Supplement to Large Vector Auto Regressions for Multi-Layer Spatially Correlated Time Series

Rodrigue Nguoyep

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

Supplement A: Simulations

1 Simulations

1.1 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. In the generative models described below for one and two layers, spatial dependencies are generated by selection of nearest spatial neighbours. This nearest neighbour dependency is not part of the VAR model described in Section 3 of the paper.

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
 $\forall i \in \{1, \dots, K\}, B_{ii}^{[1]} \sim \text{Uniform}(a, b)$.
 - Randomly select the C_i closest neighbors to the site s_i of time series $\{Y_t^{[i]}\}$ $C_i \sim \text{Binomial}(T_{neighbors}, P_{neighbors})$ where $T_{neighbors}$ is the maximum possible number of neighbors sites selected and $P_{neighbors}$ 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: $\forall j \in S_i, B_{ij}^{[1]} = B_{ii}^{[1]} * \exp(-\delta \|s_i - s_j\|)$. Note that δ 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:
 $\forall i, j \in \{1, \dots, K\}^2, B_{ij}^{[l]} = l^\eta B_{ij}^{[1]}$, with $l > 1$ and $\eta < 1$.
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 $\mathbb{B}$ and error covariance matrix Σ .

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: $T_{neighbors} = 5$.
- The probability for the generation of influential neighbors for each site: $P_{neighbors} = 0.8$.
- 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 $\lambda_1 = \{1, 10, 20, \dots, 100\}$, so

λ_1 varies by increments of 10, and $\lambda_3 \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$. After finding the parameter (λ_1, λ_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, $(\lambda_1, \lambda_2) \in \{1, 10, 20, \dots, 100\}^2$, 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:

$TP = \frac{\#[(i,j):\hat{B}_{ij} \neq 0 \text{ and } B_{ij} \neq 0]}{\#[(i,j):B_{ij} \neq 0]}$, the true positive rate measuring the ability of a model to capture non null VAR coefficients .

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

1.2 Simulation results

Figures 1 and 2 summarize our findings. In Sub-figures 1(a) and 2(a), we report the true positive rates of SMTSE, of the lasso and of the spatial lasso; in Sub-figures 1(b) and 2(b), we report the true negative rates; and in Sub-figures 1(c) and 2(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 covariance.
- As we increase the number of lags, the true positive rates decrease for all the methods.
- When the error covariance is weak ($\rho = 0.1$), 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 ($\rho = 0.7$), 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

$$\text{RMSE}_i^{(h)} = \left[\frac{1}{(50)} \sum_{t=300-h}^{350-h} \left(\hat{Y}_{t+h|I(t)}^{[i]} - Y_{t+h}^{[i]} \right)^2 \right]^{\frac{1}{2}}.$$

Figure 3 presents 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 B 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 1: Performance metrics for one layer experiment

Figure 2: Performance metrics for two layers experiment

Figure 3: Prediction mean squared error for one and two layers experiment

Supplement B: Additional prediction results for the simulation Study

We show additional root mean squared errors (RMSEs) for the simulated models described in Section A of the supplement. For instance, under the simulation setup, we had $\rho = 0.1$

Simulation	OLS + AIC	OLS + BIC	SMTSE
1 $VAR(2), \rho = 0.1$	2.00	1.08	2.06
2 $VAR(2), \rho = 0.7$	2.00	1.9	2.08
3 $VAR(1), \rho = 0.1$	1.00	1.00	1.00
4 $VAR(1), \rho = 0.4$	1.00	1.00	1.00

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

and the number of lags $P = 2$. We presented the results of the prediction performance for simulated models with 2 lags in the manuscript. In Figures 4, 5 and 6, we show the results when the number of lags used to fit the models is greater than the true number of lags. Under all the simulation settings, the ordinary least squares (OLS) has the worst performance. More importantly as we increase the number of lags, the performance of the OLS worsens because of overfitting. On the other hand, the SMTSE performance is slightly better than the performance of other regularized methods considered in these simulations.

Figure 4: RMSE for one layer simulations under settings 1 and 2.

We also show similar results for the two layers experiments, presented under simulation settings 3 and 4.

Supplement C: VAR matrices of case study

In Figures 7 and 8, we present the plots of the VAR coefficient matrices for the lasso, the spatial lasso and the OLS with 1 and 2 lags.

- The OLS VAR coefficient matrices are dense and lead to models that are not interpretable.
- The lasso selects many non null coefficients with small magnitude and it is unable to

Figure 5: RMSE for two layer simulations under setting 3.

Figure 6: RMSE for two layer simulations under setting 3.

provide sparse models. The model resulting from the lasso is not highly interpretable for all the time series. For instance, the lasso is unable to identify the fact that the employment time series shouldn't affect the number of building permit issued.

- The spatial lasso fitted with one lag yields a model with higher sparsity than the lasso with one lag. The model is also able to capture the spatial effects, since most of the influential coefficients are around the diagonals of the matrices. The spatial lasso with two lags exhibit similar properties and it also captures partially the lack of effect of employment on the number of building permits.

Figure 7: VAR matrix coefficients for employment and building permit time series

Supplement D: Algorithm for the VAR coefficients

For any layer $J \in \{1, \dots, L\}$, we define $Y_{D_J} \in \mathbb{R}^{(T-P) \times K_J}$ 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 below.

$$\begin{aligned} \min_{B_{D_J}} \text{Tr} \left[\frac{1}{T-P} (Y_{D_J} - \mathbb{X} B_{D_J})' (Y_{D_J} - \mathbb{X} B_{D_J}) \tilde{\Omega}_J \right] + \\ \lambda_1 \sum_{i=a_J}^{b_J} \sum_{p=1}^P \sum_{k=1}^M p^\alpha e^{\gamma \|s_i - s_k\|} |B_{ik}^{(p)}| + \lambda_2 \sum_{i=a_J}^{b_J} \sum_{\substack{l=1 \\ l: i \notin D_l}}^L \|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}} \end{aligned} \quad (1)$$

Let $S = \mathbb{X}'\mathbb{X}$ and $H^{[J]} = \mathbb{X}'Y_{D_J}\Omega_J$, For convenience we write $B_{ri} = B_{ij}^{(p)}$ where $(r = (p-1)M + j)$. For each layer we define the sets $A_l = \{a_l, \dots, b_l, \dots, (P-1)M + a_l, \dots, (P-1)M + b_l\}$. For any value $r \in \{1, \dots, M\}$ we can recover p and j by $p = \lfloor \frac{r}{M} \rfloor + 1$ and $j = (r \bmod M) + 1$, If $j \in D_l$ then $r \in A_l$.

In the following we just write $H^{[J]}$ as H and we assume that for the column of interest $B_{:,i}$ is in the layer J

Algorithm 1:

At the n^{th} iteration, we do the following.

- Set the estimate $\hat{B}_{D_I}^{(n)} \Leftarrow \hat{B}_{D_I}^{(n-1)}$
- For each column $B_{i,\cdot}$ of B_{D_I} . {
 - For each layer $l \in \{1, \dots, L\}$ {

(a) Lasso with 2 lags

(b) Spatial lasso with 2 lags

(c) OLS with 2 lags

Figure 8: VAR matrix coefficients for employment and building permit time series

- * Test if we only have one layer ($L = 1$) {
 - For each $r \in A_l$ {
 - Compute $U_{ri} = \sum_{q=1}^{PM} \sum_{m=1}^{K_J} S_{rq} \hat{B}_{qm}^{(n)} \tilde{\Omega}_{mi}$
 - Update $B_{ri}^{(n)}$ with the minimizer of the function (1) along this coordi-

nate direction as in Rothman et. al (2010)

$$\hat{B}_{ri}^{(n)} \Leftarrow \text{sign} \left(\hat{B}_{ri}^{(n)} + \frac{H_{ri} - U_{ri}}{S_{rr}\tilde{\Omega}_{ii}} \right) \left(\left| \hat{B}_{ri}^{(n)} + \frac{H_{ri} - U_{ri}}{S_{rr}\tilde{\Omega}_{ii}} \right| - \frac{\lambda_1 p^\alpha e^{\|s_i - s_j\|}}{S_{rr}\tilde{\Omega}_{ii}} \right)_{(2)}^+$$

with $(x)_+ = \max(0, x)$

}

}

* Else {

· For each $r \in A_l$ {

Compute $U_{ri} = \sum_{q=1}^{PM} \sum_{m=1}^{K_J} S_{rq} \hat{B}_{qm}^{(n)} \tilde{\Omega}_{mi}$

Compute

$$a_{ri} = \left[-U_{ri} + H_{ri} + \sum_{k \in A_l} S_{rk} \hat{B}_{ki}^{(n)} \tilde{\Omega}_{ii} \right]$$

Compute

$$\hat{t}_r = \begin{cases} \frac{a_{ri}}{\lambda_1 p^\alpha e^{\gamma \|s_i - s_j\|}} & \text{if } \left| \frac{a_{ri}}{\lambda_1 p^\alpha e^{\gamma \|s_i - s_j\|}} \right| \leq 1, \\ \text{sign} \left(\frac{a_{ri}}{\lambda_1 p^\alpha e^{\gamma \|s_i - s_j\|}} \right) & \text{if } \left| \frac{a_{ri}}{\lambda_1 p^\alpha e^{\gamma \|s_i - s_j\|}} \right| > 1 \end{cases}$$

}

· Compute $Q(\hat{t})$

$$Q(\hat{t}) = \frac{1}{\lambda_2^2 K_l} \sum_{r \in A_l} \left(\frac{a_{ri}}{p^\alpha e^{\gamma \|s_i - s_j\|}} - \lambda_1 \hat{t}_r \right)^2$$

· If $Q(\hat{t}) \leq 1$ set $\hat{B}_{iD_l}^{(n)} \Leftarrow \mathbf{0}$

· Else if $(Q > 1)$ {

For each $r \in A_l$ {

Compute $C_{ri} = H_{ri} - U_{ri} + S_{rr} \hat{B}_{ri}^{(n)} \tilde{\Omega}_{ii}$

If $|C_{ri}| \leq \lambda_1 p^\alpha e^{\|s_i - s_j\|}$ set $\hat{B}_{ri}^{(n)} \Leftarrow 0$

Else If $|C_{ri}| > \lambda_1 p^\alpha e^{\|s_i - s_j\|}$ {

$$\hat{B}_{ri}^{(n)} \Leftarrow \underset{Z \in \mathbb{R}}{\text{argmin}} \Theta_1 * Z^2 + \Theta_2 * Z + \lambda_1 \Theta_3 * |Z| + \lambda_2 \sqrt{K_l} (\Theta_4 + \Theta_5 * Z^2)^{\frac{1}{2}}$$

Where

$$\Theta_1 = S_{rr}\tilde{\Omega}_{ii}$$

$$\Theta_2 = 2 * \left(U_{ri} - H_{ri} - S_{rr}B_{ri}\tilde{\Omega}_{ii} \right) = -2C_{ri}$$

$$\Theta_3 = p^\alpha e^{\gamma\|s_i-s_j\|} \text{ where } p = \lfloor \frac{r}{M} \rfloor + 1, \text{ p is the current lag}$$

$$\Theta_4 = \left(\|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}}^2 - p^{2\alpha} e^{2\gamma\|s_i-s_j\|} (B_{ri})^2 \right)$$

$$\Theta_5 = p^{2\alpha} e^{2\gamma\|s_i-s_j\|}$$

}

}

}

}

}

}

- If $\sum_{j,k} \left| \hat{B}_{jk}^{(n)} - \hat{B}_{jk}^{(n-1)} \right| < \epsilon \sum_{j,k} \left| \hat{B}_{jk}^{Ridge} \right|$ stop, otherwise start new loop

Supplement E: Derivation of the algorithm for the one layer model

For a one layer model with a fixed covariance matrix $\tilde{\Omega}$, we try to minimize the objective function $f(\mathbb{B}, \tilde{\Omega})$.

$$f(\mathbb{B}, \tilde{\Omega}) = Tr \left[(Y - X\mathbb{B})^T (Y - X\mathbb{B}) \tilde{\Omega} \right] + \lambda_1 (T - P) \sum_{i=1}^M \sum_{p=1}^P \sum_{k=1}^M p^\alpha e^{\gamma\|s_i-s_k\|} |B_{ik}^{(p)}|$$

Let $S = X^T X$ and $H = X^T Y \tilde{\Omega}$. We also use $B_{ij}^{(p)} = B_{ri}$ with $j = (r \bmod M) + 1$ and $p = \lfloor \frac{r}{M} \rfloor + 1$. We also define $U_{ri} = \sum_{q=1}^{PM} \sum_{m=1}^M S_{rq} B_{qm} \tilde{\Omega}_{mi}$.

Since we have an l_1 -norm we need to derive the directional derivatives associated with each

parameter in the VAR model.

$$\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}^+} = U_{ri} - H_{ri} + \lambda_1 * (T - P) * p^\alpha * e^{-\gamma \|s_i - s_j\|} \text{ if } B_{ri} > 0$$

$$\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}^-} = U_{ri} - H_{ri} - \lambda_1 * (T - P) * p^\alpha * e^{-\gamma \|s_i - s_j\|} \text{ if } B_{ri} < 0$$

If we define the current estimate as $B_{ri}^{(n)}$. $B_{ri}^{(n+1)}$ is a optimal if and only if, the directional derivatives $\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}} \geq 0$ in all directions. To update the directional derivatives we can use the formula below.

If the solution is such that $B_{ri} < 0$

$$\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}^+} = S_{rr} B_{ri}^{(n+1)} \tilde{\Omega}_{ii} - S_{rr} B_{ri}^{(n)} \tilde{\Omega}_{ii} + U_{ri} - H_{ri} - \lambda_1 * (T - P) * p^\alpha * e^{-\gamma \|s_i - s_j\|}$$

$$\frac{\partial f(B, \tilde{\Omega})}{\partial B_{ri}^+} > 0 \text{ if and only if } B_{ri}^{(n+1)} > \hat{B}_{ri}^* + \frac{\lambda_1 * (T - P) * p^\alpha * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}}$$

$$\text{where } \hat{B}_{ri}^* = B_{ri}^{(n)} + \frac{H_{ri} - U_{ri}}{S_{rr} \tilde{\Omega}_{ii}}$$

If $\hat{B}_{ri}^* + \frac{\lambda_1 * (T - P) * p^\alpha * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}} < 0$, we update the B_{ri} by using :

$$\hat{B}_{ri}^{(n+1)} = \hat{B}_{ri}^* + \frac{\lambda_1 * (T - P) * p^\alpha * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}}.$$

If $\hat{B}_{ri}^* + \frac{\lambda_1 * (T - P) * p^\alpha * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}} \geq 0$ the only solution that satisfies the constraint $B_{ri} < 0$ is $\hat{B}_{ri}^{(n+1)} = 0$.

So we have that

$$\begin{aligned} \hat{B}_{ri}^{(n+1)} &= \min \left[0, \hat{B}_{ri}^* + \frac{\lambda_1 * (T - P) * p^\alpha * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}} \right] \\ &= \text{Sign} \left(\hat{B}_{ri}^* \right) \max \left[0, |\hat{B}_{ri}^*| - \frac{\lambda_1 * (T - P) * p^\alpha * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}} \right] \end{aligned}$$

If the solution is such that $B_{ri} < 0$ we use a similar reasoning to show that

$$\begin{aligned}\hat{B}_{ri}^{(n+1)} &= \min \left[0, \hat{B}_{ri}^* - \frac{\lambda_1 * (T - P) * p^\alpha * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}} \right] \\ &= \text{Sign} \left(\hat{B}_{ri}^* \right) \max \left[0, |\hat{B}_{ri}^*| - \frac{\lambda_1 * (T - P) * p^\alpha * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}} \right]\end{aligned}$$

So we showed that to update each coefficient in the VAR matrix, we can simply use cyclical coordinate descent $\hat{B}_{ri}^{(n+1)} = \text{Sign} \left(\hat{B}_{ri}^* \right) \left(|\hat{B}_{ri}^*| - \frac{\lambda_1 * (T - P) * p^\alpha * e^{\gamma \|s_i - s_j\|}}{S_{rr} \tilde{\Omega}_{ii}} \right)_+$

Supplement F: Derivation of the algorithm for multiple layers model

In a multi-layer setting, with a block diagonal error covariance matrix. If we focus on layer $J \in \{1, \dots, L\}$, we want to solve the following problem.

$$\min_{B_{D_J}} f(B_{D_J}, \tilde{\Omega}_J) = \text{Tr} \left[(Y_{D_J} - \mathbb{X} B_{D_J})' (Y_{D_J} - \mathbb{X} B_{D_J}) \tilde{\Omega}_J \right] +$$

$$\lambda_1 (T - P) \sum_{i=a_J}^{b_J} \sum_{p=1}^P \sum_{k=1}^M p^\alpha e^{\gamma \|s_i - s_k\|} |B_{ik}^{(p)}| + \lambda_2 (T - P) \sqrt{(K_J)} \sum_{i=a_J}^{b_J} \sum_{\substack{l=1 \\ l: i \notin D_l}}^L \|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}}$$

If we focus on layer $l \in \{1, \dots, L\}$ and on column i then The subgradient equations are given by:

$$\frac{\partial f(B_{D_J}, \tilde{\Omega}_J)}{\partial B_{ri}} = U_{ri} - H_{ri} + \lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j\|} t_r + \lambda_2 (T - P) (K_J)^{\frac{1}{2}} g_r$$

As in supplement A, each equation above corresponds to a VAR coefficient $B_{ij}^{(p)} = B_{ri}$. Where r is such that $j = (r \bmod M) + 1$, $p = \lfloor \frac{r}{M} \rfloor + 1$ and $j \in (a_l, \dots, b_l)$ with a_l and b_l are defined in subsection 2.3 of the manuscript .

We also have that $r \in A_l = \{a_l, \dots, b_l, \dots, (P-1)M + a_l, \dots, (P-1)M + b_l\}$ $\mathbf{g}$ is the vector that contains all the (g_r) values.

$$g_r = \frac{p^{2\alpha} e^{2\gamma \|s_i - s_j\|} B_{ij}^{(p)}}{\|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}}} \text{ if } \|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}} \neq 0$$

$$\|\mathbf{g} W^{-\frac{1}{2}}\| \leq 1 \text{ if } \|B_{iD_l}\|_{\tilde{\Delta}_l^{[i]}} = 0$$

$$\text{where } W = I_P \otimes \begin{bmatrix} e^{2\gamma \|s_i - s_1^{[l]}\|} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & e^{2\gamma \|s_i - s_{K_l}^{[l]}\|} \end{bmatrix} \text{ and } W^{-\frac{1}{2}} = I_P \otimes \begin{bmatrix} \frac{1}{e^{\gamma \|s_i - s_1^{[l]}\|}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{e^{\gamma \|s_i - s_{K_l}^{[l]}\|}} \end{bmatrix}$$

Additionally, we have that:

$$t_r = \text{sign}(B_{ri}) \text{ if } B_{ri} \neq 0$$

$$t_r \in [-1, 1] \text{ if } B_{ri} = 0$$

A necessary and sufficient conditions for $B_{iD_l} = 0$ is that $\forall r \in A_l$:

$$U_{ri} - H_{ri} - \sum_{k \in A_l} S_{rk} B_{ki} \tilde{\Omega}_{ii} + \lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j\|} t_r + \lambda_2 (T - P) (K_J)^{\frac{1}{2}} g_r = 0$$

has a solution with $\|\mathbf{g}W^{-\frac{1}{2}}\| \leq 1$ and $t_r \in [-1, 1]$.

We define $a_{ri} = [-U_{ri} + H_{ri} + \sum_{k \in A_l} S_{rk} B_{ki} \tilde{\Omega}_{ii}]$, this implies that if the necessary and sufficient condition is satisfied.

$$\forall r \in A_l, a_{ri} = \lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|} t_r + \lambda_2 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2} (K_J)^{\frac{1}{2}} g_r$$

We can solve the system of equations resulting from the necessary and sufficient conditions by minimizing.

$$Q(\mathbf{t}) = \|\mathbf{g}W\|_2 = \frac{1}{\lambda_2^2 K_l} \sum_{j=1}^{K_l} \sum_{p=1}^P \left[\frac{a_{ri}}{(T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}} - \lambda_1 t_r \right]^2$$

We note that the minimum is $J(t) = 0$, it implies that:

$$\begin{aligned} \frac{\partial Q(t)}{\partial t_r} &= -2 \frac{\lambda_1}{\lambda_2^2 K_l} \sum_{j=1}^{K_l} \sum_{p=1}^P \left[\frac{a_{ri}}{(T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}} - \lambda_1 t_r \right] \\ \frac{\partial Q(t)}{\partial t_r} &= 0 \text{ if } t_r = \frac{a_{ri}}{\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}} \end{aligned}$$

But we need to have $t_r \in [-1, 1]$. So if $\left| \frac{a_{ri}}{\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}} \right| < 1$ then

$$t_r = \frac{a_{ri}}{\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}}.$$

On the other end,

if $\frac{a_{ri}}{\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}} > 1$, then $t_r = 1$, and

if $\frac{a_{ri}}{\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}} < -1$ then $t_r = -1$. So we have that:

$$\hat{t}_r = \begin{cases} \frac{a_{ri}}{\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j\|_2}} & \text{if } \left| \frac{a_{ri}}{\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j\|_2}} \right| \leq 1, \\ \text{sign} \left(\frac{a_{ri}}{\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j\|_2}} \right) & \text{if } \left| \frac{a_{ri}}{\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j\|_2}} \right| > 1 \end{cases}$$

Now that $\min J(t)$ is found at $\hat{t}$, we compute $Q(\hat{t})$. If $Q(\hat{t}) < 1$ then the necessary and sufficient conditions stated above is satisfied with $Q(\hat{t}) = \|\mathbf{g}W\|_2 < 1$, this implies that $B_{iD_l} = 0$.

If $Q(\hat{t}) > 1$, then some coefficients $B_{ri} \neq 0$ in B_{iD_l} . We have to solve for each coefficients B_{ri} with $r \in A_l$.

If $B_{ri} \geq 0$

$$\frac{\partial f(B, \tilde{\Omega}_l)}{\partial B_{ri}} = U_{ri} - H_{ri} + \lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2} + \lambda_2 (T - P) p^{2\alpha} e^{2\gamma \|s_i - s_j^{[l]}\|_2} K_l \frac{B_{ri}}{\|B_{iD_l}\|_2}$$

To update the partial derivative with respect to the B_{ri} coordinate, we use the formula below.

$$\begin{aligned} \frac{\partial f(B, \tilde{\Omega}_l)}{\partial B_{ri}} = & U_{ri} + S_{rr} \hat{B}_{ri}^{(n+1)} \tilde{\Omega}_{ii} - S_{rr} \hat{B}_{ri}^{(n)} \tilde{\Omega}_{ii} - H_{ri} + \lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2} + \\ & \frac{\lambda_2 (T - P) p^{2\alpha} e^{2\gamma \|s_i - s_j^{[l]}\|_2} K_l}{\|B_{iD_l}\|_2} \hat{B}_{ri}^{(n+1)} \end{aligned} \quad (1)$$

$$C_{ri} = H_{ri} - U_{ri} - S_{rr} \hat{B}_{ri}^{(n)} \tilde{\Omega}_{ii}$$

$$\hat{B}_{ri}^{(n)} = 0 \text{ and } \frac{\partial f(B, \tilde{\Omega}_l)}{\partial B_{ri}} > 0 \text{ if and only if } C_{ri} < \lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}.$$

The same reasoning can be applied to the case where $B_{ri} \leq 0$. This leads to $\hat{B}_{ri}^{(n)} = 0$ and $\frac{\partial f(B, \tilde{\Omega}_l)}{\partial B_{ri}} > 0$ if and only if $C_{ri} > -\lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}$.

So if $|C_{ri}| < \lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}$ we set $\hat{B}_{ri}^{(n+1)} = 0$.

if $|C_{ri}| > \lambda_1 (T - P) p^\alpha e^{\gamma \|s_i - s_j^{[l]}\|_2}$, we need to minimize the one dimensional function given by.

Acknowledgments

This research was supported by the National Science Foundation Grant CMMI-0954283 and the research funding provided by Predictix. The authors thank to Dr. Loren Williams for initiating and motivating this research idea and to Dr. Ming Yuan for helpful insights on the methodological aspects of this research.

References

- [1] Anderson, T. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, **22**: 327-351.
- [2] Bańbura, M., Giannone, D., Reichlin, L. (2010). Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, **25**: 71-92.
- [3] Bedrick, E. and Tsai, C. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**: 226-231.
- [4] Breiman, L. (1995) Better subset regression using the nonnegative garrote, *Technometrics*, **4**:373-384.
- [5] Brüggeman, R., Lütkepol, H. (2001). Lag selection in subset VAR models with an application to U.S. monetary system. In: *Friedmann,R., Knüppel, L., Lütkepol, H., Econometric Studies- A festschrift in honour of Joachim Frohn. LIT, Münster*,pp.107-128.
- [6] Bühlmann P. and Sarah Van de Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. *Heidelberg: Springer Verlag*.
- [7] Cripps E., Carter C. and Kohn R. (2005). Variable Selection and Covariance Selection in Multivariate Regression Models. *Handbook of Statistics*, **25**: 519-552.
- [8] Davis, R. Zang, P. and Zheng T. (2012). Sparse Vector Autoregressive Modeling *arXiv:1207.0520*.
- [9] De Luna, X., and Genton, M.G. (2005). Predictive spatio-temporal models for spatially sparse environmental data. *Statistica Sinica*, **15**: 547-568.
- [10] Hill, E., Fogli, A. and Perri F. (2012). Geography of the Great Recession. *NBER working paper No. 18447*.
- [11] Friedman,J., Hastie, T., Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*,**9**: 432-441.
- [12] Friedman,J., Hastie, T., Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso, *arXiv:1001.0736*.
- [13] Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression, *Biometrika*, **84**: 707-716.
- [14] Huang, H., Hsu, N., Theobald D.M., Breidt F.J. (2010) Spatial lasso with the application to GIS model selection, *Journal of Computational and Graphical Statistics*, **19**(4): 963-983.

- [15] Hsu, N.J., Hung, H.L. and Chang V.M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis*, **52**: 3645-3657.
- [16] Izenman A.J. (1975). Reduced-rank regression for the multivariate linear regression model. *Journal of Multivariate Analysis*, **5**(2): 248-264.
- [17] Krolzig, H.-M., Hendry, D.F. (2001). Computer automation of general linear regression model selection procedures, *Journal of Economic Dynamics and Control*, **25**: 831-866.
- [18] Lozano, A., Li, H., Niculescu-Mizil, A., Liu, Y. Perlich C., Hosking, J., Abe, N. (2009). Spatial-temporal causal modeling for climate change attribution, *IBM T.J. Watson Research Center*.
- [19] Lütkepohl, H. (2005). New introduction to multiple time series analysis, *Heidelberg: Springer Verlag*.
- [20] Obonzinski, G., Wainwright, M.J. and Jordan M.I., (2008). Union support recovery in high-dimensional multivariate regression. *Communication, Control and Computing*, 46th Annual Allerton Conference : 21-26.
- [21] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D., Pollack, J., Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, **4**: 53-77.
- [22] Penm, J.H.W., Terrell, R.D., (1982). On the recursive fitting of the subset autoregressions. *Journal of Time Series Analysis* **3**: 969 - 983.
- [23] Penm, J.H.W., Terrell, R.D., (1984). Multivariate subset autoregressive modeling with zero constraints for detecting causality. *Journal of Econometrics*, **24**: 311 - 330.
- [24] Reinsel, G., Velu, R. (1998). Multivariate reduced-rank regression: Theory and applications, *New York: Springer*.
- [25] Roecker,E. (1991). Prediction error and its estimation for subset selected models, *Technometrics*, **33** (4): 459-468.
- [26] Rothman, A., Levina, E., Zhu, J., (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, **19**(4): 947-962.
- [27] Sims C.A. (1980). Macroeconomics and reality. *Econometrica*. **48**(1): 1-48.
- [28] Song S., Bickel P. (2011). Large vector autoregressions, arXiv:1106.3915v1.
- [29] Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, **6**, 461-464.
- [30] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, series B*, **58**(1): 267-288.

- [31] B.A., Venables, W.N., and Wright, S.J. (2005). Simultaneous variable selection. *Technometrics*, **47**(3): 349-363.
- [32] Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society, series B*, **68**(1): 49-67.
- [33] Yuan, M., Ekici, A., Lu, Z., and Monteiro, R., (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of Royal Statistical Society, series B*, **69**(3): 329-346.
- [34] Zou, H., Hastie, T. and Tibshirani, T. (2007). On the "degrees of freedom" of the Lasso *Annals of Statistics*, **35**: 2173-2192.