, we use the latent space setup introduced in Section 3.2.2 to interpret findings from the reduced-rank covariance estimation.

Stock returns from S&P 500. In the first example, the data consist of daily returns of $K = 55$ stocks in S&P 500 and the stocks come from 4 sectors: *energy*, *industry*, *finance* and *technology*. The returns are calculated as the logarithm of the ratio between two consecutive daily close prices from the 252 trading days in 2006.

Figure 3.1 displays the first 60 observations of the return series.

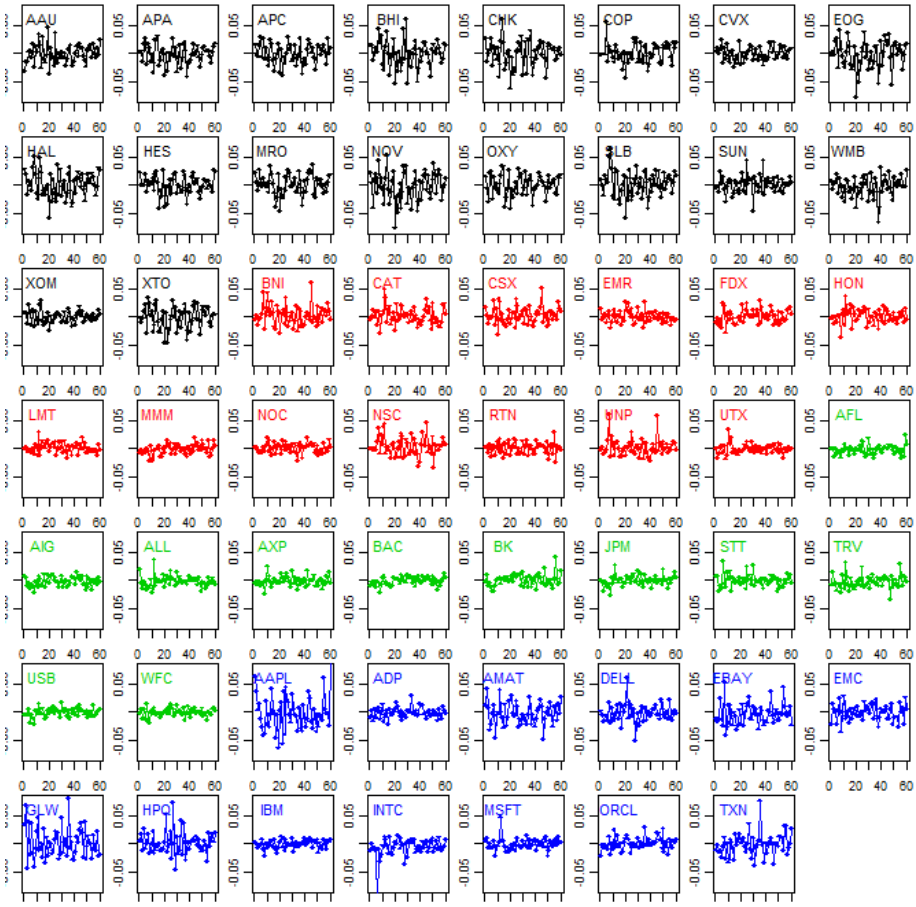


Figure 3.1: The first 60 observations of the return series. The color indicates the sector each stock belongs to: *energy* (black), *industry* (red), *finance* (green), *technology* (blue).

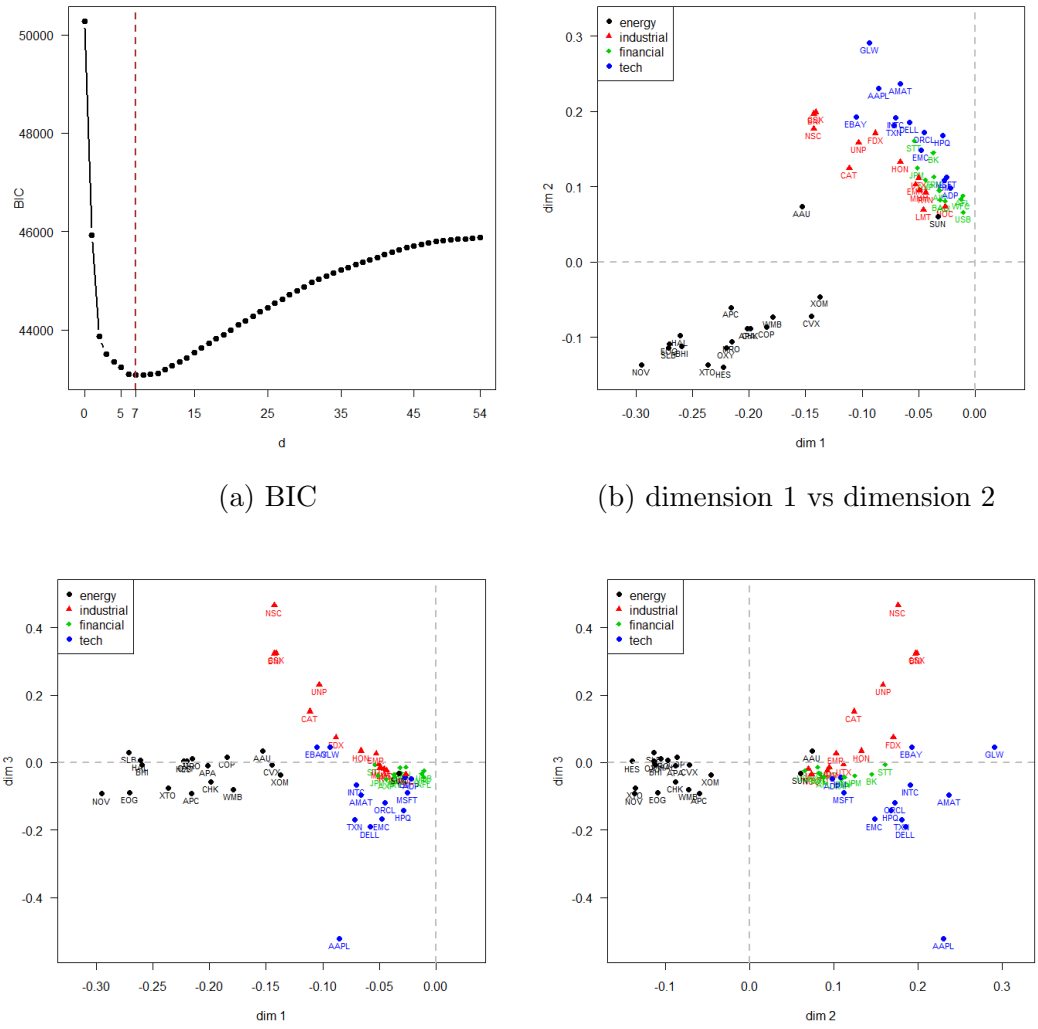
Our interest is to describe the pattern of contemporaneous dependence between returns of the 55 stocks. For this purpose, we apply the reduced-rank covariance estimator in the VAR modeling of the 55-dimensional return series. We use the 2-step method (the first scenario described in Section 3.2.2) to fit a VAR model

with unconstrained AR coefficients and reduced-rank noise covariance matrix. In particular, we first fit an unconstrained VAR(1) model to the 55-dimensional return series, where the autoregression order 1 is selected from $\{0, 1, 2, 3\}$ according to BIC. Then we obtain the reduced-rank covariance estimator based on the residuals from the fitted autoregression. We choose the reduced-rank d from $\{0, 1, \ldots, 54\}$ and panel (a) in Figure 3.2 displays the BIC curve as d varies, which shows that the minimum BIC leads to $d = 7$. In other words, the contemporaneous dependence structure between the 55 stock return series can be well represented in a 7-dimensional latent space.

Panels (b), (c) and (d) in Figure 3.2 explore the 7-dimensional latent space by displaying the layouts of the 55 stocks in its first 3 dimensions, where the color indicates the sector each stock belongs to. Panel (b) corresponds to the first 2 dimensions of the latent space, from which we can observe a "clustering" phenomenon of the 55 stocks in these 2 dimensions. Specifically, the within-sector contemporaneous dependence is most noticeable among stock returns from the *energy* sector, since those *energy* stocks are positioned close to each other while far away from the origin of the latent space. We also observe that the *energy* stocks hold opposite signs along the second dimension against stocks from the *industry*, *finance* and *technology* sectors. It means that returns of the *energy* stocks are negatively contemporaneously related to stock returns from the other 3 sectors. On the contrary, the within-sector contemporaneous dependence is much weaker among stock returns from the *finance* sector, since those stocks are positioned close to the original of the latent space. Moreover, panel (b) also shows that the first 2 dimensions provide information for separating the *energy* sector from the other 3 sectors, but not for distinguishing among the *industry*, *finance* and *technology* stocks. One exception is that there also exists separation between the *industry* and the *technology* sectors. This separation becomes more noticeable after

we take into account the third dimension of the latent space. From panels (c) and (d), both of which display the third dimension along the vertical direction, we can see that the third dimension is informative for separating the *industry* from the *technology* stocks, while it has little power for distinguishing between the *energy* and the *finance* sectors. For the diagnostic check, Figure 3.3 displays the auto-correlation (ACF) and cross-correlation functions (CCF) among the first 5 marginals of the estimated latent variable $\hat{\delta}_t$ as in (3.9), which exhibits little significant auto- or cross- correlation. In fact, we observe little significant auto- or cross- correlation among all 7 marginals of $\hat{\delta}_t$. This observation is consistent with the assumption of the reduced-rank covariance model.

The use of the reduced-rank covariance estimator in the VAR modeling of the returns series also provides improvement over the situation where the unrestricted covariance estimator is employed for the noise covariance matrix. Here the unrestricted covariance estimator refers to the sample covariance matrix of the residuals from the fitted autoregression and it corresponds to the case when $d = K-1 = 54$ in a reduced-rank estimator. Below we will show that, at least for this example, reducing the effective dimension of the noise covariance estimator in large dimensional VAR models also has benefits. From (3.12), we can see that the AR coefficient estimates from the two unconstrained VAR(1) models with $d = 54$ and $d = 7$ are identical. Due to the selection of the reduced-rank d, however, these two VAR(1) models are different in two aspects: the confidence intervals of AR coefficient estimates and the forecast mean squared error, both of which are explored below. Figure 3.4 displays the confidence intervals of the $3025(= 55^2)$ AR coefficient estimates from the two VAR(1) models with $d = 54$ and $d = 7$, respectively. The solid curve shows the 3025 AR coefficient estimates in ascending order, which are identical between the two VAR(1) models. Each vertical line indicates ± 1.96 the corresponding standard error,



(a) BIC

(b) dimension 1 vs dimension 2

(c) dimension 1 vs dimension 3

(d) dimension 2 vs dimension 3

Figure 3.2: Panel (a): The BIC curve as the reduced-rank d varies from 0 to 54. Panels (b), (c) and (d): Layouts of the 55 stocks in the first 3 dimensions of the latent space. The color indicates the sector each stock belongs to: *energy* (black), *industry* (red), *finance* (green), *technology* (blue). The dashed lines show the axes of the latent space.

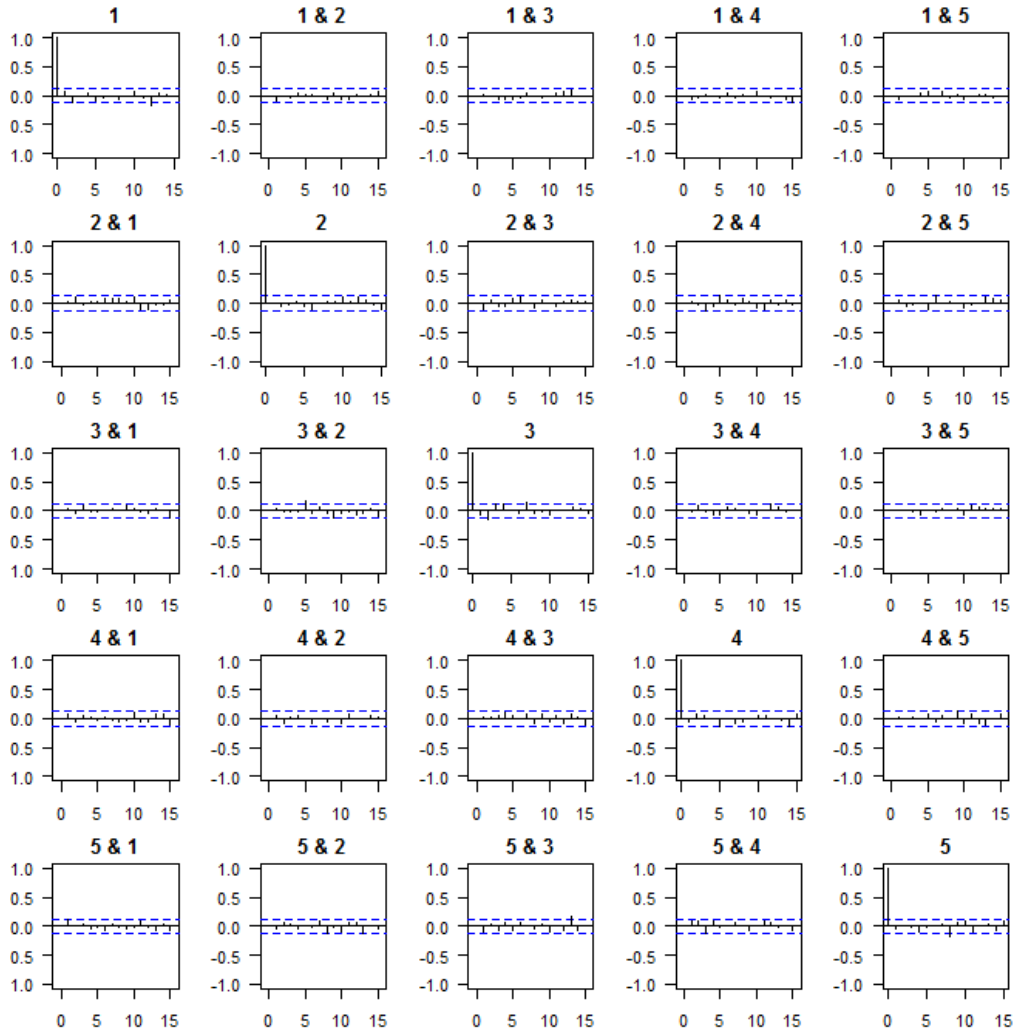


Figure 3.3: The ACF and CCF plots of the first 5 marginals of the estimated latent variable $\hat{\delta}_t$.

where the standard error is computed using (2.14). From Figure 3.4 we can see that reducing the complexity of the noise covariance estimator in VAR models can lead to narrower confidence intervals for AR coefficient estimates as compared to using the unrestricted covariance estimator. The narrower confidence intervals help to better discover significant temporal relationships in the VAR model. We also compare the forecast performance of the two VAR(1) models. The mean squared error (MSE) matrix of 1-step forecast of a VAR(p) model with *estimated* AR matrices $\hat{A}_1, \ldots, \hat{A}_p$ is defined as

$$\text{fMSE}(1) := \mathbb{E}(Y_{t+1} - \sum_{k=1}^{p} \hat{A}_k \hat{Y}_t(1-k))(Y_{t+1} - \sum_{k=1}^{p} \hat{A}_k \hat{Y}_t(1-k))', \tag{3.14}$$

where $\hat{Y}_t(k) := Y_t$ for $k \leq 0$. We use the results in Appendix 3.4.2 to approximate the forecast MSE matrix (3.14) using the estimated AR parameters $\hat{A}_1, \ldots, \hat{A}_p$ and noise covariance matrix $\hat{\Sigma}_Z$. Figure 3.5 displays the difference between the approximate 1-step forecast MSE matrices from the two VAR(1) models with $d = 54$ and $d = 7$ (i.e., the approximate fMSE(1) from $d = 54$ minus the approximate fMSE(1) from $d = 7$) . The solid lines indicate the 4 sectors. We can see from Figure 3.5 that the reduced-rank covariance estimator gives smaller forecast MSE than the unrestricted covariance estimator. The reduction in 1-step forecast MSE is most significant for stocks from the *energy* and the *technology* sectors, which correspond to the first and the fourth blocks in Figure 3.5.

Temperatures in southeast China. This example is concerned with monthly temperature series of $K = 7$ cities in southeast China [2] from January 1988 to December 1998 (Pan and Yao 2008), which gives $T = 132$ observations. Figure 3.6 displays the first 36 observations of the monthly temperature series.

[2]The 7 cities are *Anqing, Dongtai, Hangzhou, Hefei, Huoshan, Nanjing* and *Shanghai*.

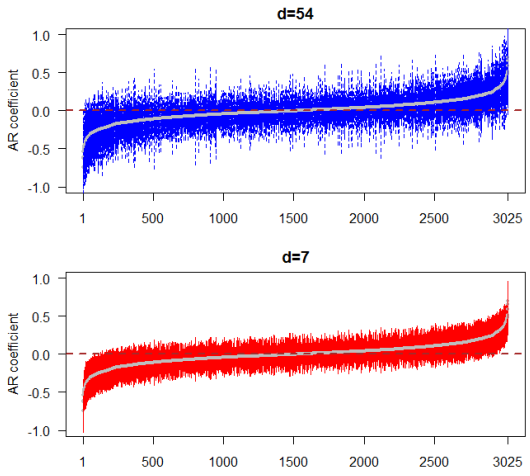


Figure 3.4: Comparison of the confidence intervals of the AR coefficient estimates between the two VAR(1) models with $d = 54$ and $d = 7$. The solid curve shows the $3025 (= 55^2)$ AR coefficient estimates in ascending order. Each vertical line indicates ± 1.96 the corresponding standard error.

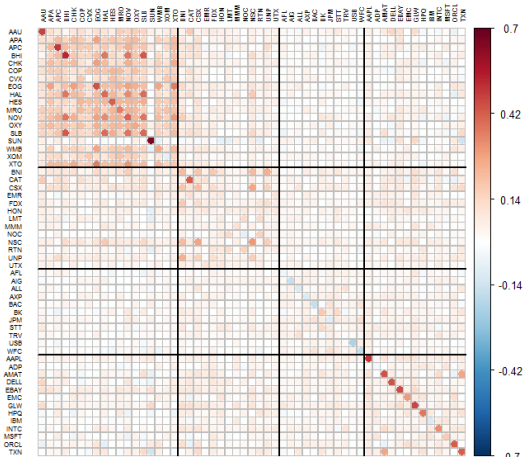


Figure 3.5: Display of the difference between the approximate 1-step forecast MSE matrices of the two VAR(1) models when $d = 54$ and $d = 7$. The solid lines indicate the 4 sectors.

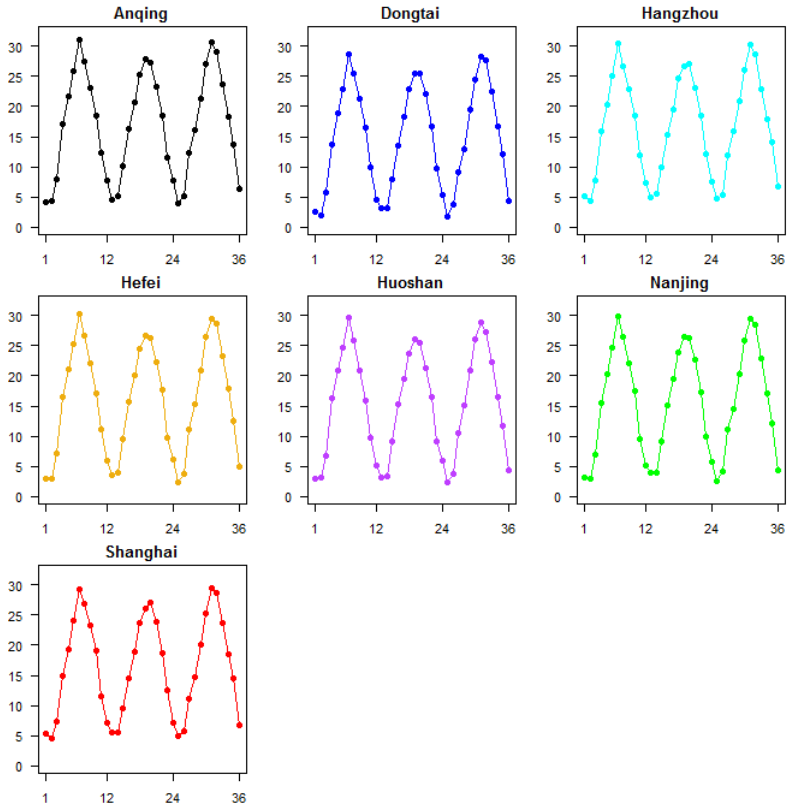
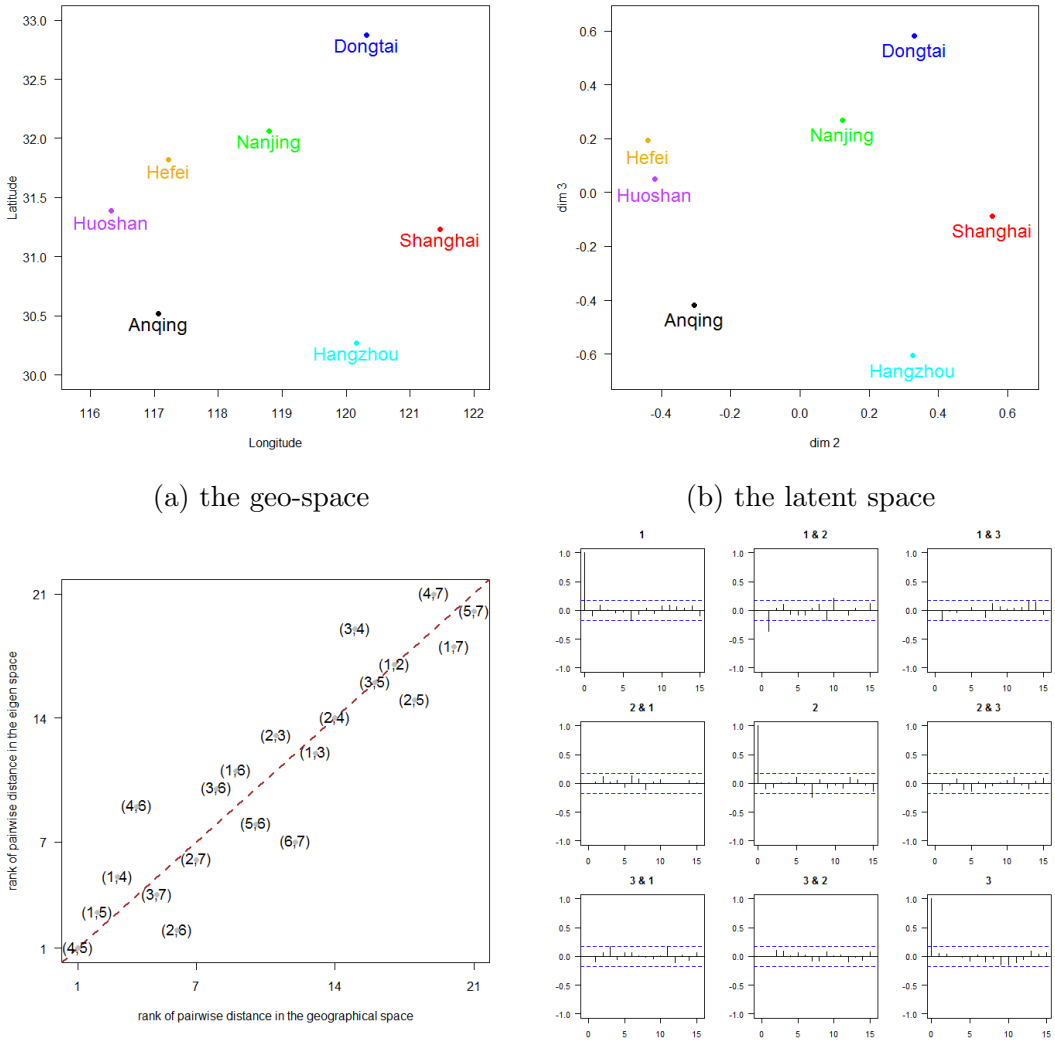


Figure 3.6: The first 36 observations of the temperature series.

We are interested in the contemporaneous dependence structure between the 7 cities' temperature movements. For this purpose, we apply the reduced-rank covariance estimator in the VAR modeling of the 7-dimensional temperature series. We use the iterative procedure (the second scenario described in Section 3.2.2) to fit a VAR model with constrained AR coefficients and reduced-rank noise covariance matrix. Specifically, for each $d \in \{0, 1, \ldots, 6\}$, we use the 2-stage approach introduced in Chapter 2 to determine the zero constraints on the AR coefficients according to minimum BIC. In applying the 2-stage approach, the order of autoregression p is selected

from $\{0, 1, \ldots, 8\}$. Then we choose the reduced-rank d from $\{0, 1, \ldots, 6\}$ according to minimum BIC as well. We finally obtain an order-1 VAR model with 29 non-zero AR coefficients and reduced-rank $d = 3$. The selection of $d = 3$ suggests that the core structure of contemporaneous dependence between the 7 cities' temperatures can be represented in a 3-dimensional latent space.

To obtain insight about this 3-dimensional latent space, we compare the 7 cities' actual geographical locations with their positions in the estimated latent space. The findings are summarized in Figure 3.7. Panels (a) and (b) in Figure 3.7 display the 7 cities' geographical locations (longitude vs latitude) and latent positions (dimension 2 vs dimension 3), respectively. The most noticeable aspect is the similarity between the layouts of the 7 cities in these two spaces. In addition, panel (c) compares the rank of pairwise distance among the 7 cities in the geographical space with that in the latent space. The correlation coefficient between the two sets of ranks is as much as 0.92. The above findings from the reduced-rank covariance estimator suggest that geographical layout is an important factor in explaining the contemporaneous dependence between the 7 cities' temperature movements. This finding is obviously not unexpected since neighboring cities are likely to share similar meteorological and geological conditions, which will impact the temperature within a region. Here we emphasize that *no* geographical information is provided to our model. The latent positions, as given by the rows of $\hat{U}$ (3.5), are purely discovered by the reduced-rank covariance estimation in the VAR modeling of the temperature data. For model diagnostics, panel (d) of Figure 3.7 displays the ACF and CCF among the 3 marginals of the estimated latent variable $\hat{\delta}_t$ as in (3.9). We can see that, with few exceptions, neither the auto-correlation nor cross-correlation is significant, which is consistent with model assumptions.



(a) the geo-space

(b) the latent space

(c) pairwise distance

(d) ACF and CCF

Figure 3.7: Panels (a) and (b): Locations of the 7 cities in the actual geographical space and the estimated latent space, respectively. Panel (c): Ranks of pairwise distance among the 7 cities in the geographical space (x-axis) and in the latent space (y-axis). The numbers stand for: 1-Anqing, 2-Dongtai, 3-Hangzhou, 4-Hefei, 5-Huoshan, 6-Nanjing and 7-Shanghai. Panel (d): The ACF and CCF plots of the 3 marginals of the estimated latent variable $\hat{\delta}_t$.

3.4 Appendix to Chapter 3

3.4.1 Proof of Proposition 1 in Section 3.2.1

Proof of Proposition 1. Notice that the K eigenvalues of $\Sigma_Z = U\Lambda U^{'} + \sigma^2 I_K$ are $\lambda_1 + \sigma^2, \ldots, \lambda_d + \sigma^2, \sigma^2, \ldots, \sigma^2$, so the $-\frac{2}{T}$log-likelihood (3.3) becomes

$$\begin{aligned} -\frac{2}{T}\log L(U, \Lambda, \sigma^2) &= \log|\Sigma_Z| + \mathrm{tr}(\Sigma_Z^{-1} S) \\ &= (K-d)\log(\sigma^2) + \sum_{i=1}^{d}\log(\lambda_i + \sigma^2) + \mathrm{tr}(\Sigma_Z^{-1} S). \end{aligned} \tag{3.15}$$

From standard matrix results, see e.g. Schott (2004), (3.2) gives

$$\begin{aligned} \Sigma_Z^{-1} &= (U\Lambda U^{'} + \sigma^2 I_K)^{-1} \\ &= (\sigma^2 I_K)^{-1} - (\sigma^2 I_K)^{-1} U[\Lambda^{-1} + U^{'}(\sigma^2 I_K)^{-1} U]^{-1} U^{'}(\sigma^2 I_K)^{-1} \\ &= \frac{1}{\sigma^2} I_K - \frac{1}{(\sigma^2)^2} U(\mathrm{diag}\{\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_d}\} + \mathrm{diag}\{\frac{1}{\sigma^2}, \frac{1}{\sigma^2}, \ldots, \frac{1}{\sigma^2}\})^{-1} U^{'} \\ &= \frac{1}{\sigma^2} I_K + \frac{1}{\sigma^2} U \mathrm{diag}\{-\frac{\lambda_1}{\lambda_1 + \sigma^2}, -\frac{\lambda_2}{\lambda_2 + \sigma^2}, \ldots, -\frac{\lambda_d}{\lambda_d + \sigma^2}\} U^{'} \\ &= \frac{1}{\sigma^2}(I_K + U\tilde{\Lambda} U^{'}), \end{aligned} \tag{3.16}$$

where $\tilde{\Lambda} := \mathrm{diag}\{-\frac{\lambda_1}{\lambda_1+\sigma^2}, \ldots, -\frac{\lambda_d}{\lambda_d+\sigma^2}\}$. We point out that it is the assumption of the isotropic error covariance matrix $\mathrm{var}(\varepsilon_t) = \sigma^2 I_K$ that makes it possible to explicitly calculate Σ_Z^{-1} as in (3.16) and eventually leads to the analytical form of the maximum

likelihood estimator. Plugging (3.16) into (3.15), we have

$$\begin{aligned}
-\frac{2}{T}\log L(U,\Lambda,\sigma^2) &= (K-d)\log(\sigma^2)+\sum_{i=1}^{d}\log(\lambda_i+\sigma^2)+\frac{1}{\sigma^2}\mathrm{tr}[(I_K+U\tilde{\Lambda}U^{'})S] \\
&= (K-d)\log(\sigma^2)+\sum_{i=1}^{d}\log(\lambda_i+\sigma^2)+\frac{1}{\sigma^2}\mathrm{tr}(S)+\frac{1}{\sigma^2}\mathrm{tr}(U\tilde{\Lambda}U^{'}S) \\
&= (K-d)\log(\sigma^2)+\sum_{i=1}^{d}\log(\lambda_i+\sigma^2)+\frac{1}{\sigma^2}\sum_{i=1}^{K}c_i+\frac{1}{\sigma^2}\mathrm{tr}(U^{'}SU\tilde{\Lambda}).
\end{aligned} \tag{3.17}$$

Let $\hat{U}$ denote the $K\times d$ matrix whose columns consist of the d eigenvectors that correspond to the d largest eigenvalues of S as in (3.5). Since the diagonal entries of $\tilde{\Lambda}$ are negative and in increasing order, i.e.,$-\frac{\lambda_1}{\lambda_1+\sigma^2}<\ldots<-\frac{\lambda_d}{\lambda_d+\sigma^2}<0$, standard matrix results show that $\mathrm{tr}(U^{'}SU\tilde{\Lambda})$ in (3.17) is minimized by $\hat{U}$. In addition, as long as the relationship $-\frac{\lambda_1}{\lambda_1+\sigma^2}<\ldots<-\frac{\lambda_d}{\lambda_d+\sigma^2}<0$ holds, $\hat{U}$ is the minimizer regardless of the particular values of $\lambda_1,\ldots,\lambda_d$ and σ^2. If the d largest eigenvalues $c_1,\ldots,c_d$ of S are distinct, the minimizer $\hat{U}$ is unique up to column-wise reflections. Additionally, $\hat{U}$ is unique if the signs of entries in one row of $\hat{U}$ are anchored *a priori*.

Now we have $\hat{U}^{'}S\hat{U}=\mathrm{diag}\{c_1,\ldots,c_d\}$, so plugging $\hat{U}$ into (3.17) gives

$$\begin{aligned}
-\frac{2}{T}\log L(\hat{U},\Lambda,\sigma^2) &= (K-d)\log(\sigma^2)+\sum_{i=1}^{d}\log(\lambda_i+\sigma^2)+\frac{1}{\sigma^2}\sum_{i=1}^{K}c_i+\frac{1}{\sigma^2}\mathrm{tr}(\mathrm{diag}\{c_1,\ldots,c_d\}\tilde{\Lambda}) \\
&= (K-d)\log(\sigma^2)+\sum_{i=1}^{d}\log(\lambda_i+\sigma^2)+\frac{1}{\sigma^2}\sum_{i=1}^{K}c_i-\frac{1}{\sigma^2}\sum_{i=1}^{d}\frac{\lambda_i c_i}{\lambda_i+\sigma^2} \\
&= (K-d)\log(\sigma^2)+\sum_{i=1}^{d}\log(\lambda_i+\sigma^2)+\frac{1}{\sigma^2}\sum_{i=d+1}^{K}c_i+\sum_{i=1}^{d}\frac{c_i}{\lambda_i+\sigma^2}.
\end{aligned} \tag{3.18}$$

Minimizing the right-hand size of (3.18) with respect to $\lambda_1, \ldots, \lambda_d$ and σ^2, we have

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{K-d} \sum_{i=d+1}^{K} c_i; \\ \hat{\lambda}_i &= c_i - \hat{\sigma}^2, \text{ for } i = 1, \ldots, d. \end{aligned}$$

which completes the proof.

3.4.2 Approximation of MSE matrices of VAR forecasting

We give results on approximating the mean squared error (MSE) matrix for 1-step forecast of a VAR model. Let $\{Y_t\}$ be the VAR(p) process in (1.1) with $\mu = \mathbf{0}$. Then the optimal 1-step forecast with *estimated* AR coefficients $\hat{A}_1, \ldots, \hat{A}_p$ is given by

$$\hat{Y}_t(1) = \sum_{k=1}^{p} \hat{A}_k \hat{Y}_t(1-k),$$

where $\hat{Y}_t(k) := Y_t$ for $k \leq 0$. It can be shown, see e.g. Lütkepohl (1993), that the MSE matrix of the 1-step forecast $\hat{Y}_t(1)$, which is defined as

$$\text{fMSE}(1) := \mathbb{E}(Y_{t+1} - \hat{Y}_t(1))(Y_{t+1} - \hat{Y}_t(1))',$$

can be approximated by

$$\tilde{\text{fMSE}}(1) := \Sigma_Z + \Omega(1), \tag{3.19}$$

where

$$L_t := (Y_t, Y_{t-1}, \ldots, Y_{t-p+1})', \text{ for } t = 1, \ldots, T, \tag{3.20}$$

$$\Gamma_Y := \text{cov}(L_t) = \text{cov}(Y_t, Y_{t-1}, \ldots, Y_{t-p+1})', \tag{3.21}$$

$$\Omega(1) := \frac{1}{T} \sum_{t=1}^{T} \{(L_t' \Gamma_Y^{-1} L_t) \otimes \Sigma_Z\}. \tag{3.22}$$

We can see that the approximate 1-step forecast MSE matrix fM̃SE(1) (3.19) has two parts: the first part Σ_Z comes from the uncertainty inherent in the VAR model while the second part $\Omega(1)$ (3.22) accounts for the variability in the parameter estimates. We estimate the approximate 1-step forecast MSE matrix fM̃SE(1) by plugging the parameter estimates $\hat{A}_1, \ldots, \hat{A}_p$ and $\hat{\Sigma}_Z$ into (3.19). For such estimation, we need to represent the $Kp \times Kp$ covariance matrix $\Gamma_Y = \text{cov}(Y_t, Y_{t-1}, \ldots, Y_{t-p+1})'$ (3.21) in terms of $A_1, \ldots, A_p$ and Σ_Z. We derive this representation as follows. From (1.1) with $\mu = \mathbf{0}$, we can see that the Kp-dimensional process $\{L_t\}$ (3.20) satisfies the following VAR(1) recursion

$$
\begin{aligned}
& L_t = \Psi L_{t-1} + V_t \\
\Rightarrow & \begin{pmatrix} Y_t \\ Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p+1} \end{pmatrix} = \begin{pmatrix} A_1 & A_2 & \cdots & \cdots & A_p \\ I_K & 0 & \cdots & \cdots & 0 \\ 0 & I_K & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & I_K & 0 \end{pmatrix} \begin{pmatrix} Y_{t-1} \\ Y_{t-2} \\ Y_{t-3} \\ \vdots \\ Y_{t-p} \end{pmatrix} + \begin{pmatrix} Z_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad (3.23)
\end{aligned}
$$

where the $Kp \times Kp$ AR coefficient matrix Ψ in (3.23) is referred to as the *companion* matrix (Reinsel 1997)of the VAR(p) model (1.1). The covariance matrix Σ_V of the Kp-dimensional noise V_t in (3.23) is a $Kp \times Kp$ matrix of zeros except that its upper-left $K \times K$ sub-matrix is equal to Σ_Z. From (3.23), we have

$$
\Gamma_Y = \text{cov}(L_t) = \text{cov}(\Psi L_{t-1} + V_t) = \Psi \Gamma_Y \Psi' + \Sigma_V, \qquad (3.24)
$$

and (3.24) leads to

$$
\begin{aligned}
\text{vec}(\Gamma_Y) &= \text{vec}(\Psi \Gamma_Y \Psi' + \Sigma_V) \\
&= \text{vec}(\Psi \Gamma_Y \Psi') + \text{vec}(\Sigma_V) \\
&= (\Psi \otimes \Psi)\text{vec}(\Gamma_Y) + \text{vec}(\Sigma_V). \qquad (3.25)
\end{aligned}
$$

From (3.25), it follows that

$$\text{vec}(\Gamma_Y) = (I_{K^2p^2} - \Psi \otimes \Psi)^{-1}\text{vec}(\Sigma_V). \tag{3.26}$$

Since Ψ and Σ_V are defined via $A_1, \ldots, A_p$ and Σ_Z, (3.26) shows that Γ_Y can be expressed in terms of $A_1, \ldots, A_p$ and Σ_Z as well.

Chapter 4

Conclusions and Future Directions

In summary, we propose strategies for fitting large dimensional VAR models. We first introduce a 2-stage approach for estimating large VAR models where many of the AR coefficients are zero. The first stage provides initial selection of non-zero AR coefficients based on partial spectral coherence (PSC) and BIC while the second stage further refines spurious non-zero AR coefficients post first stage. The simulation result suggests that the 2-stage approach outperforms Lasso-type methods in estimating VAR models with sparse AR coefficient matrices. We also provide a method of estimating the noise covariance matrix in large dimensional VAR models. The method is based on a reduced-rank covariance model and it can reduce the effective dimension of the covariance estimator. We show by examples that applying the reduced-rank covariance estimator can give better performance of model-fitting and forecasting than using the unrestricted covariance estimator in large dimensional VAR models.

Below we list some future directions following this research.

- In the 2-stage approach, we use PSC in conjunction with BIC for initial selection of non-zero AR coefficients. From the numerical examples we have investigated,

we notice that BIC-selected models tend to exclude AR coefficients that correspond to zero PSCs. 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. It is interesting to see what is the connection between the PSC and the likelihood of a Gaussian VAR series.

- Currently the proposed 2-stage approach can be used for VAR modeling of time series not exceeding 100 dimensions, but is inappropriate for series of over 100 dimensions due to the computational burden. The computational cost primarily comes from the estimation of AR coefficients with zero-constraints imposed. Can we improve the computation efficiency of this step so that the 2-stage approach can be applied to larger dimensions ?

- We use BIC to control the complexity of both the constrained AR coefficient estimator (Chapter 2) and the reduced-rank noise covariance estimator (Chapter 3). How the result will be affected if we use other information criteria, such as AIC and AICC, or cross validations ?

Bibliography

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

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*.

Bai, J. and Ng, S. (2002), "Determining the number of factors in approximate factor models," *Econometrica*, 70, 191–221.

— (2008), "Large dimensional factor analysis," *Foundations and Trends in Econometrics*, 3, 89–163.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005), "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach," *Quarterly Journal of Economics*, 120, 387–422.

Bickel, P. J. and Levina, E. (2008), "Regularized estimation of large covariance matrices," *Annals of Statistics*, 36, 199–227.

Böhm, H. and von Sachs, R. (2009), "Shrinkage estimation in the frequency domain of multivariate time series," *Journal of Multivariate Analysis*, 100, 913–935.

Borg, I. and Groenen, P. (1997), *Modern Multidimensional Scaling: Theory and Applications*, Berlin: Springer-Verlag.

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-Verlag.

Connor, G. and Korajczyk, R. A. (1986), "Performance measurement with the arbitrage pricing theory: A new framework for analysis," *Journal of Financial Economics*, 15, 373–394.

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.

Daniels, M. J. and Kass, R. E. (2001), "Shrinkage estimators for covariance matrices," *Biometrics*, 57, 1173–1184.

Davis, R. A., Zang, P. F., and Zheng, T. (2012), "Sparse vector autoregression modeling," *Working paper*.

Demiralp, S. and Hoover, K. D. (2003), "Searching for the causal structure of a vector autoregression," *Oxford Bulletin of Economic Statistics*, 65, 745–767.

Dempster, A. P. (1972), "Covariance selection," *Biometrics*, 28, 157–175.

Dey, D. K. and Srinivasan, C. (1985), "Estimation of a covariance matrix under Stein's loss," *Annals of Statistics*, 13, 1581–1591.

Dukić, V., Lopes, H. F., and Polson, N. G. (2010), "Tracking flu epidemics using Google flu trends and particle learning," *Working paper*.

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*.

El Karoui, N. (2008), "Operator norm consistent estimation of large dimensional sparse covariance matrices," *Annals of Statistics*, 36, 2717–2756.

Engle, R. and Watson, M. (1981), "A one-factor multivariate time series model of metropolitan wage rates," *Journal of the American Statistical Association*, 76, 774–781.

Eun, C. S. and Shim, S. (1989), "International Transmission of Stock Market Movements," *Journal of Financial and Quantitative Analysis*, 24, 241–256.

Eysenbach, G. (2009), "Infodemiology: tracking flu-related searches on the web for syndromic surveillance," *AMIA: Annual Symposium Proceedings*, 244–248.

Fama, E. F. and French, K. R. (1993), "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33, 3–56.

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

Fan, J., Lv, J., and Qi, L. (2011), "Sparse high dimensional models in economics," *Annual Review of Economics*, 3, 291–317.

Fox, E. and Dunson, D. (2011), "Bayesian nonparametric covariance regression," *Arxiv preprint arXiv:1101.2017*.

Freeman, J. R., Williams, J. T., and Lin, T. (1989), "Vector autoregression and the study of politics," *American Journal of Political Science*, 33, 842–877.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

— (2010), "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33, 1–22.

Fujita, A., Sato, J. R., Garay, H. M., Yamaguchi, R., Miyano, S., Sogayar, M. C., and Ferreira, C. E. (2007), "Modeling gene expression regulatory networks with the sparse vector autoregressive model," *BMC System Biology*, 1.

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.

Granger, C. W. J. (1969), "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37, 424–438.

Haufe, S., Müller, K. R., Nolte, G., and Krämer, N. (2010), "Sparse causal discovery in multivariate time series," *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 6, 97–106.

Hoff, P. D. (2005), "Bilinear mixed-effects models for dyadic data," *Journal of the American Statistical Association*, 100, 286–295.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V., and Banavar, J. R. (2001), "Dynamic modeling of gene expression data," *Proceedings of the National Academy of Sciences*, 98, 1693–1698.

Hsu, N., Hung, H., and Chang, Y. (2008), "Subset selection for vector autoregressive processes using Lasso," *Computational Statistics and Data Analysis*, 52, 3645–3657.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance selection and estimation via penalised normal likelihood," *Biometrika*, 93, 85–98.

Hulth, A., Rydevik, G., and Linde, A. (2009), "Web queries as a source for syndromic surveillance," *PLoS ONE*, 4.

James, W. and Stein, C. (1961), "Estimation with quadratic loss," *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, 1, 361–379.

Jolliffe, I. T. (2002), *Principal Component Analysis*, New York: Springer-Verlag.

Kullback, S. and Leibler, R. A. (1951), "On information and sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

Lam, C. and Yao, Q. (2011), "Factor modeling for high-dimensional time series: inference for the number of factors," *Annals of Statistics*.

Lam, C., Yao, Q., and Bathia, N. (2011), "Estimation of latent factors for high-dimensional time series," *Biometrika*, 98, 901–918.

Lauritzen, S. L. and Wermuth, N. (1989), "Graphical models for associations between variables, some of which are qualitative and some quantitative," *Annals of Statistics*, 17, 31–57.

Lawley, D. N. (1953), "A modified method of estimation in factor analysis and some large sample results," *Uppsala Symposium on Psychological Factor Analysis*, 34–42.

Ledoit, O. and Wolf, M. (2003), "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, 10, 603–621.

— (2004), "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, 88, 365–411.

Li, W. K. and Mcleod, A. I. (1981), "Distribution of the residuals autocorrelations in multivariate ARMA time series models," *Journal of the Royal Statistical Society, B*, 43, 231–239.

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. (1993), *Introduction to Multiple Time Series Analysis*, New York: Springer-Verlag.

Lutkepohl, H. (1993), "Testing for causation between two variables in higher dimensional VAR models," *Studies in Applied Econometrics, Physica, Heidelberg*, 75–91.

Moneta, A. (2004), "Graphical models for structural vector autoregressions," *Working paper*.

Opgen-Rhein, R. and Strimmer, K. (2007), "Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process," *BMC Bioinformatics*, 8.

Pan, J. and Yao, Q. (2008), "Modeling multiple time series via common factors," *Biometrika*, 95, 365–379.

Polgreen, P. M., Chen, Y., Pennock, D. M., and Forrest, N. D. (2008), "Using internet searches for influenza surveillance," *Clinical Infectious Diseases*, 47, 1443–1448.

Reale, M. and Wilson, G. T. (2001), "Identification of vector AR models with recursive structural errors using conditional independence graphs," *Statistitcal Methods and Applications*, 10, 49–55.

Reinsel, G. C. (1997), *Elements of Multivariate Time Series Analysis*, New York: Springer.

Schäfer, J. and Strimmer, K. (2005), "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, 4, 1175–1189.

Schott, J. R. (2004), *Matrix Analysis for Statistics*, New York: Wiley.

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.

Sims, C. A. (1980), "Macroeconomics and reality," *Econometrica*, 48, 1–48.

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.

Stein, C. (1975), "Estimation of a covariance matrix," *Reitz Lecture, IMS-ASA Annual Meeting*.

Stock, J. H. and Watson, M. W. (2001), "Vector autoregressions," *Journal of Economic Perspectives*, 15, 101–115.

— (2006), "Forecasting with many predictors," *Handbook of Economic Forecasting*, 1, 515–554.

Tao, M., Wang, Y., Yao, Y., and Zou, J. (2011), "Large volatility matrix inference via combining low-frequency and high-frequency approaches," *Journal of the American Statistical Association*, 106, 1025–1040.

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

Tipping, M. E. and Bishop, C. M. (1999), "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B*, 61.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., and Canales-Rodríguez, E. (2005), "Estimating brain functional connectivity with sparse multivariate autoregression," *Philosophical Transactions of the Royal Society B*, 360, 969–981.

Wilson, G. T. (2010), "Atmospheric CO_2 and global temperatures: the strength and nature of their dependence." *Working paper*.

Zellner, A. (1970), "Estimation of regression relationships containing unobservable independent variables," *International Economic Review*, 11, 441–454.