Econometrics and Statistics 10 (2019) 120–133

Contents lists available at ScienceDirect

Econometrics and Statistics

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

Joint estimation of multiple network Granger causal models

A. Skripnikov[a,b,*], G. Michailidis[b]

[a] Department of Mathematics, University of Houston, Houston, TX, 77054, USA
[b] Department of Statistics, University of Florida, 102 Griffin-Floyd Hall P.O. Box 118545, Gainesville, FL 32611, USA

ARTICLE INFO

Article history:
Received 27 June 2017
Revised 9 April 2018
Accepted 4 August 2018
Available online 23 August 2018

Keywords:
Alternating direction method of multipliers
Factor covariance
Generalized fused lasso
Sparse estimation
Vector autoregression

ABSTRACT

Joint regularized modeling framework is presented for the estimation of multiple Granger causal networks. High-dimensional network Granger models focus on learning the corresponding causal effects amongst a large set of distinct time series. They are operationalized through the formalism of Vector Autoregressive Models (VAR). The latter represent a popular class of time series models that has been widely used in applied econometrics and finance. In particular, the setting of the same set of variables being measured on different entities over time is considered (e.g. same set of economic indicators for multiple US states). Moreover, the covariance structure of the error term is assumed to exhibit low rank structure which can be recovered by a factor model. The framework allows to account for both sparsity and potential similarities between the related networks by introducing appropriate structural penalties on the transition matrices of the corresponding VAR models. An alternating directions method of multipliers (ADMM) algorithm is developed for solving the underlying joint estimation optimization problem. The performance of the joint estimation method is evaluated on synthetic data and illustrated on an application involving economic indicators for multiple US states[1].

1. Introduction

There has been a lot of recent interest in modeling and analysis of high-dimensional time series data. Application areas include brain fMRI data (Song et al., 2011), financial portfolio selection (Fan et al., 2011), gene regulatory network inference (Michailidis and d'Alché Buc, 2013), macroeconomic time series forecasting and structural analysis (Bańbura et al., 2010), just to name a few. Their common characteristic is the large number of time series and their cross-relationships being analyzed compared to the number of time points available, thus leading to a high-dimensional setting where the number of model parameters exceeds the number of available time points (samples).

In many cases, the temporal dynamics of the data under consideration can be captured through the framework of *Granger causality* (Granger, 1969; Dahlhaus and Eichler, 2003). Time series $\{X_t\}$ is said to Granger-cause time series $\{Y_t\}$ if past values of $\{X_t\}$ can be used to better forecast future values of $\{Y_t\}$. Extending it to the case of more than two time series introduces a concept of *network Granger causality*, which can be represented through Vector Autoregressive models (VAR) (Basu et al., 2015b). This class of models enables us to account for each time series' own temporal dynamics, as well as temporal linear cross-dependencies amongst them.

* Corresponding author at: Department of Statistics, University of Florida, 7901 Cambridge St, Apt 22, Houston, TX, 77054, USA.
E-mail addresses: usdandres1@gmail.com (A. Skripnikov), gmichail@umich.edu (G. Michailidis).

[1] Code and data are available as online supplement.

https://doi.org/10.1016/j.ecosta.2018.08.001

Another framework that focuses on the contemporaneous dependence (correlations) amongst a large number of time series is that of Dynamic Factor Models (DFM) (Stock and Watson, 2011). In this case, their correlation structure is capture through a low-dimensional factor model, where the factors exhibit VAR dynamics. This model has been widely used in macroeconomics (Stock and Watson, 2016) and extensively studied from a theoretical viewpoint; see Bai et al. (2008) and references therein.

However, in presence of a large number of time series and relative few time points, it is necessary to incorporate regularizing *sparsity assumptions* to estimate the parameters of the VAR model as described in detail in Basu et al. (2015a). This has led to a novel body of work that examines both computational approaches and theoretical properties of sparse VAR models. The DFM exhibits better scaling, since it allows the number of time series to be of the order of the time points available; for a discussion on obtaining consistent estimates of the DFM parameters see Bai et al. (2008). Hence, depending on whether the focus of the analysis is on lead-lag temporal (VAR) or contemporaneous (DFM) relationships, the practitioner has access to tools to carry out estimation of the parameters of the model of interest, even in high dimensional settings.

However, in many applications, one has to deal with related sets of multivariate time series. As a motivating example, consider the data analyzed in Section 6. It considers a number of employment and economic indicator variables for four US states that exhibit similarities regarding their economic infrastructure: Pennsylvania, Michigan, Ohio and Illinois. In particular, they share a strong manufacturing base, a fairly large agricultural sector, a strong presence in banking, education and health services, and access to the Great Lakes waterways. At the same time, they also exhibit differences due to specific conditions, like the developed financial industry in Chicago, or the strong and sustained presence of the coal, oil and gas industries in Pennsylvania. Hence, it is desirable to account for potential similarities, while leaving room for estimation of individual features of each state. Within the DFM framework one could adapt the model posited in Hallin and Liška (2011) that incorporates common factors and idiosyncratic components that would correspond to the time series of each state. However, there is nothing analogous in the VAR setting. Thus, the main aspect of this paper is to introduce a modeling framework that enables *joint estimation* of multiple related VAR models in order to borrow strength across them, instead of estimating each model separately. The problem of joint estimation has received attention in the literature recently, primarily focusing on the estimation of multiple graphical models. The approach leverages various penalties that encourage both sparsity and similarity for parameters of multiple related models; see for example, the hierarchical penalty used in Guo et al. (2011), the group lasso penalty in Ma and Michailidis (2016), or mixed norm penalties in Cai et al. (2015). In our particular case, the *fused lasso* penalty (Tibshirani et al., 2011) is employed as it imposes similarity assumption across the VAR models while leaving room for idiosyncratic relationships within each model.

We consider p-variable stationary time series for K related entities (e.g. economies, households, etc.): $\{X_k^t = (X_{1k}^t, \ldots, X_{pk}^t)' \in \mathbb{R}^p,\ t = 0, \ldots, T\},\ k = 1, \ldots, K$. The corresponding VAR model of lag order D, denoted as *VAR(D)*, is given by

$$X_k^t = A_k^1 X_k^{t-1} + \cdots + A_k^D X_k^{t-D} + \epsilon_k^t,\ \epsilon_k^t \sim N(0, \Sigma_k), \quad t = D, \ldots, T, \quad k = 1, \ldots, K, \tag{1}$$

where $A_k^d = (a_k^d)_{ij} \in \mathbb{R}^{p \times p}$ is the transition matrix for entity k, $d = 1, \ldots, D,\ k = 1, \ldots, K$. Matrices $\{A_k^d, d = 1, \ldots, D\}$ represent a Granger causal network in the following sense: if $a_{k,ij}^d \neq 0$ for at least one $d = 1, \ldots, D$, we say that variable j Granger-causes variable i, implying that values of time series j are useful to better forecast future values of time series i, for entity k. Meanwhile, $\epsilon_k^t = (\epsilon_{1k}^t, \ldots, \epsilon_{pk}^t)' \sim N(0, \Sigma_k)$, and the covariance matrix $\Sigma_k \in \mathbb{R}^{p \times p},\ \Sigma_k \succ 0$, allows for additional contemporaneous dependence between the p variables under consideration within entity k. The standard assumption is that Σ_k is diagonal and thus no extra dependence is allowed (Lütkepohl, 2005). In the high dimensional setting, (Basu et al., 2015a; Lin and Michailidis, 2017) allow for a general covariance matrix Σ_k, assuming that it possesses a sparse inverse. Then, a joint estimation procedure is introduced for obtaining sparse estimates of both (A_k, Σ_k^{-1}).

In this work due to the similarity of the K entities under study, we assume that Σ_k exhibits a low rank structure, stemming from a factor model formulation of the error process $\{\epsilon_k^t,\ t = 1, \ldots, T\}$. This assumption implies that the relationships between the covariation of the p error processes within entity k can be explained by a smaller number L_k of common underlying factors. Such factor models are widely used in econometric and finance applications, such as forecasting bond yields (Diebold and Li, 2006), modeling interest rates (Rudebusch, 2010), forecasting macroeconomic indicators (Stock and Watson, 2002). While past work focused on observed time series, we use the factor model on the error processes themselves. To avoid identifiability issues for the parameters of the VAR model given in 1 as discussed in Deistler et al. (2011) and Anderson et al. (2012) we further assume sparsity in the covariance matrix of the idiosyncratic component in the factor model posited for Σ_k.

Hence, the main contributions of this work are the introduction of the joint modeling framework for estimating multiple related network Granger causal models, including the development of an alternating direction method of multipliers (ADMM) algorithm to implement appropriate penalties, and the incorporation of factor models to estimate the error covariances for each of those VAR models. The remainder of the paper is organized as follows: in Section 2 the detailed modeling framework is introduced along with the objective function that enables us to estimate the model parameters. Section 3 provides details on the the estimation procedure, such as the algorithm for error covariance estimation, the employed ADMM algorithm for implementing generalized sparse fused optimization, and tuning parameter selection. Sections 4 and 5 describe the results of applying our joint modeling approach to synthetic and real data, respectively, and how they compare to the alternative separate estimation procedure. Finally, some concluding remarks are drawn in Section 6.

2. Joint model estimation problem formulation

We develop the joint estimation procedure over K related VAR models comprising of the same set of p variables and sampled over the same T time points. The $VAR(D)$ model for a single entity was first described in (1) in Section 1. Let $X_k^t = (X_{1k}^t, \ldots, X_{pk}^t)^\top$ denote the vector containing the data for p time series at time $t = 0, \cdots, T$ for entity $k = 1, \cdots, K$, and let $\epsilon_k^t = (\epsilon_{1k}^t, \ldots, \epsilon_{pk}^t)'$ be the corresponding error term for time t. Further, $\{A_k^d,\ d = 1, \ldots, D\}$ denotes the set of all D transition matrices for the kth $VAR(D)$ model, and Σ_k is the covariance matrix of the error term. We assume that $\{A_k^d,\ d = 1, \ldots, D\}$ are *sparse* in order to accommodate high-dimensional scaling, and that the error covariance matrix is low rank, whose structure is dictated by an L_k-rank factor model. Finally, all K models are assumed to be stable, namely that $\det(\mathcal{A}_k(z)) \neq 0$ on the unit circle $\{z \in \mathbb{C} : |z| = 1\}$, where $\mathcal{A}_k(z) = I_p - \sum_{t=1}^D A_k^t z^t$; for a discussion see Basu et al. (2015a).

It is well known that a VAR model can be expressed in regression form as (see Lütkepohl, 2005, with a detailed description also given in Appendix A.1):

$$W_k = Z_k \beta_k + \underline{\epsilon}_k, \underline{\epsilon}_k \sim N(O, \tilde{\Sigma}_k),\ \ k = 1, \ldots, K, \tag{2}$$

where $\beta_k \in \mathbb{R}^{Dp^2}$ corresponds to the vector contained all elements of all the $p \times p$ transition matrices $\{A_k^d,\ d = 1, \ldots, D\}$ appropriately arranged. In case of a known error covariance matrix $\tilde{\Sigma}_k$ the optimization criterion for obtaining sparse estimates of β_k is given by the standard lasso formulation (Tibshirani, 1996):

$$\min_{\beta_k} ||\tilde{\Sigma}_k^{-1/2}(W_k - Z_k\beta_k)||_2^2 + \lambda_1^k |\beta_k|_1,\ \ k = 1, \ldots, K. \tag{3}$$

Note that a larger value of the tuning parameter λ_1^k leads to sparser Granger causal effects. There are a number of available algorithms in the literature for solving (3), including a cyclic coordinate descent type algorithm (Friedman et al., 2010). Further, the consistency of the lasso estimates of β_k under high dimensional scaling is established in Basu et al. (2015a).

Next, we present the joint model formulation for K entities. First, we write all K VAR models in a compact form as

$$\mathbf{X}^t = \mathbf{A}^1\mathbf{X}^{t-1} + \cdots + \mathbf{A}^D\mathbf{X}^{t-D} + \boldsymbol{\epsilon}^{t-D},\ \ \boldsymbol{\epsilon}^{t-D} \sim N(\mathbf{O}, \boldsymbol{\Sigma}),\ \ t = D, .., T, \tag{4}$$

where $\mathbf{X}^t = (X_1^t, \ldots, X_K^t)^\top \in \mathbb{R}^{Kp}$, $\mathbf{A}^d \in \mathbb{R}^{Kp \times Kp}$ - block-diagonal matrix with k^{th} block equal to A_k^d, $d = 1, \ldots, D$; $\boldsymbol{\epsilon}^{t-D} = (\epsilon_1^{t-D}, \ldots, \epsilon_K^{t-D})^\top \in \mathbb{R}^{Kp}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{Kp \times Kp}$ - block-diagonal matrix with k^{th} block equal to $\Sigma_k \in \mathbb{R}^{p \times p}$, $k = 1, \ldots, K$; $\mathbf{O}^p \in \mathbb{R}^{Kp \times Kp}$ denotes a zero-matrix.

We also assume that there exist similarities between the transition matrices A_k, in addition to being sparse. Note that while sparsity implies a relatively low number of active Granger causal effects, similarity further implies that there are *shared* such effects across the K models. The operationalization of this sharing effect will be accomplished through an appropriate penalty term.

Analogously to the single VAR model, we can transition from (4) to the following standard regression formulation (for details on $\boldsymbol{W}$, $\boldsymbol{Z}$, $\underline{\boldsymbol{\epsilon}}$ and $\tilde{\boldsymbol{\Sigma}}$ refer to Appendix A.1):

$$\mathbf{W} = \mathbf{Z}\boldsymbol{\beta} + \underline{\boldsymbol{\epsilon}},\ \ \underline{\boldsymbol{\epsilon}} \sim N(\mathbf{O}, \tilde{\boldsymbol{\Sigma}}), \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)' \in \mathbb{R}^{K(Dp^2)}$, with $\beta_k \in \mathbb{R}^{Dp^2}$ denoting the vector containing the elements of all the $p \times p$ transition matrices $\{A_k^d,\ d = 1, \ldots, D\}$. We posit the following *generalized sparse fused lasso* optimization criterion that achieves joint estimation of K network Granger causal models:

$$\min_{\boldsymbol{\beta}} ||\tilde{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{W} - \mathbf{Z}\boldsymbol{\beta})||_2^2 + \lambda_1 \sum_{k=1}^K |\beta_k|_1 + \lambda_2 \sum_{k>m} |\beta_k - \beta_m|_1. \tag{6}$$

The above formulation has a regularization term comprising of two components: the first is the standard lasso one, regulated by the tuning parameter λ_1 that controls the sparsity level of the K transition matrices A_k; the second corresponds to fused lasso penalty that encourages similarities between the corresponding elements of the K transition matrices and is regulated by the tuning parameter λ_2. Note that in the absence of the fused lasso term, optimization criterion (6) reduces to estimating separately the parameters of the K models.

Note that the fused lasso has been used in penalized regression settings before (Tibshirani et al., 2011). However, the focus was on a single regression model and the objective was to encourage joint sparsity across variables that exhibit a natural ordering (e.g. adjacent frequencies in mass spectrometry applications). On the other hand, we use the fused lasso in a novel way to achieve joint sparsity across the parameters of K VAR models.

Solving (6) is a more involved problem, vis-a-vis the separate sparse estimation procedure, due to the fused lasso penalty term. Next, we outline the key steps of the estimation procedure.

- **Step 1.** Given the low-rank structure, we first estimate the error covariance matrices $\hat{\Sigma}_k, k = 1, \ldots, K$, to use as plug-in estimates in (6). Additional details are given in Section 3.1 and Appendix A.2.
- **Step 2.** Solve (6) using an ADMM algorithm, introduced in Section 3.2, to obtain the transition matrix estimates $\hat{A}_k$. Selecting the tuning parametersλ_1 and λ_2 is discussed in Section 3.3.

3. Estimation procedure

We discuss next the key elements in our joint estimation procedure.

3.1. Factor covariance matrix estimation

While the transition matrices $\{A_k^d,\ d = 1, \ldots, D,\ k = 1, \cdots, K\}$ in (1) capture the cross-temporal effects of the p variables in a VAR model, the error covariance $\Sigma_k \in \mathbb{R}^{p\times p}$ captures their contemporaneous dependence. There have been numerous assumptions made on this quantity in the literature: the simplest is that $\Sigma = \sigma^2 I_p$, where $I_p \in \mathbb{R}^{p\times p}$ denotes the identity matrix of order p (Lütkepohl, 2005), while Basu et al. (2015a); Lin and Michailidis (2017) examined the case of an arbitrary Σ with a sparse inverse. As explained in the introductory section, given the focus on K related entities, we use a factor modeling approach for Σ.

Factor models are widely used for covariance estimation in finance and econometrics with the purpose of dimension reduction (Bernanke et al., 2004; Fama and French, 1993). Such models assume that the relationship between the p variables under consideration can be explained by a small number $L \ll p$ of common underlying factors, which gives rise to an *L-factor model*. A very popular example in the literature is the Fama-French three-factor model for stock returns. It posits that their associations can be modeled based on the following underlying factors: market risk, difference in stock returns between big and small companies of that market, difference in stock returns between companies with high and low price-to-book ratio. In our case, we assume that the relationship between the p error processes for the same entity k is driven by few unobserved common factors.

Specifically, the p-dimensional error processes $\{\epsilon_k^t,\ t = 1, \ldots, T\}$ for the VAR model given in (1) is assumed to be generated from the following L-factor model:

$$\epsilon_k^t = \Lambda_k F_k^t + \epsilon_{k,U},\ \epsilon_{k,U} \sim N(0, \Sigma_{k,U}),\ \mathrm{cov}(F_k^t) = I_{L_k},\ t = 0, \ldots, T, \quad k = 1, \cdots, K \tag{7}$$

where $F_k^t \in \mathbb{R}^L$ - $L_k \times 1$ vector of L_k factor values at time t (not observed), $\Lambda_k \in \mathbb{R}^{p\times L_k}$ - $p \times L_k$ matrix of factor loadings, $\Sigma_{k,U} \in \mathbb{R}^{p\times p}$ - $p \times p$ matrix with a sparse inverse (idiosyncratic component). This representation leads to

$$\mathrm{Cov}(\epsilon_k^t, \epsilon_k^t) = \Sigma_k = \Lambda_k \Lambda_k' + \Sigma_{k,U}. \tag{8}$$

It can be seen that instead of directly estimating $\frac{p(p-1)}{2}$ elements for Σ_k, we only have to estimate the $p \times L_k$ elements of the factor loading matrix Λ_k of dimensions $p \times L_k$ and those of a sparse (diagonal, in most cases) inverse of $\Sigma_{k,U}$, which approximately adds up to $p(L_k + 1)$ effective parameters. Considering that $L_k \ll p$, this leads to a substantial dimension reduction. The following procedure, with selected key steps borrowed from the POET estimator (Fan et al., 2013), will be used to calculate the factor model estimates of the error covariance matrix Σ_k *separately* for each entity k:

Error covariance estimation procedure

- **Step 0.** Initialize $\hat{\Sigma}_k = I_p$.
- **Step 1.** Plug $\Sigma_k = \hat{\Sigma}_k$ into optimization criterion (3), solve the latter to get the sparse estimate $\hat{\beta}$ (the criterion for selecting the sparsity tuning parameter will be described in Section 3.3).
- **Step 2.** Obtain residuals $\underline{\hat{\epsilon_k}} = W - Z\hat{\beta}$ and calculate their empirical covariance matrix $\hat{\Sigma}_{k\epsilon_k}$.
- **Step 3*.** Perform an eigenvalue decomposition for $\hat{\Sigma}_{k,\epsilon_k}$, identify $\widehat{L_k}$ large eigenvalues for $\widehat{L_k}$-factor model. Construct the estimate $\hat{\Lambda}_k \in \mathbb{R}^{p\times\widehat{L_k}}$ of Λ_k from (8) with those eigenvalues and corresponding eigenvectors.
- **Step 4*.** Use $\hat{\Sigma}_{k,\epsilon_k} - \hat{\Lambda}_k \hat{\Lambda}_k'$ as 'data' to get the estimate $\hat{\Sigma}_{k,U}$ of $\Sigma_{k,U}$ from (8) using a graphical lasso procedure from Friedman et al. (2008) (which works well under our assumption of a sparse inverse of $\Sigma_{k,U}$).
- **Step 5.** Calculate the factor model estimate: $\hat{\Sigma}_k = \hat{\Lambda}_k \hat{\Lambda}_k' + \hat{\Sigma}_{k,U}$.
- **Step 6.** Return to Step 1 and repeat the whole procedure until Step 5 and stop (simulation studies showed that iterating more than once does not produce significant improvement).

* Note that further details for steps 3 and 4 are given in Appendix A.2.

Remark 1. We assume that each entity k has its own idiosyncratic error component ϵ_k^t and thus we estimate its corresponding factor structure separately. Another possibility that imposes a much stringent contemporaneous dependence structure is to assume that all errors components *share* the same factor model; i.e. $\Lambda_k = \Lambda$ for all $k = 1, \cdots, K$. However, this is a strong assumption that may fail to hold in many applications. An interesting intermediate alternative is to assume that $\Lambda_k \sim \Lambda_\ell$ for $\ell \neq k$; namely, assume that the loadings of the factor models are related, in an analogous manner to the posited assumption on the transition matrices of the k entities. Nevertheless, it is an open issue the best way to operationalize this similarity relationship and constitutes a topic of future work.

3.2. Estimation of transition matrices

To solve the optimization criterion (6) for an arbitrary choice of values (λ_1, λ_2), we introduce an alternating directions method of multipliers (ADMM) algorithm (Boyd et al., 2011). Criterion (6) can be rewritten in the following form:

$$\min_{\boldsymbol{\beta}} ||C - D\boldsymbol{\beta}||_2^2 + \lambda_1|\boldsymbol{\beta}|_1 + \lambda_2|L\boldsymbol{\beta}|_1, \tag{9}$$

where $C = \tilde{\boldsymbol{\Sigma}}^{-1/2}\mathbf{W}$, $D = \tilde{\boldsymbol{\Sigma}}^{-1/2}\mathbf{Z}$, $L\boldsymbol{\beta} \equiv (\beta_1 - \beta_2, \ldots, \beta_1 - \beta_K, \beta_2 - \beta_3, \ldots, \beta_{K-1} - \beta_K)' \in \mathbb{R}^{\binom{K}{2}p^2}$. Subsequently, criterion (9) is representable in the form of a constrained optimization task:

$$\begin{cases} \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{min}\ f(\boldsymbol{\beta}) + g(\boldsymbol{\gamma}), \\ L\boldsymbol{\beta} = \boldsymbol{\gamma}, \end{cases} \tag{10}$$

where $f(\boldsymbol{\beta}) = ||C - D\boldsymbol{\beta}||_2^2 + \lambda_1|\boldsymbol{\beta}|_1$ is a convex function of $\boldsymbol{\beta} \in \mathbb{R}^{Kp^2}$ and $g(\boldsymbol{\gamma}) = \lambda_2|\boldsymbol{\gamma}|_1$ is a convex function of $\boldsymbol{\gamma} \in \mathbb{R}^{\binom{K}{2}p^2}$. As stated in Boyd et al. (2011), for convex functions f and g oen can devise an ADMM algorithm that guarantees numerical convergence to a local minimum for optimization task (10). Such an algorithm breaks the initial minimization problem (9) into a set of simpler convex optimization tasks, which take the form of the following update rules:

$$\begin{cases} \boldsymbol{\beta}^{(k+1)} = \underset{\boldsymbol{\beta}}{argmin}(f(\boldsymbol{\beta}) + \frac{\rho}{2}||L\boldsymbol{\beta} - \boldsymbol{\gamma}^{(k)} + \mathbf{u}^{(k)}||_2^2), \\ \boldsymbol{\gamma}^{(k+1)} = \underset{\boldsymbol{\gamma}}{argmin}(g(\boldsymbol{\gamma}) + \frac{\rho}{2}||L\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\gamma} + \mathbf{u}^{(k)}||_2^2), \\ \mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + L\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\gamma}^{(k+1)}, \end{cases} \tag{11}$$

where ρ is the step-size term and $\mathbf{u} \in \mathbb{R}^{\binom{K}{2}p^2}$ is the augmented update vector that keeps track of the constraint $L\boldsymbol{\beta} = \boldsymbol{\gamma}$ in (10). While there is literature available on theoretically robust choices for ρ (Ghadimi et al., 2015), we pick the value heuristically after numerical experimentation (more on that given in the supplement). The first optimization task above can be cast into the form of a standard lasso optimization problem by completing the square with respect to $\boldsymbol{\beta}$ and can be efficiently solved with a cyclic coordinate descent algorithm (Tibshirani et al., 2011). The second optimization task has a closed-form solution given by $\boldsymbol{\gamma}^{(k+1)} = s_{\lambda_2/\rho}(L\boldsymbol{\beta}^{(k+1)} + \mathbf{u}^{(k)})$, where $s : \mathbb{R}^{\binom{K}{2}p^2} \to \mathbb{R}^{\binom{K}{2}p^2}$ corresponds to the soft thresholding operator from Donoho (1995). Convergence diagnostics of the proposed algorithm are provided in the supplement.

3.3. Tuning parameter selection

Next, we discuss strategies for selecting the values of the tuning parameters in (3) and (6). Note that for the separate estimation procedure, one simply has to pick the sparsity parameter for each model, which can be done either through cross-validation or by using a heuristic AIC/BIC criterion, akin to similar strategies used in lasso regularized regression problems.

In our numerical experimentation, for the separate estimation method we use a *corrected Akaike Information Criterion* (*AICc*) to pick the estimate from the full solution path for the standard lasso problem (3). The criterion is given by

$$AICc(\lambda_1) = n\log(||W - Z\hat{\beta}_{\lambda_1}||_2^2/n) + 2\ d_{\lambda_1}, \tag{12}$$

where n - length of W, $d_{\lambda_1} = \frac{l+(l+1)(l+2)}{(n-l-2)}$, l - number of non-zero elements in $\hat{\beta}_{\lambda_1}$.

For the joint estimation method, we first set $\lambda_2 \equiv 0$ in criterion (6) and use the AICc criterion of the form (12) to pick the sparsity parameter value $\hat{\lambda}_1$. Subsequently, we set $\lambda_1 = \hat{\lambda}_1$ and perform a grid search for (6) using the following *distinct Bayesian Information Criterion* (*BIC.dist*) to pick the value for the fused lasso parameter λ_2:

$$BIC.dist(\lambda_1, \lambda_2) = n\log(||\boldsymbol{W} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{\lambda_1,\lambda_2}||_2^2/n) + \log(n)\ d_{\lambda_1,\lambda_2}, \tag{13}$$

where n - length of W, d_{λ_1,λ_2} corresponds to the number of distinct non-zero elements in $\hat{\boldsymbol{\beta}}_{\lambda_1,\lambda_2}$.

4. Performance evaluation

We assess the performance of the proposed joint VAR model on simulated data and compare it to that obtained from implementing separate VAR models across the K entities introduced in (Basu et al., 2015a). The performance metrics employed are (i) false positive and false negative rates, as well as the Matthews correlation coefficient (a summary measure of the latter two corresponding to their geometric mean), (ii) accuracy of the estimates of the transition matrices, captured by the normalized Frobenius norm difference between the true transition matrices and their estimates, and (iii) one-step mean forecasting error. The definitions of these quantities are given next (where $\hat{A}^{(k)} = (\hat{a}_{i,j}^{(k)}) \in \mathbb{R}^{p\times p}$ denotes the estimate of $A^{(k)} = (a_{i,j}^{(k)}) \in \mathbb{R}^{p\times p}$, $k = 1, \ldots, K$):

- Matthews correlation coefficient (MC) - geometric mean of false positive (FP) and false negative rates (FN). MC values near 1.0 indicate better estimates of matrix support.

$$MC = \frac{TP \times TN\ -\ FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}},$$

where

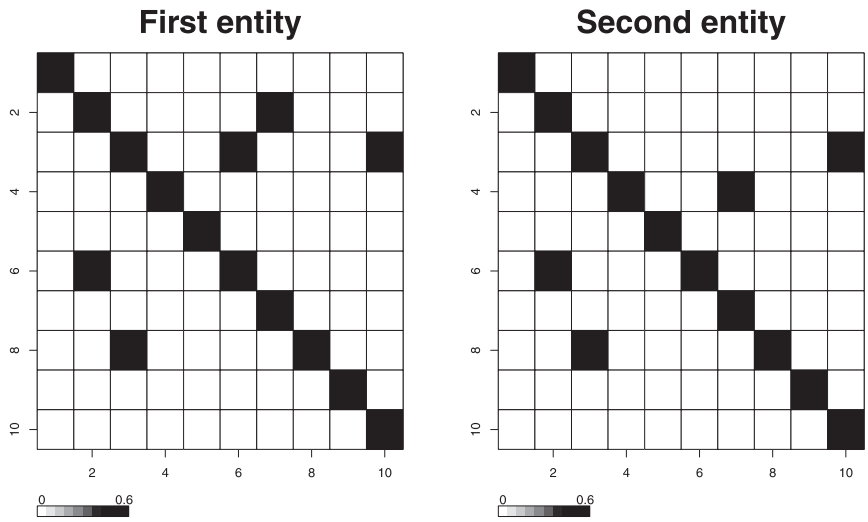

Fig. 1. Generated VAR(1) transition matrices for two entities for the case of $A_1 \sim A_2$.

$$FP = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum\limits_{1\le i<j\le p} I(a_{i,j}^{(k)} = 0, \hat{a}_{i,j}^{(k)} \neq 0)}{\sum\limits_{1\le i<j\le p} I(a_{i,j}^{(k)} = 0)}, \qquad TN = 1 - FP$$

$$FN = \frac{1}{K}\sum_{k=1}^{K}\frac{\sum\limits_{1\le i<j\le p} I(a_{i,j}^{(k)} \neq 0, \hat{a}_{i,j}^{(k)} = 0)}{\sum\limits_{1\le i<j\le p} I(a_{i,j}^{(k)} \neq 0)}, \qquad TP = 1 - FN.$$

- normalized Frobenius difference (NFD) in magnitudes of elements for the estimate and the true matrix. Smaller NFD values point to better accuracy ($||.||_2$ - classic L^2-norm).

$$NFD = \frac{1}{K}\frac{\sum_{k=1}^{K}||A_k - \hat{A}_k||_2^2}{\sum_{k=1}^{K}||A_k||_2^2} = \frac{\sum\limits_{k=1}^{K}\sum\limits_{1\le i<j\le p}(a_{i,j}^{(k)} - \hat{a}_{i,j}^{(k)})^2}{\sum\limits_{k=1}^{K}\sum\limits_{1\le i<j\le p}(a_{i,j}^{(k)})^2}.$$

- one-step mean squared forecast error (MSFE): after training the model on first $T-1$ time points (out of T available), we compare the forecasted values with the actual values for time point T. Lower values of MSFE indicate better forecasting performance. Let $\mathbf{Y} = (Y_1, \ldots, Y_{Kp})' \in \mathbb{R}^{Kp}$ denote a vector of p observed values at time point T combined over K models, $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_{Kp})' \in \mathbb{R}^{Kp}$ - vector of p forecasted values at time point T combined over K models. Then,

$$MSFE = \sum_{i=1}^{Kp}(\hat{Y}_i - Y_i)^2/Kp.$$

Next, we describe the data generation mechanism used in our numerical experiments. Here we focus on 1) *VAR*(1) model, as the value $D = 1$ is the least demanding from a computational standpoint and we do not emphasize the issue of lag selection in this work, 2) the $K = 2$ case, since all the key findings hold for larger values of K. We consider two versions of the transition matrices: (i) identical A_1 and A_2, where we expect that the joint procedure will leverage the implicit larger sample size and provide better estimates than the separate procedure, and (ii) A_1 and A_2 are not identical, but share common patterns as illustrated in Fig. 1 below.

Further, the spectral radius (maximum absolute eigenvalue of the transition matrices $A_k k = 1, 2$ that captures the degree of temporal dependence across all time series; for more details see discussion in (Basu et al., 2015a)) is set to 0.6 and 0.8 (the latter results given in Appendix A.3) and all diagonal elements are non-zero, which is commonly the case in many economic applications. The number of time series considered per entity is $p = 10,\ 20,\ 30$ and the density (percentage of non-zero *non-diagonal elements*) is set to 6, 3 and 2%, respectively. Finally, a measure of signal-to-noise ratio (SNR), defined as $\max_{i,j}|A_{i,j}|/sd(\{X_i^t, t = 1, \ldots, T\})$, is set to 2 throughout.

Next, we describe the error covariance generating mechanism, that follows the approach discussed in Fan et al. (2011). The loadings matrix is given by $\Lambda_{p\times L} = (b)_{ij},\ b_{ij} \sim N(0,1), 1 \le i \le p, 1 \le j \le L$. Further, we considered the following number

Table 1
Results of a simulation study comparing joint (J) and separate (S) estimation methods.

Setting	Method	MSFE	FP	FN	MC	NFD
p=10 T=30	S	0.25 (0.15)	0.12 (0.05)	0.09 (0.06)	0.79 (0.08)	0.64 (0.13)
$A_1 \equiv A_2$	J	0.24 (0.14)	0.03 (0.02)	0.01 (0.03)	0.96 (0.03)	0.35 (0.08)
p=10 T=30	S	0.25 (0.17)	0.11 (0.05)	0.09 (0.06)	0.8 (0.08)	0.59 (0.12)
$A_1 \sim A_2$	J	0.23 (0.14)	0.04 (0.03)	0.03 (0.03)	0.93 (0.04)	0.39 (0.07)
p=20 T=40	S	0.37 (0.21)	0.04 (0.01)	0.04 (0.03)	0.92 (0.03)	0.48 (0.05)
$A_1 \equiv A_2$	J	0.35 (0.2)	0.01 (0)	0 (0.01)	0.99 (0.01)	0.28 (0.05)
p=20 T=40	S	0.36 (0.2)	0.04 (0.01)	0.06 (0.06)	0.9 (0.06)	0.46 (0.06)
$A_1 \sim A_2$	J	0.35 (0.19)	0.01 (0.01)	0.02 (0.03)	0.96 (0.03)	0.32 (0.06)
p=30 T=50	S	0.46 (0.21)	0.02 (0)	0.02 (0.02)	0.96 (0.02)	0.47 (0.04)
$A_1 \equiv A_2$	J	0.44 (0.21)	0 (0)	0 (0)	1 (0)	0.32 (0.04)
p=30 T=50	S	0.51 (0.25)	0.02 (0)	0.05 (0.08)	0.94 (0.08)	0.42 (0.06)
$A_1 \sim A_2$	J	0.49 (0.24)	0 (0)	0.03 (0.03)	0.97 (0.03)	0.32 (0.06)

Note: Pairs of rows correspond to the same setting (e.g. rows 1–2 correspond to $p = 10, T = 30, A_1 \equiv A_2$, rows 3–4 correspond to $p = 10, T = 30, A_1 \sim A_2$ etc). Means and standard deviations are shown for performance metrics discussed at the top of Section 4, with 100 replicates per each setting.

of factors in different settings: $L = 2$ factors for settings with $p = 10$, $L = 3$ for $p = 20$, $L = 4$ for $p = 30$, respectively. Finally, Σ_U is generated as a diagonal matrix, thus ensuring its positive definiteness, and a measure of signal-to-noise ratio, defined as $\frac{\sum_{\lambda_i \in eig(\Lambda\Lambda')} \lambda_i}{\sum_{\lambda_i^U \in eig(\Sigma_U)} \lambda_i^U}$, where *eig* denotes the eigenvalues of the corresponding matrix, is set to 2. Table 1 gives the results for different settings with spectral radius set to 0.6; additional results can be found in Appendix A.3.

The superior performance of the joint approach in terms of overall accuracy can be easily seen by examining the NFD column, which consistently shows a 25–40% improvement over separate method within each setting. Meanwhile, the Matthews coefficient, false positive and false negative rates demonstrate that the joint approach estimates the presence/absence of Granger causal effects across all settings. As for the forecasting performance, the MSFE values appear to be slightly smaller for the joint method overall, but generally comparable. The latter finding can be explained by the propensity of the separate estimates to include more non-zero elements, which leads to overfitting, but also benefits forecasting. In this setting, we are dealing with a sparse model, which the joint approach estimates accurately, while not compromising the forecasting performance. When the spectral radius is 0.8, the broad patterns in overall accuracy (NFD) and estimation of present/absent Granger causal effects hold, while the joint method also exhibits a clearer edge in forecasting (see Table A.7 in Appendix A.3). This could be attributed to the decrease in effective sample size for the separate method due to the presence of additional autocorrelation implied by a higher value for the spectral radius. On the other hand, the joint method by borrowing information across related models compensates for this decrease.

5. Applicationto economic indicators for multiple US states

We illustrate the joint estimation method on time series data of various economic indicators from the following US states: Pennsylvania (PA), Michigan (MI), Ohio (OH) and Illinois (IL). The joint modeling is appropriate due to their similarly large industrial and manufacturing base; however, we also expect these four states to exhibit some differences, due to, amongst other reasons, the presence of a strong financial industry in Chicago, IL and a diversified service sector in the metropolitan Philadelphia, PA area.

The data were obtained from the Federal Reserve Board of St. Louis website (FRED). For each state under consideration, the data set contains seasonally-adjusted monthly time series for 18 economic indicators, spanning the period from December, 2009 to December, 2015, for a total of 70 time points. The period selected corresponds to the post-2008 financial crisis; according to the National Bureau of Economic Research, June 2009 marks the end of the recession as a consequence of the crisis. The variables under study reflect employment data (employee total, average hourly earnings, average weekly working hours) for different sectors (Construction, Education/Health, Financial Activities, Manufacturing, Goods Producing) along with the total of non-farm employees, the leading index and the unemployment rate. The objective is to identify Granger causal effects (cross-dependencies) that are common across all four states, as well as the state specific ones. A detailed description of the variables is given in Table A.8 of Appendix A.4.

We considered different sets of variables in our modeling. The smallest model (Model I) only included the leading index, the unemployment rate, total non-farm employment and employee totals for the five sectors, yielding a total of eight variables. Model IIa was augmented by including the average hourly earnings for the five sectors (13 variables in total), Model IIb replaced the average hourly earnings with the average weekly hours, while Model III included both hourly earnings and weekly hours indicators (18 variables in total). For obtaining both joint and separate estimates of the parameters of the network Granger causality model we used the estimation procedure described in Section 3. In the first step, we separately used factor models for estimating the error covariance matrices for each of the four states, and the results can be seen in Table 2 below. Model I yields a three factor model, while models IIa/b and III yield five and six factor models, respectively,

Table 2
Estimated number $\hat{L}$ of factors in L-factor model for error covariance of each state.

Model setting	p	PA	IL	OH	MI
Model I	8	3	3	3	3
Model IIa	13	5	5	5	5
Model IIb	13	4	5	5	5
Model III	18	5	6	6	6

Table 3
Forecasting errors for separate and joint estimates in the econometric time series study.

Model setting	S	Rep	Univariate AR(1)	Separate	Joint
Model I	30	40	0.65 (0.26)	0.32 (0.19)	0.30 (0.18)
Model IIa	40	30	0.92 (0.23)	0.42 (0.10)	0.44 (0.12)
Model IIb	40	30	0.77 (0.27)	0.30 (0.15)	0.29 (0.14)
Model III	50	20	0.75 (0.17)	0.37 (0.10)	0.37 (0.11)

Note: Means and deviations of MSFE for one-step ahead forecasts of four considered models.

for most states. Therefore, we see a considerable dimension reduction in each case, with estimated number of factors $\widehat{L} \ll p$, where p - number of variables in the model.

Although our focus is on estimation and interpretation of Granger causal effects for the different states under consideration, we nevertheless assess the forecasting performance of the separate and joint estimation methods, since it provides a direct measure of comparison. To obtain multiple measurements of forecast errors we used the following rolling-window strategy. We fix the window size to S months and use both methods to estimate the parameters of the respective models for the period $t = 1, \cdots, S$ and forecast the period $S+1$. Then, we shift the estimation period to $t = 2, \cdots, S+1$ and forecast the period $S+2$, and continue until the period $t = T-S, \cdots, T-1$ that provides forecasts for the period $T = 70$. This approach produces a sequence of $T-S$ forecasts and their deviation from the true values, which allows us to obtain multiple measurements of the one-step mean squared forecasting error (MSFE, introduced in Section 4) for both the separate and joint estimation methods for all four models. The results for the four models, together with the window size S and the number of forecasts, are depicted in the Table 3 below, while the univariate AR(1) model forecasts were used as the benchmark. It can be seen that the performance of both estimation methods is very similar, with a marginal advantage for the joint approach, but both outperforming the standard AR(1) model. Hence, building large size VAR models is beneficial as discussed in the economics literature (see (Bernanke et al., 2005; Bańbura et al., 2010) and references therein).

Next, we discuss the results obtained based on the two estimation methods. In order to overcome the lack of uncertainty measures (which is still the case for most high-dimensional regularized models) for the obtained Granger causal effects estimates, we use a form of *stability selection* (Meinshausen and Bühlmann, 2010) to obtain the more pertinent effects. Specifically, we accumulate multiple estimates of the VAR model using both the separate and joint estimation methods based on the rolling window strategy previously described. The resulting *stability matrices* for both approaches will indicate the proportion of times a Granger causal effect shows up in $T-W$ estimates for the corresponding state within the corresponding model. In Figs. 2 and 3, we provide these results for Model IIb for the four states.

It can be seen that the joint estimation method provides sparser, but more stable estimates. For example, the separate estimation procedure has many Granger causal effects with proportions ranging between 0.4–0.6, which points to the fact that these effects are weaker. On the other hand, the joint method provides more consistent estimates, especially when it comes to the autocorrelation coefficients (diagonal elements). Further, the separate estimates are very dense and hard to interpret, while the joint method provides estimates with a more parsimonious structure, which improves our ability to interpret the results. Setting a threshold of 0.5 for interpretation purposes, we see that the leading indicator index exhibits a strong effect on a number of other variables including the unemployment rate, as well as various employee totals and/or average hourly earnings. It should be noted that there is higher commonality between PA and IL, and OH and MI, which can be interpreted as these two pairs of states having more similar economic fundamentals between them than with members of the other pair. Other notable strong effects detected are as follows: for IL, goods producing employee total impacts the unemployment rate, education/health employee total affects the non-farm employee total, hourly earnings in construction Granger-cause the hourly earnings in goods producing. For OH, hourly earnings for goods producing impact that of manufacturing, and hourly earnings for education/health affect that of finance. For MI, hourly earnings for construction Granger-cause those for goods producing, while hourly earnings for manufacturing impact those for both goods producing and construction. Appendix contains summary tables on both the effects shared among multiple states (Table A.9), and those that are state-specific (Table A.10).

In summary, the joint estimation method provides us with the following features shared across the states: (i) strong estimates for the autocorrelation effects, (ii) hourly earnings in construction appear to impact future hourly earnings in

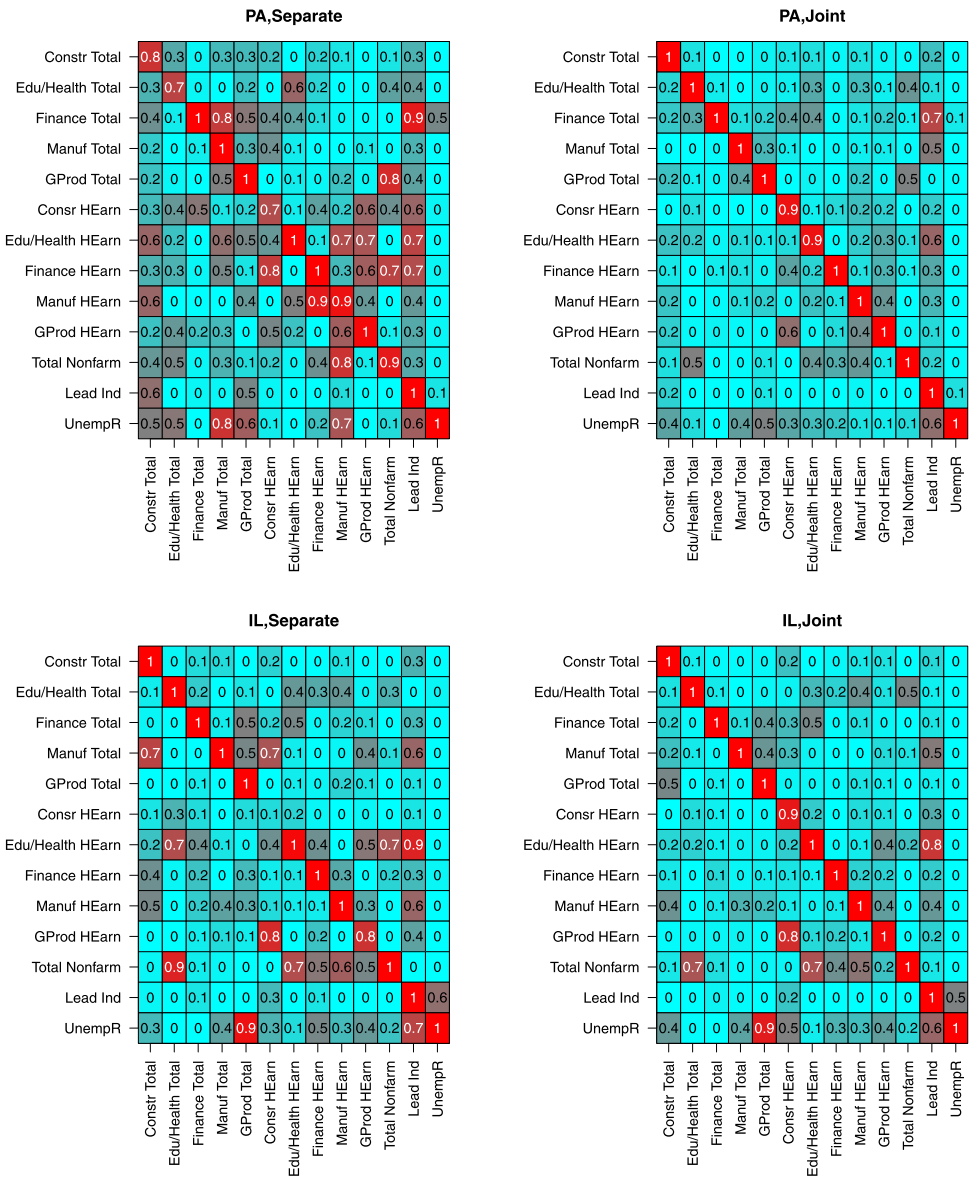

Fig. 2. Proportion of detected Granger effects for Model IIb for the states of PA (top row) and IL (bottom row) for the separate (left) and joint (right) estimation methods. Variable names correspond to shortened descriptions in Table A.8 of Appendix A.4.

goods producing for three states (all but OH), (iii) employee total in education/health sector impacts employee total in non-farm sector, (iv) the lead index has a temporal effect on hourly earnings in education/health sector for IL and OH (also present in PA and MI, but not as strong). Another aspect is that for all the states there are moderate Granger-causal effects from various economic indicators on the unemployment rate. On the other hand, the method also identifies some state specific effects, summarized next. In PA the lead index Granger-causes employee total in finance, in IL hourly earnings in education/health have an effect on employee total in non-farm, in OH hourly earnings in education/health impact finance earnings, earnings in goods producing affect those in manufacturing, and the lead index impacts construction earnings. Finally, in MI the hourly earnings in manufacturing appear to affect earnings in both goods producing and construction.

6. Concluding remarks

In this paper, we examined the problem of jointly estimating multiple *related* network Granger causal models, assuming that the corresponding transition matrices are sparse. The latter would allow us to use the procedure when the number of time series under consideration (and hence the number of parameetrs) exceeds the number of time points available. Further, we assume that the error covariance matrix, that captures contemporaneous dependence between the time series,

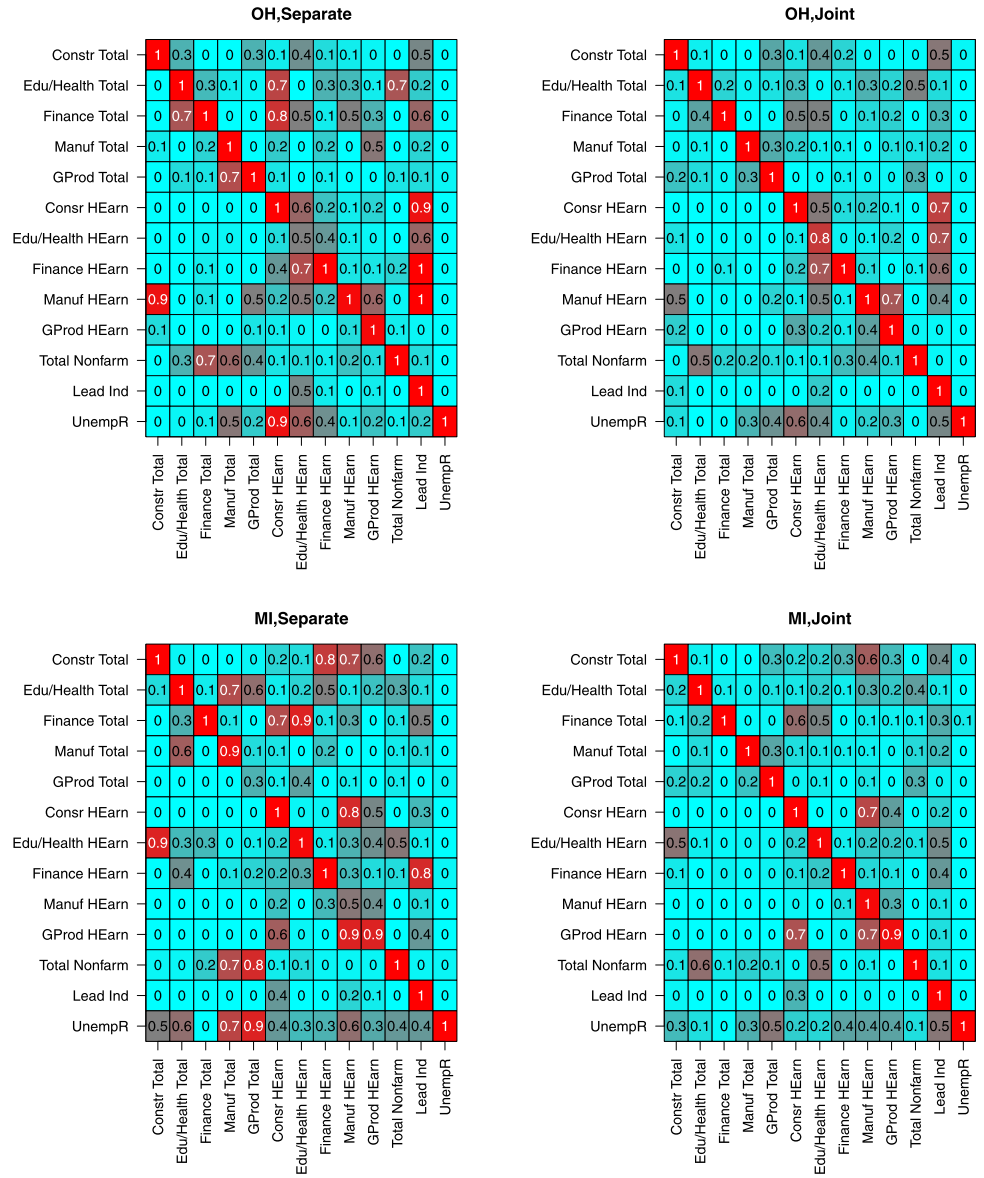

Fig. 3. Proportion of detected Granger effects for Model IIb for the states of OH (top row) and MI (bottom row) for the separate (left) and joint (right) estimation methods. Variable names correspond to shortened descriptions in Table A.8 of Appendix A.4.

is low-rank and can be efficiently approximated by a factor model - a common assumption in macroeconomic and financial applications. Numerical results based on synthetic data show that the joint method, with effective tuning of the parameters controlling the degree of both sparsity and similarity of the estimates across entities, clearly outperforms its counterparts obtained from estimating a Granger causal model for each entity separately. Finally, we illustrate the merits of the proposed modeling framework by jointly estimating key economic indicators of four US states, known to have similar economical infrastructure. The joint approach led to more parsimonious results that were much easier to interpret in comparison with the separate method, while maintaining good forecasting performance.

Acknowledgments

The authors would like to thank the Editor and two anonymous referees for many constructive comments and suggestions. This work was supported in part by a UF Graduate Fellowship and by a National Science Foundation grant DMS-1632730.

Appendix A

A.1. Transition from VAR(D) to standard regression formulation

Basu et al. (2015a) propose the following sequence of transformations to transition from (1) to (2) for a single entity (dropping k from the notation):

$$\underbrace{\begin{pmatrix} (X^T)' \\ \dots \\ (X^D)' \end{pmatrix}}_{\tilde{Y}_{(T-D+1)\times p}} = \begin{pmatrix} \underbrace{\begin{matrix} (X^{T-1})' \dots (X^{T-D})' \\ \dots \quad \dots \quad \dots \\ (X^{D-1})' \quad \dots \quad (X^0)' \end{matrix}}_{X_{(T-D+1)\times Dp}} \end{pmatrix} \underbrace{\begin{pmatrix} (A^1)' \\ \dots \\ (A^D)' \end{pmatrix}}_{\mathbf{A}_{(Dp)\times p}} + \underbrace{\begin{pmatrix} (\epsilon^T)' \\ \dots \\ (\epsilon^D)' \end{pmatrix}}_{\tilde{\epsilon}_{(T-D+1)\times p}},$$

$$\mathrm{vec}(Y) \quad = \mathrm{vec}(XA) + \mathrm{vec}(\tilde{\epsilon}) = (I \otimes X)\mathrm{vec}(A) + \mathrm{vec}(\tilde{\epsilon}),$$

and

$$\underbrace{Y}_{(T-D+1)p\times 1} = \underbrace{Z}_{(T-D+1)p\times Dp^2} \underbrace{\beta}_{Dp^2\times 1} + \underbrace{\tilde{\epsilon}_j}_{(T-D+1)p\times 1}, \quad \tilde{\epsilon}_j \sim N(O, \Sigma \otimes I_{T-D+1}). \tag{.1}$$

A.2. Details of theerror covariance estimation procedure

Step 3 details. The largest eigenvalues retained were picked in the following manner: after calculating the average of all eigenvalues for the empirical covariance matrix, any eigenvalue above that average was deemed as "large". This approach tends to overestimate the true number of factors by one, which is preferred to missing out on the important factors. Alternative procedures that were examined exhibit a higher proportion of instances estimating the number of underlying factors correctly, but also have a considerably higher number of cases of underestimating them (method of sharp drop-off points) or cases with too many factors included (total variance explained). Moreover, our approach demonstrates good performance with respect to total variance explained, consistently accounting for about 70–90% of the variance, while picking the most important factors only. The performance of the approach employed was tested extensively in multiple simulation settings with 100 replicates for each setting. The number of factors L considered and numerical results are given in Table A.4 below. Multiple settings are obtained by varying the signal-to-noise ratio (SNR, as defined in Section 4), the number of variables p per entity, the number of time points T and that of factors L. We show that the technique is robust to increases in SNR.

After identifying $\widehat{L}$ large eigenvalues $\lambda_1, \dots, \lambda_{\widehat{L}}$, we set $\hat{\Lambda}_{p\times\widehat{L}} = V_{p\times\widehat{L}} \times diag(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_{\widehat{L}}})$, where $V_{p\times\widehat{L}} = [v_1, \dots, v_{\widehat{L}}]$, with v_i being the eigenvector corresponding to eigenvalue λ_i, $i = 1, \dots, \widehat{L}$.

Step 4 details. The graphical lasso procedure (Friedman et al., 2008) takes an empirical covariance matrix as input and outputs a sparse estimate for the inverse of that matrix. In our case, the input matrix corresponds to $\hat{\Sigma}_\epsilon - \hat{\Lambda}\hat{\Lambda}'$ and the output will be the sparse $\hat{\Sigma}_U^{-1}$ matrix. Then, we simply take inverse of $\hat{\Sigma}_U^{-1}$ as the estimate $\hat{\Sigma}_U$ of Σ_U. Tuning parameter value for graphical lasso optimization criterion was set to $\log(p)/T$ as a conventional choice discussed in Janková et al. (2015).

Table A.1
Simulation study results for estimating the number of factors in an L-factor model.

Setting	$\widehat{L} = L$ or $L+1$(%)	$\widehat{L} \gg L$ (%)	$\widehat{L} < L$ (%)	$\sum_{i=1}^{\widehat{L}} \lambda_i / \sum_{i=1}^{p} \lambda_i$
SNR = 2				
p=10, T=30, L=2	99	1	0	0.82
p=20, T=40, L=3	99	1	0	0.87
p=30, T=50, L=4	93	7	0	0.88
SNR = 3				
p=10, T=30, L=2	100	0	0	0.81
p=20, T=40, L=3	100	0	0	0.86
p=30, T=50, L=4	92	8	0	0.88

Note: Percentage of estimates $\widehat{L}$ that are (i) exactly correct or overestimating L by one, (ii) overestimating L by two or more, (iii) underestimating L. Last column–proportion of variance explained by $\hat{L}$ factors.

A.3. Additionalsimulation results

We re-ran our simulation study for the same generated data, but addressing the issue of result sensitivity to the tuning parameter selection. In particular, taking into account the estimates for full grid of λ_1 and λ_2 values (instead of just those selected via criteria from Section 3.3), we calculated

- the mean squared forecasting error (MSFE);
- the cumulative area under the ROC curve (AUROC);
- the mean normalized Frobenius norm (NFD).

In the table below, we present the aforementioned metrics averaged across all replicates, alongside their standard deviations. It can be seen that the joint method exhibits a small edge over the separate one in forecasting and structure estimation (MSFE and AUROC, respectively), while completely dominating it in the evaluation of the effect magnitudes (NFD).

Moreover, to judge the performance of our method in case of higher degree of dissimilarity, below we provide the results for simulation settings with all the off-diagonal non-zero elements (6, 3, 2% for $p = 10, 20$ and 30, respectively) being in different positions for the two matrices. First table corresponds to the metrics for single selected estimates (according to criteria from Section 3.3), while the second one describes the aggregate metrics across full grids of tuning parameter values. Both the forecasting and structure estimation performance appear to be comparable for the two methods, with joint method still coming out on top in effect magnitude estimation.

Table A.2
Results of a simulation study comparing joint (J) and separate (S) estimation methods in terms of forecasting (MSFE), area under the classification curve (AUROC) and Frobenius norm (Frob).

Setting	Method	MSFE	AUROC	NFD
p=10 T=30	S	0.36 (0.2)	0.93 (0.03)	0.98 (0.1)
$A_1 \equiv A_2$	J	0.32 (0.19)	0.96 (0.02)	0.73 (0.06)
p=10 T=30	S	0.4 (0.3)	0.93 (0.03)	1 (0.11)
$A_1 \sim A_2$	J	0.36 (0.28)	0.95 (0.02)	0.78 (0.08)
p=20 T=40	S	0.4 (0.19)	0.95 (0.02)	1.06 (0.08)
$A_1 \equiv A_2$	J	0.37 (0.16)	0.96 (0.01)	0.81 (0.05)
p=20 T=40	S	0.41 (0.23)	0.96 (0.01)	1.11 (0.1)
$A_1 \sim A_2$	J	0.36 (0.19)	0.96 (0.01)	0.82 (0.05)
p=30 t=50	S	0.44 (0.16)	0.96 (0.01)	1.1 (0.1)
$A_1 \equiv A_2$	J	0.39 (0.15)	0.96 (0.01)	0.78 (0.04)
p=30 t=50	S	0.46 (0.23)	0.96 (0.02)	1.15 (0.22)
$A_1 \sim A_2$	J	0.42 (0.23)	0.94 (0.02)	0.84 (0.06)

Table A.3
Results of a simulation study comparing joint (J) and separate (S) estimation methods for highly dissimilar matrices in terms of forecasting (MSFE), area under the classification curve (AUROC) and Frobenius norm (Frob).

Setting	Method	MSFE	AUROC	Frob
p=10 t=30	S	0.34 (0.18)	0.91 (0.05)	1.18 (0.34)
$A_1 \neq A_2$	J	0.31 (0.17)	0.91 (0.04)	0.86 (0.14)
p=20 t=40	S	0.42 (0.24)	0.92 (0.06)	1.51 (0.6)
$A_1 \neq A_2$	J	0.4 (0.23)	0.9 (0.05)	1 (0.27)
p=30 t=50	S	0.38 (0.17)	0.93 (0.06)	1.46 (0.57)
$A_1 \neq A_2$	J	0.36 (0.17)	0.91 (0.04)	0.95 (0.21)

Table A.4
Simulation study for joint (J) and separate (S) methods, spectral radius 0.8.

Setting	Method	MSFE	FP	FN	MC	NFD
p=10 T=30	S	0.45 (0.3)	0.13 (0.06)	0.02 (0.03)	0.85 (0.07)	0.5 (0.12)
$A_1 \equiv A_2$	J	0.42 (0.28)	0.04 (0.03)	0 (0)	0.96 (0.03)	0.25 (0.05)
p=10T=30	S	0.48 (0.34)	0.12 (0.04)	0.02 (0.03)	0.87 (0.06)	0.44 (0.08)
$A_1 \sim A_2$	J	0.45 (0.31)	0.05 (0.04)	0.01 (0.02)	0.94 (0.04)	0.29 (0.07)
p=20 T=40	S	0.66 (0.32)	0.04 (0.02)	0.01 (0.02)	0.95 (0.02)	0.38 (0.06)
$A_1 \equiv A_2$	J	0.61 (0.3)	0.01 (0.01)	0 (0.01)	0.99 (0.01)	0.22 (0.04)
p=20 T=40	S	0.7 (0.35)	0.04 (0.01)	0.01 (0.02)	0.95 (0.02)	0.36 (0.05)
$A_1 \sim A_2$	J	0.65 (0.32)	0.02 (0.01)	0.01 (0.01)	0.98 (0.01)	0.25 (0.04)
p=30 t=50	S	0.92 (0.42)	0.02 (0.01)	0 (0.01)	0.98 (0.01)	0.36 (0.03)
$A_1 \equiv A_2$	J	0.86 (0.4)	0 (0)	0 (0)	1 (0)	0.24 (0.02)
p=30t=50	S	0.96 (0.44)	0.02 (0)	0.01 (0.01)	0.98 (0.01)	0.32 (0.04)
$A_1 \sim A_2$	J	0.91 (0.42)	0.01 (0)	0 (0.01)	0.99 (0.01)	0.24 (0.03)

Table A.5
Set of abbreviations used in FRED data set.

Abbreviation*	Description	Units
ILCONS	Employee Total: Construction in IL	Thousands of Persons
ILEDUH	Employee Total: Education and Health in IL	Thousands of Persons
ILFIRE	Employee Total: Financial Activities in IL	Thousands of Persons
ILMFG	Employee Total: Manufacturing in IL	Thousands of Persons
ILNA	Employee Total: Total Non-Farm in IL	Thousands of Persons
ILSLIND	Leading Index for IL	Percent
ILUR	Unemployment Rate in IL	Percent
SMS17000000600000001	Employee Total: Goods Producing in IL	Thousands of Persons
SMU170000006000000002SA	WeeklyH: Goods Producing in IL	Hours
SMU170000006000000003SA	HEarn: Goods Producing in IL	Dollars per Hour
SMU170000020000000002SA	WeeklyH: Construction in IL	Hours
SMU170000020000000003SA	HEarn: Construction in IL	Dollars per Hour
SMU170000030000000002SA	WeeklyH: Manufacturing in IL	Hours
SMU170000030000000003SA	HEarn: Manufacturing in IL	Dollars per Hour
SMU170000055000000002SA	WeeklyH: Financial Activities in IL	Hours
SMU170000055000000003SA	HEarn: Financial Activities in IL	Dollars per Hour
SMU170000065000000002SA	WeeklyH: Education and Health Services in IL	Hours
SMU170000065000000003SA	HEarn: Education and Health Services in IL	Dollars per Hour

Note: WeeklyH - average weekly hours of all employees in respective sector (e.g. Construction), HEarn - average hourly earnings of all employees, IL - Illinois.

Table A.7 below shows simulation results for spectral radius of 0.8, rest of the settings and notations are the same as described in Section 4.

A.4. Description and abbreviations of variables extracted from the FRED website

Below we provide variable abbreviations together with their description (for the state of Illinois) for the data from Section 5.

* There might be some inconsistencies in abbreviations across the states.

A.5. Summary tables for shared and state-specific effects.

Below follows the table summarizing numbers of effects shared by multiple states, alongside with some select examples of those effects.

On the other hand, the table below describes the effects that were specific to certain states. Ohio showed the largest number of distinguished temporal effects (8), while the other three states had a few of their own as well.

Table A.6
Numbers and examples of effects shared by multiple states.

Shared by	# of effects	Examples of effects
All four states	3	Lead Index → UnempR, Lead Index → Edu/Health Total, ..
At least three states	6	GProd Total → UnempR, Constr HEarn → GProd HEarn, ..
At least two states	11	Constr HEarn → UnempR, Manuf Total → Lead Ind, ..

Table A.7
Numbers and examples of effects specific to a particular state.

State	# of effects	Examples of effects
PA	2	Total Nonfarm → GProd Total, Lead Ind → Finance Total, ..
IL	3	Constr Total → GProd Total, Manuf HEarn → Total Nonfarm, ..
OH	8	GProd HEarn → Manuf HEarn, Lead Ind → Finance HEarn,..
MI	4	Constr Total → Edu/Health HEarn, Manuf HEarn → GProd HEarn, ..

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2018.08.001.

References

Anderson, B.D., Deistler, M., Chen, W., Filler, A., 2012. Autoregressive models of singular spectral matrices. Automatica 48 (11), 2843–2849.
Bai, J., Ng, S., et al., 2008. Large dimensional factor analysis. Found. Trends® Econ. 3 (2), 89–163.
Bańbura, M., Giannone, D., Reichlin, L., 2010. Large Bayesian vector auto regressions. J. Appl. Econ. 25 (1), 71–92.
Basu, S., Michailidis, G., et al., 2015a. Regularized estimation in sparse high-dimensional time series models. Ann. Stat. 43 (4), 1535–1567.
Basu, S., Shojaie, A., Michailidis, G., 2015. Network Granger causality with inherent grouping structure. J. Mach. Learn. Res. 16 (1), 417–453.
Bernanke, B.S., Boivin, J., Eliasz, P., 2004. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. Technical Report. National Bureau of Economic Research.
Bernanke, B.S., Boivin, J., Eliasz, P., 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. Q. J. Econ. 120 (1), 387–422.
Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. 3 (1), 1–122.
Cai, T.T., Li, H., Liu, W., Xie, J., 2015. Joint estimation of multiple high-dimensional precision matrices. Ann. Stat. 38, 2118–2144.
Dahlhaus, R., Eichler, M., 2003. Causality and Graphical Models in Time Series Analysis. Oxford Statistical Science Series, pp. 115–137.
Deistler, M., Filler, A., Funovits, B., 2011. AR systems and ar processes: the singular case. Commun. Inf. Syst. 11 (3), 225–236.
Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. J. Econ. 130 (2), 337–364.
Donoho, D.L., 1995. De-noising by soft-thresholding. IEEE Trans. Inf. Theory 41 (3), 613–627.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Finan. Econ. 33 (1), 3–56.
Fan, J., Liao, Y., Mincheva, M., 2011. High dimensional covariance matrix estimation in approximate factor models. Ann. Stat. 39 (6), 3320.
Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 75 (4), 603–680.
Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1.
Ghadimi, E., Teixeira, A., Shames, I., Johansson, M., 2015. Optimal parameter selection for the alternating direction method of multipliers (admm): quadratic problems. IEEE Trans. Autom. Control 60 (3), 644–658.
Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econ. J. Econ. Soc. 37 (3), 424–438.
Guo, J., Levina, E., Michailidis, G., Zhu, J., 2011. Joint estimation of multiple graphical models. Biometrika 98 (1), 1–15.
Hallin, M., Liška, R., 2011. Dynamic factors in the presence of blocks. J. Econ. 163 (1), 29–41.
Janková, J., Van De Geer, S., et al., 2015. Confidence intervals for high-dimensional inverse covariance estimation. Electron. J. Stat. 9 (1), 1205–1229.
Lin, J., Michailidis, G., 2017. Regularized estimation and testing for high-dimensional multi-block vector-autoregressive models. J. Mach. Learn. Res. 18 (1), 4188–4236.
Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer Science & Business Media.
Ma, J., Michailidis, G., 2016. Joint structural estimation of multiple graphical models. J. Mach. Learn. Res. 17 (166), 1–48.
Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 72 (4), 417–473.
Michailidis, G., d'Alché Buc, F., 2013. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. Mathematical biosciences 246 (2), 326–334.
Rudebusch, G.D., 2010. Macro-finance models of interest rates and the economy. The Manchester School 78 (s1), 25–52.
Song, S., Zhan, Z., Long, Z., Zhang, J., Yao, L., 2011. Comparative study of svm methods combined with voxel selection for object category classification on fmri data. PloS One 6 (2), e17191.
Stock, J.H., Watson, M., 2011. Dynamic factor models. In: Andersen, M.P., Clements, D.F., Hendry (Eds.), Oxford Handbook on Economic Forecasting. Oxford University Press, Oxford.
Stock, J.H., Watson, M.W., 2002. Forecasting using principal components from a large number of predictors. J. Am. Stat. Assoc. 97 (460), 1167–1179.
Stock, J.H., Watson, M.W., 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In: Handbook of Macroeconomics, 2. Elsevier, pp. 415–525.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodological) 58 (1), 267–288.
Tibshirani, R.J., Taylor, J.E., Candes, E.J., Hastie, T., 2011. The Solution Path of the Generalized Lasso. Stanford University.