Empirical Bayes approach to shrinkage estimation for vector autoregressive models

by

Namgil Lee, Hyemi Choi, Sung-Ho Kim

BK21 Research Report

11 - 01

April 1, 2011

DEPARTMENT OF MATHEMATICAL SCIENCES

The second stage of BK21 Fostering A World Class Talent

KAIST 한국과학기술원 Korea Advanced Institute of Science and Technology

Empirical Bayes Approach to Shrinkage Estimation for Vector Autoregressive Models

Namgil Lee
KAIST, Daejeon, South Korea
Hyemi Choi
Chonbuk National University, Jeonju, South Korea
Sung-Ho Kim
KAIST, Daejeon, South Korea

Summary.
A vector autoregressive(VAR) process is a multivariate version of time series. When it involves a number of variables which are too many for the length of the process, we encounter computational obstacles which include an issue of singular matrix. Several methods were proposed in literature to handle high-dimensional sparse data problems, most of which are however based on the iid-observations assumption. We propose in this paper a Bayesian approach for modeling a VAR process. A main idea in the approach is that we apply Bayesian methods for estimating the coefficient parameters of the VAR model by imposing priors on the coefficients of the model and the variance of the noise which are instrumental for computational feasibility and estimate stability. The shrinkage parameter which is deemed as a hyper-parameter is then sought for under some optimality conditions by applying a variation of cross-validation. The proposed method is compared favorably with other methods known in literature through simulated data and it is also applied to real world data from systems biology. The model structure from the real world data was simpler by the proposed method with at least as good efficiency when compared with other methods.

Keywords: covariance stationarity; high-dimensional data; likelihood function; parameterized cross validation; score function; separation strategy; shrinkage hyper-parameters

1. Introduction

In multivariate data analysis, high-dimensional sparse data problems are commonplace nowadays. As data collection technologies improve, so the dimensionality of data explodes far beyond the number of observations or the data size, which is conspicuous in biology and medicine among others. In addition to the sparsity of data, when the iid-assumption has no ground for the data, computational load usually adds up in the estimation process. We will address an issue of statistical learning with high-dimensional sparse time series data.

A vector autoregressive (VAR) model is useful for representing the inter-relationship among a set of random variables which are autoregressive. Let $\mathbf{y}_t = [y_{t1}, \ldots, y_{td}]', t =$

Address for correspondence: Sung-Ho Kim, Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon, 305-701, South Korea
Email: sung-ho.kim@kaist.edu

$1, \ldots, T$, denote $(d \times 1)$ vectors of time series variables. A VAR model of order p is represented by

$$\mathbf{y}_t = \sum_{k=1}^{p} A_k \mathbf{y}_{t-k} + \mathbf{c} + \varepsilon_t, \tag{1}$$

where $A_k, k = 1, \ldots, p$, are $(d \times d)$ coefficient matrices, $\mathbf{c}$ is a $(d \times 1)$ vector, and ε_t is a $(d \times 1)$ noise vector process with mean zero and covariance matrix V, i.e., $\mathrm{E}[\varepsilon_t] = \mathbf{0}$, and $\mathrm{E}[\varepsilon_t \varepsilon_\tau'] = V$ if $t = \tau$ and 0 if $t \neq \tau$.

The VAR models became popular for analyzing dynamic behavior of economic and financial time series by Sims (1980). Due to flexibility and generality in specifying the correlations between past and future realizations of the variables, VAR models have plenty of applications to forecasting (Sims, 1982; Doan et al., 1984). VAR models are also useful for structural inference. Once the coefficients of a VAR model are estimated, the underlying causal network can be inferred by inspecting the nonzero coefficients of the model, which imply Granger causality between variables (Granger, 1969; Sims, 1972).

In this work we focus on improving accuracy of estimation of the coefficients which are given in matrices $A_k, k = 1, \ldots, p$, especially in case that the number of observations, T, is not large enough relative to the dimensionality of data, d. The ordinary least squares (OLS) method is a standard way for estimating the coefficients provided that T is large enough. When T is smaller than d, however, the OLS is of no use. This may be one of the reasons that the VAR model has not been widely used in research areas such as systems biology and functional magnetic resonance imaging (fMRI). Data sets from these areas usually consist of up to tens or hundreds of thousands of variables while the number of observed time points, T, is often at most one hundred.

The ridge regression method, which is a kind of penalized least squares methods, is an alternative to the OLS method when the data size is not large enough. The ridge regression was introduced by Hoerl and Kennard (1970) to deal with the difficulties caused by the correlations among the predictors. In the ridge regression, we consider a parameter related to the penalty term in addition to the parameters involved in the regular regression model. This parameter is often called the regularization parameter or the shrinkage parameter, and it is used to achieve some level of shrinkage of estimates. A value of the parameter is chosen at an initial stage of the ridge regression analysis. Literature abounds concerning the shrinkage parameter of the ridge regression. For example, Golum et al. (1979) describe the generalized cross validation for determining the shrinkage parameter of the ridge regression. But no known method of estimating the shrinkage parameter will guarantee smaller values of the mean squared error (MSE) than the OLS. Moreover, a wild use of the shrinkage parameter without any assumptions on the coefficients would keep us from statistically meaningful interpretation.

In general, a *shrinkage estimator* refers to an estimator that incorporates a factor of shrinkage into estimation. Stein (1956) developed a sort of explicit shrinkage estimators, and it was improved by James and Stein (1961). The James-Stein type shrinkage estimators were shown to always achieve lower MSE than the OLS estimator. This improvement motivated lots of developments of shrinkage estimators for other classes of parameters including covariance matrices. Especially, Ledoit and Wolf (2004) proposed a linear shrinkage estimator with a uniformly minimum MSE for covariance matrices. They showed that the proposed estimator is invertible and well-conditioned for large dimensional covariance matrices, which means that inverting it does not amplify estimation errors even if the dimension is large compared with the sample size.

Schäfer and Strimmer (2005b) further exploited the Ledoit and Wolf (2004) lemma for analytic calculation of the optimal shrinkage level, and suggested a nonparametric shrinkage method for covariance matrices. The nonparametric shrinkage method adopted the so-called separation strategy in which we express a covariance matrix in terms of variance and correlation matrices. A $(n \times n)$ covariance matrix, Σ, is expressed as

$$\Sigma = D^{\frac{1}{2}} R D^{\frac{1}{2}} \tag{2}$$

where D is the diagonal matrix with diagonal entries $(\Sigma)_{ii}, i = 1, \ldots, n$, and R is the $(n \times n)$ correlation matrix. The nonparametric shrinkage method can be applied for estimation of two shrinkage parameters. One is for the estimation of the variance matrix, D, and the other for the estimation of correlation matrix, R. The nonparametric shrinkage method is computationally convenient and it does not assume any specific underlying distribution. Also its theoretical approach to minimizing the MSE is appropriate for data with a small number of observations.

Barnard et al. (2000) further investigated the separation strategy from the perspective of Bayesian analysis. They noted that the primary motivation for the separation strategy is its flexibility and directness. First, it enables us to deal with individual components of variances. For example, in a Bayesian analysis, we can assess marginal distributions for the variance components and deal with tails of the distributions of individual components. This flexible assessment of individual variance components are especially useful when each variable of the data has different scale units, e.g., height, weight, speed, and so on. Second, we gain computational convenience by dealing with D and R separately. In the absence of reliable knowledge of the dependence structure, it is better to deal with D and R separately than to blindly use such models as the inverse-Wishart distribution. Third, since we don't have much a priori information to distinguish among the entries $(R)_{ij}, i \neq j$, the prior distribution for $(R)_{ij}$ should be invariant to permutations of indices, i.e., $(R)_{ij}$'s are exchangeable in the sense of de Finetti (1972). Finally, most practitioners are more comfortable about thinking in terms of variances and correlations rather than in terms of the spectral decomposition of Σ. The nonparametric shrinkage method was applied for estimating autocovariance matrices and coefficient matrices for high-dimensional VAR processes (Opgen-Rhein and Strimmer, 2007c) and for analyzing high-dimensional data from systems biology (Opgen-Rhein and Strimmer, 2007a,b,c; Schäfer and Strimmer, 2005b).

Despite the aforementioned advantages, however, there are two pitfalls in applying the nonparametric shrinkage method as far as VAR models are concerned. First, independence is assumed among the observations in the data, where nonzero autocovariances are apparent in the data. So the estimated shrinkage parameters are prone to bias and inconsistency. Second, it attempts to estimate the autocovariance matrices while the purpose of estimation is to obtain VAR model coefficients. Since obtaining VAR model coefficients from the estimated autocovariance matrices involves matrix inversion, the estimates are vulnerable to noisy fluctuation in the data. And the number of parameters in the autocovariance matrices is larger than that in the coefficient matrices, which is an additional computational burden.

On the other hand, it is known that shrinkage is implicit in Bayesian inference. That is, the use of a prior distribution makes the maximum likelihood estimator (MLE) to shrink according as the priors implicate (Carlin, 2009; Koop and Korobilis, 2010). A simple choice for prior distributions in a Bayesian VAR model analysis is the natural conjugate prior, i.e., normal and Wishart distributions. Alternative prior distributions include the Minnesota prior by Doan et al. (1984) and Litterman (1986). But these priors require their parameters to be specified beforehand.

A noninformative prior is often adopted in Bayesian inference as an alternative choice to typical prior distributions. It is useful when no reliable prior information concerning the model parameters exists. The uniform prior and the Jeffreys prior (Jeffreys, 1961) are of this kind. However, care must be exercised in making inferences from a Bayesian analysis with noninformative priors when the data size is not large enough. Sun and Ni (2004) noted that a uniform prior on the VAR model coefficients leads to the MLE which is the same as the OLS estimates.

As is well known, we apply empirical Bayes approach to estimate hyper-parameters. Akaike (1980) and Morris (1983) showed examples of the maximum marginal likelihood estimators (MMLE) for empirical Bayes approach, but the MMLE approach still requires enough number of data observations. Akaike (1978) developed an evaluation procedure of the prior distribution in a Bayesian analysis, which is based on the Kullback-Leibler (KL) divergence. This method can be applied when the true distribution is known, and the suggested expected predictive distribution is not easy to calculate in many cases. However, this approach gives rise to the notion that the prior distributions can be evaluated through posterior distributions and that the KL divergence can be used for the evaluation.

In this paper we propose empirical Bayes approaches to shrinkage estimation of VAR model coefficients. We first decompose the VAR model coefficient matrices into variance and correlation components. This decomposition differs from that of the nonparametric shrinkage method in the sense that the nonparametric shrinkage method decomposes the autocovariance matrices while we decompose the model coefficient matrices in the proposed method. In this way we can take advantage of the separation strategy while avoiding the pitfalls mentioned above. Herein the variance components and the correlation components are estimated separately. We modified the nonparametric shrinkage method by incorporating the dependency structure of the autocovariance that is latent in the data into the estimation process of variance components. As for estimating the correlation components, we applied a Bayesian method and obtained conditional posterior distributions for making inferences on the correlation components.

The shrinkage parameter involved in the estimation of the correlation components is one of the hyper-parameters. We know that selection of prior distributions affects estimation especially when the data are of a small size. Therefore, we present analysis results with novel score functions with a view to derive an optimal value of the shrinkage parameter. Moreover, we propose a computational methodology which is called the parameterized cross-validation (PCV) for finding the optimal value of the shrinkage parameter. The PCV method was developed motivated from an earlier work by Koo, Lee, and Kil (2008), and it is effective especially for the data with few observations.

We evaluated the proposed method using simulated data sets. The experimental results demonstrate that the proposed method performs better than existing methods. And we applied the proposed method to real world data from systems biology, which ends up with causal networks which are far sparser than those by Opgen-Rhein and Strimmer (2007c).

This paper is organized in six sections. In Section 2 we describe the nonparametric shrinkage method of Schäfer and Strimmer (2005b) and Opgen-Rhein and Strimmer (2007c) which was inspirational for our work. In Section 3 we suggest the decomposition of VAR model coefficients into variance components and correlation components, and propose a modification to the nonparametric shrinkage method for the estimation of the variance components. In Section 4 we present a Bayesian approach for the shrinkage estimation of the correlation components. We then analyze novel score functions to obtain an optimal value of the shrinkage parameter. Moreover the PCV method is described in search of the

optimal value of the shrinkage parameter. In Section 5 we demonstrate experimental results based on simulated data sets and real world data sets. Conclusions are given in Section 6.

2. Preliminaries

Let $\mathbf{x}_t = [\mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p}]', t = p+1, \ldots, T$, denote $(dp \times 1)$ vectors of predictors. Then the VAR model (1) is re-expressed as

$$\mathbf{y}_t = \Phi' \mathbf{x}_t + \mathbf{c} + \varepsilon_t, \tag{3}$$

where

$$\Phi = [A_1, \ldots, A_p]'$$

represents a $(dp \times d)$ coefficient matrix, and ε_t is a $(d \times 1)$ noise vector process with mean zero and covariance matrix V. We denote the mean-corrected data matrices corresponding to $\mathbf{x}_t$ and $\mathbf{y}_t$ by X and Y, respectively, that is,

$$X = \begin{bmatrix} \mathbf{x}'_{p+1} - \bar{\mathbf{x}}' \\ \vdots \\ \mathbf{x}'_T - \bar{\mathbf{x}}' \end{bmatrix}, \quad Y = \begin{bmatrix} \mathbf{y}'_{p+1} - \bar{\mathbf{y}}' \\ \vdots \\ \mathbf{y}'_T - \bar{\mathbf{y}}' \end{bmatrix}$$

where $\bar{\mathbf{x}} = \frac{1}{T-p} \sum_{t=p+1}^{T} \mathbf{x}_t$ and $\bar{\mathbf{y}} = \frac{1}{T-p} \sum_{t=p+1}^{T} \mathbf{y}_t$ are sample mean vectors.

The OLS estimate of Φ is given by

$$\widehat{\Phi} = (X'X)^{-1} X'Y.$$

The OLS estimate cannot be calculated when the number of observations, T, is small relative to the dimensionality, d. The nonparametric shrinkage (NS) method was proposed by Schäfer and Strimmer (2005b) for inference on large-scale covariance matrices, and Opgen-Rhein and Strimmer (2007c) applied it for estimation of VAR model coefficients.

For notational convenience, let $\mathbf{z}_t = [\mathbf{x}'_t, \mathbf{y}'_t]'$ denote the $((dp+d) \times 1)$ data vectors and $Z = [X, Y]$ denote the corresponding mean-corrected data matrix. Then the sample covariance matrices, $\widehat{S}_1 = \frac{1}{T-p-1} X'X$ and $\widehat{S}_2 = \frac{1}{T-p-1} X'Y$, are submatrices of $\widehat{S} = \frac{1}{T-p-1} Z'Z$, and $\widehat{\Phi}$ is represented by $\widehat{\Phi} = \left(\widehat{S}_1\right)^{-1} \widehat{S}_2$.

Moreover, suppose that the vector process $\mathbf{y}_t$ is covariance-stationary, that is, its first and second order moments are independent of the time t. Let $\mu_{\mathrm{z}} = \mathrm{E}[\mathbf{z}_t]$ and

$$\Sigma = \mathrm{E}\left[(\mathbf{z}_t - \mu_{\mathrm{z}})(\mathbf{z}_t - \mu_{\mathrm{z}})'\right].$$

Σ is often referred to as the 0th autocovariance matrix of $\mathbf{z}_t$ (Hamilton, 1994). And we define $\mu_{\mathrm{x}} = \mathrm{E}[\mathbf{x}_t]$, $\mu_{\mathrm{y}} = \mathrm{E}[\mathbf{y}_t]$, $\Sigma_1 = \mathrm{E}\left[(\mathbf{x}_t - \mu_{\mathrm{x}})(\mathbf{x}_t - \mu_{\mathrm{x}})'\right]$, and $\Sigma_2 = \mathrm{E}\left[(\mathbf{x}_t - \mu_{\mathrm{x}})(\mathbf{y}_t - \mu_{\mathrm{y}})'\right]$. Then Σ_1 and Σ_2 are submatrices of Σ. Under the assumption that Σ is positive definite, Σ_1 is also positive definite and the diagonal entries of Σ are all positive. Then Φ and Σ are related as in

$$\Phi = \Sigma_1^{-1} \Sigma_2. \tag{4}$$

Once a shrinkage estimate $\widehat{S}^*$ of Σ is obtained, we denote the submatrices of $\widehat{S}^*$ corresponding to $\widehat{S}_1$ and $\widehat{S}_2$ of $\widehat{S}$ by $\widehat{S}_1^*$ and $\widehat{S}_2^*$, respectively. Then the shrinkage estimate $\widehat{\Phi}^*$ of Φ is determined from $\widehat{S}^*$ by

$$\widehat{\Phi}^* = \left(\widehat{S}_1^*\right)^{-1} \widehat{S}_2^*.$$

The NS method derives James-Stein type shrinkage estimators (Stein, 1956; James and Stein, 1961) to obtain $\widehat{S}^*$. Specifically, let $\hat{s}_{ij}$ denote the (i,j) entry of $\widehat{S}$. Then it is expressed by sample variances and sample correlations as

$$\hat{s}_{ij} = \hat{r}_{ij}\sqrt{\hat{s}_{ii}\hat{s}_{jj}}.$$

$\widehat{S}^*$ is obtained by shrinking the sample variances and the sample correlations, respectively. The sample variances are shrunken toward their median as in

$$\hat{s}^*_{ii} = \lambda_{\mathrm{v}}\hat{s}_{\mathrm{med}} + (1-\lambda_{\mathrm{v}})\hat{s}_{ii}, \quad i = 1, \ldots, dp+d, \tag{5}$$

where $0 \leq \lambda_{\mathrm{v}} \leq 1$ is a shrinkage parameter and $\hat{s}_{\mathrm{med}}$ is the median of the sample variances, i.e., $\hat{s}_{\mathrm{med}} = \mathrm{median}(\hat{s}_{11}, \ldots, \hat{s}_{dp+d,dp+d})$. $\hat{s}_{\mathrm{med}}$ is taken as the shrinkage target of the estimates of the variances, $\sigma_{ii}, i = 1, \ldots, dp+d$, where σ_{ii} denotes the variance of the ith component of $\mathbf{z}_t$, i.e., σ_{ii} is the (i,i) entry of Σ. We will also use the median as the shrinkage target for our proposed method. Contrary to the sample variances, the sample correlations are shrunken, in the NS method, toward zero as

$$\hat{r}^*_{ij} = \begin{cases} (1-\lambda)\hat{r}_{ij}, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases} \tag{6}$$

where $0 \leq \lambda \leq 1$ is a shrinkage parameter. Finally, the (i,j) entry of $\widehat{S}^*$ is obtained by

$$\hat{s}^*_{ij} = \hat{r}^*_{ij}\sqrt{\hat{s}^*_{ii}\hat{s}^*_{jj}}.$$

Schäfer and Strimmer (2005b) describe how the NS method determines the shrinkage parameters λ_{v} and λ. In this section we only describe how λ_{v} is determined. λ is determined in the same way.

Schäfer and Strimmer (2005b) consider the mean of the sum of squared error losses as a cost function:

$$R(\lambda_{\mathrm{v}}) = \mathrm{E}\left[\sum_{i=1}^{dp+d}(\hat{s}^*_{ii} - \sigma_{ii})^2\right]. \tag{7}$$

Under the assumption that the first two moments of the distributions of $\hat{s}_{ii}$ and $\hat{s}_{\mathrm{med}}$ exist, the cost function is expanded as follows:

$$\begin{aligned} R(\lambda_{\mathrm{v}}) \quad = \quad & \sum_{i=1}^{dp+d} \big\{\lambda_{\mathrm{v}}^2\mathrm{Var}(\hat{s}_{\mathrm{med}}) + (1-\lambda_{\mathrm{v}})^2\mathrm{Var}(\hat{s}_{ii}) \\ & +2\lambda_{\mathrm{v}}(1-\lambda_{\mathrm{v}})\mathrm{Cov}(\hat{s}_{\mathrm{med}}, \hat{s}_{ii}) \\ & + \left(\lambda_{\mathrm{v}}\mathrm{E}[\hat{s}_{\mathrm{med}} - \hat{s}_{ii}] + \mathrm{Bias}(\hat{s}_{ii})\right)^2\big\}. \end{aligned}$$

The optimal value of λ_{v} is obtained by minimizing this function. From $\mathrm{d}R/\mathrm{d}\lambda_{\mathrm{v}} = 0$, we have

$$\lambda^*_{\mathrm{v}} = \frac{\sum_{i=1}^{dp+d}\{\mathrm{Var}(\hat{s}_{ii}) - \mathrm{Cov}(\hat{s}_{\mathrm{med}}, \hat{s}_{ii}) - \mathrm{Bias}(\hat{s}_{ii})\mathrm{E}[\hat{s}_{\mathrm{med}} - \hat{s}_{ii}]\}}{\sum_{i=1}^{dp+d}\mathrm{E}[(\hat{s}_{\mathrm{med}} - \hat{s}_{ii})^2]}. \tag{8}$$

For the sake of simplicity, Schäfer and Strimmer (2005b) make two assumptions. First, they assume that the observations in the data are independent and identically distributed,

so that $\mathrm{Bias}(\hat{s}_{ii})$ is zero. Second, they assume that $\hat{s}_{\mathrm{med}}$ is almost constant, so that $\mathrm{Cov}(\hat{s}_{\mathrm{med}}, \hat{s}_{ii})$ is near zero. These assumptions yield the optimal value of λ_{v} given by

$$\lambda_{\mathrm{v}}^{*} = \frac{\sum_{i=1}^{dp+d} \mathrm{Var}(\hat{s}_{ii})}{\sum_{i=1}^{dp+d} \mathrm{E}[(\hat{s}_{\mathrm{med}} - \hat{s}_{ii})^2]}. \tag{9}$$

Concerning expression (9), Schäfer and Strimmer (2005b) made some interpretations with regard to the optimal value of the shrinkage parameter. The smaller the variance of $\hat{s}_{ii}$ gets, the smaller λ_{v}^{*} becomes; and the smaller the mean squared difference between $\hat{s}_{\mathrm{med}}$ and $\hat{s}_{ii}$ is, the larger λ_{v}^{*} becomes.

Schäfer and Strimmer (2005b) suggested to replace the expectation and the variance in (9) by their sample counterparts, yielding

$$\hat{\lambda}_{\mathrm{NS,v}}^{*} = \frac{\sum_{i=1}^{dp+d} \widehat{\mathrm{Var}}_{\mathrm{NS}}(\hat{s}_{ii})}{\sum_{i=1}^{dp+d} (\hat{s}_{\mathrm{med}} - \hat{s}_{ii})^2},$$

where $\widehat{\mathrm{Var}}_{\mathrm{NS}}(\hat{s}_{ii})$ is calculated as follows: let $w_{tii} = (z_{ti} - \bar{z}_i)^2$ and $\bar{w}_{ii} = \frac{1}{T-p} \sum_{t=p+1}^{T} w_{tii}$. Then $\hat{s}_{ii} = \frac{1}{T-p-1} \sum_{t=p+1}^{T} w_{tii}$, and the variance is further expanded under the assumption that $w_{tii}, t = p+1, \ldots, T$, are independent as

$$\begin{aligned} \widehat{\mathrm{Var}}_{\mathrm{NS}}(\hat{s}_{ii}) &= \frac{1}{(T-p-1)^2} \widehat{\mathrm{Var}}_{\mathrm{NS}} \left(\sum_{t=p+1}^{T} w_{tii} \right) \\ &= \frac{1}{(T-p-1)^2} \sum_{t=p+1}^{T} \widehat{\mathrm{Var}}_{\mathrm{NS}} \left(w_{tii} \right). \end{aligned}$$

Under the independence assumption among w_{tii}'s, the unbiased sample estimate of $\mathrm{Var}\left(w_{tii}\right)$ is given by

$$\widehat{\mathrm{Var}}_{\mathrm{NS}} \left(w_{tii} \right) = \frac{1}{T-p-1} \sum_{t=p+1}^{T} (w_{tii} - \bar{w}_{ii})^2.$$

To keep the value of $\hat{\lambda}_{\mathrm{NS,v}}^{*}$ within $(0, 1)$, we use instead $\hat{\lambda}_{\mathrm{NS,v}}^{**}$ given by $\hat{\lambda}_{\mathrm{NS,v}}^{**} = \max(0, \min(1, \hat{\lambda}_{\mathrm{NS,v}}^{*}))$.

The NS method is available when T is substantially smaller than d. However, it is under the assumption that $w_{tii}, t = p+1, \ldots, T$, are independent, which is obviously incongruous with time series data. In next section we will propose a modification to the NS method for estimation of $\mathrm{Var}(\hat{s}_{ii})$.

3. Modification to the nonparametric shrinkage method

3.1. Decomposition of VAR model coefficients

From (4), the coefficient matrix Φ is decomposed as follows.

THEOREM 1. *The following variance-correlation decomposition holds for the submatrices Σ_1 and Σ_2 of Σ:*

$$\Sigma_1 = D_1^{\frac{1}{2}} R_1 D_1^{\frac{1}{2}}, \quad \Sigma_2 = D_1^{\frac{1}{2}} R_2 D_2^{\frac{1}{2}}, \tag{10}$$

where D_1 is a $(dp \times dp)$ diagonal matrix with diagonal entries $\sigma_{ii}, i = 1, \ldots, dp$, D_2 is a $(d \times d)$ diagonal matrix with diagonal entries $\sigma_{ii}, i = dp+1, \ldots, dp+d$, and R_1 and R_2 are correlation matrices. Moreover, from (4), the coefficient matrix Φ is decomposed as

$$\Phi = D_1^{-\frac{1}{2}} \Psi D_2^{\frac{1}{2}} \tag{11}$$

where $\Psi = R_1^{-1} R_2$.

In (11), we regard Ψ as the correlation component of Φ, and (D_1, D_2) as its variance component. From Theorem 1, we get the idea that an estimate of Φ is obtained by estimating Ψ and (D_1, D_2). Then the total number of parameters to be estimated slightly increases from d^2p, for Φ, to $d^2p + dp + d$, for Ψ and (D_1, D_2). On the other hand, it is much less than $(dp + d)(dp + d + 1)/2$, for Σ, which is the number of the parameters estimated by the NS method in Opgen-Rhein and Strimmer (2007c). Moreover, this decomposition of Φ into Ψ and (D_1, D_2) takes advantage of the separation strategy mentioned in Barnard et al. (2000), that is, the flexibility and directness from the computational aspect. In subsequent sections we will propose shrinkage estimation methods based on this variance-correlation decomposition of Φ.

3.2. Estimation of variance component

In this subsection we suggest a modification to the NS method described in Section 2 for the shrinkage estimation of (D_1, D_2). Note that the diagonal entries of D_1 and D_2 are the variances of the components of $\mathbf{z}_t$, i.e., $\sigma_{ii}, i = 1, \ldots, dp + d$. Thus, the shrinkage estimators, $\hat{s}^*_{ii}$, of the variances in (5) with the optimal value, λ^*_{v}, of λ in (9) is considered for the estimation of (D_1, D_2).

In the simplification of (8) into (9), we were under two assumptions. First, we assumed that $\mathrm{Bias}(\hat{s}_{ii})$ is ignorably small. If the observations, $z_{ti}, t = p + 1, \ldots, T$, in the data were independent and identically distributed, then the bias would be zero. But this is usually violated for time series data. We rather rely on the fact that the bias decreases to zero at a fast rate as T increases (Anderson, 1971). Second, we assumed that $\mathrm{Cov}(\hat{s}_{\mathrm{med}}, \hat{s}_{ii})$ is small enough. If d gets larger, then the covariance will become smaller because the median will less be affected by a change in the value of a single sample variance. A larger covariance value implies both $\hat{s}_{ii}$ and $\hat{s}_{\mathrm{med}}$ move in a more similar direction and so the shrinkage parameter is less influential. In this respect we consider the simplified expression (9).

We propose a different way of estimating $\mathrm{Var}(\hat{s}_{ii})$ in expression (9) as

$$\hat{\lambda}^*_{\mathrm{EB,v}} = \frac{\sum_{i=1}^{dp+d} \widehat{\mathrm{Var}}_{\mathrm{EB}}(\hat{s}_{ii})}{\sum_{i=1}^{dp+d} (\hat{s}_{\mathrm{med}} - \hat{s}_{ii})^2}.$$

Since $w_{tii}, t = p + 1, \ldots, T$, are not independent, it follows that

$$\begin{aligned}
\widehat{\mathrm{Var}}_{\mathrm{EB}}(\hat{s}_{ii}) &= \frac{1}{(T-p-1)^2} \widehat{\mathrm{Var}}_{\mathrm{EB}} \left(\sum_{t=p+1}^{T} w_{tii} \right) \\
&= \frac{1}{(T-p-1)^2} \left(\sum_{t=p+1}^{T} \widehat{\mathrm{Var}}_{\mathrm{EB}} \left(w_{tii} \right) + \sum_{t=p+1}^{T} \sum_{\substack{\tau=p+1 \\ \tau \neq t}}^{T} \widehat{\mathrm{Cov}}_{\mathrm{EB}} \left(w_{tii}, w_{\tau ii} \right) \right).
\end{aligned}$$

If $w_{tii}, t = p+1, \ldots, T$, were independent, the covariance terms would become zero and only the variance terms would remain. But they are dependent on each other, so the covariance terms should not be ignored.

If we regard $\{w_{tii}\}$ as a covariance-stationary time series, the following estimates are typical choices for the variance terms and the covariance terms:

$$\widehat{\mathrm{Var}}_{\mathrm{EB}}\left(w_{tii}\right) = \frac{1}{T-p}\sum_{t=p+1}^{T}(w_{tii}-\bar{w}_{ii})^2, \tag{12}$$

$$\widehat{\mathrm{Cov}}_{\mathrm{EB}}\left(w_{tii}, w_{t+k,ii}\right) = \begin{cases} \frac{1}{T-p}\sum_{t=p+1}^{T-k}(w_{tii}-\bar{w}_{ii})(w_{t+k,ii}-\bar{w}_{ii}), & \text{for } k \geq 0 \\ \frac{1}{T-p}\sum_{t=p+1}^{T+k}(w_{t-k,ii}-\bar{w}_{ii})(w_{tii}-\bar{w}_{ii}), & \text{for } k < 0 \end{cases}. \tag{13}$$

Expressions (12) and (13) are used in literature for time series analysis (Hamilton, 1994; Wei, 2005). Wei (2005) compared the estimator, $\widehat{\mathrm{Cov}}_{\mathrm{EB}}\left(w_{tii}, w_{t+k,ii}\right)$, defined by (13) with the estimator $\frac{T-p}{T-p-k}\widehat{\mathrm{Cov}}_{\mathrm{EB}}\left(w_{tii}, w_{t+k,ii}\right)$ from the perspective of bias and variance of estimator. As for bias, both of the estimators are biased estimators. But the bias of the latter increases faster than that of the former as $\mathrm{Var}(\bar{w}_{ii})$ increases. So if there are only a small number of observations, the former estimator is preferred. As for variance, the variance of the latter is larger than that of the former, and the difference depends on k. So if k is large, the former estimator is preferred. Wei (2005) pointed out that the former estimator was shown to have smaller MSEs in some cases.

4. Empirical Bayes approach to estimation of correlation components

In this section we suggest a Bayesian approach for the shrinkage estimation of $\Psi = R_1^{-1}R_2$. Note that in Section 2 the shrinkage estimators, $\hat{r}^*_{ij}$, of the correlations are defined in (6) with a shrinkage parameter λ. This can be rewritten by using matrices as follows. Let $\widehat{R}$ denote the $((dp+d)\times(dp+d))$ sample correlation matrix with entries $\hat{r}_{ij}$, and $\widehat{R}_1$ and $\widehat{R}_2$ denote the submatrices of $\widehat{R}$ corresponding to $\widehat{S}_1$ and $\widehat{S}_2$ of $\widehat{S}$. Then, each entry of $\widehat{R}_1$ and $\widehat{R}_2$ is shrunken toward zero except for the diagonal entries of $\widehat{R}_1$ to yield the shrinkage estimators, $\widehat{R}^*_1$ and $\widehat{R}^*_2$, as

$$\widehat{R}^*_1 = (1-\lambda)\widehat{R}_1 + \lambda I$$

and

$$\widehat{R}^*_2 = (1-\lambda)\widehat{R}_2.$$

The shrinkage estimator, $\widehat{\Psi}^*$, of Ψ is defined, based on $\widehat{R}^*_1$ and $\widehat{R}^*_2$, by

$$\widehat{\Psi}^* = (\widehat{R}^*_1)^{-1}\widehat{R}^*_2. \tag{14}$$

In this section we derive $\widehat{\Psi}^*$ from conditional posterior distributions obtained by applying a Bayesian method. Then we present analysis results with novel score functions to obtain an optimal value, λ^*, of the shrinkage parameter λ which is one of the hyper-parameters. Moreover we propose a computational methodology for calculating λ^*, especially for the data with few observations.

4.1. Bayesian approach for shrinkage estimation

From model (3) we get

$$\mathbf{y}_t^{\mathrm{s}} = \Psi' \mathbf{x}_t^{\mathrm{s}} + \mathbf{c}^{\mathrm{s}} + D_2^{-\frac{1}{2}} \varepsilon_t, \tag{15}$$

where $\mathbf{y}_t^{\mathrm{s}} = D_2^{-\frac{1}{2}} \mathbf{y}_t$, $\mathbf{x}_t^{\mathrm{s}} = D_1^{-\frac{1}{2}} \mathbf{x}_t$, and $\mathbf{c}^{\mathrm{s}} = D_2^{-\frac{1}{2}} \mathbf{c}$ denote the standardized vectors. In this section, we assume that the variance component, (D_1, D_2), is known.

Suppose that the noise vectors ε_t in (15) are iid with the multivariate normal distribution with mean zero and covariance matrix $V = \sigma^2 D_2$. Then, from (15), the likelihood function of $\left(\Psi, \sigma^2, \mathbf{c}^{\mathrm{s}}\right)$ is

$$\begin{aligned} L(\Psi, \sigma^2, \mathbf{c}^{\mathrm{s}}) &= \prod_{t=p+1}^{T} \mathrm{N}_d(\mathbf{y}_t | \Phi' \mathbf{x}_t + \mathbf{c}, \sigma^2 D_2) \\ &= \prod_{t=p+1}^{T} \left[\frac{1}{(2\pi)^{d/2} |\sigma^2 D_2|^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left\| D_2^{-\frac{1}{2}} (\mathbf{y}_t - \Phi' \mathbf{x}_t - \mathbf{c}) \right\|^2 \right\} \right] \\ &= \left(\frac{1}{2\pi}\right)^{d(T-p)/2} \left(\frac{1}{|D_2|}\right)^{(T-p)/2} \left(\frac{1}{\sigma^2}\right)^{d(T-p)/2} \\ &\quad \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=p+1}^{T} \left\| \mathbf{y}_t^{\mathrm{s}} - \Psi' \mathbf{x}_t^{\mathrm{s}} - \mathbf{c}^{\mathrm{s}} \right\|^2 \right\}. \end{aligned} \tag{16}$$

We apply a noninformative prior distribution of $\mathbf{c}^{\mathrm{s}}$. That is,

$$\pi(\mathbf{c}^{\mathrm{s}}) \propto m_1 > 0.$$

The likelihood function of (Ψ, σ^2) is obtained as in

$$L(\Psi, \sigma^2) = \int L(\Psi, \sigma^2, \mathbf{c}^{\mathrm{s}}) \pi(\mathbf{c}^{\mathrm{s}}) \mathrm{d}\mathbf{c}^{\mathrm{s}}.$$

Since we have

$$\begin{aligned} \sum_{t=p+1}^{T} \left\| \mathbf{y}_t^{\mathrm{s}} - \Psi' \mathbf{x}_t^{\mathrm{s}} - \mathbf{c}^{\mathrm{s}} \right\|^2 &= (T-p) \left\| \mathbf{c}^{\mathrm{s}} - (\bar{\mathbf{y}}^{\mathrm{s}} - \Psi' \bar{\mathbf{x}}^{\mathrm{s}}) \right\|^2 \\ &\quad + \sum_{t=p+1}^{T} \left\| \mathbf{y}_t^{\mathrm{s}} - \Psi' \mathbf{x}_t^{\mathrm{s}} - (\bar{\mathbf{y}}^{\mathrm{s}} - \Psi' \bar{\mathbf{x}}^{\mathrm{s}}) \right\|^2 \end{aligned}$$

where $\bar{\mathbf{y}}^{\mathrm{s}} = \frac{1}{T-p} \sum_{t=p+1}^{T} \mathbf{y}_t^{\mathrm{s}}$ and $\bar{\mathbf{x}}^{\mathrm{s}} = \frac{1}{T-p} \sum_{t=p+1}^{T} \mathbf{x}_t^{\mathrm{s}}$ are the sample mean vectors, we get

$$\begin{aligned} L(\Psi, \sigma^2) &\propto m_1 \left(\frac{1}{2\pi}\right)^{d(T-p-1)/2} \left(\frac{1}{|D_2|}\right)^{(T-p)/2} \left(\frac{1}{\sigma^2}\right)^{d(T-p-1)/2} \left(\frac{1}{T-p}\right)^{d/2} \\ &\quad \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=p+1}^{T} \left\| \mathbf{y}_t^{\mathrm{s}} - \Psi' \mathbf{x}_t^{\mathrm{s}} - (\bar{\mathbf{y}}^{\mathrm{s}} - \Psi' \bar{\mathbf{x}}^{\mathrm{s}}) \right\|^2 \right\}. \end{aligned} \tag{17}$$

The MLE of (Ψ, σ^2) is obtained by maximizing the likelihood function (17), which is given in the closed form as

$$\widehat{\Psi} = \left((X^{\mathrm{s}})' X^{\mathrm{s}}\right)^{-1} (X^{\mathrm{s}})' Y^{\mathrm{s}}$$

and

$$\hat{\sigma}^2 = \frac{1}{T-p}\left\|Y^{\mathrm{s}} - X^{\mathrm{s}}\,\widehat{\Psi}\right\|^2$$

when the inverse exists, where X^{s} and Y^{s} are the mean-corrected data matrices defined by

$$X^{\mathrm{s}} = \begin{bmatrix} (\mathbf{x}^{\mathrm{s}}_{p+1})' - (\bar{\mathbf{x}}^{\mathrm{s}})' \\ \vdots \\ (\mathbf{x}^{\mathrm{s}}_{T})' - (\bar{\mathbf{x}}^{\mathrm{s}})' \end{bmatrix}, \quad Y^{\mathrm{s}} = \begin{bmatrix} (\mathbf{y}^{\mathrm{s}}_{p+1})' - (\bar{\mathbf{y}}^{\mathrm{s}})' \\ \vdots \\ (\mathbf{y}^{\mathrm{s}}_{T})' - (\bar{\mathbf{y}}^{\mathrm{s}})' \end{bmatrix},$$

and the matrix norm $\|A\|$ for an $(m \times n)$ matrix A is defined by

$$\|A\| = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{\mathrm{trace}(A'A)}\,.$$

Note that if we replace (D_1, D_2) with the sample variances, $(\widehat{D}_1, \widehat{D}_2)$, then $\widehat{\Psi}$ is re-expressed in terms of the sample correlation matrices as

$$\widehat{\Psi} = \widehat{R}_1^{-1}\widehat{R}_2 = \left(\frac{1}{T-p-1}(X^{\mathrm{s}})'X^{\mathrm{s}}\right)^{-1}\frac{1}{T-p-1}(X^{\mathrm{s}})'Y^{\mathrm{s}}, \tag{18}$$

and $\widehat{\Psi}^* = \widehat{\Psi}$ when $\lambda = 0$.

$\widehat{\Psi}^*$ can be obtained by the following Bayesian procedure. We impose prior distributions for the parameters Ψ and σ^2. The shrinkage parameter λ is regarded as a hyper-parameter for the prior distributions. As for the prior of Ψ, we take the multivariate normal distribution given by

$$\pi_1(\,\Psi|\sigma^2, \lambda) = \prod_{i=1}^{d}\prod_{j=1}^{dp} \mathrm{N}\left(\,\Psi_{ij}\,\middle|\,0, \frac{(1-\lambda)\sigma^2}{\lambda(T-p-1)}\right), \quad 0 < \lambda < 1, \tag{19}$$

where Ψ_{ij} is the (i,j) entry of Ψ. This prior reflects our view-point that the coefficients, $\{\,\Psi_{ij}\}$, are dispersed around 0 and exchangeable in the sense of de Finetti (1972). Actually, the assumption of the zero prior mean is already reflected in the shrinkage estimator (14).

We take for the prior of σ^2 an inverse gamma distribution given by

$$\pi_2(\sigma^2) = \mathrm{IG}(\sigma^2|\alpha, \beta), \quad \alpha, \beta > 0. \tag{20}$$

The inverse gamma distribution is known to be a conjugate prior of the normal distribution.

Now we are ready to calculate the conditional posterior of $(\,\Psi, \sigma^2)$ given λ.

THEOREM 2. *With the likelihood given in (17) and the priors in (19) and (20), the following holds:*

(a) The conditional posterior distribution of Ψ given σ^2 and λ is the multivariate normal distribution given by

$$\pi_1(\,\Psi|\sigma^2, \lambda; \{\mathbf{y}_t\}_{t=1}^{T}) = \prod_{i=1}^{d} \mathrm{N}_{dp}(\,\Psi_i|\,\widehat{\Psi}_i^*, \widehat{K}^*(\sigma^2, \lambda))$$

where Ψ_i and $\widehat{\Psi}_i^*$ represent the ith column vectors of Ψ and $\widehat{\Psi}^*$, respectively, and

$$\widehat{K}^*(\sigma^2,\lambda)=\sigma^2\left((X^{\mathrm{s}})'X^{\mathrm{s}}+\frac{\lambda(T-p-1)}{1-\lambda}I\right)^{-1}.$$

That is, the conditional posterior mean is the same as the shrinkage estimator, $\widehat{\Psi}^*$, with the shrinkage parameter λ.

(b) *The conditional posterior distribution of σ^2 given λ is the inverse gamma distribution given by*

$$\pi_2(\sigma^2|\lambda;\{\mathbf{y}_t\}_{t=1}^T)=\mathrm{IG}(\sigma^2|\hat{\alpha}^*,\hat{\beta}^*(\lambda)),$$

where

$$\hat{\alpha}^*=\alpha+\frac{d(T-p-1)}{2}$$

and

$$\begin{aligned}\hat{\beta}^*(\lambda) &= \beta+\frac{1}{2}\mathrm{trace}\left((Y^{\mathrm{s}})'Y^{\mathrm{s}}-(Y^{\mathrm{s}})'X^{\mathrm{s}}\left((X^{\mathrm{s}})'X^{\mathrm{s}}+\frac{\lambda(T-p-1)}{1-\lambda}I\right)^{-1}(X^{\mathrm{s}})'Y^{\mathrm{s}}\right)\\ &= \beta+\frac{1}{2}\left(\left\|Y^{\mathrm{s}}-X^{\mathrm{s}}\widehat{\Psi}^*\right\|^2+\frac{\lambda(T-p-1)}{1-\lambda}\left\|\widehat{\Psi}^*\right\|^2\right).\end{aligned}$$

(c) *The marginal likelihood of λ is given by*

$$\begin{aligned}L(\lambda) &= \int\int L(\Psi,\sigma^2)\pi_1(\Psi|\sigma^2,\lambda)\pi_2(\sigma^2)\mathrm{d}\Psi\mathrm{d}\sigma^2\\ &\propto \frac{1}{(\hat{\beta}^*(\lambda))^{\hat{\alpha}^*}}\left(\frac{\lambda}{1-\lambda}\right)^{d^2p/2}\left|(X^{\mathrm{s}})'X^{\mathrm{s}}+\frac{\lambda(T-p-1)}{1-\lambda}I\right|^{-d/2}. \qquad (21)\end{aligned}$$

PROOF. See Appendix A.

In search for an optimal value of λ, maximizing the marginal likelihood of λ is one of the typical methods in Bayesian analysis . From (21), we have

$$\frac{2}{d}\log L(\lambda)=const.+dp\log\lambda-\frac{2\hat{\alpha}^*}{d}\log\hat{\beta}^*(\lambda)-\log\left|(1-\lambda)\widehat{R}_1+\lambda I\right| \qquad (22)$$

where $\widehat{R}_1$ is the sample correlation matrix as in (18). Some properties of the maximum marginal likelihood estimator (MMLE), $\hat{\lambda}$, are summarized from (22) as follows:

(a) The term $dp\log\lambda$ implies that $\hat{\lambda}$ increases as dp increases.
(b) We can show that $\hat{\beta}^*(\lambda)$ is an increasing function of λ as follows:

$$2\left(\hat{\beta}^*(\lambda)-\beta\right)=\mathrm{trace}\left((Y^{\mathrm{s}})'Y^{\mathrm{s}}-(Y^{\mathrm{s}})'X^{\mathrm{s}}\left((X^{\mathrm{s}})'X^{\mathrm{s}}+\frac{\lambda(T-p-1)}{1-\lambda}I\right)^{-1}(X^{\mathrm{s}})'Y^{\mathrm{s}}\right)$$

and

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{trace}\left((Y^{\mathrm{s}})'Y^{\mathrm{s}}-(Y^{\mathrm{s}})'X^{\mathrm{s}}\left((X^{\mathrm{s}})'X^{\mathrm{s}}+\theta I\right)^{-1}(X^{\mathrm{s}})'Y^{\mathrm{s}}\right)\\ &\quad=\sum_{i=1}^{d}(Y_i^{\mathrm{s}})'X^{\mathrm{s}}\left((X^{\mathrm{s}})'X^{\mathrm{s}}+\theta I\right)^{-2}(X^{\mathrm{s}})'Y_i^{\mathrm{s}}\\ &\quad\geq 0\end{aligned}$$

where Y_i^{s} is the ith column vector of Y^{s}. Since $2\hat{\alpha}^*/d = 2\alpha/d + T - p - 1$ is increasing in T, the term, $-\frac{2\hat{\alpha}^*}{d}\log\hat{\beta}^*(\lambda)$, enforces $\hat{\lambda}$ to decrease as T increases.

(c) The smaller T is, the more the eigenvalues of $\widehat{R}_1$ become zero when $T < dp$. Hence a larger $\hat{\lambda}$ is consequential for a smaller T.

But the MMLE is unreliable when T is small relative to d. This point of concern leads us to another way of searching for an optimal value of λ.

4.2. Empirical Bayes analysis for shrinkage parameter

In this section we present analysis results based on some criteria regarding the prior and posterior distributions in search of an optimal value of λ.

4.2.1. Bounded variance principle

In (19), we presumed that the prior mean for each Ψ_{ij} is zero. However, if the prior variance is too large, the prior mean loses its influence. In this respect, when T is small relative to d, it is desirable that we bound the variance of the prior distribution from above.

From the prior distributions, $\pi_1(\Psi|\sigma^2,\lambda)$ and $\pi_2(\sigma^2)$, in (19) and (20), we can derive that

$$\begin{aligned}\mathrm{Var}(\Psi_{ij}|\lambda) &= \mathrm{Var}(\mathrm{E}[\Psi_{ij}|\sigma^2,\lambda]|\lambda) + \mathrm{E}\left[\mathrm{Var}(\Psi_{ij}|\sigma^2,\lambda)|\lambda\right] \\ &= \frac{1-\lambda}{\lambda(T-p-1)}\frac{\beta}{\alpha-1}\end{aligned}$$

where α and β are the hyper-parameters of $\pi_2(\sigma^2)$ in (20). Therefore, for some constant $m_2 > 0$, the inequality, $\mathrm{Var}(\Psi_{ij}|\lambda) \le m_2$, implies that an optimal value, λ^*, of λ satisfies

$$\lambda^* \ge \frac{1}{m_3(T-p-1)+1} \tag{23}$$

for some constant $m_3 > 0$.

4.2.2. Analysis based on Kullback-Leibler divergence

In this section we will consider a score function of λ based on the so-called Kullback-Leibler (KL) divergence. The KL divergence between two probability densities f and g is defined by

$$D_{\mathrm{KL}}(f||g) = \int f\log\frac{f}{g}\ . \tag{24}$$

As is well known, the KL divergence is nonnegative with $D_{\mathrm{KL}}(f||g) = 0$ if and only if $f = g$.

We define $(\Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*})$ as the value of $(\Psi, \sigma^2, \mathbf{c}^{\mathrm{s}})$ that minimizes the KL divergence between the true distribution $f(\mathbf{x}_t, \mathbf{y}_t) = f(\mathbf{y}_t|\mathbf{x}_t)f(\mathbf{x}_t)$ and the distribution $g(\mathbf{x}_t, \mathbf{y}_t|\,\Psi, \sigma^2, \mathbf{c}^{\mathrm{s}}) = g(\mathbf{y}_t|\,\Psi, \sigma^2, \mathbf{c}^{\mathrm{s}}, \mathbf{x}_t)f(\mathbf{x}_t)$ from our probability model. We suppose that $(\Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*})$ is the unique minimizer of the KL divergence, which can be verified by using the true VAR model in (15) with parameters $(\Psi, V, \mathbf{c}^{\mathrm{s}})$ and the density function $g(\mathbf{y}_t|\,\Psi, \sigma^2, \mathbf{c}^{\mathrm{s}}, \mathbf{x}_t)$ in (16).

A basic result from a Bayesian analysis is that the posterior distribution $\pi(\Psi, \sigma^2, \mathbf{c}^{\mathrm{s}}|\,\{\mathbf{y}_t\}_{t=1}^T)$ becomes concentrated about $(\Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*})$ as T increases (Gelman et al., 1995).

This limiting distribution is usually represented by the Dirac delta function or the Dirac measure (Aliprantis and Burkinshaw, 1998). The Dirac delta function $\delta(\Psi|\Psi^*)$ is loosely defined by

$$\delta(\Psi|\Psi^*) = \begin{cases} +\infty, & \text{if } \Psi = \Psi^* \\ 0, & \text{if } \Psi \neq \Psi^* \end{cases}$$

and it satisfies that

$$\int_{\mathrm{R}^{dp\times d}} \delta(\Psi|\Psi^*)\mathrm{d}\Psi = 1.$$

Then $\delta(\Psi|\Psi^*)$ represents the limit of a posterior density as $T \to \infty$. The Dirac delta function is rigorously defined as a probability measure. Let Δ_{Ψ^*} be the distribution on $\mathrm{R}^{dp\times d}$ such that for any subset $A \subset \mathrm{R}^{dp\times d}$, $\Delta_{\Psi^*}(A) = 1$ if $\Psi^* \in A$ and 0 otherwise. Then Δ_{Ψ^*} represents the limit of a posterior distribution as $T \to \infty$. The Lebesgue integral with respect to Δ_{Ψ^*} satisfies

$$\int_{\mathrm{R}^{dp\times d}} f(\Psi)\Delta_{\Psi^*}(\mathrm{d}\Psi) = f(\Psi^*)$$

for every compactly supported continuous function $f(\Psi)$. A common abuse of notation is to define the integral of $\delta(\Psi|\Psi^*)$ against a continuous function $f(\Psi)$ by

$$\int_{\mathrm{R}^{dp\times d}} f(\Psi)\delta(\Psi|\Psi^*)\mathrm{d}\Psi = f(\Psi^*).$$

We aim to find a value λ^* of λ for which the distribution $f(\Psi, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda^*, \mathbf{x}_{p+1})$ converges to the Dirac delta function $\delta(\Psi|\Psi^*)$ at a fast rate. To assess the quality of a value of λ, we consider the so-called cross entropy between $\delta(\Psi|\Psi^*)$ and $f(\Psi, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda, \mathbf{x}_{p+1})$. The cross entropy between two probability densities f and g is defined by

$$H(f, g) = -\int f \log g.$$

The KL divergence (24) is re-expressed as

$$\begin{aligned} D_{\mathrm{KL}}(f||g) &= \int f \log f - \int f \log g \\ &= const. + H(f, g). \end{aligned}$$

If f is a fixed reference distribution, minimization of the KL divergence between f and g is equivalent to minimization of the cross entropy.

By using the Jensen's inequality, we can calculate the upper bound of the expectation of the cross entropy between $\delta(\Psi|\Psi^*)$ and $f(\Psi, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda, \mathbf{x}_{p+1})$ as in

THEOREM 3. *Based on the density function (16) with* $(\Psi, \sigma^2, \mathbf{c}^{\mathrm{s}}) = (\Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*})$, *it follows that*

$$\begin{aligned} &\mathrm{E}\left[H(\delta(\Psi|\Psi^*), f(\Psi, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda, \mathbf{x}_{p+1}))\middle|\ \Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*}\right] \\ &\quad \leq -\log m_4 - \frac{d^2p}{2}\log\theta(\lambda) \\ &\qquad + \left(\alpha + \frac{d(T-p-1)+d^2p}{2}\right)\log\left(\beta + \frac{d(T-p-1)}{2}\sigma^{*2} + \frac{1}{2}\theta(\lambda)\,\|\Psi^*\|^2\right) \end{aligned} \quad (25)$$

where m_4 *is a constant that does not depend on* $(\Psi, \sigma^2, \lambda)$, *and* $\theta(\lambda) = \lambda(T-p-1)/(1-\lambda)$.

PROOF. See Appendix A.

Let $Q(\lambda; \Psi^*, \sigma^{*2})$ denote the right-hand side of the inequality in (25). The optimal value, λ^*, of λ is obtained by minimizing $Q(\lambda; \Psi^*, \sigma^{*2})$ with respect to λ. From $\mathrm{d}Q/\mathrm{d}\lambda = 0$, we have

$$\lambda^* = \frac{d^2 p}{G(\Psi^*, \sigma^{*2})(T-p-1) + d^2 p} \tag{26}$$

where

$$G(\Psi^*, \sigma^{*2}) = \frac{\|\Psi^*\|^2 (2\alpha + d(T-p-1))}{2\beta + d(T-p-1)\sigma^{*2}}.$$

Moreover, if $d(T-p-1)$ is relatively large, then α and β are ignorable and we get the approximation as

$$G(\Psi^*, \sigma^{*2}) \approx \frac{\|\Psi^*\|^2}{\sigma^{*2}}. \tag{27}$$

If the entries of Ψ^* are bounded by a constant M, i.e., $\left|\Psi^*_{ij}\right| \leq M$, then we have $\|\Psi^*\|^2 \leq M^2 d^2 p$, which implies $G(\Psi^*, \sigma^{*2}) = O(d^2)$.

The results (26) and (27) suggest that the optimal value of λ has a parametric form as follows.

COROLLARY 4. *The optimal value, λ^*, of λ has the following parametric form:*

$$\lambda^* = \frac{d^2 p}{\nu(T-p-1) + d^2 p} \tag{28}$$

where $\nu > 0$ is a constant depending only on the parameter (Ψ, V) in (15), and it satisfies $\nu = O(d^2)$.

In conclusion, results (23) and (26) with (27) are congruent with Corollary 4. The larger the number of observations gets, the smaller the optimal value of λ becomes. Also, the larger the dimensionality gets, the closer the optimal value of λ gets to 1. Moreover (26) suggests that if $\|\Psi^*\|$ is large, then it means that the variables in the data are highly correlated, which ends up with a small optimal value of λ. On the other hand, if σ^{*2} is large, it implies that the data are contaminated with a large amount of noise, which ends up with a large optimal value of λ. The interpretation of the optimal value of λ is at least in tune with our intuition and that described by Schäfer and Strimmer (2005b).

4.3. Parameterized cross validation

In this section we present a computational method in search of an optimal value λ^* of the shrinkage parameter λ. This method is referred to as the PCV method which was first proposed by Koo, Lee, and Kil (2008) for estimating the parameterized form of risk bounds for prediction models. We use the PCV method for estimating the parameterized form of the optimal shrinkage parameter.

The conventional cross-validation (CV) methods, including the k-fold CV method, have widely been used in regression and time series model selection (Seber and Lee, 2003; Tsay, 2005). When they are applied to model selection for time series data, the data pairs of inputs and outputs are considered together, e.g., $(\mathbf{x}^{\mathrm{s}}_t, \mathbf{y}^{\mathrm{s}}_t), t = p+1, \ldots, T$. A part of the data pairs, called the training set, is used to train a model, while the other part of the

data pairs, called the validation set, is used to evaluate the performance of the model. CV methods select the model which is optimal under given criteria by using the validation set. This is critical in case that there is a small number of observations. The PCV method selects the model which is optimal for the whole data (Koo, Lee, and Kil, 2008).

The PCV method is carried out as follows. In Corollary 4 we suggested the parameterized form of λ^* for VAR models. Especially, the term ν in (28) is a constant with respect to T. From (28) we have

$$\frac{\lambda^*}{1-\lambda^*} = \frac{d^2p}{\nu(T-p-1)}.$$

Let $\gamma = -\log\nu$ and $\phi(\lambda) = \log(\lambda/(1-\lambda))$. Then

$$\gamma = \phi(\lambda^*) + \log\left(\frac{T-p-1}{d^2p}\right).$$

First, we consider the k-fold CV for estimating γ. We randomly partition the given data pairs $(\mathbf{x}_t^{\mathrm{s}}, \mathbf{y}_t^{\mathrm{s}}), t = p+1, \ldots, T$, into k sets of nearly equal sizes. Suppose the ith set of pairs is selected as the validation data for evaluating a model, and the remaining $k-1$ sets are used as the training data. From the training data, we can get the sample mean vectors $\bar{\mathbf{x}}^{\mathrm{s}}(i)$ and $\bar{\mathbf{y}}^{\mathrm{s}}(i)$. Then, after forming the mean-corrected data matrices $X^{\mathrm{s}}(i)$ and $Y^{\mathrm{s}}(i)$, we can calculate the coefficient matrix $\widehat{\Psi}^*(i)$ as in (14) for $0 < \lambda \leq 1$. By using the validation data, $(\mathbf{x}_t^{\mathrm{s,val}}(i), \mathbf{y}_t^{\mathrm{s,val}}(i)), t = 1, \ldots, M_i$, we can find the optimal value of λ that minimizes the prediction error defined by

$$\sum_{t=1}^{M_i} \left\| \mathbf{y}_t^{\mathrm{s,val}}(i) - \bar{\mathbf{y}}^{\mathrm{s}}(i) - (\widehat{\Psi}^*(i))'(\mathbf{x}_t^{\mathrm{s,val}}(i) - \bar{\mathbf{x}}^{\mathrm{s}}(i)) \right\|^2.$$

Let $\hat{\lambda}^*(i)$ denote the optimal value of λ for the ith validation set. The ith estimate of γ is calculated by

$$\hat{\gamma}(i) = \phi(\hat{\lambda}^*(i)) + \log\left(\frac{N_i - 1}{d^2p}\right)$$

where N_i is the number of data pairs in the training data.

Next, we determine the optimal value $\hat{\lambda}^*_{\mathrm{EB}}$ based on the k-fold CV results. Let us define

$$\hat{\gamma}^* = \frac{1}{k}\sum_{i=1}^{k} \hat{\gamma}(i) \tag{29}$$

$$= \frac{1}{k}\sum_{i=1}^{k} \phi(\hat{\lambda}^*(i)) + \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{N_i - 1}{d^2p}\right). \tag{30}$$

Since $N_i, i = 1, \ldots, k$, are nearly equal among themselves, the second term of (30) may be regarded as a constant. The first term therein is the arithmetic mean of the logit. Finally, $\hat{\gamma}^*$ in (29) is used as an estimate of $-\log\nu$, and thus $\hat{\lambda}^*_{\mathrm{EB}}$ is obtained from (28) by

$$\hat{\lambda}^*_{\mathrm{EB}} = \frac{d^2p}{\exp\{-\hat{\gamma}^*\}(T-p-1) + d^2p}.$$

5. Experiments

We conducted a series of experiments with simulated data sets and real world data sets. We examined several methods of estimating VAR model coefficients such as ordinary least squares (OLS), ridge regression (Ridge), nonparametric shrinkage method (NS), and the empirical Bayesian shrinkage method (EB) which is proposed in this paper. We compared their performance in the context of parameter estimation and structure learning. We then applied the EB method to the real world data set and built a VAR model along with some interpretations.

5.1. Estimation based on simulated data sets

We generated multivariate time series data from VAR models of order $p \in \{1, 2, 3\}$ with dimension $d \in \{5, 10, 20, 40, 80, 160\}$ and number of observations $T \in \{10, 20, 40, 80, 160\}$. The covariance matrix for a noise vector was set to $V = I$. Among d^2p coefficients in the coefficient matrix Φ, only d coefficients were set to nonzero values and the other coefficients zero. The values of the nonzero coefficients were drawn uniformly from the set, $[-1, -0.2] \cup [0.2, 1]$. For notational convenience, we will denote by VAR(p, d) a VAR model of order p and dimension d.

We generated 30 data sets from each of the VAR models for one experiment. For each of the 30 data sets for given p, d, and T, we estimated the coefficient matrix Φ and calculated an estimate, $\widehat{MSE}$, of the MSE of $\widetilde{\Phi}$, $\mathrm{E}\left[\left\|\Phi - \widetilde{\Phi}\right\|^2\right]$, given by

$$\widehat{MSE} = \frac{1}{30}\sum_{l=1}^{30}\left\|\Phi - \widetilde{\Phi}(l)\right\|^2 = \frac{1}{30}\sum_{l=1}^{30}\sum_{i=1}^{dp}\sum_{j=1}^{d}\left(\Phi_{ij} - \widetilde{\Phi}(l)_{ij}\right)^2$$

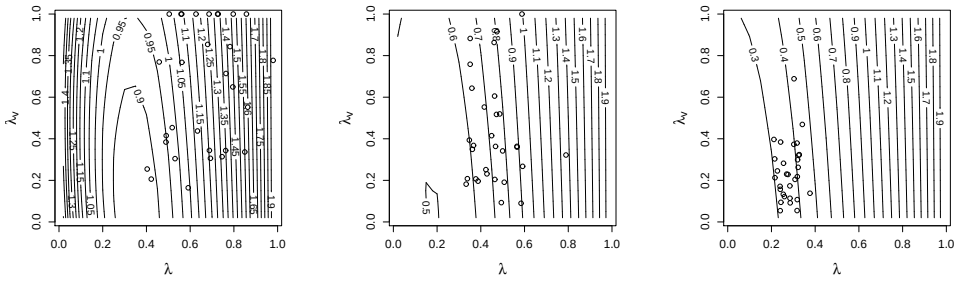
where Φ is the true coefficient matrix and $\widetilde{\Phi}(l)$ is the estimate of Φ based on the lth data set.

Figs. 1 and 2 display the contours of the $\widehat{MSE}$, and scatter plots of 30 $(\hat{\lambda}^*, \hat{\lambda}^*_{\mathrm{v}})$ pairs obtained by the NS method. Data were generated from a VAR$(1, 5)$ model for Fig. 1, and a VAR$(1, 40)$ model for Fig. 2. The number of observations is $T = 20$, 40, and 80 for panels (a), (b), and (c), respectively.

The contours of Figs. 1 and 2 show that each $\widehat{MSE}$ surface is convex with a unique minimum. The $\widehat{MSE}$ surfaces clearly indicate that smaller shrinkage parameters should be used for larger T, whereas larger shrinkage parameters should be used for larger d. Moreover, since the $\widehat{MSE}$ contours are elliptic and vertically prolate, we can see that the $\widehat{MSE}$'s are less sensitive to $\hat{\lambda}^*_{\mathrm{v}}$ than $\hat{\lambda}^*$.

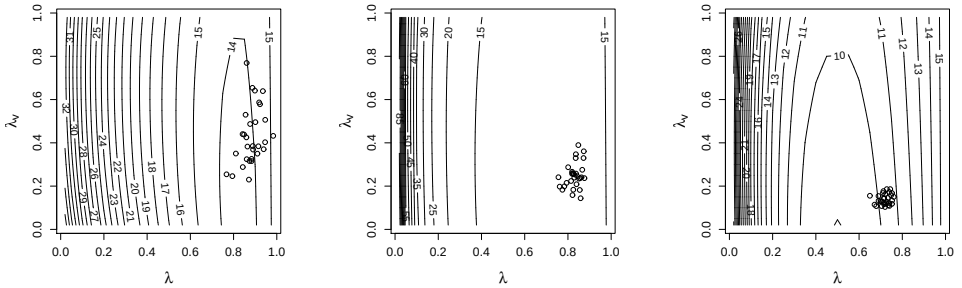
Figs. 3 and 4 are counterparts of Figs. 1 and 2, respectively, that are based on the estimate results from the EB method. We can see in the four figures that the 30 points of $(\hat{\lambda}^*, \hat{\lambda}^*_{\mathrm{v}})$ from the EB method appear closer to the minimum point of the $\widehat{MSE}$ surface than those from the NS method. This phenomenon is more apparent in Figs. 5 and 6 which show boxplots of the 30 estimates of λ by the NS and the EB method, respectively. We can also see in the four figures, Figs. 1, 2, 3, and 4, that the points, $(\hat{\lambda}^*, \hat{\lambda}^*_{\mathrm{v}})$, are scattered over a wider area as T gets smaller for fixed p and d. The same phenomenon is displayed by boxplots for the values, $\hat{\lambda}^*$, in Figs. 5 and 6.

We compared $\widehat{MSE}$ values between the four methods, the OLS, Ridge, NS, and EB methods. The comparison is displayed in Fig. 7 for VAR$(1, d), d = 5, 10, 20, 40$. We can see



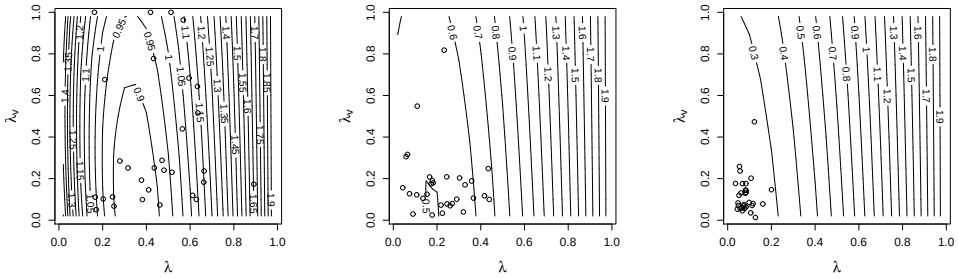
(a) $p = 1, d = 5, T = 20$ (b) $p = 1, d = 5, T = 40$ (c) $p = 1, d = 5, T = 80$

Fig. 1. Contours of $\widehat{MSE}$'s and scatter plots of $(\hat{\lambda}^*, \hat{\lambda}^*_{\mathrm{v}})$ pairs obtained by the NS method. Data were generated from a VAR$(1, 5)$ model, and the numbers of observations are $T = 20$, 40, and 80 for (a), (b), and (c), respectively.



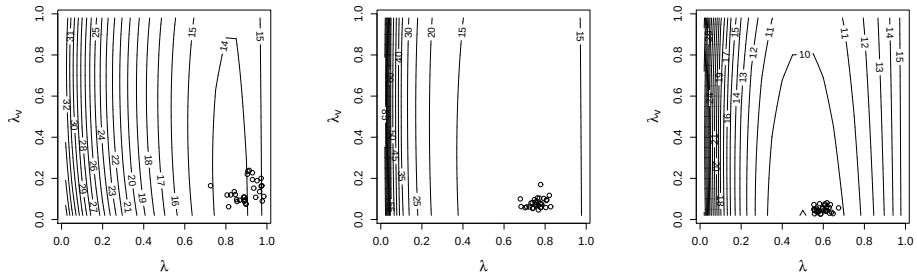
(a) $p = 1, d = 40, T = 20$ (b) $p = 1, d = 40, T = 40$ (c) $p = 1, d = 40, T = 80$

Fig. 2. Contours of $\widehat{MSE}$'s and scatter plots of $(\hat{\lambda}^*, \hat{\lambda}^*_{\mathrm{v}})$ pairs obtained by the NS method. Data were generated from a VAR$(1, 40)$ model, and the numbers of observations are $T = 20$, 40, and 80 for (a), (b), and (c), respectively.



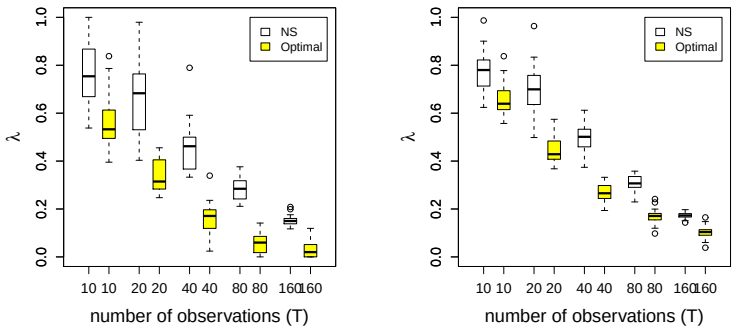
(a) $p = 1, d = 5, T = 20$ (b) $p = 1, d = 5, T = 40$ (c) $p = 1, d = 5, T = 80$

Fig. 3. Contours of $\widehat{MSE}$'s and scatter plots of $(\hat{\lambda}^*, \hat{\lambda}^*_{\mathrm{v}})$ pairs obtained by the EB method. Data were generated from a VAR$(1, 5)$ model, and the numbers of observations are $T = 20$, 40, and 80 for (a), (b), and (c), respectively.

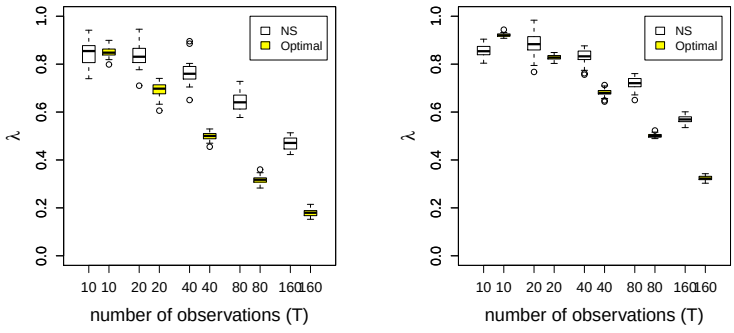


(a) $p = 1, d = 40, T = 20$ (b) $p = 1, d = 40, T = 40$ (c) $p = 1, d = 40, T = 80$

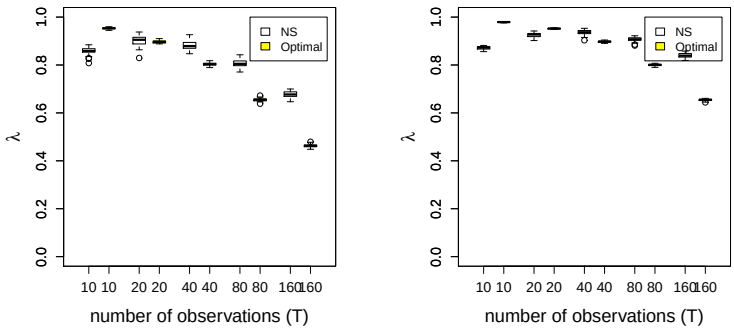
Fig. 4. Contours of $\widehat{MSE}$'s and scatter plots of $(\hat{\lambda}^*, \hat{\lambda}^*_{\mathrm{v}})$ pairs obtained by the EB method. Data were generated from a VAR$(1, 40)$ model, and the numbers of observations are $T = 20$, 40, and 80 for (a), (b), and (c), respectively.



(a) $p = 1, d = 5$ (b) $p = 1, d = 10$

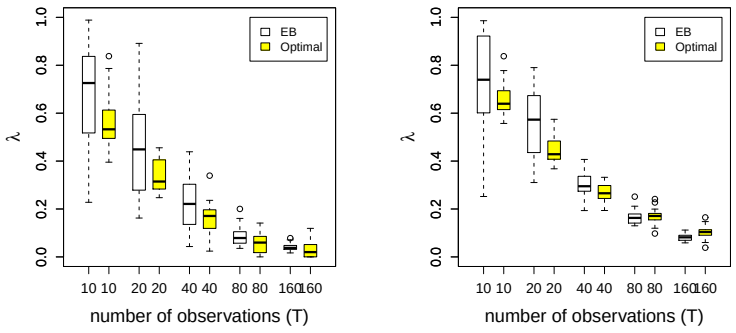


(c) $p = 1, d = 20$ (d) $p = 1, d = 40$

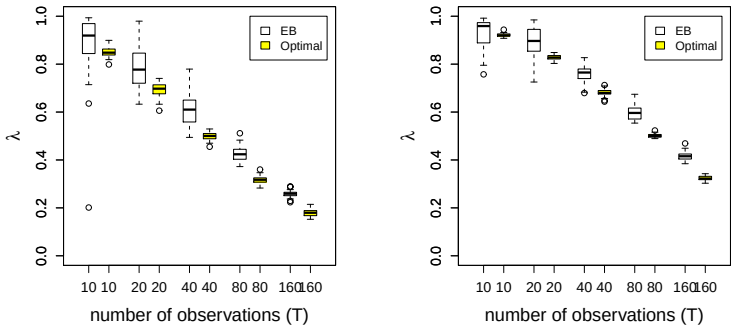


(e) $p = 1, d = 80$ (f) $p = 1, d = 160$

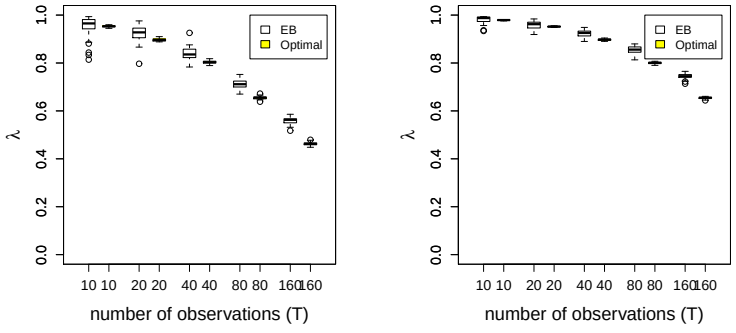
Fig. 5. Box plots of the $\hat{\lambda}^*$ values obtained by the NS method, and box plots of the optimal λ values at which minimum $\widehat{MSE}$ is achieved. Data were generated from VAR$(1, d)$ models with $d = 5$, 10, 20, 40, 80, and 160 for (a), (b), (c), (d), (e), and (f), respectively.



(a) $p = 1, d = 5$ (b) $p = 1, d = 10$

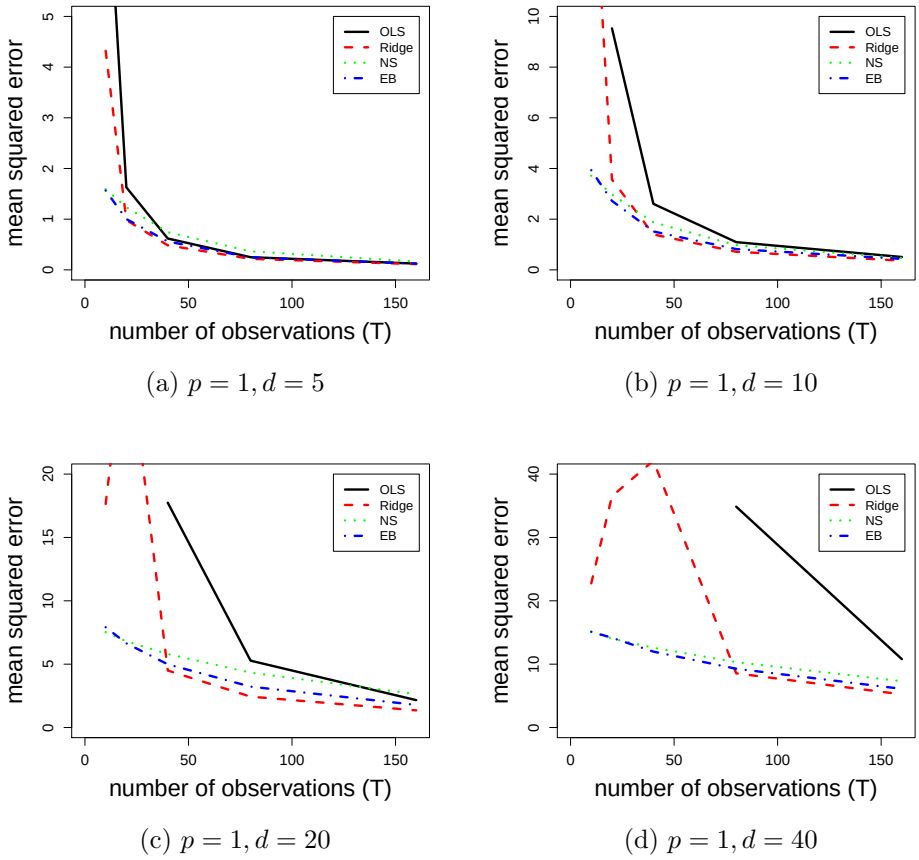


(c) $p = 1, d = 20$ (d) $p = 1, d = 40$



(e) $p = 1, d = 80$ (f) $p = 1, d = 160$

Fig. 6. Box plots of the $\hat{\lambda}^*$ values obtained by the EB method, and box plots of the optimal λ values at which minimum $\widehat{MSE}$ is achieved. Data were generated from VAR$(1, d)$ models with $d = 5$, 10, 20, 40, 80, and 160 for (a), (b), (c), (d), (e), and (f), respectively.



(a) $p = 1, d = 5$ (b) $p = 1, d = 10$

(c) $p = 1, d = 20$ (d) $p = 1, d = 40$

Fig. 7. The $\widehat{MSE}$'s by the four methods, the OLS, Ridge, NS, and EB methods. Data were generated from VAR$(1, d)$ models with $d = 5$, 10, 20, and 40 for (a), (b), (c), and (d), respectively.

in the figure that the Ridge method produces $\widetilde{\Phi}$'s that are unstable when T is relatively small and that the NS method yields larger $\widehat{MSE}$ values than the EB method. We can see analogous results for VAR$(2, d)$ and VAR$(3, d)$ which are displayed in Figs. 17 and 18 in Appendix B.

The comparison of the $\widehat{MSE}$ values between the NS and the EB method is summarized in Fig. 8 via boxplots of the 30 log ratios of $\widehat{MSE}$ values from as many simulated data sets. The log ratio from the lth simulated data set is obtained by

$$logratio(l) = \log\left(\frac{\widehat{MSE}_{\mathrm{NS}}(l)}{\widehat{MSE}_{\mathrm{EB}}(l)}\right),$$

where $\widehat{MSE}_{\mathrm{NS}}$ denotes the $\widehat{MSE}$ by the NS method and analogously for $\widehat{MSE}_{\mathrm{EB}}$. The figure shows a larger $\widehat{MSE}$ values by the NS method over a wider range than the EB method. The discrepancy between the two methods diminishes as d increases, which one

can anticipate from the results displayed in Figs. 5 and 6.

5.2. Structure inference based on simulation data from a VAR(1,40)

A VAR model is a causal model whose model structure is determined by its nonzero coefficients. As indicated in (1), the causal relationship between any two variables is defined as follows: the jth variable $y_{\cdot j}$ does not Granger-cause the ith variable $y_{\cdot i}$ if the coefficient matrices satisfy $(A_1)_{ij} = \cdots = (A_p)_{ij} = 0$ (Granger, 1969; Sims, 1972; Hamilton, 1994). In a graphical representation of a VAR model structure, each node corresponds to a variable and each directed edge corresponds to the Granger-causality between the connected variables, the arrow heading from a causal node to its effect node.

A typical way of checking Granger-causality is to conduct an F test of the null hypothesis $H_0 : (A_1)_{ij} = \cdots = (A_p)_{ij} = 0$ after estimating coefficient matrices by the OLS (Seber and Lee, 2003). However the OLS is not appropriate for data with T not large enough for d and p, and the test should be conducted repeatedly for each i and j, which accompanies another computational burden.

Therefore it is preferred to run a statistical test using partial correlations instead of coefficient matrices in case of data sets with small T and large d (Schäfer and Strimmer, 2005a; Opgen-Rhein and Strimmer, 2007c). A partial correlation $\mathrm{corr}(y_{\cdot i}, x_{\cdot j}|x_{\cdot \mathrm{rest}})$ between two variables $y_{\cdot i}$ and $x_{\cdot j}$ represents the correlation between the two variables conditioned on the rest of the predictor variables. It is shown in Whittaker (1990) that in the multivariate normal linear regression model, $\Phi_{ji} = 0$ if and only if $\mathrm{corr}(y_{\cdot i}, x_{\cdot j}|x_{\cdot \mathrm{rest}}) = 0$.

Specifically, the partial correlations are directly related to the VAR coefficients (Whittaker, 1990) as follows. First of all, the VAR model is described by

$$\begin{aligned} y_{\cdot i} &= \Phi_{1i}x_{\cdot 1} + \cdots + \Phi_{dp,i}x_{\cdot dp} + c_i + \varepsilon_{\cdot i} \\ &= \Phi_{ji}x_{\cdot j} + \sum_{k=1,k\neq j}^{dp} \Phi_{ki}x_{\cdot k} + c_i + \varepsilon_{\cdot i} \end{aligned}$$

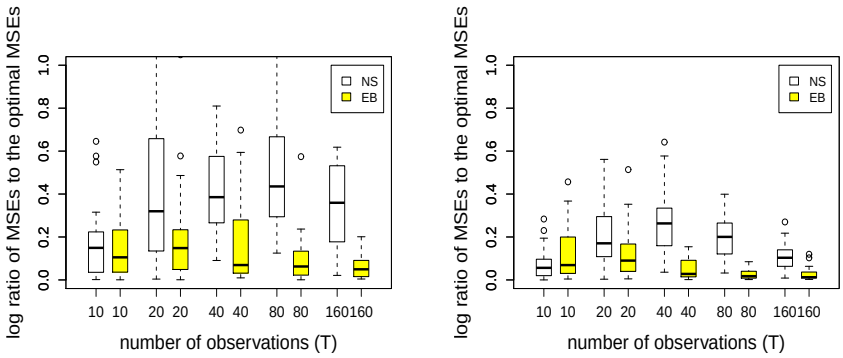
for $i = 1, \ldots, d$. Next, for each (i, j), by switching the role of $y_{\cdot i}$ and $x_{\cdot j}$, that is, letting $x_{\cdot j}$ be the response variable, the model is expressed as

$$x_{\cdot j} = \Theta_{ji}y_{\cdot i} + \sum_{k=1,k\neq j}^{dp} \Theta_{ki}x_{\cdot k} + v_j + \eta_{\cdot j}$$

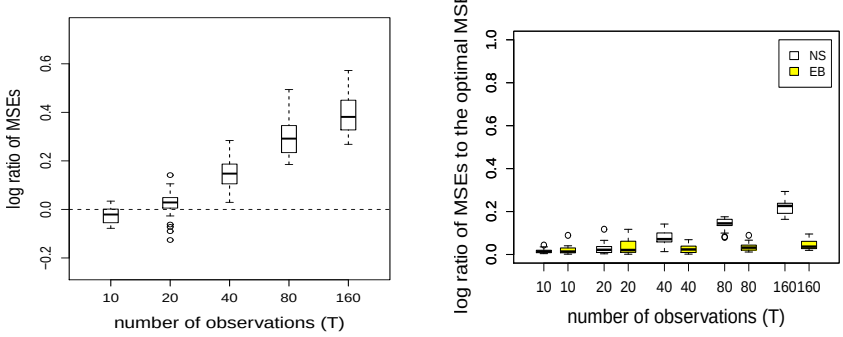
where Θ is a $(dp \times d)$ coefficient matrix, v_j is a scalar, and $\eta_{\cdot j}$ is a noise random variable with mean zero. Finally, the estimated partial correlation is obtained from the estimated coefficient matrices $\widetilde{\Phi}$ and $\widetilde{\Theta}$ by

$$\widehat{\mathrm{corr}}(y_{\cdot i}, x_{\cdot j}|x_{\cdot rest}) = \mathrm{sign}\left(\widetilde{\Theta}_{ji}\right)\sqrt{\widetilde{\Phi}_{ji}\widetilde{\Theta}_{ji}}\ .$$

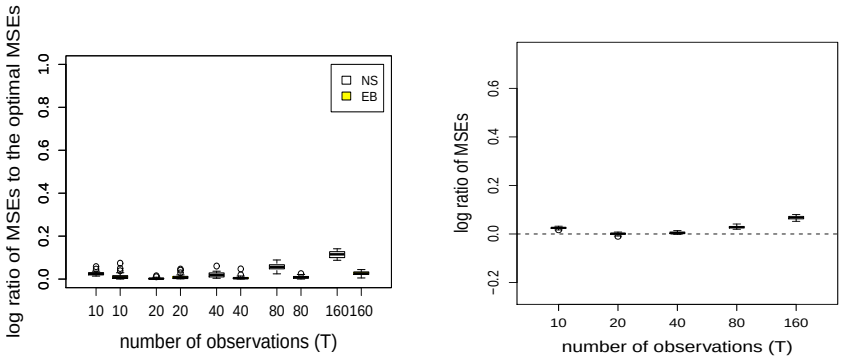
For simulation experiment, we generated multivariate time series data from a VAR$(1, 40)$ model with numbers of observations $T \in \{10, 20, 40, 80, 160\}$. The covariance matrix for a noise vector was set as $V = I$. The number of nonzero coefficients in the coefficient matrix Φ was set at 450. The values of the nonzero coefficients were drawn uniformly from the set, $[-0.5, -0.1] \cup [0.1, 0.5]$. Fig. 9 displays the subgraph with the 30 strongest-intensity



(a) $p = 1, d = 5$ (b) $p = 1, d = 10$



(c) $p = 1, d = 20$ (d) $p = 1, d = 40$



(e) $p = 1, d = 80$ (f) $p = 1, d = 160$

Fig. 8. Box plots of the log ratios of the $\widehat{MSE}$'s to compare the NS method and the EB method. Data were generated from VAR$(1, d)$ models with $d = 5$, 10, 20, 40, 80, and 160 for (a), (b), (c), (d), (e), and (f), respectively.

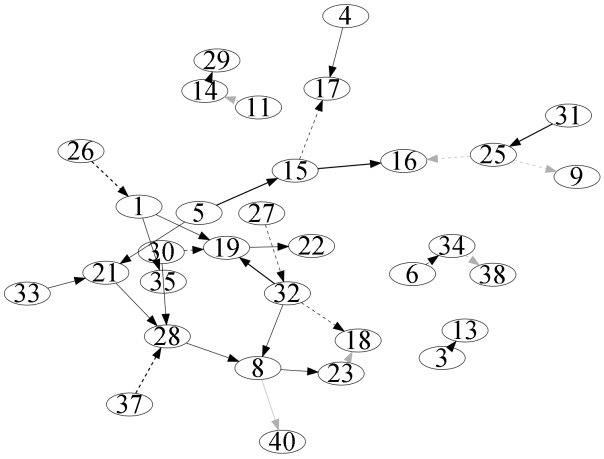


Fig. 9. The subgraph with the 30 strongest-intensity edges of an actual VAR(1,40) model which is used for a simulation experiment, where solid and dotted lines indicate positive and negative coefficients, respectively.

edges of the VAR(1,40) network, where solid and dotted lines indicate positive and negative coefficients, respectively.

We generated 30 data sets from each of the VAR models for one experiment. For each of the 30 data sets for given p, d, and T, we applied the NS and the EB methods to obtain $(\hat{\lambda}^*, \hat{\lambda}^*_{\mathrm{v}})$. After estimating the VAR model coefficients, we calculated the partial correlations. Since there are $d^2p = 1600$ coefficients in Φ, we obtained $d^2p = 1600$ estimated partial correlations for each of the NS and the EB methods.

The estimated partial correlations represent the significance of the causal relationship between each pair of variables, based on which a statistical test is conducted in search of nonzero entries of the coefficient matrix. We can compare the NS and the EB methods by evaluating the estimated partial correlations through the receiver operating characteristic (ROC) curves. Partial correlations were estimated from each of the 30 data sets with $p = 1$, $d = 40$, and $T = 40$ by each of the EB and the NS method, and the ROC curves were created as in Fig. 10. We can see in the figure that the EB method performs far better than the NS method.

Fig. 11 (a) displays the precision score by each of the two estimation methods, which is the proportion of the correctly selected edges out of the 450 edges whose true coefficients are nonzero. These correctly selected edges are also called true positives. It is apparent in the figure that the EB method performs better than the NS method.

Fig. 11 (b) and (c) display the difference in the average number of true positives in more detail. In Fig. 11 (b), the difference increases up to around 50 edges. This number is more than 10% of the total number of actual nonzero coefficients. To see the significance of the difference, we conducted paired t-tests for testing $H_0 : p^{\mathrm{EB}} = p^{\mathrm{NS}}$ against $H_1 : p^{\mathrm{EB}} > p^{\mathrm{NS}}$ where p^{EB} and p^{NS} are the true positive rates of the EB and the NS methods, respectively. The p-values of the test are shown by the green dotted line in Fig. 11 (b) which are obtained based on 30 data sets generated for each VAR model.

Fig. 11 (c) shows the difference in the true positives in more detail. We classified the model coefficients according to their absolute values into one of the four intervals: $[0.1, 0.2)$,

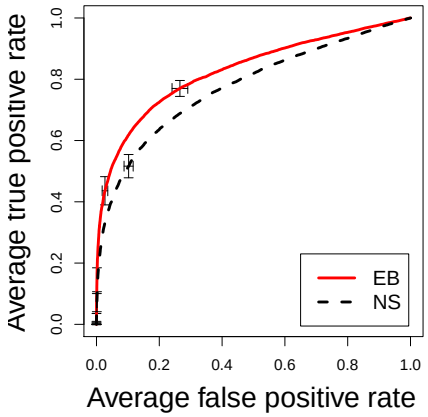


Fig. 10. ROC curves by the NS and the EB method using simulated data with $T = 40$ from a VAR(1,40) model.

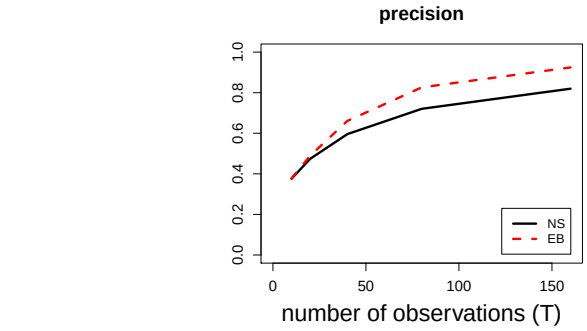
$[0.2, 0.3)$, $[0.3, 0.4)$, and $[0.4, 0.5]$. We investigated the difference in the true positives for each interval. We can see in the figure that the difference in the true positive rate gets relatively higher for the interval, $[0.1, 0.2)$, among the four intervals as T gets larger. This may be interpreted as a higher efficiency of edge detection for the EB method in comparison with the NS method.

5.3. Structure inference based on simulation data from a VAR(1,400)

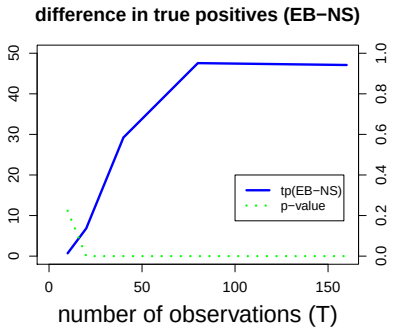
In the preceding subsection, we compared the performance of the EB method with others, in particular with the NS method, for relatively small VAR models where the number of nonzero Φ_{ij}'s is equal to the dimensionality, d, of the model. In this subsection, we will compare the NS and EB methods in the context of structure learning using a larger model, VAR$(1, 400)$ with lots of nonzero Φ_{ij}'s.

A VAR model is a causal model whose model structure is determined by its nonzero coefficients. As indicated in (1), the causal relationship between any two variables is defined as follows: the jth variable $y_{\cdot j}$ does not Granger-cause the ith variable $y_{\cdot i}$ if the coefficient matrices satisfy $(A_1)_{ij} = \cdots = (A_p)_{ij} = 0$ (Granger, 1969). In a graphical representation of a VAR model structure, each node corresponds to a variable and each directed edge corresponds to the Granger-causality between the connected variables, the arrow heading from a causal node to its effect node.

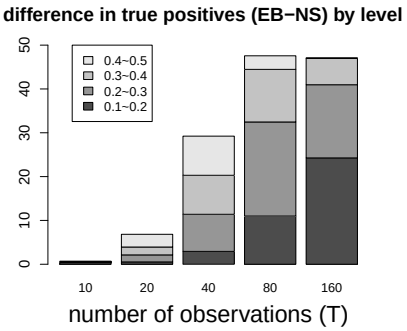
For checking the Granger-causality, it is preferred to run a statistical test using partial correlations instead of conducting F tests which are based on the OLS estimates (Seber and Lee, 2003) in case of data sets with small T and large d (Schäfer and Strimmer, 2005a). Let $\mathrm{corr}(y_{\cdot i}, x_{\cdot j}|x_{\cdot \mathrm{rest}})$ denote the partial correlation between two variables $y_{\cdot i}$ and $x_{\cdot j}$. It is shown in Whittaker (1990) that in the multivariate normal linear regression model, $\Phi_{ji} = 0$ if and only if $\mathrm{corr}(y_{\cdot i}, x_{\cdot j}|x_{\cdot \mathrm{rest}}) = 0$. For the calculation of the partial correlations from shrinkage estimates, see Schäfer and Strimmer (2005a).



(a)



(b)



(c)

Fig. 11. Results by the NS and the EB methods in the selection of the edges of VAR networks. Data were generated from a VAR$(1, 40)$ model with 450 nonzero coefficients.

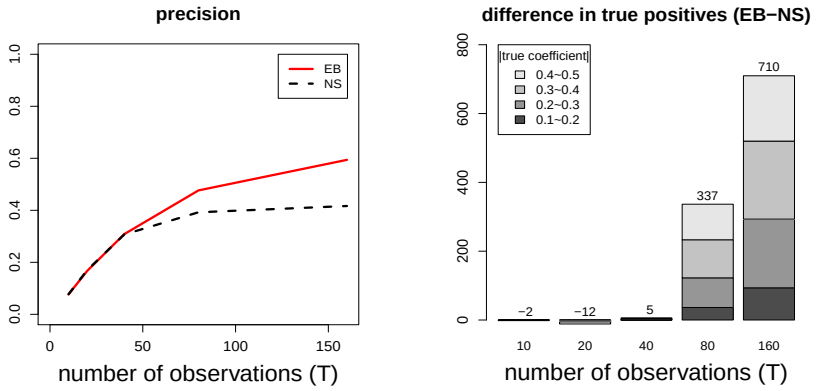


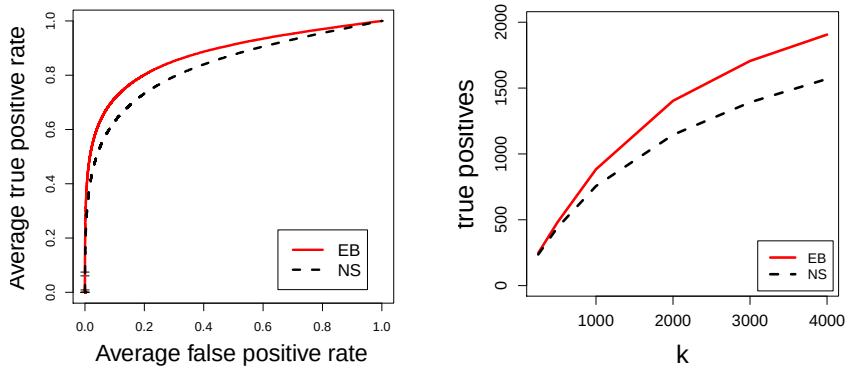
Fig. 12. Results by the EB and NS methods in the selection of the edges of VAR networks. Data were generated from a VAR$(1, 400)$ model.

For simulation experiments, we generated multivariate time series data from a VAR$(1, 400)$ model with $T \in \{10, 20, 40, 80, 160\}$. The covariance matrix for a noise vector was set to $V = 0.01^2 I$. The number of nonzero Φ_{ij} was set at 4000, whose values were drawn uniformly from the set, $[-0.5, -0.1] \cup [0.1, 0.5]$. For each of the 30 data sets generated for given p, d and T, we applied the NS and EB methods to obtain $(\hat{\lambda}^*, \hat{\lambda}^*_{\mathrm{v}})$ and $d^2p = 160000$ estimated partial correlations.

Fig. 12 (a) displays the precision score by each of the two estimation methods, which is the proportion of the correct edges out of the 4000 edges which are selected in accordance with the absolute values of the estimated partial correlations. The correctly selected edges are called the true positives. It is apparent in the figure that the EB method performs far better than the NS method when $T \geq 80$ and more or less at the same level when $T \leq 40$.

Fig. 12 (b) displays the difference in the average number of true positives between the EB and NS methods in more detail, where the difference increases up to 710 edges when $T = 160$ which is about 18% out of 4000 edges. Moreover, we classified the model coefficients according to their absolute values into one of the four intervals: $[0.1, 0.2)$, $[0.2, 0.3)$, $[0.3, 0.4)$, and $[0.4, 0.5]$, and we investigated the difference in true positives for each interval. In the figure the difference for the interval $[0.1, 0.2)$ grows relatively faster among the four intervals as T gets larger. For instance, the difference for the interval $[0.1, 0.2)$ increased from 36 at $T = 80$ to 94 at $T = 160$ while the difference for the interval $[0.4, 0.5)$ increased from 103 at $T = 80$ to 190 at $T = 160$. This can be interpreted as a higher efficiency of edge detection for the EB method in comparison with the NS method.

The NS and EB methods were compared through the receiver operating characteristic (ROC) curves as in Fig. 13 (a) and through the partial sum of true positives as in Fig. 13 (b) based on the estimated partial correlations for data with $T = 80$. We denote by seq_{NS} the sequence of the edges which are arranged in the order of the absolute values of the estimated partial correlations from large to small that are obtained by the NS method, and similarly for seq_{EB}. The partial sum of true positives are the number of the correct edges out of the first k edges in seq_{NS} (or seq_{EB}). We will denote the partial sum for the first k edges in seq_{NS} by $\tau_{NS}(k)$ and similarly for $\tau_{EB}(k)$.



(a) (b)

Fig. 13. (a) The ROC curves based on the estimated partial correlations and (b) partial sums of the true positives out of the first k edges in seq_{NS} and in seq_{EB}. Data were generated from a VAR$(1, 400)$ model with $T = 80$.

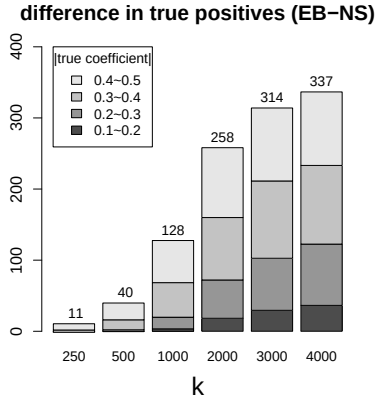


Fig. 14. $\tau_{EB}(k) - \tau_{NS}(k)$ for $k = 250, 500, 1000(1000)4000$. Data were generated from a VAR$(1, 400)$ model with $T = 80$.

Table 1. The numeric version of Fig. 14. The values are the mean and the standard deviation of $\tau_{EB}(k) - \tau_{NS}(k)$ based on the 30 iterations of experiment.

Intervals of true coefficient	$\tau_{EB}(k) - \tau_{NS}(k)$											
	k =250		500		1000		2000		3000		4000	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$[0.1, 0.2)$	-0.7	1.8	-0.4	2.6	3.4	5.1	18.3	6.3	29.5	9.0	36.4	11.3
$[0.2, 0.3)$	-0.8	3.8	2.5	4.3	16.3	7.2	53.9	12.4	73.2	17.4	86.2	17.9
$[0.3, 0.4)$	3.3	4.9	14.0	6.8	48.6	11.1	87.8	15.2	108.7	19.7	110.6	19.2
$[0.4, 0.5]$	8.8	6.1	23.6	8.5	59.2	13.2	98.2	19.7	102.6	17.2	103.4	17.4
$[0.1, 0.5]$	10.6	6.9	39.7	12.1	127.6	20.3	258.1	40.5	314.0	53.5	336.6	54.1

The graph in Fig. 13 (b) is zoomed in Figure 14 whose numeric version is Table 1. Note in the graph that $(\tau_{EB}(k) - \tau_{NS}(k))/k = 0.04, 0.08, 0.13, 0.13, 0.10, 0.08$ for $k = 250, 500, 1000(1000)4000$, respectively. This result is noteworthy. For instance, that $(\tau_{EB}(k) - \tau_{NS}(k))/k = 0.13$ when $k = 1000$ means that 13% or 128 more actual edges are found among the first 1000 edges in seq_{EB} than in seq_{NS}. In other words, the NS method chose wrongly 128 more edges than the EB method, whose actual coefficients are zero, in the first 1000 of seq_{NS}. What is more interesting is that among the 128 false positive edges by the NS method, 108 edges are as to the actual edges whose absolute values of coefficient are not smaller than 0.3. We see more differences in the true positives between the two methods where the strengths of the actual edges are relatively higher. This phenomenon is more or less the same for the other k values, as we can check in both Figure 14 and Table 1. This result is obtained when $T = 80$. An analogy of this result is also expected when $T = 160$ as is indicated in Figure 12 (b). Note that, when $T = 160$, $(\tau_{EB}(4000) - \tau_{NS}(4000))/4000 = 710/4000 = 0.18$ which is considerably large in favor of the EB method.

We had similar results as above for data from VAR(1,400) models with other large numbers of non-zero Φ_{ij}'s. For instance, when the number of non-zero Φ_{ij}'s is 2000 with $T = 80$, $\tau_{EB}(2000) - \tau_{NS}(2000) = 221$; when the number of non-zero Φ_{ij}'s is 6000 with $T = 80$, $\tau_{EB}(k) - \tau_{NS}(k) = 252, 1013$, respectively, for $k = 2000, 6000$. As far as structure learning is concerned, the simulation experiment strongly indicates that the EB method performs at least as good as the NS method.

5.4. Inference for VAR model based on real world data

The real world data we used for the experiment are obtained from an examination study on the synthesis and functions of enzymes of starch metabolism in leaves of *Arabidopsis thaliana* (Smith et al., 2004). Microarray analysis of diurnal changes in the starch transcriptome was carried out in the examination study. Opgen-Rhein and Strimmer (2007c) have downloaded original data of 22,814 probes and 11 time points for each of the two biological replicates from experiment no. 60 of the NASCArrays repository (Craigon et al., 2004) and preprocessed it to obtain the data with a subset of 800 genes. The data set descriptions are available in the package *GeneNet* in R programming language.

We estimated the VAR coefficients and the corresponding partial correlations using both of the EB and the NS methods. The estimates of $(\lambda, \lambda_{\mathrm{v}})$ are $(\hat{\lambda}^*_{\mathrm{EB}}, \hat{\lambda}^*_{\mathrm{EB,v}}) = (0.866, 0.013)$ and $(\hat{\lambda}^*_{\mathrm{NS}}, \hat{\lambda}^*_{\mathrm{NS,v}}) = (0.141, 0.035)$ by the EB and the NS methods, respectively. Note that $\hat{\lambda}^*_{\mathrm{EB}} \gg \hat{\lambda}^*_{\mathrm{NS}}$ for the data with $T = 22$ and $d = 800$.

The local FDR algorithm described by Strimmer (2008) was applied to the estimated partial correlations to select VAR networks. A VAR network with 2266 edges connecting 510 nodes was recommended by the EB method. Fig. 15 depicts the subgraph with 150 edges connecting 59 nodes. Opgen-Rhein and Strimmer (2007c) applied the NS method and recommended a VAR network with 7381 edges connecting 707 nodes using the local FDR algorithm. We notice that the EB method ended up with a VAR model of a smaller size. This result seems mostly due to $\hat{\lambda}^*_{\mathrm{EB}}$ and $\hat{\lambda}^*_{\mathrm{NS}}$, i.e., a larger shrinkage on the cross-autocorrelations results in a sparser model structure.

The difference in $\hat{\lambda}^*$ between the two methods is reflected in the difference in the empirical distribution of the partial correlations between the two methods as displayed in Fig. 16. We can see in the figure that more estimates of the partial correlations are clustered near

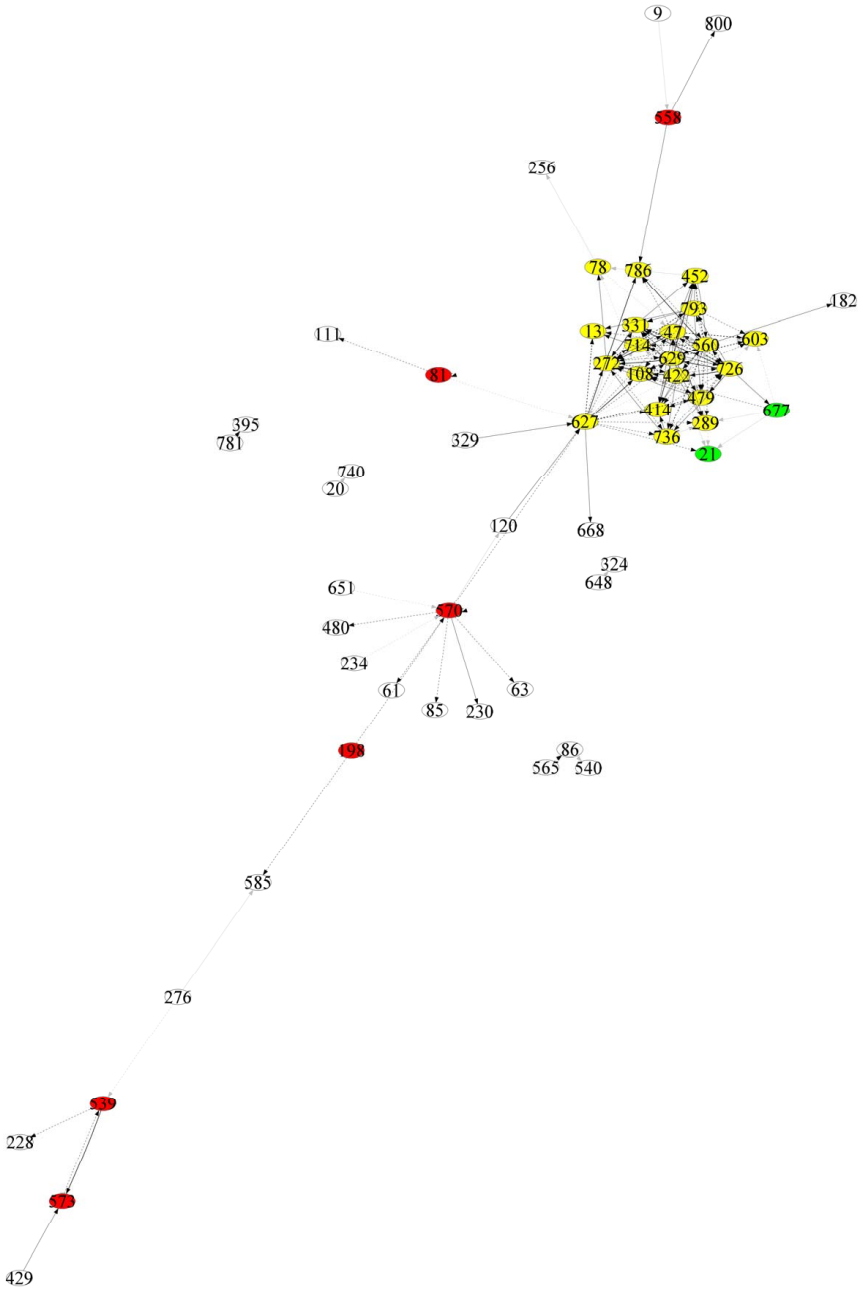


Fig. 15. Subgraph with 150 edges obtained by the EB method. The solid and dotted lines indicate positive and negative partial correlation coefficients, respectively, and the line intensity denotes their strength. Two green nodes are newly added to the "yellow" web nodes which are found in Opgen-Rhein and Strimmer (2007c).

zero by the EB method than by the NS method. The empirical distribution is approximated by a mixture distribution

$$f(r) = \eta_0 f_0(r;\kappa) + (1-\eta_0) f_{\mathrm{A}}(r), \quad 0 < \eta_0 < 1,$$

where f_0 is called a null distribution and f_{A} the alternative distribution. The distribution of a sample (partial) correlation coefficient, say r, is derived under the normality assumption in Hotelling (1953). When $\mathrm{corr}(y_{\cdot i}, x_{\cdot j}|x_{\cdot \mathrm{rest}}) = 0$, the distribution of r is given by

$$f_0(r;\kappa) = (1-r^2)^{(\kappa-r)/2} \frac{\Gamma(\kappa/2)}{\pi^{1/2}\Gamma((\kappa-1)/2)}, \quad -1 \le r \le 1,$$

where κ is the degree of freedom. The variance of the sample partial correlation coefficient equals the inverse of κ, i.e. $\mathrm{Var}(r) = 1/\kappa$. The local FDR algorithm estimates κ and η_0 from the estimated partial correlations. We can see in Fig. 16 that the EB method has a larger κ value which implies a smaller variance of the null distribution, and a larger η_0 value which implies a smaller fraction, $1-\eta_0$, for the nonzero partial correlations. Especially, η_0 directly affects the number of edges of a VAR network through the following local area-based FDR score (Strimmer, 2008)

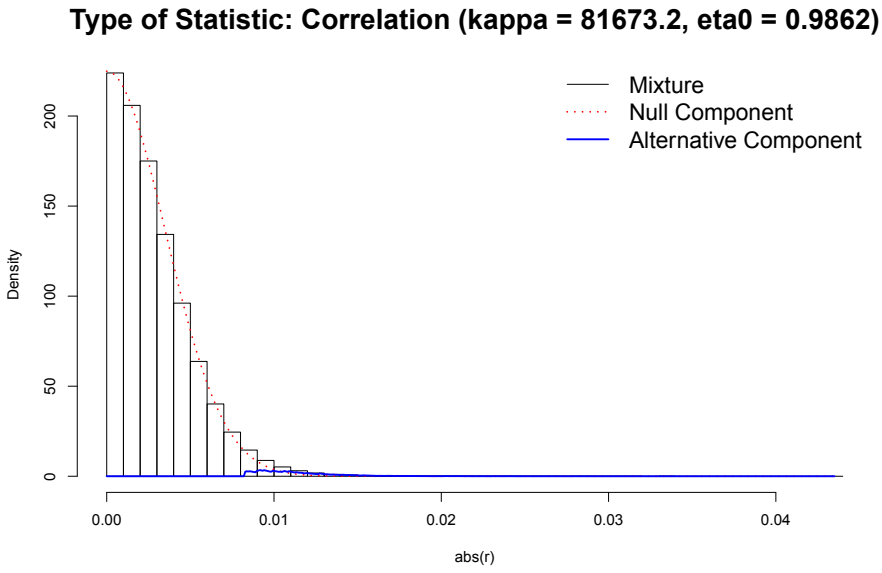
$$\mathrm{fdr}(r) = \Pr(\text{zero partial correlation}|r) = \frac{\eta_0 f_0(r;\kappa)}{f(r)}.$$

An edge in a VAR network is selected if $\mathrm{fdr}(r) \le 0.2$. Hence a larger η_0 value results in a sparser model structure.

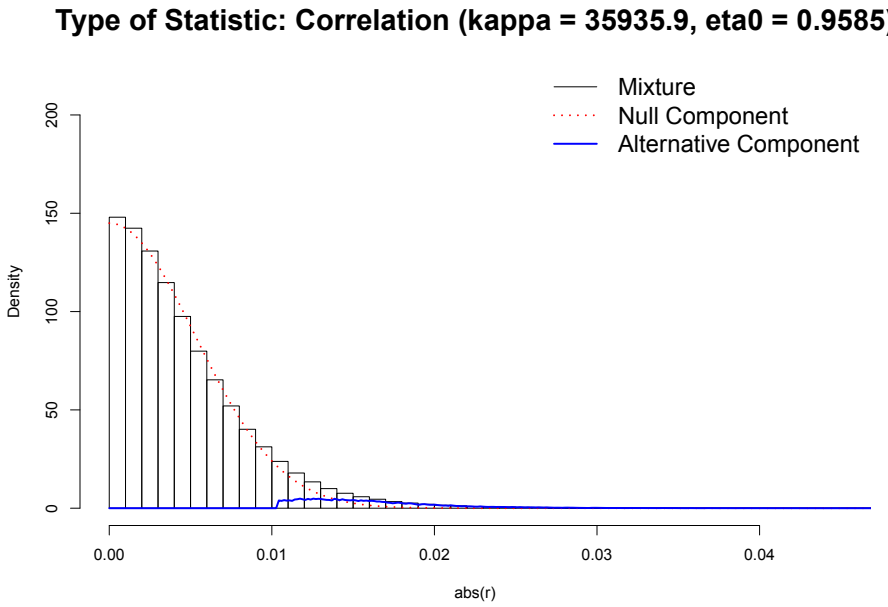
In Opgen-Rhein and Strimmer (2007c), the subgraph of 150 edges of the VAR network obtained by the NS method (call it NS-VAR network) has hub nodes colored in red and a web of highly connected genes colored in yellow. In the VAR network of Fig. 15 (call it EB-VAR network), the nodes are colored in the same manner as in Opgen-Rhein and Strimmer (2007c), and we found nodes 21 and 677 added to the "yellow" web nodes of Opgen-Rhein and Strimmer (2007c). These nodes are colored green in Fig. 15. When comparing the two VAR networks, it is worthwhile to note that the "yellow" nodes in the web are connected by far more edges in the EB-VAR network than the NS-VAR network, which is concomitant with fewer edges outside the densely intra-connected web in the EB-VAR network. As a matter of fact we could check in Figs. 19 and 20 in Appendix B that, out of 100 strongest edges, 84 edges were found in the "yellow" web in the EB-VAR network while they are 46 edges in the NS-VAR network. Besides that, all the 20 "yellow" web nodes appeared in the EB-VAR network plus the 2 "green" nodes while they are just 18 in the other network.

We present the EB-VAR network of 300 edges in Fig. 21 in Appendix B in order to help readers have a better understanding of the model structure. Since the edges are selected in the order of the edge intensity, we can see that the interrelationships among the nodes in the "yellow" web plus the two green nodes are relatively higher in the full model. The 150-edge and the 300-edge EB-VAR networks look quite different, which is just a matter of order of node-appearance in the network according to the edge intensity.

Since the VAR networks have directed edges, all genes are classified either as genes having mostly outgoing edges or as genes having mostly incoming edges. Genes having mostly outgoing edges are likely to be key regulatory genes. Among the hub nodes in Fig. 21 in Appendix B, the following nodes have mostly outgoing edges: node 570, an AP2 transcription factor, node 81, a DNA-directed RNA polymerase, node 539, a transcription



(a)



(b)

Fig. 16. Empirical distributions of the partial correlations obtained (a) by the EB method and (b) by the NS method, respectively, for the data considered in Section 5.4. The dotted line depicts the null distribution and the solid line the alternative distribution. The empirical distribution depicted by the histogram is approximated by a mixture of the null and alternative distributions. κ and η_0 are explained in the text.

factor, node 328, an ATPase, and node 783, a RNA methyltransferase. On the other hand, the following nodes seem to have mostly incoming edges: node 573, an unknown protein, and node 679, an unknown protein.

Although the EB-VAR and the NS-VAR networks share structural similarity very much, there are some noteworthy differences as well. Some nodes have many connections in the VAR network of the NS method, but not in the network of the EB method, and vice versa. Such differences will be examined in future works in cooperation with relevant biologists.

6. Conclusion

A main idea in the proposed method is that we search for an optimal value, $\hat{\lambda}^*$, of the shrinkage parameter under a Bayesian framework where the shrinkage parameter is set as a hyper-parameter for the priors on the parameters of the VAR model. The value $\hat{\lambda}^*$ is then selected so that the prediction error is minimized. The selection is made through the parameterized cross-validation. We assume that the shrinkage parameter is bounded away from zero in forming priors on the VAR coefficients but it does no harm to the estimation since the shrinkage parameter moves along with the sample size as compromising counterparts during the searching process of $\hat{\lambda}^*$. As the sample size increases, $\hat{\lambda}^*$ tends to decrease under the formal framework imposed by the priors, the chosen VAR model, and the optimization criterion.

The proposed method or the EB method for short is compared favorably with the existing methods including the OLS, the Ridge, and the nonparametric shrinkage (NS) method using simulated data sets. When compared with the NS method, the EB method performed better in the context of the mean squared error and the ROC curve whether the sample size is smaller than the minimum required sample size for a given model or not. It is worthwhile to note that the EB method was more powerful in detecting edges of weaker intensities (i.e., a smaller absolute value of the coefficient) as the sample size increases.

As for the analysis of the real world data, the EB method is shown to be very conservative in comparison with the NS method considering that the EB method suggests a VAR network of 510 nodes with 2266 edges while the network is of 707 nodes with 7381 edges for the NS method. The node-edge ratios are 4.4 and 10.4 for the EB and the NS method, respectively. The VAR network by the NS method had 2.4 times as many edges as the network by the EB method. Results of the simulation experiment strongly indicates a higher conservative tendency by the EB method than by the NS method when the sample size is far smaller than the dimensionality of data. While being conservative, the EB method seems to detect such nodes as 21 and 677 as new member nodes of the densely connected "yellow" web which was formed by the NS method.

We applied the proposed method for building a VAR model of a pre-specified order p. The method can also be used for selecting a best VAR model by comparing, for different values of p, values of such an evaluator as the prediction error sum of squares which is used in the parameterized cross-validation of section 4.3. The VAR model with the smallest value of the evaluator would be the one to recommend as the most appropriate. The algorithm of the proposed method is written in R programming language and available upon request to the authors.

A. Appendix A: Proofs

A.1. Proof of Theorem 2

First, we calculate the joint density function $f(\Psi, \sigma^2, \mathbf{y}_{p+1}, \ldots, \mathbf{y}_T | \lambda, \mathbf{x}_{p+1})$ as follows:

$$\begin{aligned} & f(\Psi, \sigma^2, \mathbf{y}_{p+1}, \ldots, \mathbf{y}_T | \lambda, \mathbf{x}_{p+1}) \\ & \quad = L(\Psi, \sigma^2)\pi(\Psi|\sigma^2, \lambda)\pi(\sigma^2) \\ & \quad \propto \left(\frac{1}{\sigma^2}\right)^{\alpha + d(T-p-1)/2 + d^2 p/2 + 1} \left(\frac{\lambda}{1-\lambda}\right)^{d^2 p/2} \exp\left\{-\frac{\beta}{\sigma^2}\right\} \\ & \qquad \times \exp\left\{-\frac{1}{2\sigma^2}\left(\|Y^{\mathrm{s}} - X^{\mathrm{s}}\Psi\|^2 + \frac{\lambda(T-p-1)}{1-\lambda}\|\Psi\|^2\right)\right\}. \end{aligned} \tag{31}$$

Since $(X^{\mathrm{s}})'Y^{\mathrm{s}} = \left((X^{\mathrm{s}})'X^{\mathrm{s}} + \frac{\lambda(T-p-1)}{1-\lambda} I\right)\widehat{\Psi}^*$, we get

$$\begin{aligned} & \|Y^{\mathrm{s}} - X^{\mathrm{s}}\Psi\|^2 + \frac{\lambda(T-p-1)}{1-\lambda}\|\Psi\|^2 \\ & \quad = \sum_{i=1}^{d}\left(\|Y_i^{\mathrm{s}} - X^{\mathrm{s}}\Psi_i\|^2 + \frac{\lambda(T-p-1)}{1-\lambda}\|\Psi_i\|^2\right) \\ & \quad = \sum_{i=1}^{d}\left((Y_i^{\mathrm{s}})'Y_i^{\mathrm{s}} - (Y_i^{\mathrm{s}})'X^{\mathrm{s}}\left((X^{\mathrm{s}})'X^{\mathrm{s}} + \frac{\lambda(T-p-1)}{1-\lambda}I\right)^{-1}(X^{\mathrm{s}})'Y_i^{\mathrm{s}}\right) \\ & \qquad + \sum_{i=1}^{d}\left(\Psi_i - \widehat{\Psi}_i^*\right)'\left((X^{\mathrm{s}})'X^{\mathrm{s}} + \frac{\lambda(T-p-1)}{1-\lambda}I\right)\left(\Psi_i - \widehat{\Psi}_i^*\right) \\ & \quad = 2\left(\hat{\beta}^*(\lambda) - \beta\right) \\ & \qquad + \sigma^2 \sum_{i=1}^{d}\left(\Psi_i - \widehat{\Psi}_i^*\right)'\left(\widehat{K}^*(\sigma^2, \lambda)\right)^{-1}\left(\Psi_i - \widehat{\Psi}_i^*\right). \end{aligned}$$

Therefore

$$\begin{aligned} & f(\Psi, \sigma^2, \mathbf{y}_{p+1}, \ldots, \mathbf{y}_T | \lambda, \mathbf{x}_{p+1}) \\ & \quad \propto \left(\frac{1}{\sigma^2}\right)^{\alpha + d(T-p-1)/2 + d^2 p/2 + 1} \left(\frac{\lambda}{1-\lambda}\right)^{d^2 p/2} \exp\left\{-\frac{\hat{\beta}^*(\lambda)}{\sigma^2}\right\} \\ & \qquad \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{d}\left(\Psi_i - \widehat{\Psi}_i^*\right)'\left(\widehat{K}^*(\sigma^2, \lambda)\right)^{-1}\left(\Psi_i - \widehat{\Psi}_i^*\right)\right\} \\ & \quad = \frac{\Gamma(\hat{\alpha}^*)}{(\hat{\beta}^*(\lambda))^{\hat{\alpha}^*}}\left(\frac{2\pi\lambda}{1-\lambda}\right)^{d^2 p/2}\left|(X^{\mathrm{s}})'X^{\mathrm{s}} + \frac{\lambda(T-p-1)}{1-\lambda}I\right|^{-d/2} \\ & \qquad \times \mathrm{IG}(\sigma^2 | \hat{\alpha}^*, \hat{\beta}^*(\lambda)) \prod_{i=1}^{d} \mathrm{N}_{dp}(\Psi_i | \widehat{\Psi}_i^*, \widehat{K}^*(\sigma^2, \lambda)). \end{aligned} \tag{32}$$

Next, the conditional posterior of (Ψ, σ^2) given λ is obtained by

$$\begin{aligned}\pi(\Psi, \sigma^2|\lambda; \{\mathbf{y}_t\}_{t=1}^T) &= \frac{f(\Psi, \sigma^2, \mathbf{y}_{p+1}, \ldots, \mathbf{y}_T|\lambda, \mathbf{x}_{p+1})}{\int\int f(\Psi, \sigma^2, \mathbf{y}_{p+1}, \ldots, \mathbf{y}_T|\lambda, \mathbf{x}_{p+1}) \mathrm{d}\Psi \mathrm{d}\sigma^2} \\ &= \mathrm{IG}(\sigma^2|\hat{\alpha}^*, \hat{\beta}^*(\lambda)) \prod_{i=1}^{d} \mathrm{N}_{dp}(\Psi_i|\widehat{\Psi}_i^*, \widehat{K}^*(\sigma^2, \lambda)),\end{aligned}$$

which is the product of the conditional posterior of σ^2 and the conditional posterior of Ψ. The marginal likelihood of λ is

$$L(\lambda) = \int\int L(\Psi, \sigma^2)\pi(\Psi|\sigma^2, \lambda)\pi(\sigma^2)\mathrm{d}\Psi\mathrm{d}\sigma^2,$$

which is easily calculated from (32).

A.2. Proof of Theorem 3

Let $\theta(\lambda) = \lambda(T-p-1)/(1-\lambda)$. We begin with the joint density function (31):

$$\begin{aligned}&f(\Psi, \sigma^2, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda, \mathbf{x}_{p+1}) \\ &\quad = m_5 \left(\frac{1}{\sigma^2}\right)^{\alpha+d(T-p-1)/2+d^2p/2+1} \left(\frac{\theta(\lambda)}{T-p-1}\right)^{d^2p/2} \exp\left\{-\frac{\beta}{\sigma^2}\right\} \\ &\qquad \times \exp\left\{-\frac{1}{2\sigma^2}\left(\|Y^{\mathrm{s}} - X^{\mathrm{s}}\Psi\|^2 + \theta(\lambda)\|\Psi\|^2\right)\right\}\end{aligned}$$

where m_5 is a constant that does not depend on $(\Psi, \sigma^2, \lambda)$. If we integrate this with respect to σ^2, then, through the density of Gamma distribution,

$$\begin{aligned}&f(\Psi, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda, \mathbf{x}_{p+1}) \\ &\quad = \int f(\Psi, \sigma^2, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda, \mathbf{x}_{p+1})\mathrm{d}\sigma^2 \\ &\quad = m_4\theta(\lambda)^{\frac{d^2p}{2}} \left(\beta + \frac{1}{2}\|Y^{\mathrm{s}} - X^{\mathrm{s}}\Psi\|^2 + \frac{1}{2}\theta(\lambda)\|\Psi\|^2\right)^{-\alpha-d(T-p-1)/2-d^2p/2} \qquad (33)\end{aligned}$$

where m_4 is a constant that does not depend on $(\Psi, \sigma^2, \lambda)$.

The cross entropy between $\delta(\Psi|\Psi^*)$ and $f(\Psi, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda, \mathbf{x}_{p+1})$ is

$$\begin{aligned}H(\delta, f) &= -\int \delta(\Psi|\Psi^*) \log f(\Psi, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda, \mathbf{x}_{p+1})\mathrm{d}\Psi \\ &= -\log f(\Psi^*, \{\mathbf{y}_t\}_{t=p+1}^T|\lambda, \mathbf{x}_{p+1}). \qquad (34)\end{aligned}$$

Let $\varepsilon_t^{\mathrm{s}} = \mathbf{y}_t^{\mathrm{s}} - (\Psi^*)'\mathbf{x}_t^{\mathrm{s}} - \mathbf{c}^{\mathrm{s}*}$. The term $\|Y^{\mathrm{s}} - X^{\mathrm{s}}\Psi^*\|^2$ is expanded as

$$\begin{aligned}\|Y^{\mathrm{s}} - X^{\mathrm{s}}\Psi^*\|^2 &= \sum_{t=p+1}^{T} \left\|\varepsilon_t^{\mathrm{s}} - \frac{1}{T-p}\sum_{\tau=p+1}^{T}\varepsilon_\tau^{\mathrm{s}}\right\|^2 \\ &= \sum_{t=p+1}^{T}\|\varepsilon_t^{\mathrm{s}}\|^2 - \frac{1}{T-p}\left\|\sum_{\tau=p+1}^{T}\varepsilon_\tau^{\mathrm{s}}\right\|^2 .\end{aligned}$$

Based on the density function (16), we note that

$$\varepsilon_t^{\mathrm{s}} | (\, \Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*}, \mathbf{x}_t^{\mathrm{s}}) \sim \mathrm{N}_d(\mathbf{0}, \sigma^{*2} I) \ .$$

Therefore it follows that

$$\mathrm{E}\left[\left\| \varepsilon_t^{\mathrm{s}} \right\|^2 \middle| \ \Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*} \right] = \mathrm{E}\left[\mathrm{E}\left[\left\| \varepsilon_t^{\mathrm{s}} \right\|^2 \middle| \ \Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*}, \mathbf{x}_t^{\mathrm{s}} \right] \right] = d\sigma^{*2},$$

and

$$\begin{aligned}
&\mathrm{E}\left[\left\| \sum_{\tau=p+1}^{T} \varepsilon_\tau^{\mathrm{s}} \right\|^2 \middle| \ \Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*} \right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left\| \sum_{\tau=p+1}^{T-1} \varepsilon_\tau^{\mathrm{s}} \right\|^2 + 2(\varepsilon_T^{\mathrm{s}})' \sum_{\tau=p+1}^{T-1} \varepsilon_\tau^{\mathrm{s}} + \left\| \varepsilon_T^{\mathrm{s}} \right\|^2 \middle| \ \Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*}, \{\mathbf{x}_t^{\mathrm{s}}\}_{t=p+1}^{T} \right] \right] \\
&= \mathrm{E}\left[\left\| \sum_{\tau=p+1}^{T-1} \varepsilon_\tau^{\mathrm{s}} \right\|^2 \middle| \ \Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*} \right] + 0 + d\sigma^{*2} \\
&= d(T-p)\sigma^{*2} \ .
\end{aligned}$$

Thus,

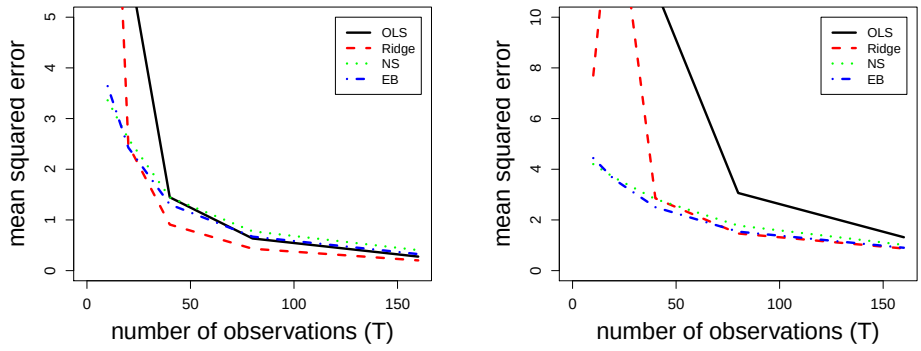
$$\begin{aligned}
&\mathrm{E}\left[\left\| Y^{\mathrm{s}} - X^{\mathrm{s}} \Psi^* \right\|^2 \middle| \ \Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*} \right] \\
&\quad = \mathrm{E}\left[\sum_{t=p+1}^{T} \left\| \varepsilon_t^{\mathrm{s}} \right\|^2 - \frac{1}{T-p} \left\| \sum_{\tau=p+1}^{T} \varepsilon_\tau^{\mathrm{s}} \right\|^2 \middle| \ \Psi^*, \sigma^{*2}, \mathbf{c}^{\mathrm{s}*} \right] \\
&\quad = d(T-p)\sigma^{*2} - d\sigma^{*2} \\
&\quad = d(T-p-1)\sigma^{*2} \ . \qquad (35)
\end{aligned}$$

From the Jensen's inequality, we get

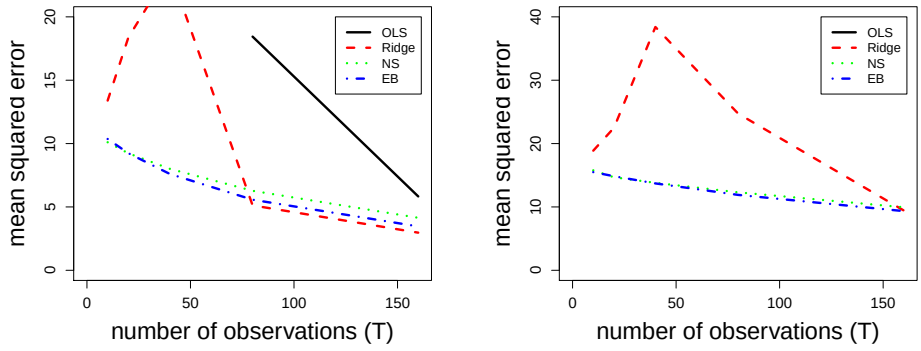
$$\begin{aligned}
&\mathrm{E}\left[\log\left(\beta + \frac{1}{2} \left\| Y^{\mathrm{s}} - X^{\mathrm{s}} \Psi^* \right\|^2 + \frac{1}{2}\theta(\lambda) \left\| \Psi^* \right\|^2 \right) \right] \\
&\quad \leq \log\left(\beta + \frac{1}{2}\mathrm{E}\left[\left\| Y^{\mathrm{s}} - X^{\mathrm{s}} \Psi^* \right\|^2 \right] + \frac{1}{2}\theta(\lambda) \left\| \Psi^* \right\|^2 \right) \ . \qquad (36)
\end{aligned}$$

And from (33), (34), (35) and (36), we get the result in Theorem 3.

B. Appendix B: Supplemental figures

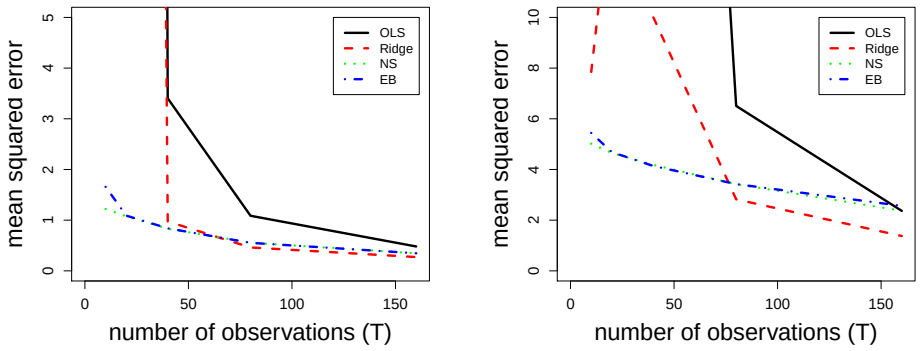


(a) $p = 2, d = 5$ (b) $p = 2, d = 10$

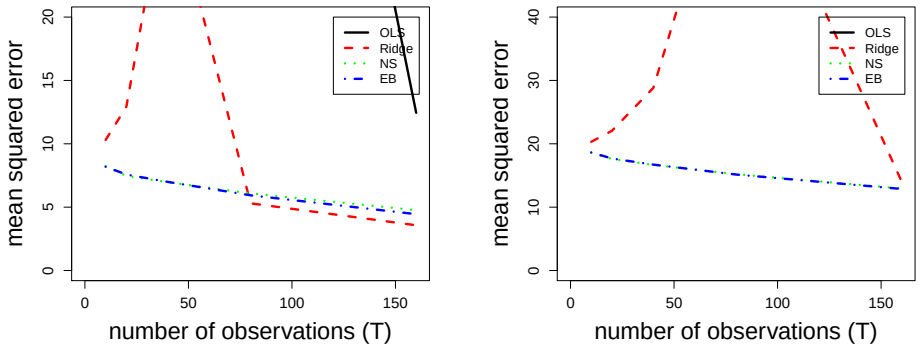


(c) $p = 2, d = 20$ (d) $p = 2, d = 40$

Fig. 17. The $\widehat{MSE}$'s by the four methods, the OLS, Ridge, NS, and EB methods. Data were generated from VAR$(2, d)$ models with $d = 5$, 10, 20, and 40 for (a), (b), (c), and (d), respectively.



(a) $p = 3, d = 5$ (b) $p = 3, d = 10$



(c) $p = 3, d = 20$ (d) $p = 3, d = 40$

Fig. 18. The $\widehat{MSE}$'s by the four methods, the OLS, Ridge, NS, and EB methods. Data were generated from VAR$(3, d)$ models with $d = 5$, 10, 20, and 40 for (a), (b), (c), and (d), respectively.

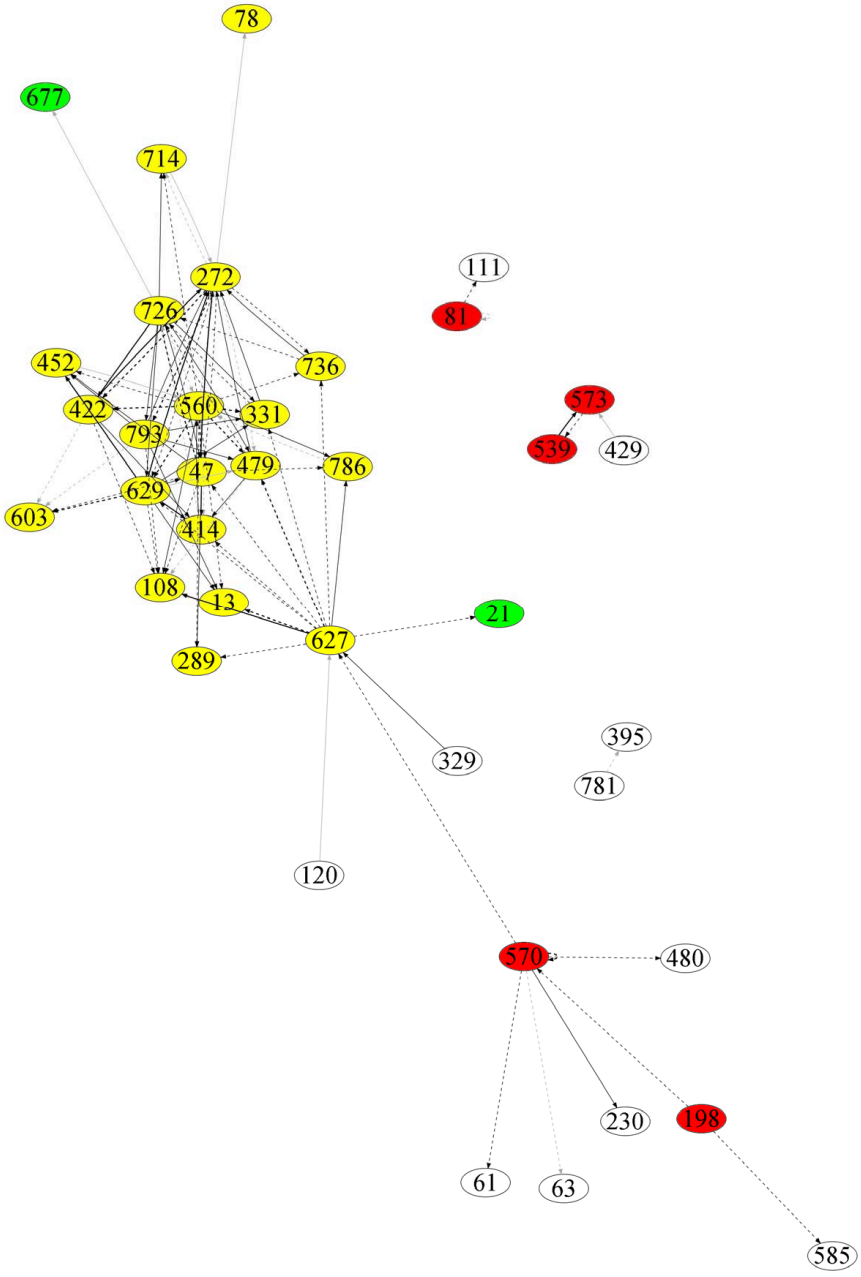


Fig. 19. Subgraph with the 100 strongest-intensity edges obtained by the EB method. The solid and dotted lines indicate positive and negative partial correlation coefficients, respectively, and the line intensity denotes their strength.

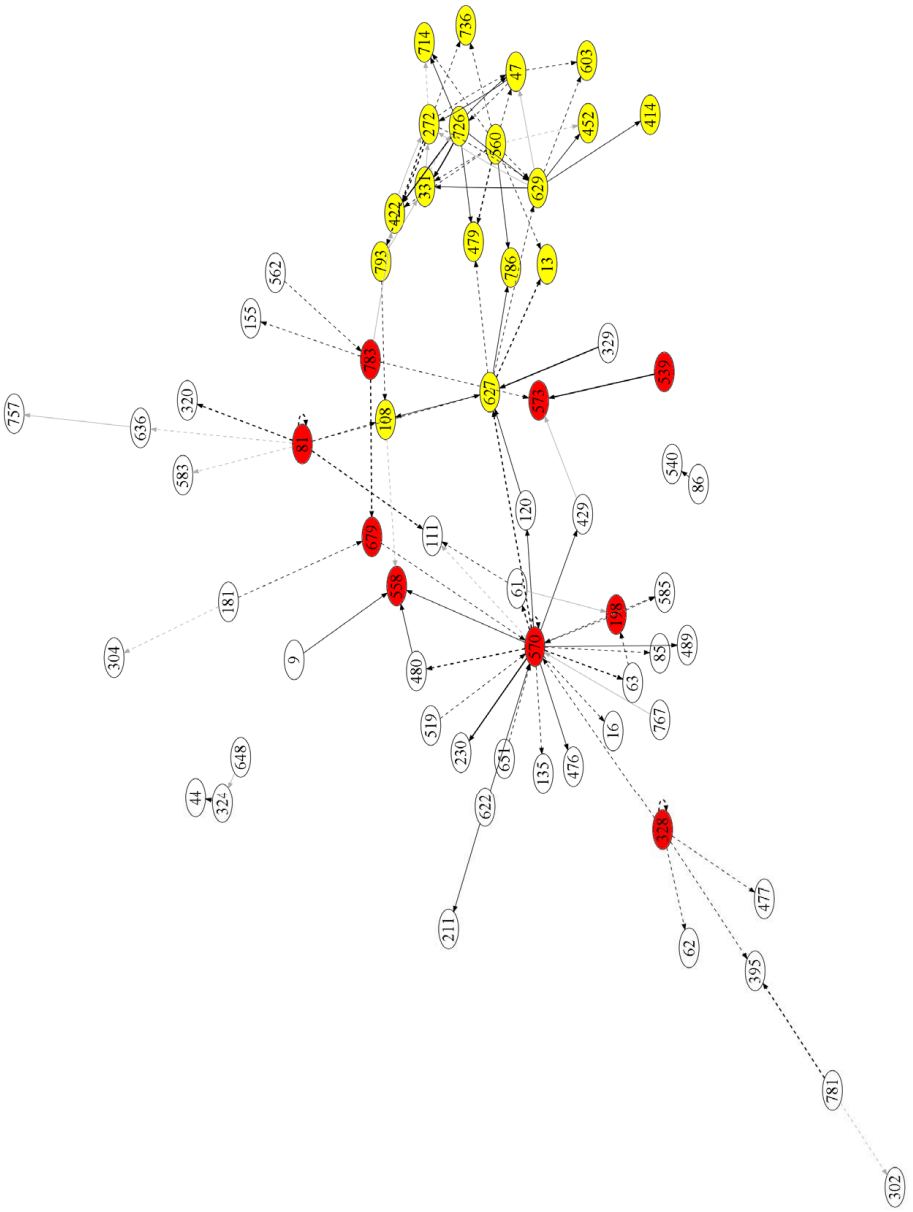


Fig. 20. Subgraph with the 100 strongest-intensity edges obtained by the NS method. The solid and dotted lines indicate positive and negative partial correlation coefficients, respectively, and the line intensity denotes their strength.

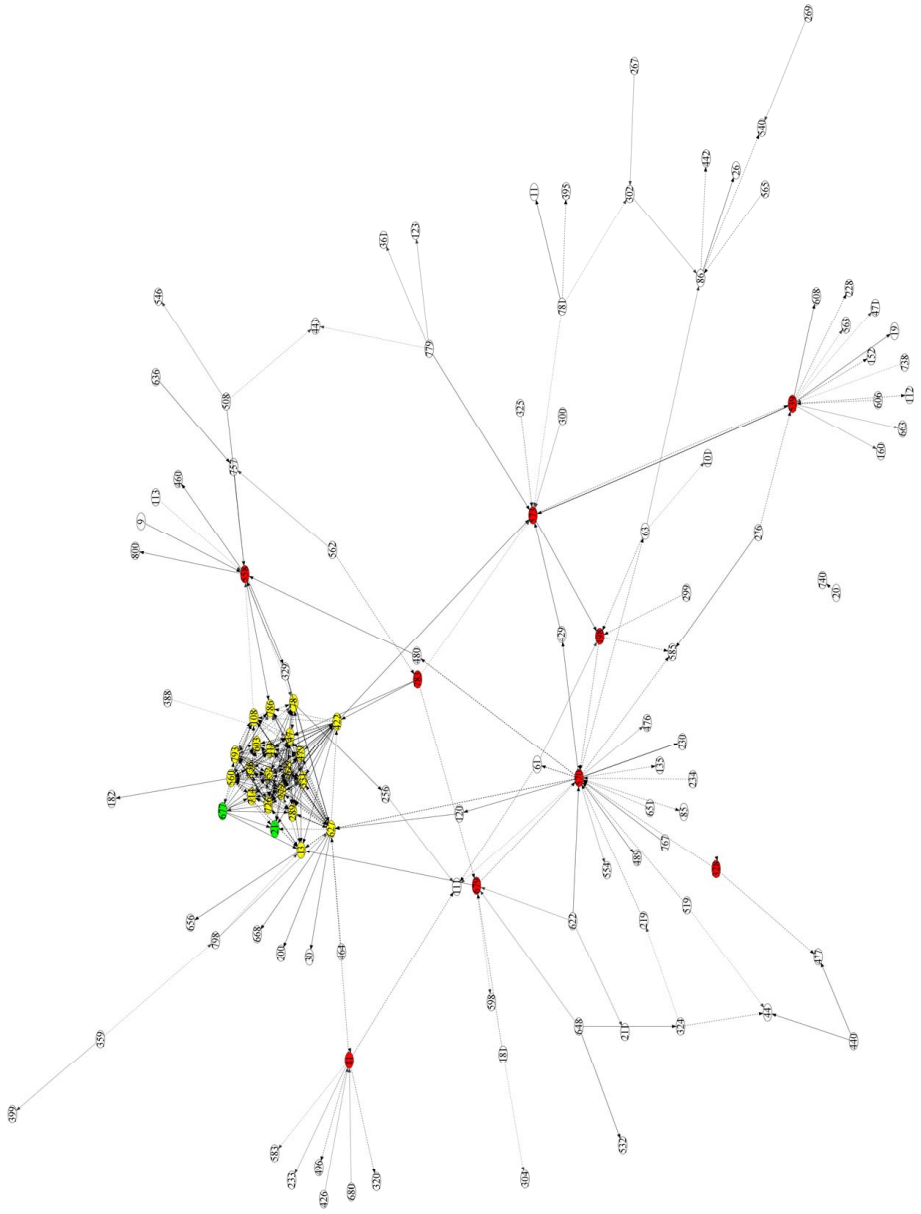


Fig. 21. Subgraph with the 300 strongest-intensity edges obtained by the EB method. The solid and dotted lines indicate positive and negative partial correlation coefficients, respectively, and the line intensity denotes their strength.

References

Akaike, H. (1978) A new look at the Bayes procedure. *Biometrika*, **65**, 53–59.

Akaike, H. (1980) Likelihood and the Bayes procedure. In *Bay Statistics* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and M. Smith), 143–166, Valencia: University Press.

Aliprantis, C. D. and Burkinshaw, O. (1998) *Principles of Real Analysis*, 3rd edn. San Diego: Academic Press.

Anderson, T. W. (1971) *The Statistical Analysis of Time Series*. New York: John Wiley & Sons.

Barnard, J., McCulloch, R. and Meng, X.-L. (2000) Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, **10**, 1281–1311.

Carlin, B. P. and Louis, T. A. (2009) *Bayesian Methods for Data Analysis*, 3rd edn. Boca Raton: CRC Press.

Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J. and May, S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Research*, **32**, D575–D577. `http://affymetrix.arabidopsis.info/`

de Finetti, B. (1972) *Probability, Induction, and Statistics*. New York: Wiley.

Doan, T., Litterman, R. and Sims, C. A. (1984) Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, **3**(1), 1–100.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995) *Bayesian Data Analysis*. London: Chapman and Hall.

Golub, G. H., Heath, M. and Wahba, G. (1979) Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**(2), 215–223.

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

Hamilton, J. D. (1994) *Time Series Analysis*. Princeton, New Jersey: Princeton University Press.

Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Hotelling, H. (1953) New light on the correlation coefficient and its transforms. *J. R. Statist. Soc.* B, **15**, 193–232.

James, W. and Stein, C. (1961) Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 361–379. Berkeley: University of California Press.

Jeffreys, H. (1961) *Theory of Probability*, 3rd edn. New York: Oxford University Press.

687 Koo, I., Lee, N., and Kil, R. M. (2008) Parametrized cross-validation for nonlinear regression
688 models. *Neurocomputing*, **71**, 3089–3095.

689 Koop, G. and Korobilis, D. (2010) Bayesian multivariate time series methods for empirical
690 macroeconomics. *Foundations and Trends in Econometrics*, **3**(4), 267–358.

691 Ledoit, O. and Wolf, M. (2004) A well conditioned estimator for large-dimensional covari-
692 ance matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

693 Litterman, R. (1986) Forecasting with Bayesian vector autoregressions-Five years of expe-
694 rience. *Journal of Business and Economic Statistics*, **4**, 25-38.

695 Morris, C. N. (1983) Parametric empirical Bayes inference: Theory and applications, *J.*
696 *Am. Statist. Ass.*, **78**, 47–55.

697 Opgen-Rhein, R. and Strimmer, K. (2007a) Accurate ranking of differentially expressed
698 genes by a distribution-free shrinkage approach. *Statistical Applications in Genetics and*
699 *Molecular Biology*, **6**:9.

700 Opgen-Rhein, R. and Strimmer, K. (2007b) From correlation to causation networks: a
701 simple approximate learning algorithm and its application to high-dimensional plant gene
702 expression data. *BMC Systems Biology*, **1**:37.

703 Opgen-Rhein, R. and Strimmer, K. (2007c) Learning causal networks from systems biology
704 time course data: an effective model selection procedure for the vector autoregressive
705 process. *BMC Bioinformatics*, **8**(Suppl 2):S3.

706 Schäfer, J. and Strimmer, K. (2005a) An empirical Bayes approach to inferring large-scale
707 gene association networks. *Bioinformatics*, **21**(6), 754–764.

708 Schäfer, J. and Strimmer, K. (2005b) A shrinkage approach to large-scale covariance matrix
709 estimation and implications for functional genomics. *Statistical Applications in Genetics*
710 *and Molecular Biology*, **4**(1):32.

711 Seber, G. A. F. and Lee, A. J. (2003) *Linear Regression Analysis*, 2nd edn. Hoboken, New
712 Jersey: John Wiley & Sons.

713 Sims, C. A. (1972) Money, income, and causality. *The American Economic Review*, **62**(4),
714 540–552.

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

716 Sims, C. A. (1982) Policy analysis with econometric models. *Brookings Papers on Economic*
717 *Activity*, **13**, 107–164.

718 Smith, S. M., Fulton, D. C., Chia, T., Thorneycroft, D., Chapple, A., Dunstan, H., Hylton,
719 C., Zeeman, S. C. and Smith, A. M. (2004) Diurnal changes in the transcriptom encoding
720 enzymes of starch metabolism provide evidence for both transcriptional and posttran-
721 scriptional regulation of starch metabolism in Arabidopsis leaves. *Plant Physiology*, **136**,
722 2687–2699.

723 Stein, C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal
724 distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics*
725 *and Probability* (ed J. Neyman), **1**, 197–206. Berkeley: University of California Press.

726 Strimmer, K. (2008) fdrtool: a versatile R package for estimating local and tail area-based
727 false discovery rates. *Bioinformatics*, **24**(12), 1461–1462.

728 Sun, D. and Ni, S. (2004) Bayesian analysis of vector-autoregressive models with noninfor-
729 mative priors. *J. Statist. Plannng Inf.*, **121**, 291–309.

730 Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.*
731 B, **58**, 267–288.

732 Tsay, R. S. (2005) *Analysis of Financial Time Series*, 2nd edn. Hoboken, New Jersey: John
733 Wiley & Sons.

734 Wei, W. W. (2005) *Time Series Analysis: Univariate and Multivariate Methods*, 2nd edn.
735 Addison-Wesley.

736 Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. New York: John
737 Wiley & Sons.