Large Vector Auto Regression for Multi-Layer Spatially Correlated Time Series

Rodrigue Ngueyep  
H. Milton Stewart School of Industrial Systems and Engineering  
Georgia Institute of Technology  
rngueyep3@gatech.edu

Nicoleta Serban  
H. Milton Stewart School of Industrial Systems and Engineering  
Georgia Institute of Technology  
nserban@isye.gatech.edu

Abstract. One of the most commonly used methods for modeling multivariate time series is the Vector Autoregressive Model (VAR). VAR is generally used to identify lead, lag and contemporaneous relationships describing Granger causality within and between time series. In this paper, we investigate VAR methodology for analyzing data consisting of multi-layer time series which are spatially interdependent. When modeling VAR relationships for such data, the dependence between time series is both a curse and a blessing. The former because it requires modeling the between time series correlation or the contemporaneous relationships which may be challenging when using likelihood-based methods. The latter because the spatial correlation structure can be used to specify the lead-lag relationships within and between time series, within and between layers. To address these challenges, we propose a $L_1 \backslash L_2$ regularized likelihood estimation method. The lead, lag and contemporaneous relationships are estimated using an efficient algorithm that exploits sparsity in the VAR structure, accounts for the spatial dependence and models the error dependence. We assess the performance of the proposed VAR model and compare it with existing methods within a simulation study. We also consider a case study to illustrate the applicability of our method.

1 Introduction

Analyzing multivariate time series is a common statistical problem in several fields, such as economics and environmental sciences. One of the most commonly used methods for modeling multivariate time series is the Vector Autoregressive Model (VAR) introduced by Sims (1980). Generally, VAR has been used to identify Granger causal relationships between
variables which vary over time. The primary focus is on the lead and lag effects between
time series but often contemporaneous relationships provide additional information about
how variables are related to each other over a period of time.

In this paper, we investigate VAR methodology for analyzing data consisting of spatially interdependent multi-layer time series. Specific examples are from the field of industrial economics such as multiple time-varying economic indicators such as state level employment rates in the construction industry and the number of building permits issued for new homes observed at the county or even the census tract level within a state or nationally; from the field of industrial engineering such as multiple turbines installed at different geographic locations for which time-varying wind speed and generated power are recorded; and from the field of environmental sciences such as multiple measurements observed at different stations as often generated by environmental and climatological studies. In many of these examples, one layer corresponds to a different measurement or indicator. Specifically, the observed time series data are

\[ Y_{t,k}^{[J]} = Y_{t,s_k^{[J]}}, \text{ with } k \in \{1, \ldots, K_J\} \text{ and } J \in \{1, \ldots, L\} \]

where \( t \) is the time unit and \( L \) is the number of layers, typically small and in our simulations and application its maximum value is 2. \( K_J \) represents the total number of sites in the \( J^{th} \) layer and \( s_1^{[J]}, \ldots, s_{K_J}^{[J]} \) are the spatial units or locations for the time series in the \( J^{th} \) layer. They are recorded as coordinates (latitude and longitude) of each spatial site and are used to map the pair (site number, layer), for example \( s_3^{[1]} \) is the latitude and longitude of the 3rd site in layer 1. Each time series \( Y_{t,k}^{[J]} \) can be influenced by observations from

- **Own lags:** \( Y_{t-p,k}^{[J]} \) for \( p = 1, \ldots, P \), where \( P \) is the maximum lag considered in the study.
- **Lags of neighboring time series within the same layer \( J \):** for example, observations of the time series \( Y_{t-p,k'}^{[J]} \) for \( p = 1, \ldots, P \) located at a site \( s_{k'}^{[J]} \) such that site \( s_{k'}^{[J]} \) is close to \( s_k^{[J]} \).
- **Lags of neighboring time series within layers other than layer \( J \):** for example, observations of the time series \( Y_{t-p,k}^{[J']} \) in layer \( J' \) at site \( s_{k}^{[J']} \), \( J \neq J' \) and \( s_{k}^{[J']} \) is close to \( s_k^{[J]} \).

Therefore, each time series can be influenced by observations within its layer or outside its layer. We assume that for each influential layer, the set of sites that affect a targeted time series is most likely restricted to a close spatial neighborhood of the site of observation of
the targeted time series.

Figure 1: Illustration of spatial representation of time series data with two layers, with targeted site (spatial site 21 in layer 1)

To better explain our methodology, we consider the example illustrated in Figure 1. Assume that we have two layers \((L = 2)\); the crosses correspond to the sites of time series in layer 1, the circles correspond to the sites of time series in layer 2, and we are interested in predicting the time series at site 21 in layer 1 \(\{Y_{t,21}^{[1]}\}\). Based on the proximity of their sites, time series \(\{Y_{t,5}^{[1]}\}, \{Y_{t,8}^{[1]}\}, \{Y_{t,15}^{[1]}\}\) in layer 1 contribute to within layer effects (relationships showed by solid lines in figure 1). The time series \(\{Y_{t,12}^{[2]}\}, \{Y_{t,22}^{[2]}\}\) in layer 2 will be responsible for cross layer effects (dashed lines in figure 1). The anticipated influence of \(\{Y_{t,12}^{[2]}\}\) on the response time series \(\{Y_{t,21}^{[1]}\}\) will be more important than the influence of \(\{Y_{t,22}^{[2]}\}\) on the same response time series, because the coordinates \(s_{12}^{[2]}\) of site 12 in layer 2 is closer to the coordinates \(s_{21}^{[1]}\) of the target site 21 in layer 1 than the coordinates \(s_{22}^{[2]}\) of site 22 in layer 2.

In certain settings, the dynamics of a time series can be approximated by a linear function of its own lags and the lags of influential time series. This reduces to a VAR model with complex lead and lag relationships, which often results in a model with a higher dimensionality than the number of observations. On the other hand, only a small number of lead and lag relationships are expected to be significant. Employing methods which account for this sparsity will allow estimation of a high dimensional VAR model. In order to take advantage
of the sparsity in the relationships within and between time series, our method uses regularization penalties that are functions of the lags of the time series within the same layer and from different layers. In specifying the regularization penalties, we assume that closer information in time has more relevance. When time series are spatially correlated, we also assume that closer information in space has more relevance. Therefore, the relationships of one targeted time series to other time series are increasingly penalized with higher temporal lags and at higher spatial distance.

The primary contribution of our paper is a method for identifying lead and lag relationships between a large number of time series that also exhibit strong contemporaneous spatial dependence. Additionally, the method allows for estimating sparse relationships within and between layers of time series. To the best of our knowledge, this is the first and only approach that accounts simultaneously for the large dimensionality of the problem, the temporal dependence among time series, the spatial dependence present in the errors and the layer group effects. The second contribution is an efficient algorithm that can be used to solve the optimization problem in regularized selection approaches for models similar to the one proposed in this paper.

The remainder of the paper is organized as follows. In Section 2, we review the literature on model selection with a focus on VAR modeling and we motivate the general approach introduced in this paper. In Section 3, we describe the spatial VAR model applied to one layer followed by its extension to a multi-layer setting. Section 4 introduces the estimation procedure and explains the computational algorithms used to fit the model. Simulation studies are carried out in Section 5. In Section 6, we analyze the relationship between state level employment rates in the construction industry and the number of building permits issued for new homes. We conclude with insights in the application of the proposed methodology in Section 7. Some technical details and simulation results are deferred to the Supplemental Material.
2 Background and Motivation

The analysis of multivariate time series, and particularly VAR, has been extensively covered in the statistical, computer science and econometrics literature, but most of the existing methods fail to jointly perform model selection and estimation in a high dimensional setting, meaning when the number of time series is large relative to the sample size. In the context of VAR, variable selection reveals statistically significant relationships within and between time series. Variable selection is critical because a large number of time series implies a large number of potential lead and lag relationships.

One straightforward methodology consists in regressing each time series onto the others separately resulting in multiple regressions, one for each time series. This approach often produces inefficient coefficient estimates due to the large model dimensionality as compared to the sample size of each time series potentially leading to poor forecasting due to overfitting. This challenging aspect has been highlighted in other existing studies (Roecker, 1991; Breiman, 1995).

Alternatively, one could consider variable selection within a multivariate regression model. Variable selection tools based on information criteria have been developed for multivariate regression by Bedrick and Tsai (1994), Fujikoshi and Satoh (1997) among others. Because of the high computational cost, these methods are not used to select the best model among all possible subset structures. Instead, these methods rely on greedy search algorithms, for example, top-down and bottom-up approaches, that are highly unstable, path dependent and suboptimal (Krolzig and Hendry, 2001; Penm and Terrell, 1984). An alternative approach to multivariate regression is to reduce the dimensionality of the predictors - in the VAR context, the predictors consist of lead and lag relationships between time series - using factor analysis. Related work includes reduced-rank regression methods (Anderson, 1951; Izenman, 1975; Reinsel and Velu, 1998) and the Factor Estimation and Selection (FES) proposed by Yuan et. al (2007). For these methods, because the set of predictors is reduced to a few important principal factors, the interpretation of the Granger causal relationships is difficult. Some papers have proposed to use a bayesian approach for the estimation of multivariate regression, for example Cripps et al. (2005) perform variable selection and covariance selection in multivariate regression models.
The emergence of regularized estimation methods such as the Lasso by Tibshirani (1996) has led to the development of regularized sparse estimation schemes for multivariate regression. For example, Turlach et al. (2005) perform model selection using a $L_\infty$-regularization scheme applied to all the coefficients related to a predictor. Obozinski et al. (2008) apply the $L_1\backslash L_2$ regularization for union support recovery, and Peng et al. (2010) introduce a $L_1\backslash L_2$ penalization method for identification of “master” predictors in a multivariate regression. In a more recent study, Rothman et al. (2010) introduce joint estimation of the regression parameters and the covariance of errors by $L_1$ regularized log likelihood. But their approach does not apply to time series data as it does not allow for modeling the serial correlation within time series. Song and Bickel (2011) propose to impose lag-dependence in the regularization penalties to estimate a large VAR model. While this method accounts for serial dependence in the data, it doesn’t include the effects due to contemporaneous correlation present in the errors. Davis et al. (2012) propose a 2-stage approach for estimating sparse VAR (sVAR) models. Their method uses partial spectral coherence with BIC to select non-zero AR coefficients. But their methodology does not explicitly take into consideration the spatial correlation in the errors.

Although the research studies discussed above are a leap from the more traditional VAR modeling, they are still limited in their application. Particularly, they do not simultaneously select lead, lag and contemporaneous relationships within and between time series. The lead & lag relationship selection performance worsens when we do not account for the spatial correlation in the errors. Moreover, existing approaches do not readily extend to data observed for multiple measurements (e.g. humidity, precipitation and temperature) often called layers (Huang et al., 2010).

To address these limitations, we use a $L_1\backslash L_2$—regularized likelihood method to select the temporal lags and spatial sites that influence a targeted time series. Specifically, $L_1$ regularization is used for selecting individual time series effects while $L_2$ regularization is used for selecting entire layers viewed as group effects similar to the sparse group lasso introduced by Friedman et al. (2010). For example, if we are interested in finding the effect of other layers on a time series at a targeted location, the time lag- and spatial distance-weighted regularization associated with the $L_2$ penalty will perform group selection between the layers.
This regularization identifies whether entire layers are not relevant, meaning that all time series in the layer will have no effect on a targeted time series. Since group lasso doesn’t yield within group sparsity, to identify the most influential neighborhood for the selected layers, we therefore apply a temporal lag- and spatial distance-weighted $L_1$-regularization. This penalization approach allows for selection of parsimonious models resulting in efficient parameter estimation and accuracy of time series prediction.

Moreover, to incorporate contemporaneous (spatial) dependence, we propose to use a penalized log-likelihood scheme since it allows the estimation of the covariance matrix of the errors. A similar idea is applied by Rothman et. al (2010) in the context of multivariate regression. Within the multi-layer time series framework, we assume that there is no cross-layer contemporaneous dependence. This assumption allows us to use a divide-and-conquer algorithm to simultaneously solve $L$ optimization problems of smaller size, therefore, reducing the computational effort.

The estimation procedure of our model consists of alternatively solving for the VAR coefficients and solving for the inverse covariance matrix of the errors. To estimate the VAR coefficients, we solve the $L_1 \setminus L_2$-regularization likelihood by providing an algorithm that uses block coordinate descent. To solve for the inverse covariance matrix, we use a spatially weighted graphical lasso method as introduced by Friedman et. al (2008). Details about the model and the estimation algorithm are provided in the next two sections.

3 The Model

3.1 The VAR model

The model of interest in this paper is the Vector Autoregressive model of order $P$ denoted $VAR(P)$. We assume there are $K$ time series that are centered (no intercept),

$$Y_t = B_1 Y_{t-1} + \cdots + B_P Y_{t-P} + V_t$$

with time observed on a regular grid where $Y_t = (Y_{t,1}, \cdots, Y_{t,K})'$ is a $K \times 1$ vector and $Y_{t,k}$ is the observation of the $k^{th}$ time series $\{Y_{t,k}\}$ at time $t$. $B_p$ is a fixed $(K \times K)$ coefficient matrix
for \( p = 1, \ldots, P \) and \( \mathbf{V}_t = (V_{t,1}, \cdots, V_{t,K})' \) is a \((K \times 1)\) vector of error terms. We assume that the error terms \( \mathbf{V}_t \) follow a multivariate normal distribution \( N(0, \Sigma) \) and that they are independently and identically distributed. We also assume that the VAR is stationary.

The equation in (1) can be expressed as a multivariate regression model

\[
\begin{bmatrix}
\mathbf{Y}'_T \\
\vdots \\
\mathbf{Y}'_p \\
\end{bmatrix}
= 
\begin{bmatrix}
\mathbf{Y}'_{T-1} & \cdots & \mathbf{Y}'_{T-P} \\
\vdots & \cdots & \vdots \\
\mathbf{Y}'_{p-1} & \cdots & \mathbf{Y}'_0 \\
\end{bmatrix}
\begin{bmatrix}
\mathbf{B}'_1 \\
\vdots \\
\mathbf{B}'_p \\
\end{bmatrix}
+ 
\begin{bmatrix}
\mathbf{V}'_T \\
\vdots \\
\mathbf{V}'_p \\
\end{bmatrix}
\]

which is equivalent to

\[
\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{V} \tag{2}
\]

A common method for the estimation of \( \mathbf{B} \) is conditional maximum log-likelihood, where the conditional variables are the lagged time series. The goal is to minimize the negative log-likelihood Gaussian function.

\[
g(\Omega, \mathbf{B}) = \text{Tr} \left[ \frac{1}{(T - P)} (\mathbf{Y} - \mathbf{X}\mathbf{B})'(\mathbf{Y} - \mathbf{X}\mathbf{B}) \Omega \right] - \log(|\Omega|) \text{ where } \Omega = \Sigma^{-1}. \tag{3}
\]

### 3.2 The spatial VAR with one layer

We now assume that each component of an observation at time \( t \), \( \mathbf{Y}_t = (Y_{t,1}, \cdots, Y_{t,K})' \), corresponds to a response recorded at each of \( K \) different spatial units with coordinates \( s_k \) with \( k \in \{1, \cdots, K\} \) and \( s_i \in \mathbb{R}^d \). We use the notation \( \mathbf{Y}_t = (Y_{t,1} = Y_{t,s_1}, \cdots, Y_{t,K} = Y_{t,s_K})' \), where \( Y_{t,s_k} \) is the observation of the variable of interest at time \( t \) and at spatial unit \( s_k \). In this setting, the precision matrix \( \Sigma^{-1} \) of the error terms \( \mathbf{V}_t = (V_{t,s_1}, \cdots, V_{t,s_K}) \) has a certain degree of sparsity. In particular we assume that time series observed at sites that are far from each other are more likely to have a null entry in the precision matrix. Beyond this assumption, we do not make other structural assumptions such as isotropy or parametric
shape. The resulting one-layer VAR model becomes

\[
\begin{bmatrix}
Y_{t,1} \\
\vdots \\
Y_{t,K}
\end{bmatrix} =
\begin{bmatrix}
B_{11}^{(1)} & \cdots & B_{1K}^{(1)} \\
\vdots & \ddots & \vdots \\
B_{K1}^{(1)} & \cdots & B_{KK}^{(1)}
\end{bmatrix}
\begin{bmatrix}
Y_{t-1,1} \\
\vdots \\
Y_{t-1,K}
\end{bmatrix} + \cdots +
\begin{bmatrix}
B_{11}^{(P)} & \cdots & B_{1K}^{(P)} \\
\vdots & \ddots & \vdots \\
B_{K1}^{(P)} & \cdots & B_{KK}^{(P)}
\end{bmatrix}
\begin{bmatrix}
Y_{t-P,1} \\
\vdots \\
Y_{t-P,K}
\end{bmatrix} +
\begin{bmatrix}
V_{t,1} \\
\vdots \\
V_{t,K}
\end{bmatrix}
\tag{4}
\]

For any \( p \in \{1, \ldots, P\} \) and any \( k, k' \in \{1, \ldots, K\} \), \( B_{kk'}^{(p)} \) measures the effect of the observation \( Y_{t-p,k'} \) at the spatial location \( s_{k'} \) at a past time \( t-p \) on the observation \( Y_{t,k} \) at the location \( s_k \).

### 3.3 The spatial VAR with multiple layers

The model described in Section 3.2 can be generalized to a setting with more than one layer. For instance, a typical geostatistical study involves the joint modeling of two economic indicators, unemployment and house prices, across counties in the US. In such a study, one might arbitrarily set employment rate to be the first layer and the house prices the second layer. More generally, assume that there are \( L \) layers, and that for each layer \( J \in \{1, \ldots, L\} \), observations are acquired at spatial units \( \{s_1^{[J]}, \ldots, s_{K_J}^{[J]}\} \) at discrete times \( t \in \{0, \ldots, T\} \). For the layer \( J \), we have \( K_J \) time series \( \{Y_{t,a_1^{[J]}}, \ldots, Y_{t,s_{K_J}^{[J]}}\} \). The observation spatial units are not necessarily the same across layers. Given that the notation in \( \{Y_{t,a_1^{[J]}}, \ldots, Y_{t,s_{K_J}^{[J]}}\} \) involves many forms of indices, we re-express all the time series under the form \( Y_t^{[\text{ind}]} \) where \( \text{ind} \) is an index unique to each time series. In what follows, we use the set of indices \( D_J = (a_J, a_J + 1 \cdots, b_J) \) where \( a_J = \sum_{j=1}^{J-1} K_j + 1 \) and \( b_J = \sum_{j=1}^{J} K_j \). The time series within the first layer \( (J = 1) \) correspond to time series with indices in \( D_1 = (1, \ldots, K_1) \) where \( a_1 = 1 \) and \( b_1 = K_1 \). The time series within the second layer \( (J = 2) \) correspond to time series with indices in \( D_2 = (K_1 + 1, \ldots, K_1 + K_2) \) where \( a_2 = K_1 + 1 \) and \( b_2 = K_1 + K_2 \). Generally, the vector of time series within the \( J^{th} \) layer becomes \( \left(Y_t^{[a_J]}, \ldots, Y_t^{[b_J]}\right) \). The total number of time series in the model is \( M = \sum_{j=1}^{L} K_j \). We apply the same transformation to the indices associated with the sites, so that, we can interchangeably use \( s_1^{[J]} \) and \( s_{a_J} \) to denote the site where the first time series \( Y_t^{[a_J]} \) in layer \( J \) is observed.

When we consider multiple layers the coefficient matrix \( \mathbb{B} \) in (2) becomes \( \mathbb{B} = \left[B_{D_1,}, \cdots, B_{D_L,}\right] \).
where $\mathbf{B}_{D,J}$ represents all the coefficients that affect the observations in layer $J$:

$$
\mathbf{B}_{D,J} = \begin{bmatrix}
B_{a_1}^{(1)} & \cdots & B_{a_M}^{(1)} & \cdots & B_{a_1}^{(P)} & \cdots & B_{a_M}^{(P)} \\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
B_{b_1}^{(1)} & \cdots & B_{b_M}^{(1)} & \cdots & B_{b_1}^{(P)} & \cdots & B_{b_M}^{(P)}
\end{bmatrix}^T \in \mathbb{R}^{(PM) \times K_J} \tag{5}
$$

For any column $\mathbf{B}_i$ of $\mathbb{B}$, if $i \in \{a_J, \cdots, b_J\}$ then $\mathbf{B}_i$ is a column of the matrix $\mathbf{B}_{D,J}$.

$$
\mathbf{B}_i = \left(\mathbf{B}_i^{(1)}', \cdots, \mathbf{B}_i^{(P)}'\right)' \in \mathbb{R}^{(PM) \times 1}
$$

$\mathbf{B}_i^{(p)}$ is the sub-column of $\mathbf{B}_i$ that contains the coefficients associated with lag order $p$:

$$
\mathbf{B}_i^{(p)} = \begin{cases}
(B_{i1}^{(p)}, \cdots, B_{iK_1}^{(p)}), \cdots, B_{ia_J}^{(p)}, \cdots, B_{ib_J}^{(p)}, \cdots, B_{ia_L}^{(p)}, \cdots, B_{ib_L}^{(p)}',
\end{cases}
$$

1st layer effect

within layer effect

Lag $p$ effect

Lth layer effect

Next we introduce the estimation method used to estimate the VAR coefficients and the covariance matrix of the errors. We also describe the algorithms used to obtain these estimates.

4 Estimation Algorithm

4.1 The methodology

In this section, we introduce the estimation method for one layer data followed by a description of how it extends to multiple layer data.

4.1.1 One layer sparse estimation

Spatio-temporal data exhibit statistical features that can be exploited to improve the efficiency of the model parameter estimates. Given the high-dimensional nature of the estimation problem, we need to impose some sparsity inducing constraints on the VAR coefficients $\mathbb{B}$ and potentially on the precision matrix $\Omega$. As assumed in Bańbura et. al (2010) and
Song et al. (2011), more recent temporal lags should be more predictive than the more distant lags. The second assumption, usually stated as the First Law of Geography, is that the observations collected at more distant spatial sites should be less influential on the observations collected at the site of interest. Given these constraints and assuming that the precision matrix $\Omega$ is known $\Omega = \tilde{\Omega}$, we solve the following optimization problem

$$\hat{B} = \arg\min_{B \in \mathbb{R}^{PK \times K}} \left[ g(\tilde{\Omega}, B) + \lambda_1 \sum_{p=1}^{P} \sum_{i=1}^{K} \sum_{k=1}^{K} P^\alpha e^{\gamma \|s_i - s_k\|} |B_{ik}^{(p)}| \right]$$

where $\lambda_1$ is a penalty parameters and $\alpha, \gamma$ are lag and distance weight parameters respectively that are always strictly positive. The optimization problem in (6) is convex, since it is the sum of a convex objective function $g(\tilde{\Omega}, B)$ and of a convex penalty. As in Song and Bickel (2011), we account for the lag effect by penalizing more heavily the coefficients associated with observations that are more distant in time. Additionally, we account for the spatial effect by using penalties weighted by a function that depends on the distance function $e^{\gamma \|s_i - s_k\|}$, a similar idea is used in Lozano et al. (2010). For example, if we consider the lagged $p$ time series $\{Y_{t-p}^u\}$ and $\{Y_{t-p}^v\}$ influencing the targeted time series $\{Y_{t}^i\}$, the penalty on the term $B_{iu}^{[p]}$ is higher than the penalty on the term $B_{iv}^{[p]}$ if $\|s_i - s_u\| > \|s_i - s_v\|$. To account for the lag and spatial effects we can use other penalty functions, for instance, $f(p) = (1 + \log(p))^{\alpha}$ or $f(p) = \exp(p)^{\alpha}$ for lag functions, in place of $p^\alpha$. In this paper, we do not suggest that the penalty functions we chose are optimal. Identifying the optimal functions would considerably increase the number of tuning parameters.

4.1.2 Multi-layer sparse estimation

As presented in Section 3.3, the $J^{th}$ layer is identified by the index set

$$D_J = \{a_J = \sum_{j=1}^{J-1} K_j + 1, \ldots, b_J = \sum_{j=1}^{J} K_j\}.$$
Let $B_i$ of $\mathbb{B}$ be the column of coefficients corresponding to the time series $\{Y^{[i]}_t\}$ in layer $J$, meaning $i \in D_J$. The terms in column $B_i$, can be rearranged in the following manner

$$B_i = \{B_{iD_1}, \ldots, B_{iD_l}, \ldots, B_{iD_L}\}$$  \hspace{1cm} (7)

Each set of coefficients $B_{iD_l}$ in (7) represents the effect from time series in the $l^{th}$ layer on the time series of interest $\{Y^{[i]}_t\}$. For any layer $l \in \{1, \ldots, L\}$

$$B_{iD_l} = \begin{pmatrix} B_{iD_l}^{(1)} & \cdots & B_{iD_l}^{(P)} \end{pmatrix}$$  \hspace{1cm} (8)

The regularization scheme we propose for the estimation of the column $B_i$, is the following

$$\lambda_1 \sum_{p=1}^P \sum_{k=1}^M P^\alpha \epsilon^{\gamma||s_i-s_k||} |B_{ik}^{(p)}| + \lambda_2 \sum_{l=1}^L \|B_{iD_l}\|_{\Delta^{[i]}_l}$$  \hspace{1cm} (9)

with $B_i$, such that $i \in D_J$

$$\|B_{iD_l}\|_{\Delta^{[i]}_l} = K_l \begin{pmatrix} B_{iD_l}^{(1)} & \cdots & B_{iD_l}^{(P)} \end{pmatrix} \left( P \otimes \Delta^{[i]}_l \right) \begin{pmatrix} B_{iD_l}^{(1)} \\ \vdots \\ B_{iD_l}^{(P)} \end{pmatrix}$$

where

$$\Delta^{[i]}_l = \begin{bmatrix} e^{2\gamma||s_i-s_{a_1}||} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & e^{2\gamma||s_i-s_{b_l}||} \end{bmatrix} \quad \text{and} \quad P = \begin{bmatrix} (1)^{2\alpha} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & P^{2\alpha} \end{bmatrix}$$

The first term in (9) is similar to the penalty in (6) for the one layer experiment. This penalty captures the within layer sparsity effect. As explained in Section 4.1.1, coefficients corresponding to the lead and lag effects for the time series closer in time and space are less
penalized. The second term in (9) uses the group lasso penalty introduced by Yuan et al (2006). We apply the group penalty to all the layers, except to the $J^{th}$ layer that contains the coefficients $B_{iD_J}$ linked to the time series $\{Y_i^{[i]}\}$ (with $i \in D_J$). By imposing the group sparsity scheme, we will select only the layers that have lag effects on the response time series $\{Y_t^{[i]}\}$. The $J^{th}$ layer, which is not penalized by the group sparsity norm, is always selected, suggesting that we assume that own layer effect is always present.

In (9), we introduce the norm applied to the vector of coefficients $B_{iD_i}$; this norm is used for specifying the between-layer penalty. The matrices $\Delta^{[i]}$ and $P$ in (9) are designed to account for three important statistical features of the data. First, the term $K_l$ quantifies the size of the group since it measures the number of observation spatial units. Consequently, the model applies heavier penalties on layers with more spatial units. Second, the lag and distance weights serve the same purpose as in the one layer case described in (6). Third, a layer with a set of distant sites should be penalized as a group higher than if the sites are nearby. This idea is conveyed through the use of the weight matrix $\tilde{\Delta}^{[i]}$.

This penalization scheme induces group-wise and within-layer sparsity. The group-wise penalty allows assessment of between-layer lead and lag relationships. The within-layer penalty will select the influential spatial units within the layers which have a lag influence on the variable of interest $Y_t^{[i]}$. If we assume that the precision matrix is known, then the problem we solve is the following:

$$\min_B \left[ g(\tilde{\Omega}, B) + \lambda_1 \sum_{i=1}^M \sum_{p=1}^P \sum_{k=1}^M p^\alpha e^{\gamma s_i - s_k} \|B^{(p)}_{ik}\| + \lambda_2 \sum_{i=1}^M \sum_{l=1}^L \|B_{iD_l}\|_{\tilde{\Delta}^{[i]}} \right]$$

Equation (10) is obtained by applying the penalties in (9) to each of the $i^{th}$ columns of the matrix $B$. The summation of the within layer penalty is over all layers and over all spatial units within the layers. The between layer penalty is applied to all the layers but not to the layer containing the targeted time series.
4.1.3 Estimation of the precision matrix with spatial structure

We assume that the spatial covariance matrix is block diagonal, i.e. we only have within layer spatial dependence. Assuming that pairs of time series sampled from distant sites within the same layer are independent, we model each layer covariance using a distance weighted graphical lasso method. This idea was suggested by Friedman et al. (2008). We specify the amount of regularization to depend on the distance between two targeted time series. If we assume that the regression coefficients are known \( \mathbb{B} = \tilde{\mathbb{B}} \) the following optimization problem is solved for each layer \( J \).

\[
\widehat{\Omega}_J = \arg\min_{\Omega_J \in \mathbb{R}^{K_J \times K_J}; \Omega_J \succeq 0} g \left( \tilde{\mathbb{B}}_{D_J}, \Omega_J \right) + \lambda_3 \sum_{i=a_J}^{b_J} \sum_{k \neq i} e^{-|s_i - s_k|} |\Omega_{ik}| \tag{11}
\]

With the submatrix, \( \mathbb{B}_{D_J} = [\mathbb{B}_{a_J}, \cdots, \mathbb{B}_{b_J}] \) and \( \Omega_J \) is the precision matrix associated with the \( J^{th} \) layer.

The second alternative method one could consider for the estimation of the precision matrix is to use parametric spatial covariance function. This method could be used in cases where the precision matrix is not sparse.

4.1.4 Joint Estimation of the VAR coefficients and the precision matrix

To jointly estimate \( \mathbb{B} \) and \( \Omega \) in a multi-layer and spatial setting, we apply the idea introduced by Rothman et al. (2011), which reduces to penalized likelihood estimation including all regularization schemes introduced for estimation of the VAR coefficients and the covariance matrix. If we have no cross-layer spatial dependence in the errors, \( \mathbb{B} \) and \( \Omega \) are estimated by minimizing the \( L_1/L_2 \) regularized negative log-likelihood function \( g \). We can decompose the large optimization problem in \( L \) optimization problems of smaller size. The \( J^{th} \) layer
optimization problem becomes

$$\left( \tilde{B}_{D_J}, \tilde{\Omega}_J \right) = \arg\min_{\Omega_J \in \mathbb{R}^{K_J \times K_J}, \Omega_J \geq 0} g\left( B_{D_J}, \Omega_J \right)$$

$$+ \lambda_1 \sum_{i=a_J}^{b_J} \sum_{p=1}^{P} \sum_{k=1}^{M} p^\alpha e^\gamma \|s_i - s_k\| \left| B_{ik}^{(p)} \right| + \lambda_2 \sum_{i=a_J}^{b_J} \sum_{l=1}^{L} \| B_{iD_l} \| \tilde{\Delta}^{[i]}$$

$$(12)$$

$$+ \lambda_3 \sum_{i=a_J}^{b_J} \sum_{k \neq i} \|s_i - s_k\| \left| \Omega_{ik} \right|$$

The problem presented in (12) is not convex, but we can alternatively solve for $B_{D_J}$, with $\Omega_J$ fixed at $\tilde{\Omega}_J$ as in (10), and solve for $\Omega_J$ with $B_{D_J}$ as in (11). In the next section, we introduce the algorithms for solving these two convex optimization problems.

4.2 Computational algorithms

The algorithm for the estimation of the VAR coefficients borrows the idea from the block cyclical coordinate descent applied to sparse group lasso in a technical report by Friedman et al. (2010). The algorithm used for estimating the precision matrix is a modified graphical lasso introduced by Friedman et al. (2008).

4.2.1 Algorithm for the VAR coefficients

For any layer $J \in \{1, \cdots, L\}$, we define $Y_{D_I} \in \mathbb{R}^{(T-P) \times K_J}$ the set of time series within the $J^{th}$ layer. If we assume that the precision matrix is set at $\tilde{\Omega}_J$, we need to solve a problem similar to problem (10):

$$\min_{B_{D_J}} \text{Tr} \left[ \frac{1}{T-P} (Y_{D_J} - \bar{X} B_{D_J})' (Y_{D_J} - \bar{X} B_{D_J}) \tilde{\Omega}_J \right]$$

$$+ \lambda_1 \sum_{i=a_J}^{b_J} \sum_{p=1}^{P} \sum_{k=1}^{M} p^\alpha e^\gamma \|s_i - s_k\| \left| B_{ik}^{(p)} \right| + \lambda_2 \sum_{i=a_J}^{b_J} \sum_{l=1}^{L} \| B_{iD_l} \| \tilde{\Delta}^{[i]}$$

$$(13)$$
If we have just one layer the algorithm used is similar to the MRCE of Rothman and Levina (2011). If we have more than one layer the algorithm used is inspired from the sparse group lasso of Friedman et al. (2010). We visit each column of the matrix $B_{D,l}$, and apply a cyclical group coordinate descent procedure to all the coefficients within each column $B_{iD_l}$ associated with layers $l$ such that $l \neq J$. Further, for all the group of coefficients selected in the previous step, we again apply a cyclical coordinate descent to identify the non-null coefficients within the selected groups. The details and the derivations of the algorithm are presented in the supplemental material.

4.2.2 Algorithm for the joint estimation of the VAR Coefficients and the precision matrix

The algorithm used to solve problem (12) is the following:

For $\lambda_1$ and $\lambda_2$

- Set $\hat{B}^{(0)} = 0$ and use graphical lasso to solve L problems (11) $\tilde{\Omega}_J^{(0)} = \tilde{\Omega}_J \left( \hat{B}^{(0)}_{D,J} \right)$
- For each $l \in 1, \ldots, L$ compute $\hat{B}^{(m+1)}_{D,l} = \hat{B}_{D,l} \left( \tilde{\Omega}_l^{(m)} \right)$ by solving problem (10) with algorithm for VAR coefficients.
- Compute $\tilde{\Omega}_J^{(m+1)} = \tilde{\Omega}_J \left( \hat{B}^{(m+1)}_{D,J} \right)$ by using graphical lasso to solve (11)
- If $\sum_{j,k} \left| \hat{B}^{(m+1)}_{jk} - \hat{B}^{(m)}_{jk} \right| < \epsilon \sum_{j,k} \left| \hat{B}^{\text{Ridge}}_{jk} \right|$ stop, otherwise start new loop
- $\hat{B}^{\text{Ridge}}$ is the solution of the VAR obtained by using a ridge regression for each time series.

4.2.3 Selection of tuning parameters

As for any regularization method, achieving a satisfactory performance in terms of model selection and parameter estimation requires proper selection of the penalty parameters $\lambda = (\lambda_1, \lambda_2, \lambda_3)$. Additionally in our method, we need to assess the importance of the distance effects and the lag effects parameters, $\alpha$ and $\gamma$ respectively. For this, we employ a computationally efficient approach, we use the Bayesian Information Criterion (BIC) intro-
duced by Schwarz (1978) that minimizes

$$BIC = -2 \log L(\hat{B}_{\lambda,\alpha,\gamma}) + \log(T)df$$

where $\hat{B}_{\lambda,\alpha,\gamma}$ is the estimator associated with the tuning parameters $\lambda, \alpha$ and $\gamma$, $L(\hat{B})$ is the maximum likelihood of the VAR model and $df$ is the number of degrees of freedom approximated by the number of non-zero estimated parameters. Zhou et al. (2007) finds that if in a regression setting the rank of a design matrix is equal to the number of predictors then the degrees of freedom of the lasso is well approximated by the number of non null coefficients. The BIC criterion allows the determination of the optimal lag for each layer.

The use of BIC for non-convex regularized likelihood is advocated by Bühlmann and Van De Geer (2011) as a simple and computationally convenient method. However, there is no rigorous justification for the use of BIC in the context of regularized non-convex likelihood to date.

For comparison purposes, we also analyze the performance of the lasso and a modified lasso scheme that accounts for the distance between the sites. To select the penalization parameters for these two methods, we use the rolling prediction scheme used by Song and Bickel (2011) for consistency with the existing relevant papers.

5 Simulations

5.1 Simulations setup

We assess the performance of the method using two simulation experiments. We herein refer to our method as SMTSE (Sparse Multivariate Time Series Estimation). In the first experiment, we assume that the time series are observed for one type of measurement, i.e. one-layer data. In the second experiment, we generate time series from two distinct layers. For each experiment, we evaluate the model selection performance by assessing how well an estimation method captures the sparsity in the lag relationships using metrics such as the True Positive Rate (TPR) and the True Negative Rate (TNR). We measure the estimation performance using the Frobenius norm of the difference matrix between the true VAR matrix
and the estimated VAR matrix.

**One layer simulations.** We generate the simulated set of time series as described below:

1. Randomly generate $K$ sites in a $[0,1] \times [0,1]$ square. We use a 2-dimensional uniform distribution to create the site locations.

2. Generate the VAR coefficients

   - Generate $1^{st}$ own lag coefficient for each time series
     \[ B_{ii}^{[1]} \sim \text{Uniform}(a,b). \]
   - Randomly select the $C_i$ closest neighbors to the site $s_i$ of time series $\{Y_t^{[i]}\}$
     \[ C_i \sim \text{Binomial}(T_{\text{neighbors}}, P_{\text{neighbors}}) \]
     where $T_{\text{neighbors}}$ is the maximum possible of neighbors sites selected and $P_{\text{neighbors}}$ is the probability assigned to the binomial distribution.
   - Coefficients associated with the $C_i$ closest sites of the targeted site $s_i$ are computed. We denote by $S_i$ the set that contains the index of the $C_i$ closest sites:
     \[ \forall j \in S_i, B_{ij}^{[1]} = B_{ii}^{[1]} \ast \exp (-\delta \|s_i - s_j\|). \]
     Note that $\delta$ is a term used to accentuate the decrease of coefficients associated with time series far from location $s_i$.
   - Generate the coefficients associated with lags greater than 1:
     \[ \forall i, j \in \{1, \cdots, K\}^2, B_{ij}^{[l]} = l^n B_{ij}^{[1]}, \text{with } l > 1 \text{ and } n < 1. \]

3. Set the error covariance matrix to: $\Sigma_{ij} = \rho^{|s_i - s_j|}$.

4. Simulate $K$ time series of length $T$ from VAR model with VAR coefficients $B$ and error covariance matrix $\Sigma$.

**Two-layer simulations.** To generate the two-layer simulated data, we apply a similar procedure as in the one-layer simulation experiment. Each layer consists of 25 sites. We alternate simulations in which the two layers have an effect on each other and simulations in which only layer one has an effect on layer two. Within-layer effects are always present in all simulations. The covariance matrix for the two layers experiment has a block diagonal structure, with each block defined by $\Sigma_{ij} = \rho^{|s_i - s_j|}$.
Simulation settings

Throughout all simulations we set fixed the following parameters:

- The lower and the upper bounds for the own lag coefficients: $a = -0.5$ and $b = 0.5$.
- The number of sites: $K = 25$.
- The maximum number of influential neighbors for each site: $T_{\text{neighbors}} = 5$.
- The probability for the generation of influential neighbors for each site: $P_{\text{neighbors}} = 0.8$.
- To reduce the computational cost, the temporal and spatial penalty tuning parameters are set to $\alpha = 1$ and $\gamma = 1$ for all settings.

We vary other parameters including the number of true lags and the variance of the errors. The different simulation settings are:

- **Simulation Settings 1 & 2**: Number of layers $L = 1$, lag order $P = 2$, error covariance level $\rho_1 = 0.1$ (simulation 1), $\rho_2 = 0.7$ (simulation 2). The search for the optimal regularization parameters is performed on the following grid $\lambda_1 = \{1, 10, 20, \cdots, 100\}$, so $\lambda_1$ varies by increments of 10, and $\lambda_3 \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$. After finding the parameter $(\lambda_1, \lambda_3)$ that minimize the BIC criterion, we perform a second search on a refined grid around the previous minimum.

- **Simulation Settings 3 & 4**: Number of layers $L = 2$, lag order $P = 1$, error covariance level $\rho_2 = 0.1$ (simulation 3), $\rho_4 = 0.4$ (simulation 4). For the two layers experiments, the regularization parameters are searched in the following set of values, $(\lambda_1, \lambda_2) \in \{1, 10, 20, \cdots, 100\}^2$, and $\lambda_3 \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$.

To test the performance of the SMTSE, for each simulation setup, we generate 50 different replications with time series of length $T = 300$. For each simulation setup, we apply the estimation methods assuming 2, 3 or 4 lags for one-layer simulations and 1, 2, 3 or 4 lags for two-layer simulations. We report the following metrics:

\[ TP = \frac{\#[(i,j): B_{ij} \neq 0 \text{ and } \hat{B}_{ij} \neq 0]}{\#[(i,j): B_{ij} \neq 0]} \]

the true positive rate measuring the ability of a model to capture non null VAR coefficients.

\[ TN = \frac{\#[(i,j): B_{ij} = 0 \text{ and } \hat{B}_{ij} = 0]}{\#[(i,j): B_{ij} = 0]} \]

the true negative rate measuring the ability of a model to identify null VAR coefficients.

\[ FE = \sqrt{\sum_{i,j} (B_{ij} - \hat{B}_{ij})^2} \]

the Frobenius norm error measuring the estimation error of the VAR coefficients.
5.2 Simulation results

Figures 2 and 3 summarize our findings. In Sub-figures 2(a) and 3(a), we report the true positive rates of SMTSE, of the lasso and of the spatial lasso; in Sub-figures 2(b) and 3(b), we report the true negative rates; and in Sub-figures 2(c) and 3(c), we report the Frobenius norm. The dark curves are the results obtained for the simulation settings 1 & 3, and the others are for the simulation settings 2 & 4 averaged over the 50 replications. Based on these simulations results, we find:

• SMTSE outperforms the lasso and the spatial lasso for strong and weak error variance.

• As we increase the number of lags, the true positive rates decrease for all the methods.

• When the error variance is weak, the three methods have similar performances in terms of identification of the non-null VAR coefficients. However, when the level of the error covariance increases, our method identifies the non null zero coefficients with much higher accuracy.

• Our method is not significantly sensitive to an increase in the level of the error covariance, this result validating that our estimation procedure improves the efficiency of the VAR coefficients estimates. On the other hand, the Lasso and the Spatial Lasso VAR coefficient estimates are extremely sensitive to the error covariance level since they do not model the covariance structure of the noise.

• In the two-layer setting, whether the error covariance level is strong or weak, our model performs even better (comparatively to Lasso and Spatial Lasso) than in the one layer case. This is because the group penalty excludes many false positives.

• We also study the predictive performance of all the methods considered. For each set of simulations, we use the generative models described above, this time each time series has a length $T = 350$, and we leave 50 points for out-of-sample forecasting. The $h$-step ahead forecast for time series $\{Y_t^{[i]}\}$ given all the information up to time $t$ ($I(t)$) is $\hat{Y}_{t+h|I(t)}^{[i]}$. The $h$-step ahead root mean square error for each time series is computed as

$$
\text{RMSE}_{i}^{(h)} = \left[ \frac{1}{(50)} \sum_{t=300-h}^{350-h} \left( \hat{Y}_{t+h|I(t)}^{[i]} - Y_{t+h}^{[i]} \right)^2 \right]^{\frac{1}{2}}.
$$
Figures 4 and 5 present the box plots of the accuracy of the out-of-sample prediction measured the root mean squared errors (RMSEs) for the true model, the sparse multivariate time series estimation (SMTSE), the Lasso, the Spatial Lasso (SP LASSO) and the Ordinary Least Squares (OLS). Under all the simulation settings, we observe the SMTSE has a RMSE slightly lower than the RMSE of the Lasso and the RMSE of the Spatial Lasso. The Ordinary Least Squares as expected overfits the model and yields poor out-of-sample forecasts. In Section A of the supplemental material, we report the forecasting performance of all these methods when the number of lags used for estimation is larger than the true number of lags. The proposed method remains competitive when compared to the Lasso and the Spatial Lasso, and the OLS RMSE increases due to overfitting. In Table 1, we report the lag number selected by AIC and BIC criteria under OLS and the lag number selected by our model in average over the 50 replications. We find that the lag is accurately identified using BIC in our method.

- Following the use of OLS + AIC and OLS + BIC introduced by Hsu et al. (2008), we also fitted the simulated models with OLS (results not reported here); as expected this method doesn’t introduce sparsity in the VAR coefficient matrices and the Frobenius norm error is on average significantly higher than the values obtained for the other three methods aforementioned. The true positive rate is 1, but the true negative rate is 0.
Figure 3: Performance metrics for two layers experiment

Figure 4: Prediction mean squared error for one and two layers experiment

<table>
<thead>
<tr>
<th>Simulation</th>
<th>OLS + AIC</th>
<th>OLS + BIC</th>
<th>SMTSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 VAR(2), $\rho = 0.1$</td>
<td>2.00</td>
<td>1.08</td>
<td>2.06</td>
</tr>
<tr>
<td>2 VAR(2), $\rho = 0.7$</td>
<td>2.00</td>
<td>1.9</td>
<td>2.08</td>
</tr>
<tr>
<td>3 VAR(1), $\rho = 0.1$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>4 VAR(1), $\rho = 0.4$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 1: Average of the lag selected with OLS + AIC, OLS + BIC and SMTSE for simulated VAR models over 50 replications of simulated data.
6 Application

In the case study introduced in this section, we consider the time series of construction employment (thousands of persons), and the number of new private housing units authorized by building permits in the United States. Both layers of time series are observed at the state level and on a monthly basis; they are seasonally adjusted. These time series are collected using the Geographical Economic Data (GEOFRED) of the Federal Reserve Bank of St. Louis. We first removed the time series of certain states because of missing observations. Specifically, the states of Delaware, District of Columbia, Hawaii, Maryland and Nebraska are not present in the employment dataset, the states of South Dakota and Tennessee have 38 consecutive months (January 2008 to February 2011) with missing observations. The states of District of Columbia and Hawaii are not present for the building permits time series. The observations for the two economic measurements span from April 1996 to March 2012; we leave out data from April 2012 to April 2013 for out of sample forecasting. The number of time points is \( T = 192 \) for a total of 93 time series.

Before applying the three estimation methods, we standardize all time series. To find the optimal set of regularization parameters and the optimal lags, we perform an extensive search over a grid based on the following values \( \lambda_1, \lambda_2 \in \{0.1, 10, 20, \ldots, 100\} \), \( \lambda_3 \in \{0.01, 0.1, 1, 10\} \); the lags considered for this analysis are \( P \in \{1, 2, 3\} \). The values of \( \lambda_1 \) and \( \lambda_2 \) vary by increments of 10, while the parameter \( \lambda_3 \) varies on a log-scale. Additionally, we consider \( \alpha \in \{0.1, 1\} \) to accommodate the possibility of a strong or weak temporal decay effect and \( \gamma \in \{0.1, 0.01\} \), to scale the spatial distances between the states. The optimal tuning parameters are \( \lambda_1 = 30, \lambda_2 = 1, \lambda_3 = 0.01 \) for the employment layer, \( \lambda_1 = 10, \lambda_2 = 20, \lambda_3 = 0.01 \) for the building permit layer, and for both layers the temporal tuning parameter is \( \alpha = 1 \) and the spatial tuning parameter is \( \gamma = 0.01 \).

The results from the implementation of the SMTSE are presented in Figure 5(a). The horizontal axis of this figure represents the time series \( Y_t \), while the vertical axis represents the lags. For instance, the first column contains the coefficients affecting the employment rate in the state of Connecticut. The first horizontal black separation is the first lag effects of employment, the second horizontal black separation is the first lag effects of building permits. The left half (right half) of the matrix contains the coefficients that influence the
employment rates (building permits). The vertical black line separates the employment time series from the building permit time series. The states are grouped in 10 economic regions based on their proximity to each other. The grey lines in the resulting coefficient matrix are used to delimit the economic regions.

The SMTSE finds that only one lag is needed for describing the lead and lag relationships among the employment time series, while it suggests a higher degree of persistence for the building permit time series as three lags are selected using BIC. This finding points to the fact that we can select a different number of lags for each layers. This is possible because of the divide-and-conquer approach we adopt.

For our method, the significant VAR coefficients tend to gravitate within the diagonal blocks of each economic region. In contrast, the Lasso (Figure 5(b)) tends to introduce small but non-null AR coefficients for the time series within the two layers. The spatial lasso (Figure 5(c)) is able to eliminate these small but non-null coefficients. Our method and the spatial lasso provide sparser models than the lasso.

For the SMTSE, the own lag effects are positive for all time series. The own lag coefficients for the employment time series are very high while for the building permit time series, the coefficients are smaller although still significantly greater than 0. Moreover, for the building permit time series, the own lag coefficients are slowly decreasing as the number of lags increases implying that these time series are persistent. Additionally, our model uncovers the effects of the building permits on construction employment suggesting that the number of building permits issued for new houses is a leading indicator in the housing industry. Therefore, if the number of building permits increases, it is plausible to expect a rise in the construction employment. Our model also reveals the absence of feedback effect of the employment time series on the number of building permits issued.

The lasso and the spatial lasso are both unable to provide similar results. The lasso introduces small and noisy estimates for some VAR coefficients and does not explicitly capture the lack of effect of employment on the number of building permits issued. The spatial lasso is able to remove the small and noisy VAR coefficients, but is not able to identify that employment does not lead the building permits issued. The main reason why our method properly identifies the relationships between these two layers lies in the presence of the group
penalties that uniformly remove all the coefficients associated with a potential feedback effect from employment time series. In Figure 5(a), we see that the blocks of coefficients that capture the effect of employment on building permit (on the right side of the black vertical line) are all null. But in Figures 5(b) and 5(c), we observe the presence of some coefficients of small magnitudes in these blocks. The coefficient matrices for the lasso, spatial lasso with 1 and 2 lags are presented in the supplemental material. They have a behavior similar to the their counterparts with 3 lags. We also show the coefficient matrices of the OLS with 1 and 2 lags, some of the coefficients in these matrices are very large (order of magnitude of 40 for OLS with 2 lags).

We also report the out-of-sample forecast performance of the three methods. The forecast period is from the month of April 2012 to April 2013. We use the RMSE to measure the prediction performance. As seen in Figure 6, the OLS fitted with 2 lags has the worst out-of-sample forecast for all the lag levels, and the OLS with 1 lag has the second worst performance. This poor performance can be explained by the fact that OLS commonly overfits. In contrast, the lasso generates out-of-sample forecasts that are less accurate than the forecasts resulting from the SMTSE; this is probably due to the presence of a large number of spurious VAR coefficients. The spatial lasso and the lasso have similar forecasting performance. These results imply that simply incorporating spatial distances in the lasso penalty doesn’t improve the predictions. But if we also account for the contemporaneous effects through the estimation of the precision matrix, the prediction errors become significantly smaller than the prediction errors associated with the other regularized methods and the ordinary least squares. In Table 2(a), we report the number of non null coefficients in each of the simulated models, we see that the SMTSE yields the second most sparse model and is still able to outperform the other methods in terms of prediction. Table 2(b), shows some typical computational time needed to solve the SMTSE under a very sparse ($\lambda$ large) and very dense ($\lambda$ small) settings with R, on a 1.80Ghz Intel Xeon Linux computer.

Throughout other experiments not reported here, we found that if we increase the sample size $T$, the OLS can produce predictions that are more accurate than all the regularized methods including the SMTSE. This can be explained by the fact that the $L_2$ norm of the prediction error of regularized methods such as the lasso has an upperbound that depends
Table 2: Number of non-null coefficients for fitted models and computational time for algorithm under different values for the regularization parameters and with one lag on the inverse of magnitude of the restricted eigenvalue of the matrix $\frac{X^TX}{n}$. So the OLS prediction performance could be superior (despite overfitting) to the lasso prediction performance if some compatibility conditions hold for a small restricted eigenvalue (Bühlmann et al., 2011).

Figure 5: VAR matrix coefficients for employment and building permit time series
Figure 6: The h-step ahead forecast root mean square error (RMSE) for the Sparse Multivariate Time Series Estimation (SMTSE) method, for Lasso fitted with 1 to 3 lags, Spatial Lasso fitted with 1 to 3 lags, and OLS fitted with 1 to 2 lags. Forecast period $T_0 = \text{April 2012}$ to $T_1 = \text{April 2013}$.

7 Conclusion

In this paper, we propose a $L_1/L_2$ penalized likelihood method for estimating large sparse VAR models of time series, which are spatial interdependent. The methodology explicitly accounts for sparsity, group sparsity and spatial contemporaneous correlation among the time series. We also presented algorithms for solving the $L_1/L_2$ constrained optimization problems obtained after penalizing the VAR coefficients and the error precision matrix.

We performed extensive simulations to evaluate the performance of the proposed method (SMTSE) in comparison with existing approaches. We found that the SMTSE outperforms OLS, lasso and spatial lasso in recovering sparse VAR structures, in estimating the VAR coefficients, and in forecasting future values of the time series (especially, when the time series length is smaller than the number of time series). Importantly, the identification of the sparse VAR structure improves when applying lag and distance weighted penalties to the VAR coefficients, and by penalizing VAR coefficients at the group (specified by layers) level and within groups.

Theoretical properties justifying these results are not presented in this manuscript. Wonyul et al. (2012) consider the joint estimation of a coefficient matrix $B$ and of a precision matrix $\Sigma^{-1}$ in a multivariate regression problem. They use a doubly penalized joint likelihood with penalties on entries of $B$ and $\Sigma^{-1}$. They show the consistency and sparsitency
properties of estimates obtained by alternatively solving for $B$ and $\Sigma^{-1}$. These theoretical justification could be extended to demonstrate the performance of the proposed method SMTSE.

Furthermore, the applicability of our method is illustrated by analyzing the relationship between construction employment rates and the number of new private housing units authorized by building permit. The method yields an interpretable model that matches economic intuition.

In both the simulation and application studies covered in this paper, the layers are clearly delineated. One of the reviewers however suggested that the method could be extended to applications with layers that are not necessarily clearly defined. The first step could consist of the estimation of a graphical or cluster model to identify the layers in the data.

**Supplemental Material**

Supplements to the manuscript Large Vector Autoregression for Multi-Layer Spatially Correlated Time Series include in Supplement A, additional prediction results for simulated time series; in Supplement B, the plots of VAR coefficient matrices for application; in Supplement C, the algorithm for VAR coefficients; in Supplement D, the derivation of the algorithm for one layer model; in Supplement E, the derivation of the algorithm for multiple layers model. Additionally, we also include a software package that contains a csv file with 50 simulated time series of length 300, a csv file with the distance matrices for the 50 sites of observations of the simulated time series, and the code for the SMTSE.

**References**


