Estimation of Multi-Granger Network Causal Models

Andrey Skripnikov*[1] and George Michailidis†[1]

[1]Department of Statistics, University of Florida

Abstract

Network Granger causality focuses on estimating Granger causal effects from multivariate time series and it can be operationalized through Vector Autoregressive Models (VAR). The latter represent a popular class of time series models that has been widely used in applied econometrics and finance and more recently in biomedical applications. In this work, we discuss joint estimation and model selection issues of multiple Granger causal networks. We present a modeling framework for the setting where the same variables are measured on different entities (e.g. same set of economic activity variables for related countries). The framework involves the introduction of appropriate structural penalties on the transition matrices of the respective VAR models that link the underlying network Granger models and use of factor modeling for error covariance estimation. ADMM algorithm is presented for implementation of joint optimization procedure and the model is evaluated on synthetic data.

Key words: Granger, Network, VAR, Factor model, Lasso, ADMM

1 Introduction

There has been a lot of recent interest in the modeling and analysis of high-dimensional time series data. Application areas include financial data [20] [Vyrost, 2014], medical data [12] [Flamm, 2012], brain fMRI data [17] [Song, 2010], gene regulatory network inference [15][Michailidis, 2013], macroeconomic time series forecasting and structural analysis [1][Banbura, 2010], just to name a few. Their common characteristic is the relatively large number of variables being analyzed, relative to the time points available, thus leading to a high-dimensional problem. In many cases, the temporal dynamics of the data under consideration are well captured by autoregressive models and hence the use of vector autoregressive models (VAR) enables the modeling of both the variables own temporal dynamics, as well as temporal linear cross-dependencies amongst them. VAR models are closely related to the notion of network of Granger causality as discussed in [3][Basu,Shojaie and Michailidis, JMLR 2015]. However, in the presence of a large number of variables and few time points, one needs to incorporate sparsity assumptions to estimate the parameters of the VAR model (see [2][Basu, 2015]).

*usdandres@ufl.edu

†gmichail@ufl.edu

However, in many applications one deals with multiple *related VAR models*. As a motivating application, consider the data analyzed in Section 6. It deals with a number of employment and economic indicator variables for four US states (Pennsylvania, Michigan, Ohio and Illinois) that exhibit similarities regarding their economic infrastructure with a strong manufacturing base, a fairly large agricultural sector, as well as 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 extend the modeling framework to allow for *joint estimation* of multiple related VAR models. The problem of joint estimation has received attention in the literature recently, primarily focused on the estimation of multiple graphical models that ;everaged various penalites that encouraged both sparsity and joint estimation of the parameters of the multiple models; see for example, the hierarchical penalty used in [14] [Guo et al, Biometrica 2011], the group lasso penalty in [8][Danaher et al., 2012], or mixed norm penalties in [7] [Cai, T. et al, 2015].

Next, we introduce the proposed modeling framework. We consider p stationary time series $X_k^t = (X_{1k}^t, ..., X_{pk}^t)'$ for $k = 1, ..., K$ related phenomena. Then, the corresponding VAR model with lag order 1 is given by:

$$X_k^t = A_k^1 X_k^{t-1} + \varepsilon_k^t, \quad k = 1, \cdots, K, \tag{1}$$

where the error terms follow a normal distribution; $\varepsilon_k^t \sim N(0, \Sigma_k)$. The covariance matrix Σ_k allows for additional latent contemporaneous dependence between the p variables under consideration. The standard assumption is that Σ_k is diagonal and thus no extra dependence is allowed; however, in [2][Basu, 2015], it was assumed that Σ_k is a general *sparse* covariance matrix. In this work, we assume that Σ_k is low rank, stemming from a factor model formulation of the error processes ε_k^t. Such a modeling assumption is widely used in economics and finance applications as discussed in a number of papers [9][Diebold, 2005],[16][Rudebusch, 2010],[10][Fan et al., 2011]. Although it leads to significant reduction of the parameters to be estimated, it nevrtheless poses a number of challenges in a high-dimensional setting. Finally, we employ a fused lasso penalty ([19][Tibshirani et al., 2011]) to connect the estimation of the K VAR models and presented in detail in the ensuing section.

Hence, the main contributions of this work are the development of a joint estimation modeling framework for multiple related VAR models, together with the development of fast scalable algorithms for the estimation of their parameters. The remainder of the paper is organized as follows: in Section 2, the modeling framework is introduced along with the proposed optimization algorithm. Section 3 provides details of the estimation of the low rank covariance matrices and additional algorithmic details, while Section 4 presents the performance evaluation of the proposed modeling framework applied to synthetic data and the motivation application discussed above.

The rest of the paper is organized as follows: in Chapter 2 the modeling framework is introduced along with the ADMM algorithm, in Chapter 3 we discuss aspects of error covariance estimation, Chapter 4 contains full estimation algorithm layout while Chapters 5 and 6 describe results of our modeling approach on synthetic and real data respectively. Conclusion can be found in Chapter 7.

2 Single and Joint Model Setup

We will present model setups for both single and joint estimation approaches for the case of general K number of entities. Suppose that we observe p variables per entity over T time points.

2.1 Single Model

For single model assume that $X^t = (x_1^t, ..., x_p^t)'$ is a vector of variable values at time t for one entity, $t = 1, ..., T$. A will denote $p \times p$ transition matrix, Σ - $p \times p$ error covariance matrix.

Model setup:

$$X^t = AX^{t-1} + \varepsilon^t, \varepsilon^t \sim N(O, \Sigma),\ t = 1, .., T \tag{2}$$

We assume a sparse A because the model is intended for high-dimensional cases(at least 20 variables). Next assumption is for error covariance matrix to follow a factor model. We claim that for economic/finance variable it is more reasonable to assume for error covariance to be driven by a low number of common factors rather than just having a sparse inverse. The factor model will be explained in more detail in chapter 3. Besides, for simplicity purposes, error process $\{\varepsilon^t\}$ is considered covariance-stationary and uncorrelated over time.

Problem (2) has an equivalent formulation as a standard regression problem:

$$W = Z\beta + \underline{\varepsilon}, \underline{\varepsilon} \sim N(O, \tilde{\Sigma}) \tag{3}$$

This can be achieved by letting:

- $W = (\underline{X}_1, ..., \underline{X}_p)'$, where $\underline{X}_i = (X_i^T, ..., X_i^1)', i = 1, ..., p$
- $Z_{T \times p} = (\underline{X}_1^{(-T,+0)}, ..., \underline{X}_p^{(-T,+0)})'$, where $\underline{X}_i^{(-T,+0)} = (X_i^{T-1}, ..., X_i^0)'$
- $Z = I_{p \times p} \otimes W_{T \times p}$. Let $\beta = (A_{11}, A_{12}, ..., A_{1p}, A_{21}, ..., A_{2p}, ..., A_{pp})'$ (matrix A stretched into a vector)
- $\underline{\varepsilon} = (\underline{\varepsilon}_1, ..., \underline{\varepsilon}_p)'$, where $\underline{\varepsilon}_i = (\varepsilon_i^T, ..., \varepsilon_i^1)$, $i = 1, ..., p$
- $\tilde{\Sigma} = \Sigma \otimes I_{T \times T}$

In case of a known true $\tilde{\Sigma}$ the optimization criterion is a standard lasso problem that can be solved with least angle approach or coordinate descent algorithm:

$$\min_{\beta} ||\tilde{\Sigma}^{-1/2}(W - Z\beta)||_2^2 + \lambda ||\beta||_1. \tag{4}$$

2.2 Joint Model

Let $X_i^t = X^t = (x_{i,1}^t, ..., x_{i,p}^t)'$ - vector of variable values for entity $i, i = 1, ..., K$ at time $t, t = 1, ..., T$. Model setup for K different entities:

$$\begin{pmatrix} X_1^t \\ ... \\ X_K^t \end{pmatrix} = A \begin{pmatrix} X_1^{t-1} \\ ... \\ X_K^{t-1} \end{pmatrix} + \varepsilon^t, \varepsilon^t \sim N(O, \Sigma_{Kp \times Kp}), t = 1, .., T \tag{5}$$

where $A_{Kp\times Kp}$ - block-diagonal with i^{th} block equal to a $p\times p$ matrix A_{ii},$i=1,...,K$, $\varepsilon^t=(\varepsilon_1^t,...,\varepsilon_K^t)$

For matrices $A_{ii},i=1,...,K$ we make assumptions of similarity and sparsity. By similarities we mean similar structure(positions of non-zero elements) and similar values of corresponding elements in matrices. Although we are dealing with different entities we still expect to see common patterns of relationship between p variables on those entities and there will be plenty of common zero elements because of sparsity. $\Sigma_{ii},i=1,...,K$ follow a factor model by the same logic as described for single model. And the final simplifying assumption - $\Sigma_{Kp\times Kp}$ is block-diagonal with ith block equal to a $p\times p$ matrix Σ_{ii}.

Via analogous set of assignments as for single model we can get a standard regression setup for the joint problem(equivalent to equations (6)):

$$W=Z\beta+\underline{\varepsilon},\underline{\varepsilon}\sim N(O,\tilde{\Sigma}), \tag{6}$$

where Z - block-diagonal with i^{th} block being $T\times p$ matrix Z_{ii}, $\beta=(\beta_{11},\beta_{22},...,\beta_{KK})'$ with β_{ii} being matrix A_{ii} stretched into a vector, $\underline{\varepsilon}=(\underline{\varepsilon}_1,...,\underline{\varepsilon}_K)$, $\underline{\varepsilon}_i$ - error vector for entity i. $\tilde{\Sigma}$ - block-diagonal matrix with ith block equal to $\tilde{\Sigma}_{ii}$ defined as in section 2.1.

The optimization criterion is a generalized fused lasso problem:

$$||\tilde{\Sigma}^{-1/2}(W-Z\beta)||_2^2+\sum_{i=1}^{K}\lambda_i||\beta_i||_1+\sum_{i,j\in 1,..,K}\lambda_{i,j}||\beta_{ii}-\beta_{jj}||_1, \tag{7}$$

The ADMM algorithm to solve problem (7) will be introduced in section 2.4.

2.3 Case of K=2

In that paper we emphasise the case of two entities. Let $X^t=(x_1^t,...,x_p^t)^{'}$, where x_i^t - value of variable i at time t, $Y^t=(y_1^t,...,y_p^t)^{'}$, where y_i^t - value of variable i at time t, $\varepsilon^t=(\varepsilon_1^t,...,\varepsilon_{2p}^t)^{'}$, the vector of observation errors at time t.

Model setup:

$$\begin{pmatrix}X^t\\Y^t\end{pmatrix}=A\begin{pmatrix}X^{t-1}\\Y^{t-1}\end{pmatrix}+\varepsilon^t,\varepsilon^t\sim N(O,\Sigma),t=1,..,T \tag{8}$$

where

$$A_{2p\ x\ 2p}=\begin{pmatrix}A_{11} & O_{p\times p}\\ O_{p\times p} & A_{22}\end{pmatrix}, \tag{9}$$

The standard regression setup for the joint problem(equivalent to equations (8)):

$$\begin{pmatrix}Y_1\\Y_2\end{pmatrix}=\begin{pmatrix}Z_1 & O\\ O & Z_2\end{pmatrix}\beta+\underline{\varepsilon},\underline{\varepsilon}\sim N(O,\tilde{\Sigma}), \tag{10}$$

where $\tilde{\Sigma}=\begin{pmatrix}\tilde{\Sigma}_{11} & O\\ O & \tilde{\Sigma}_{22}\end{pmatrix}$, $\beta_{2p^2\times 1}=\begin{pmatrix}\beta_{11}\\ \beta_{22}\end{pmatrix}$.

The optimization criterion would be(simplified down to just one sparsity parameter and one fusion parameter):

$$|| \begin{pmatrix} (\tilde{\Sigma}_{11})^{-1/2} Y_1 \\ (\tilde{\Sigma}_{22})^{-1/2} Y_2 \end{pmatrix} - \begin{pmatrix} (\tilde{\Sigma}_{11})^{-1/2} Z_1 \beta_{11} \\ (\tilde{\Sigma}_{22})^{-1/2} Z_2 \beta_{22} \end{pmatrix} ||_2^2$$

$$+ \lambda_1 ||\beta||_1 + \lambda_2 ||\beta_{11} - \beta_{22}||_1, \quad (11)$$

2.4 ADMM Algorithm for K entities

To solve this optimization criterion for arbitrary choice of (λ_1, λ_2) we introduce an ADMM algorithm.

The criterion (11) can be written in the following form:

$$\min_{\beta} ||C - D\beta||_2^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||L\beta||_1, \quad (12)$$

with L such that $L\beta = (\beta_{11} - \beta_{22}, \beta_{11} - \beta_{33}, ..., \beta_{11} - \beta_{KK}, \beta_{22} - \beta_{33}, ..., \beta_{22} - \beta_{KK}, ..., \beta_{K-1,K-1} - \beta_{KK})^T$.

We can represent (12) in the following form:

$$\begin{cases} \min\limits_{\beta,\gamma} f(\beta) + g(\gamma) \\ L\beta = \gamma, \end{cases},$$

where $f(\beta) = ||C - D\beta||_2^2 + \lambda_1 ||\beta||_1$ - convex function of β,
$g(\gamma) = \lambda_2 ||\gamma||_1$ - convex function of γ.

By [6][Boyd et al, 2011] ADMM algorithm with following update rules will break down our optimization problem in a set of simpler convex problems:

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

First equation is a lasso optimization problem with respect to β that can be solved with least angle approach or coordinate descent algorithm.

Second equation has closed form solution:

$$\gamma^{(k+1)} = s_{\lambda_2/\rho}(L\beta^{(k+1)} + u^{(k)}),$$
$$\text{where } s_b(a) = \begin{cases} sign(a)|a - b|, |a| > |b|, \\ 0, |a| \leq |b| \end{cases}$$

For convergence diagnostics of the algorithm please check Appendix A.

3 Error covariance estimation

We will compare three approaches for error inverse covariance estimation: graphical lasso, factor models and ensemble of the two.

3.1 Graphical lasso

Graphical lasso is a very popular approach for sparse estimation of inverse covariance matrix described in [13][Friedman et al.,2007].In my case I apply graphical lasso to estimate inverse error covariance matrices for each entity separately. For tuning parameter I pick value $\lambda = \sqrt{\frac{\log(p)}{t}}$ which is a theoretically approved choice[11][Fan et al., 2013].

3.2 Factor model

In particular, we implement a latent factor setup which assumes unobserved factors. Estimation of Σ_{11} and Σ_{22} is done separately and both procedures are identical therefore it is satisfactory to describe it for one of the classes.

The error vector is assumed to have the following structure:

$$\varepsilon^t = \Lambda F_t + \varepsilon_U,\ \varepsilon_U \sim N(0,\Sigma_U),\ cov(F_t) = I_K, t = 1,...,T$$

$$\Sigma = \Lambda\Lambda' + \Sigma_U, \tag{13}$$

where F_t - $K \times 1$ vector of unobserved factors(K - number of factors, $K < p$), Λ - $p \times K$ matrix of factor loadings, Σ_U - $p \times p$ matrix with a sparse inverse(idiosyncratic component).

In order to get estimate $\hat{\Sigma}_{11}$ we will have to get estimates $\hat{\Lambda}_{11}$ and $\hat{\Sigma}_U$ and the following algorithm will be used for that:

- *Step 1.* For single model problem set $\Sigma = I_{p\times p}$, get sparse estimate $\hat{A}$ (criterion for picking sparse estimates will be described in Chapter 4).
- *Step 2.* Get residuals $\hat{\varepsilon}_1 = W - Z\hat{A}$, calculate the number H of spiked eigenvalues for the residuals covariance matrix $\hat{\Sigma}_\varepsilon$ - it will act as an estimate of a number of latent factors for the corresponding factor model.
- *Step 3.* Do eigenvalue decomposition for this matrix: eigenvectors corresponding to spiked eigenvalues will act as columns of $\hat{\Lambda}$.
- *Step 4.* Use $\hat{\Sigma}_\varepsilon - \hat{\Lambda}\hat{\Lambda}'$ as data to get $\hat{\Sigma}_U$ through a *graphical lasso* procedure(which works well under assumption of sparse inverse).
- *Step 5.* Get the estimate $\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}' + \hat{\Sigma}_U$, take its inverse and get $\hat{\hat{\Sigma}}^{-1}$.

3.3 Ensemble of graphical lasso and factor model.

It turns out that factor models do a good job of estimating off-diagonal elements of inverse covariance while graphical lasso does better on diagonals. Therefore, for each entity I just take factor model estimate of inverse error covariance and switch it's diagonal elements to the diagonal of graphical lasso estimate of inverse error covariance for the same entity.

4 Model selection and full algorithm layout

4.1 Model selection

Here we discuss tuning parameter selection for optimization criterions of separate and joint methods. In both cases cross-validation proves way too demanding considering the size of matrices in the regression setup while BIC picks models that are too sparse to be true.

For separate method the following AIC criterion has proven to be efficient:

$$AIC(\lambda_1) = n\log(||W - Z\hat{\beta}_{\lambda_1}||_2^2/n) + 2\, df_{\lambda_1}, \tag{14}$$

where df_{λ_1} - number of distinct non-null coefficients of $\hat{\beta}_{\lambda_1}$

For joint method this particular AIC was struggling to pick up on similarities between classes even in cases of simulated data with identical transition matrices for both classes - it would pick a very small λ_2 and virtually estimate transition matrices separately. After heuristically trying out different versions of *AIC* criterion(which included running simulations and comparing estimates with the true transition matrices), the following formula yielded best results:

$$AIC(\lambda_1, \lambda_2) = n\log(||W - Z\hat{\beta}_{\lambda_1,\lambda_2}||_2^2/n) + 3\, df_{\lambda_1,\lambda_2}, \tag{15}$$

where df_{λ_1,λ_2} - number of distinct non-null coefficients of $\hat{\beta}_{\lambda_1,\lambda_2}$

4.2 Full estimation algorithm

The steps for our main estimation algorithm are the following:

- *Step 1*. Use algorithm from Chapter 3 to get $\hat{\Sigma}_{11}$ and $\hat{\Sigma}_{22}$.
- *Step 2*. Insert these estimates into optimization problems of form (5) for separate method and into optimization problem of form (8) for joint method.
- *Step 3*. Use cyclical coordinate descent algorithm(implemented in *R* package *glmnet*) to come up with solution path for λ_1 for separate optimization problems and use criterion (12) to pick the estimate
- *Step 4*. Do a sequential search of tuning parameters λ_1 and λ_2:
  - Step 4.1. Initialize λ_2 with $\hat{\lambda}_2 = 0$.
  - Step 4.2. Do a 1D grid search for λ_1 using criterion (13) with λ_2 fixed. Get $\hat{\lambda}_1$.
  - Step 4.3. Do a 1D grid search for λ_2 using criterion (13) with $\lambda_1 = \hat{\lambda}_1$ fixed. Get $\hat{\lambda}_2$. Go back to step 4.2. Repeat a couple of times.

5 Simulation study

We test the performance of our joint model on simulated data and compare the results with separate estimation approach. In particular, we look at how well was the transition

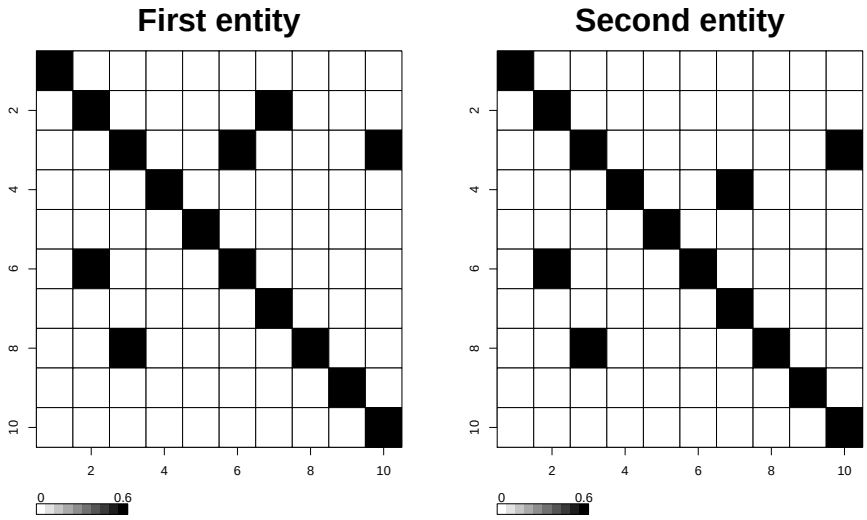


Figure 1: Generated transition matrices for first enitiy(left) and second entity(right) for the case of $A_{11} \sim A_{22}$

matrix structure estimated(positioning of non-zero elements), how close were the estimated non-zero element values to true non-zero values. First one is measured by False Positive rate, False Negative rate and Matthews Coefficient, while the second aspect is captured via Normalized Frobenius Difference between the estimate and the true matrix. Also, the predictive performance is compared by looking at one-step Mean Squared Forecasting Error. All of these measures will be described in detail in subsection 5.2.

While the main goal of our model is to estimate transition matrices, it can't really be done without proper error covariance estimation(or its inverse). Therefore we demonstrate the performance of all three approaches described in section 3 and compare the resulting Normalized Frobenius Differences.

5.1 Generation mechanism

Transition matrices A_{11} and A_{22} from (8) & (9) are generated with maximum eigenvalue 0.6 so that the resulting VAR model is stationary. We look at two cases: $A_{11} \equiv A_{22}$ and $A_{11} \sim A_{22}$(similar matrices but with more heterogeneity introduced).

For generation of transition matrices we have the following settings:

- signal to noise ratio equals $\frac{max_{i,j}|A_{i,j}|}{sd(\{X_i^t, t=1,\ldots,T\})} = 2$
- edge density of transition matrices(percentage of non-zero off-diagonal elements) A_{11} and A_{22} varies depending on number of variables: 5% for $p = 10$(about 4-5 non-zero off-diagonal elements), 3% for $p = 20$(11-12 elements), 1-2% for $p = 30$(12-18 elements)

For the case $A_{11} \sim A_{22}$ we generate matrices with certain amount of shared non-zero off-diagonal elements between A_{11} and A_{22}. Afterwards we add a certain amount of non-zero elements with randomly generated positions which leads to structural heterogeneity between the two matrices. Example can be seen in Figure 1.

To generate error covariance matrix that follows factor model($\Sigma = \Lambda\Lambda' + \Sigma_U$) we use approach introduced in [10][Cai et al.,2011]:

- $\Lambda_{p\times H} = (b)_{ji},\ b_{ji} \sim N(0,1), j \leq p, i \leq H,$
- H(number of factors) = 1
- For Σ_U - generated a diagonal matrix to make sure that the inverse is sparse and the signal of $\Lambda\Lambda'$ to Σ_U is not small(restricted it to be between 1.5 and 3)

5.2 Performance measures

Assume that $\hat{A}_{11} = (\hat{a}_{i,j})^{(1)}$ and $\hat{A}_{22} = (\hat{a}_{i,j})^{(2)}$ are the final estimates of $A_{11} = (a_{i,j})^{(1)}$ and $A_{22} = (a_{i,j})^{(2)}$ respectively. To measure the quality of those estimate we look at:

- Normalized Frobenius Difference:

$$NFD = \frac{||\hat{A}_{11} - A_{11}||_2^2 + ||\hat{A}_{22} - A_{22}||_2^2}{||A_{11} + A_{22}||_2^2}$$

- Matthews Coefficient:

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

where

$$FP = \frac{1}{2}\sum_{k=1}^{2} \frac{\sum_{1\leq i<j\leq p} I(a_{i,j}^{(k)} = 0, \hat{a}_{i,j}^{(k)} \neq 0)}{\sum_{1\leq j<j'\leq p} I(a_{i,j}^{(k)} = 0)}, \quad TN = 1 - FP$$

,

$$FN = \frac{1}{2}\sum_{k=1}^{2} \frac{\sum_{1\leq i<j\leq p} I(a_{i,j}^{(k)} \neq 0, \hat{a}_{i,j}^{(k)} = 0)}{\sum_{1\leq i<j\leq p} I(a_{i,j}^{(k)} \neq 0)}, \quad TP = 1 - FN$$

.

- 1-step Mean Squared Forecast Error. We train the model on first $T-1$ time points and check its predictive performance by calculating MSFE with the actual observation at time point T. Let $\mathbf{R} = (R_1, ..., R_{2p})'$ - vector of $2p$ observed values at next time point, $\hat{\mathbf{R}} = (\hat{R}_1, ..., \hat{R}_{2p})'$ - vector of $2p$ predicted values at next time point:

$$MSFE = \sum_{i=1}^{2p} \frac{(\hat{R}_i - R_i)^2}{2p}$$

We will be looking to minimize Normalized Frobenius Difference and maximize Matthews Coefficient.

5.3 Results

First we present the results of Σ^{-1} estimation(Normalized Frobenius Difference measure) in Figure 2 below. Methods used were pure graphical lasso(Glasso), pure factor models(Factor) and the aforementioned ensemble of the two(Our method):

Setup	Glasso	Factor	Our method
p=10 t=30(1000 replicates)	0.76(0.1)	0.65(0.12)	0.52(0.11)
p=10 t=40(1000 replicates)	0.81(0.11)	0.81(0.13)	0.64(0.12)
p=20 t=40(200 replicates)	0.76(0.08)	0.64(0.09)	0.52(0.09)
p=20 t=50(200 replicates)	0.8(0.08)	0.75(0.08)	0.6(0.08)
p=20 t=60(200 replicates)	0.82(0.09)	0.87(0.1)	0.67(0.09)
p=30 t=50(100 replicates)	0.8(0.08)	0.71(0.08)	0.59(0.08)
p=30 t=60(100 replicates)	0.82(0.07)	0.81(0.08)	0.65(0.07)
p=30 t=70(100 replicates)	0.84(0.06)	0.88(0.07)	0.69(0.06)

Figure 2: Results of Σ^{-1} estimation(Normalized Frobenius Difference) for three methods: pure glasso, pure factor models and our method(combination of the former two).

As discussed before, our method combines advantages of pure graphical lasso and factor model estimation for diagonal and off-diagonal elements of Σ^{-1} respectively. It ourperforms both pure graphical lasso and pure factor model approaches in Frobenius difference for all the cases.

$AIC(\lambda_1)$ criterion from (14) is used for tuning parameter selection in separate approach and heuristic $AIC(\lambda_1, \lambda_2)$ criterion from (15) for joint approach. Hard threshold of 0.1 was applied to the resulting estimates.

To further improve the joint estimates we applied a refitting procedure(this procedure didn't help as much with separate estimates case so we left those unaffected):

- assumed the estimated structure of transition matrices to be true;
- applied OLS estimation procedure to the reduced set of parameters(the non-zeros in initial estimates)

We ran at least 100 replicates for all cases and the results are summarised in Figure 3 below.

One can see the dominance of joint modeling approach in terms of Matthews and frobenius difference measures: it outperforms the separate approach on both accounts in all of the studied cases. If we actually break down Matthews coefficient and study false positive and false negative rates, we notice that there is a trade-off in false negatives for joint estimates in some cases. That is to be expected considering that joint estimates appear to be more sparse. Meanwhile, false positive rate performance is much better for the joint approach.

As for forecasting performance, MSFE values appear to be slightly smaller for joint method in most cases. The biggest reason why separate and joint approaches are comparable in that regard is because separate estimates are less sparse and more inclined to overfitting the data. They keep a lot of non-zero elements in the transition matrices,

Setup	Method	MSFE	FP	FN	Matthews	Frob
p=10 t=30	S	0.18(0.14)	0.26(0.1)	0.1(0.06)	0.65(0.12)	0.80(0.22)
$A_{11} \equiv A_{22}$	J	0.17(0.12)	0.09(0.1)	0.07(0.12)	0.84(0.14)	0.49(0.15)
p=10 t=40	S	0.18(0.15)	0.19(0.07)	0.04(0.04)	0.78(0.08)	0.57(0.11)
$A_{11} \sim A_{22}$	J	0.17(0.13)	0.06(0.05)	0.04(0.06)	0.90(0.07)	0.36(0.13)
p=20 t=40	S	0.18(0.12)	0.24(0.09)	0.09(0.03)	0.67(0.1)	0.87(0.25)
$A_{11} \equiv A_{22}$	J	0.15(0.08)	0.02(0.06)	0.07(0.08)	0.90(0.09)	0.33(0.12)
p=20 t=50	S	0.20(0.14)	0.14(0.06)	0.05(0.07)	0.81(0.08)	0.60(0.16)
$A_{11} \sim A_{22}$	J	0.17(0.11)	0.02(0.02)	0.09(0.13)	0.89(0.11)	0.35(0.11)
p=30 t=50	S	0.20(0.12)	0.34(0.11)	0.16(0.04)	0.51(0.13)	1.34(0.42)
$A_{11} \equiv A_{22}$	J	0.16(0.13)	0.01(0.01)	0.1(0.07)	0.89(0.07)	0.41(0.13)
p=30 t=60	S	0.19(0.11)	0.15(0.06)	0.04(0.02)	0.82(0.07)	0.67(0.13)
$A_{11} \sim A_{22}$	J	0.19(0.12)	0.01(0.00)	0.09(0.05)	0.90(0.05)	0.35(0.08)
p=30 t=70	S	0.16(0.12)	0.09(0.03)	0.02(0.01)	0.90(0.04)	0.51(0.07)
$A_{11} \sim A_{22}$	J	0.13(0.08)	0.01(0.00)	0.07(0.07)	0.92(0.06)	0.29(0.09)

Figure 3: Results of a simulation studies comparing Joint(J) and Separate(S) methods for multiple combinations of p and t. Means and standard deviations are shown over 50 replicates for one-step mean squared forecasting error(MSFE), false positive rate(FP), false negative rate(FN), Matthews coefficient(Matthews) and normalized frobenius difference(Frob).

which typically is not the case for the true underlying model. Here we know for a fact that the true model is sparse and that joint modeling approach does considerably better in terms of capturing the matrix structure.

6 Joint modeling of multivariate economic series application

Joint modeling method was applied to simultaneously model economic time series data for multiple states in United States. In particular, considering their similar industrial activity, we focused on Pennsylvania, Michigan, Ohio and Illinois. The data was taken from Federal Reserve website and consists of 14 monthly economic variables spanned over time period from December 2006 to December 2015. The variables under consideration were:

- totals of employees in five sectors(one variable for each of the sectors) - construction, education/health services, financial activities, manufacturing, goods producing;
- hourly earnings for each of the aforementioned five areas(five variables;
- totals of employees in government sector;
- total of employees in non-farm sector;
- leading index;
- unemployment rate.

Details on variable descriptions/abbreviations can be found in Appendix B.

As already mentioned, to estimate the error covariance matrix we use factor models which relies on assumption of several common underlying factors. It was demonstrated in multiple papers ([18][Stock & Watson, 2002]; [4] [Bernanke, 2004]; [5][Boivin, 2008]) that this assumption makes sense for econometric applications. The only difference is we assume that errors, not economic variables themselves, are being driven be a few common underlying factors.

In this particular application we emphasize figuring out the structure of transition matrix(positions of non-zero elements). The estimation procedure is carried out the following way: after fixing initial time point t and time period length l, the model is trained for that period and its performance tested on next time point(which is available). Then the procedure is "shifted" by one time point: make $t+1$ our initial point while keeping period l fixed, train the model for the new period, test on the next time point. Repeat until we run out of time points available in the dataset. That way a number of transition matrix estimates is accumulated and one can see how stable those are while also checking their forecasting performance.

To see how stable the estimates of transition matrices are, we summarized them into one cumulative matrix: value in each position of that matrix corresponds to the proportion of times that position contained a non-zero element in the estimate. For example, if we have 20 various estimates $\hat{A}^i$, $i = 1,..,20$ of transition matrix A, then $A^{cum}_{k,j} = \sum_{i=1}^{20} I(\hat{A}^i_{k,j} \neq 0)$.

Instead of doing a big combined estimation procedure for $K = 4$, all possible pairwise estimation procedures were performed(six of those). That way one ends up with three cumulative matrices for each state(off the pairwise comparisons with other three states). Then we simply average these three matrices to get the final joint estimate of transition matrix structure for each particular state. For comparison the separate estimation procedure was carried out for all the states and summarized in respective cumulative matrices as well.

The results for each state can be seen on Figures 4 and 5 on next two pages.

Figure 4: The cumulative transition matrices(defined in the paragraphs above) for Pennsylvania(upper row) and Ohio(lower row). Separate(left) and joint(row) estimates are presented for each state.

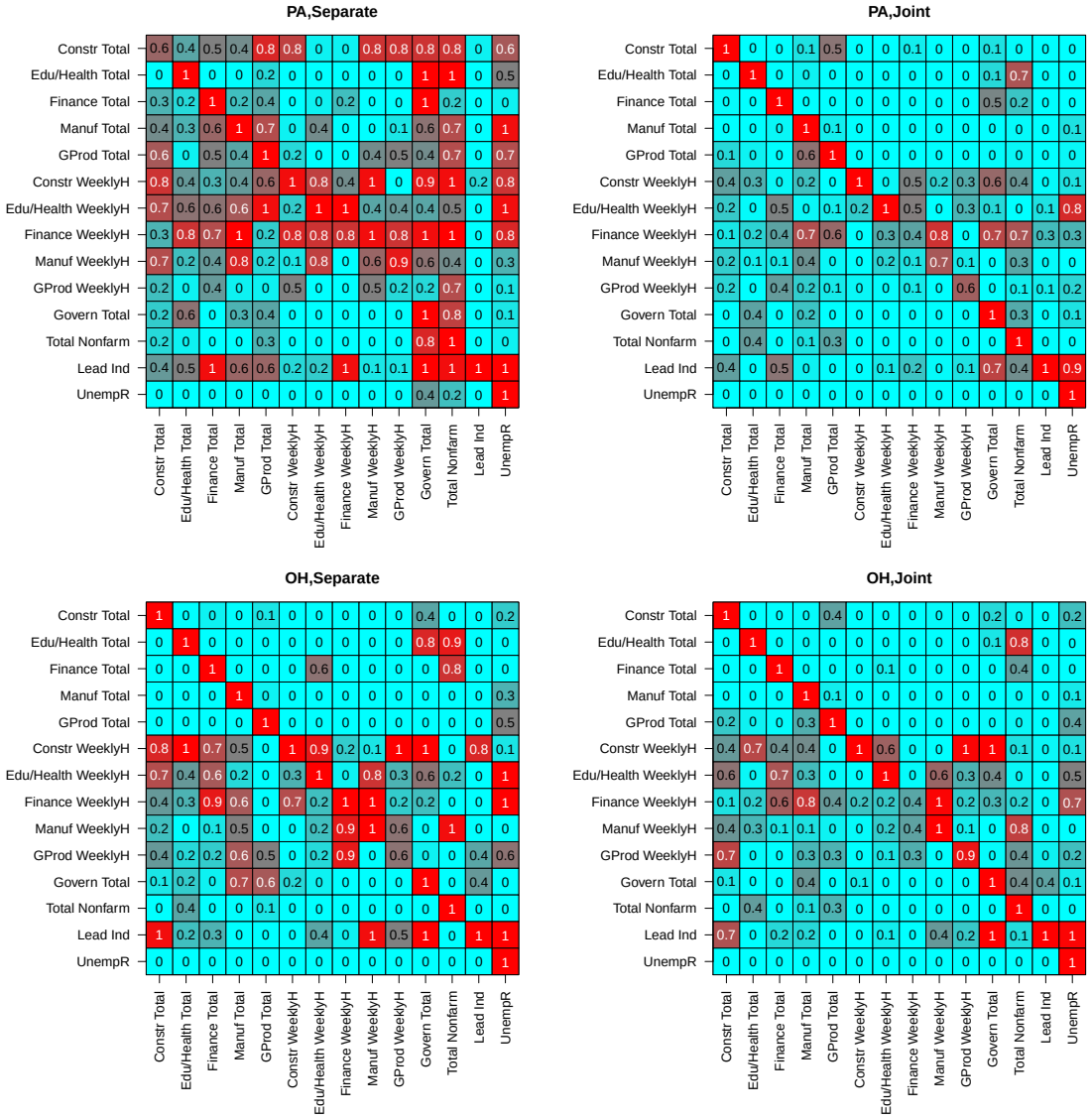
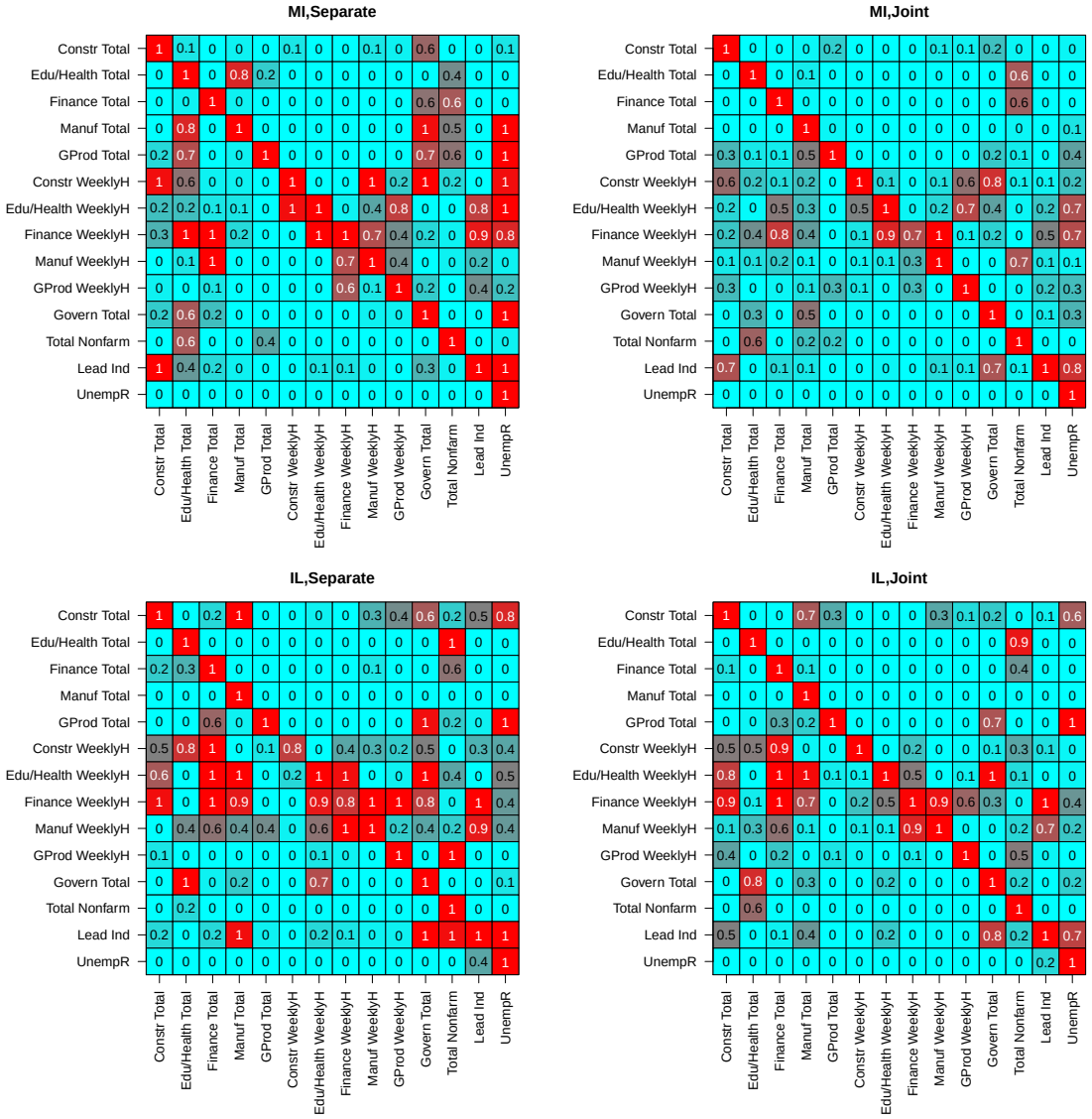


Figure 5: The cumulative transition matrices(defined in the paragraphs above) for Michigan(upper row) and Illinois(lower row). Separate(left) and joint(row) estimates are presented for each state.



One can notice a much sparser structure for joint estimates along with the fact that it picks up on diagonals more consistently than separate method. Separate method is way too dense which is indicative of overfitting. A slight advantage can be seen in terms of forecasting performance of joint method over separate(Figure 6).

Setup	Separate	Joint
14 variables, WeeklyH	0.38(0.20)	0.33(0.14)
14 variables, HourlyE	0.18(0.12)	0.16(0.11)
19 variables	0.30(0.13)	0.26(0.09)

Figure 6: Results of 1-step forecasting for all states combined for each of the setups(3 of those) and modeling approaches(separate and joint).

7 Conclusion

In this paper joint modeling approach is introduced for a problem of estimating two sparse Granger networks. It can be very advantageous when one either doesn't have enough data for separate estimation and when assumption of similarity between classes is true. Especially in case of high-dimensional sparse models joint method will almost always outperform separate method purely because of abundance of shared zero elements. One of the possible improvements in the future include extending the problem to any number H of entities. But considering that even for relatively small examples(of 20,30 variables per entity) it takes hours to run 1 replication the other area of improvement is coming up with a less computationally demanding approach.

Appendix A

Convergence diagnostics of ADMM algorithm.

Here we provide the ADMM algorithm convergence diagnostics via combination of the following four plots:

- $||\beta^{(k)} - \beta^{(k-1)}||_2^2$(denoted as $Frob(\beta^{(k)} - \beta^{(k-1)})$) against iteration number
- $||\gamma^{(k)} - L\beta^{(k)}||_2^2$(denoted as $Frob(\gamma^{(k)} - L\beta^{(k)})$) against iteration number
- $|f_{obj}^{(k)} - f_{obj}^{(k-1)}|$(denoted as $Frob(f_{obj}^{(k)} - f_{obj}^{(k-1)})$) against iteration number
- $f_{obj}^{(k)}$ against iteration number

In particular, we will demonstrate the plots for cases when it took over 40 iterations for the algorithm to converge. See the Figures 7 and 8 below (starting points for ADMM algorithm: $u^0 = (0,...,0), \gamma^0 = (0,...,0)$).

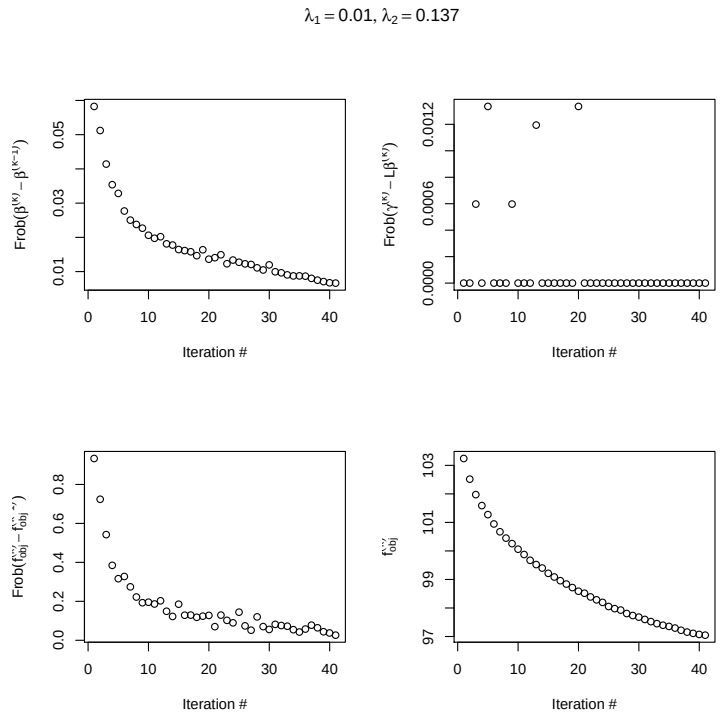


Figure 7: Diagnostics plot 1: four measures(discussed above) plotted against iteration number, fixing $\lambda_1 = 0.01, \lambda_2 = 0.137$.

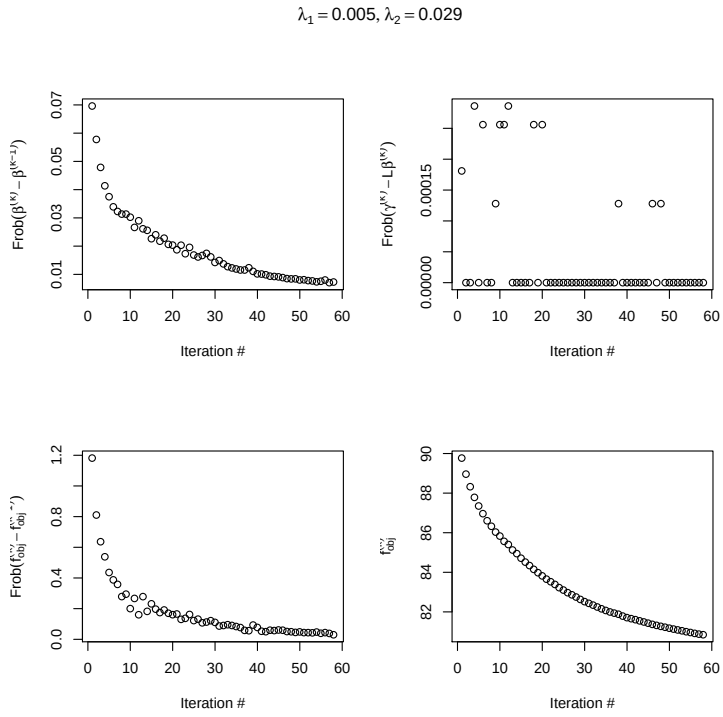


Figure 8: Diagnostics plot 2: four measures(discussed above) plotted against iteration number, fixing $\lambda_1 = 0.005, \lambda_2 = 0.029$.

One can see how the sequence of estimates $\{\beta^{(k)}\}$ stabilizes with respect to Frobenius norm(top left panel), the restriction $\gamma^{(k)} - L\beta^{(k)} = 0, k = 1, 2, ...$ of the algorithm always approximately holds(top right panel), the objective function value stabilizes(botom panels). All of that is indicative of good convergence performance of our ADMM algorithm.

Appendix B

Variable description & abbreviation on FRED website(on example of Illinois)

Abbreviation*	Description	Units
ILCONS	All Employees: Construction in Illinois	Thousands of Persons
ILEDUH	All Employees: Education and Health Services in Illinois	Thousands of Persons
ILFIRE	All Employees: Financial Activities in Illinois	Thousands of Persons
ILGOVT	All Employees: Government in Illinois	Thousands of Persons
ILMFG	All Employees: Manufacturing in Illinois	Thousands of Persons
ILNA	All Employees: Total Nonfarm in Illinois	Thousands of Persons
ILSLIND	Leading Index for Illinois	Percent
ILUR	Unemployment Rate in Illinois	Percent
SMS17000000600000001	All Employees: Goods Producing in Illinois	Thousands of Persons
SMU17000000600000002SA	Average Weekly Hours of All Employees: Goods Producing in Illinois	Hours
SMU17000000600000003SA	Average Hourly Earnings of All Employees: Goods Producing in Illinois	Dollars per Hour
SMU17000002000000002SA	Average Weekly Hours of All Employees: Construction in Illinois	Hours
SMU17000002000000003SA	Average Hourly Earnings of All Employees: Construction in Illinois	Dollars per Hour
SMU17000003000000002SA	Average Weekly Hours of All Employees: Manufacturing in Illinois	Hours
SMU17000003000000003SA	Average Hourly Earnings of All Employees: Manufacturing in Illinois	Dollars per Hour
SMU17000005500000002SA	Average Weekly Hours of All Employees: Financial Activities in Illinois	Hours
SMU17000005500000003SA	Average Hourly Earnings of All Employees: Financial Activities in Illinois	Dollars per Hour
SMU17000006500000002SA	Average Weekly Hours of All Employees: Education and Health Services in Illinois	Hours
SMU17000006500000003SA	Average Hourly Earnings of All Employees: Education and Health Services in Illinois	Dollars per Hour

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

References

[1] M. Bańbura, D. Giannone, and L. Reichlin. Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

[2] S. Basu, G. Michailidis, et al. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

[3] S. Basu, A. Shojaie, and G. Michailidis. Network granger causality with inherent grouping structure. *The Journal of Machine Learning Research*, 16(1):417–453, 2015.

[4] B. S. Bernanke, J. Boivin, and P. Eliasz. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. Technical report, National Bureau of Economic Research, 2004.

[5] J. Boivin and M. Giannoni. Global forces and monetary policy effectiveness. Technical report, National Bureau of Economic Research, 2008.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[7] T. T. Cai, H. Li, W. Liu, and J. Xie. Joint estimation of multiple high-dimensional precision matrices. *The Annals of Statistics*, 38:2118–2144, 2015.

[8] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

[9] F. X. Diebold and C. Li. Forecasting the term structure of government bond yields. *Journal of econometrics*, 130(2):337–364, 2006.

[10] J. Fan, Y. Liao, and M. Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics*, 39(6):3320, 2011.

[11] J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

[12] C. Flamm, U. Kalliauer, M. Deistler, M. Waser, and A. Graef. Graphs for dependence and causality in multivariate time series. In *System identification, environmental modelling, and control system design*, pages 133–151. Springer, 2012.

[13] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[14] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, page asq060, 2011.

[15] G. Michailidis and F. d'Alché Buc. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences*, 246(2):326–334, 2013.

[16] G. D. Rudebusch. Macro-finance models of interest rates and the economy. *The Manchester School*, 78(s1):25–52, 2010.

[17] S. Song, Z. Zhan, Z. Long, J. Zhang, and L. Yao. Comparative study of svm methods combined with voxel selection for object category classification on fmri data. *PloS one*, 6(2):e17191, 2011.

[18] J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179, 2002.

[19] R. J. Tibshirani, J. E. Taylor, E. J. Candes, and T. Hastie. *The solution path of the generalized lasso*. Stanford University, 2011.

[20] J. Vỳrost. Reflexie kvality životného prostredia, emocionálneho naladenia a kultúry sociálneho prostredia respondentov eqls 2012 ako prediktory posúdenia rizika ohrozenia kriminálnymi činmi. *Človek a spoločnost'. Internetovỳ časopis pre pôvodné teoretické a vỳskumné štúdie z oblasti spoločenskỳch vied (Man and Society. Internet journal for original, theoretical and research studies from the field of the social studies)*, 1(17), 2014.