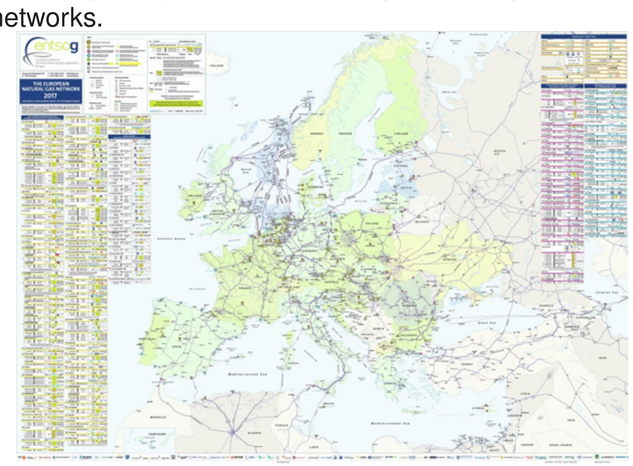
Variational Bayesian Inference for Network Autoregression Models

Wei-Ting Lai, Ray-Bing Chen, Ying Chen, Thorsten Koch

Itorduction

- Networks have emerged and become widely available in various fields such as energy transmission, logistics, transportation, and financial systems. Networks are dynamic in terms of their temporal dependence, and often have large scales.
- Tremendous growth and heterogeneity in both nodes/edges and dependence over time are the key characteristics of such networks.



- The temporal dependence of network can be represented in the vector autoregression (VAR) modeling framework. Also call the network autoregression model (NAR).
- Unlike univariate time series, the temporal dependence of a multivariate series consists of not only the serial dependence within each marginal series but also the interdependence across different marginal series.

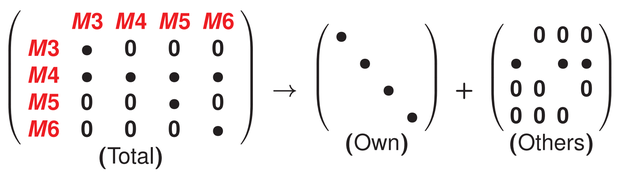
Structure Selection in Network Autoregression Model

- Consider the following NAR/VAR model with lag-$\boldsymbol{p}$,

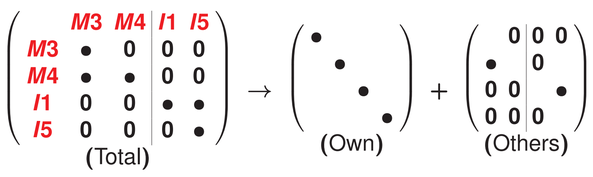
$$\boldsymbol{Y_t} = \boldsymbol{Y_{t-1}B_1} + \cdots + \boldsymbol{Y_{t-p}B_p} + \varepsilon_t, \tag{1}$$

  where $\boldsymbol{Y_t} = (\boldsymbol{y_{t,1}}, \cdots, \boldsymbol{y_{t,m}})$ is an $\mathbf{1} \times \boldsymbol{m}$ vector, $\boldsymbol{B_l}$ is a $\boldsymbol{m} \times \boldsymbol{m}$ coefficient matrix for $\boldsymbol{l} = \mathbf{1}, \cdots, \boldsymbol{p}$, and ε_t, $\boldsymbol{t} = \mathbf{1}, \ldots, \boldsymbol{T}$, are i.i.d. $\boldsymbol{MN}_{1\times m}(\mathbf{0}, \boldsymbol{\Sigma})$ random vectors.
- The total number of coefficient is $\mathcal{O}\left(\boldsymbol{m^2 p}\right)$.
- If the nodes $\boldsymbol{m}$ be large or the temporal dependence $\boldsymbol{p}$ increases, the NAR/VAR model may have an overparameterization problem.
- To overcome the overparameterization, structure sparse assumption is adopted here.
- Here we consider a Bayesian analysis approach for structured NAR/VAR models.
- Instead of MCMC algorithms which take huge computational cost, we develop a variational Bayesian method for structure selection in the NAR/VAR model.
- Following Song and Bickel [6], three structures are considered here,

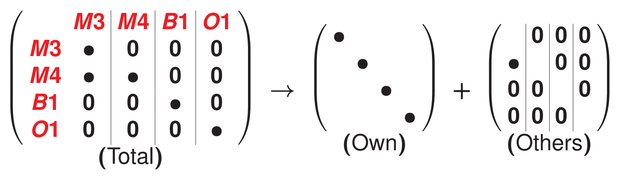
- Structure 1: Universal grouping (UG) in $\boldsymbol{B_l}$



- Structure 2: Segmentation grouping (SG) in $\boldsymbol{B_l}$



- Structure 3: No grouping (NG) in $\boldsymbol{B_l}$



- Among these three structures, it is clear that we can treat no grouping and universal grouping structures as special cases of the segmentized grouping structure.

Variational Bayesian Inference

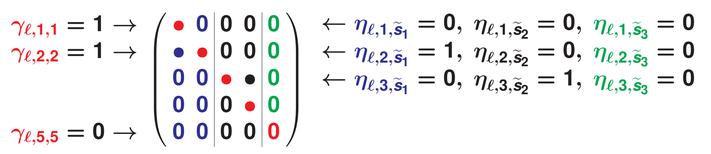
- Instead of directly generating the posterior samples, the variational Bayesian (VB) approach is used to identify the best approximation distribution of the true posterior for the further Bayesian inference.
- The Kullback-Leibler divergence (KL-divergence) is used to measure the difference among $\boldsymbol{P}$ (true posterior distribution) and $\boldsymbol{Q}$ (approximation posterior distribution). Thus given a proper assumption of $\boldsymbol{Q}$, in the variational Bayesian approach, we solve the following minimization problem,

$$\min_{\boldsymbol{Q}} \boldsymbol{KL(Q||P)} = \min_{\boldsymbol{Q}} \boldsymbol{E_Q}\left[\boldsymbol{\log\frac{Q}{P}}\right].$$

- This optimization problem can be solved via Expectation-Maximization (EM) type method.

Notation

- Take the segmentized grouping structure as an illustration.
- Let $\boldsymbol{S} = \left\{\boldsymbol{s_1}, \boldsymbol{s_2}, \cdots, \boldsymbol{s_g}\right\}$ as the overall index set for the columns in each $\boldsymbol{B_\ell}$, and let $\boldsymbol{s_k}$ denote the index set of the $\boldsymbol{k}$-th segment with size $\boldsymbol{n(s_k)}, \boldsymbol{k} = \mathbf{1}, \mathbf{2}, \cdots, \boldsymbol{g}$ and $\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{g}} \boldsymbol{n(s_k)} = \boldsymbol{m}$.
- Two indicators $\gamma_{\ell,i,i}$ and $\eta_{\ell,i,\tilde{s}_k}$ to identify the active structures in $\boldsymbol{B_\ell}$. Here $\tilde{\boldsymbol{s}}_{\boldsymbol{k}} = \boldsymbol{s_k} \setminus \{\boldsymbol{i}\}$



- Spike-and-slab prior:

$$\begin{aligned}
\boldsymbol{B_{\ell,i,i}}|\gamma_{\ell,i,i}, \sigma_B^2 &\sim \gamma_{\ell,i,i}\boldsymbol{N}(\mathbf{0}, \sigma_B^2) + (\mathbf{1} - \gamma_{\ell,i,i})\delta_{\mathbf{0}},\\
\boldsymbol{B_{\ell,i,\tilde{s}_k}}|\eta_{\ell,i,\tilde{s}_k}, \sigma_B^2 &\sim \eta_{\ell,i,\tilde{s}_k}\boldsymbol{MN}_{1\times n(\tilde{s}_k)}(\mathbf{0}, \sigma_B^2 \boldsymbol{I}_{n(\tilde{s}_k)}) + (\mathbf{1} - \eta_{\ell,i,\tilde{s}_k})\delta_{\mathbf{0}},\\
\gamma_{\ell,i,i} &\sim \mathcal{B}\boldsymbol{er}(\pi_1),\ \eta_{\ell,i\tilde{s}_k} \sim \mathcal{B}\boldsymbol{er}(\pi_2).
\end{aligned}$$

The Proposed Variational Bayesian Approach

- To simplify the model structure, we reparametrize the coefficient as the product of the coefficient and the indicator.

$$\begin{aligned}
\boldsymbol{B_{\ell,i,i}} &= \gamma_{\ell,i,i}\widetilde{\boldsymbol{B}}_{\ell,i,i},\ \widetilde{\boldsymbol{B}}_{\ell,i,i}|\sigma_B^2 \sim \boldsymbol{N}(\mathbf{0}, \sigma_B^2)\\
\boldsymbol{B_{\ell,i,\tilde{s}_k}} &= \eta_{\ell,i,\tilde{s}_k}\widetilde{\boldsymbol{B}}_{\ell,i,\tilde{s}_k},\ \widetilde{\boldsymbol{B}}_{\ell,i,\tilde{s}_k}|\sigma_B^2 \sim \boldsymbol{N}_{n(\tilde{s}_k)}(\mathbf{0}, \sigma_B^2 \boldsymbol{I}_{n(\tilde{s}_k)})
\end{aligned}$$

- The product forms $\eta_{\ell,i,\tilde{s}_k}\widetilde{\boldsymbol{B}}_{\ell,i,\tilde{s}_k}$ and $\gamma_{\ell,i,i}\widetilde{\boldsymbol{B}}_{\ell,i,i}$ are the same as the spike and slab priors.
- $\boldsymbol{\log(P(Y))} = \boldsymbol{L(q)} + \boldsymbol{KL(q||P)}$, where $\boldsymbol{L(q)} = \int \sum_\gamma \sum_\eta -\boldsymbol{q}(\widetilde{\boldsymbol{B}}, \gamma, \eta) \log\left(\frac{\boldsymbol{q}(\widetilde{\boldsymbol{B}}, \gamma, \eta)}{\boldsymbol{P}(\widetilde{\boldsymbol{B}}, \gamma, \eta, \boldsymbol{Y}|\boldsymbol{X}; \theta)}\right) \boldsymbol{d}\widetilde{\boldsymbol{B}}$.

$$\min_{\boldsymbol{q}} \boldsymbol{KL(q||P)} \Longleftrightarrow \max_{\boldsymbol{q}} \boldsymbol{L(q)}.$$

- Based on prior assumptions, the proposed variational Bayesian algorithm iterates the following steps.
  E-step: take the expectation of $\boldsymbol{L(q)}$ with respect to η, γ and $\widetilde{\boldsymbol{B}}$ to update $\boldsymbol{q}(\eta, \gamma, \widetilde{\boldsymbol{B}})$.

$$\begin{aligned}
\boldsymbol{q}(\eta, \gamma, \widetilde{\boldsymbol{B}}) = \prod_{\ell}^{\boldsymbol{p}} \prod_{i}^{\boldsymbol{m}} \prod_{k}^{\boldsymbol{g}} &\left(\phi_{\ell,i,i}\boldsymbol{N}\left(\mu_{\ell,i,i}, \boldsymbol{\Sigma}_{\boldsymbol{B}_{\ell,i,i}}\right)\right)^{\gamma_{\ell,i,i}} \left(\left(\mathbf{1} - \phi_{\ell,i,i}\right)\boldsymbol{N}(\mathbf{0}, \sigma_B^2)\right)^{(\mathbf{1}-\gamma_{\ell,i,i})}\\
&\left(\phi_{\ell,i,\tilde{s}_k}\boldsymbol{MN}_{1\times n(\tilde{s}_k)}(\mu_{\ell,i,\tilde{s}_k}, \boldsymbol{\Sigma}_{\boldsymbol{B}_{\ell,i,\tilde{s}_k}})\right)^{\eta_{\ell,i,\tilde{s}_k}} \left(\left(\mathbf{1} - \phi_{\ell,i,\tilde{s}_k}\right)\boldsymbol{MN}_{1\times n(\tilde{s}_k)}(\mathbf{0}, \sigma_B^2 \boldsymbol{I}_{n(\tilde{s}_k)})\right)^{(\mathbf{1}-\eta_{\ell,i,\tilde{s}_k})},
\end{aligned}$$

  M-step: take the derivative of $\boldsymbol{L(q)}$ with respect to $\theta = \left\{\hat{\boldsymbol{\Sigma}}, \hat{\sigma}_B^2, \hat{\pi}_1, \hat{\pi}_2\right\}$.

Simulation

UG: The true models follow the universal grouping structure with $(\boldsymbol{m}, \boldsymbol{p}) = (\mathbf{10}, \mathbf{5})$, $(\mathbf{20}, \mathbf{5})$ and $(\mathbf{50}, \mathbf{5})$, respectively. There are **72**, **145** and **355** nonzero coefficients in three cases.

SG: The segmentation structure is considered in these three simulations with $\boldsymbol{m} = \mathbf{10}$, **20**, and **50**. Let $\boldsymbol{S}_{m,g}$ denote a specific group structure with $\boldsymbol{m}$ time series for $\boldsymbol{g}$ disjoint groups. We set $\boldsymbol{S}_{10,3} = \{(\mathbf{1}, \mathbf{2}, \mathbf{3}), (\mathbf{4}, \mathbf{5}, \mathbf{6}), (\mathbf{7}, \mathbf{8}, \mathbf{9}, \mathbf{10})\}$ for $(\boldsymbol{m}, \boldsymbol{p}) = (\mathbf{10}, \mathbf{5})$, $\boldsymbol{S}_{20,4} = \{(\mathbf{1}, \ldots, \mathbf{5}), (\mathbf{6}, \ldots, \mathbf{10}), (\mathbf{11}, \mathbf{12}, \mathbf{13}), (\mathbf{14}, \ldots, \mathbf{20})\}$ for $(\boldsymbol{m}, \boldsymbol{p}) = (\mathbf{20}, \mathbf{5})$, and $\boldsymbol{S}_{50,8} = \{(\mathbf{1}, \ldots, \mathbf{5}), (\mathbf{6}, \ldots, \mathbf{10}), (\mathbf{11}, \mathbf{12}, \mathbf{13}), (\mathbf{14}, \ldots, \mathbf{20}), (\mathbf{21}, \ldots, \mathbf{30}), (\mathbf{31}, \ldots, \mathbf{35}), (\mathbf{36}, \ldots, \mathbf{40}), (\mathbf{41}, \ldots, \mathbf{50})\}$ for $(\boldsymbol{m}, \boldsymbol{p}) = (\mathbf{50}, \mathbf{5})$. There are 40, 109, and 360 nonzero coefficients, respectively.

NG: In the "no grouping" cases, there are 18, 27 and 128 nonzero coefficients for $(\boldsymbol{m}, \boldsymbol{p}) = (\mathbf{10}, \mathbf{5})$, $(\mathbf{20}, \mathbf{5})$, and $(\boldsymbol{m}, \boldsymbol{p}) = (\mathbf{50}, \mathbf{5})$ respectively.

- The simulation structures with $\boldsymbol{T} = \mathbf{301}$ in $\boldsymbol{m} = \mathbf{20}$ and **10** and $\boldsymbol{T} = \mathbf{701}$ in $\boldsymbol{m} = \mathbf{50}$ are conducted according to Chu et al. [3]. The simulation results are summarized based on 100 replications.
- For the comparison purpose, we implement VAGSA [3] and BIVAS [5].
- Initial prior π_1 and π_2 for m50NG are 0.5, other cases are 0.01.

Table: Numerical comparison results for VB-NAR, VAGSA and BIVAS for the cases with $\boldsymbol{p} = \mathbf{10}$.

VB-NAR/VAGSA/BIVAS	TPR(%)	FPR(%)	AMS	MSPE	Ave. CPU Time (sec)
m10UG-I_{10}	100/100/98	0.07/0.31/1.66	72.62/74.90/86.02	1.00/1.01/0.98	4/148/8
m10UG-Σ_{10}	100/100/97	0.06/0.28/1.79	72.51/74.59/86.59	1.00/0.83/0.94	5/138/9
m10SG-I_{10}	100/100/99	0.15/0.63/0.63	41.35/46.02/45.57	1.07/1.09/1.08	12/303/8
m10SG-Σ_{10}	100/100/98	0.13/0.26/0.74	41.17/46.02/46.16	0.86/1.09/0.90	11/303/9
m10NG-I_{10}	98/97/96	0.15/0.11/0.26	19.07/18.58/19.91	1.00/1.00/1.02	40/628/16
m10NG-Σ_{10}	99/99/96	0.11/0.08/0.33	18.86/18.55/20.55	0.89/0.88/0.89	36/653/18
m20UG-I_{20}	100/100/94	0.03/0.18/1.04	145.79/151.89/178.23	0.98/0.99/0.98	15/236/83
m20UG-Σ_{20}	100/100/97	0.02/0.14/0.63	145.56/150.54/45.57	0.95/0.95/1.00	180/237/93
m20SG-I_{20}	100/100/89	0.06/0.26/0.48	109.25/117.05/115.81	1.10/1.11/1.17	44/652/63
m20SG-Σ_{20}	100/100/99	0.06/0.21/0.79	109.22/115.20/117.20	0.90/0.90/0.99	51/672/70
m20NG-I_{20}	97/98/94	0.08/0.15/0.12	29.29/32.39/30.03	0.99/0.99/0.98	291/2183/144
m20NG-Σ_{20}	97/98/92	0.06/0.13/0.14	28.55/31.43/30.49	0.86/0.86/0.95	262/2247/151
m50UG-I_{50}	100/-/97	0.00/-/0.26	355.47/-/409.03	1.02/-/2.66	213/-/769
m50UG-Σ_{50}	100/-/98	0.00/-/0.27	355.57/-/412.43	0.87/-/2.28	239/-/731
m50SG-I_{50}	100/-/90	0.01/-/0.14	361.58/-/394.43	1.02/-/3.50	1082/-/645
m50SG-Σ_{50}	100/-/89	0.01/-/0.16	361.15/-/395.07	0.88/-/3.09	1767/-/663
m50NG-I_{50}	100/-/98	0.03/-/0.07	134.67/-/143.61	1.02/-/1.92	7860/-/1163
m50NG-Σ_{50}	92/-/98	0.03/-/0.08	123.33/-/145.00	0.89/-/1.64	58194/-/1177

Empirical Study

- Natural gas flows for the German network with 51 nodes from Oct 1st, 2013, to Sep 30th, 2015
- The segmented grouping structures based on the types of nodes that are used for the NAR analysis.
- We set the number of lags $\boldsymbol{p}$ as **14** to incorporate social dependence up to two weeks ahead.
- The data from Oct 1st, 2013, to March 31th, 2015, is used as the in-sample training data.
- We perform a one-step ahead forecast for remaining **0.5** years.

In order to illustrate the performance of the proposed VB method, in addition to the mean absolute percentage estimator (MAPE), the normalized MSE (NRMSE) for the 51 nodes is also used.

- $\boldsymbol{MAPE} = \frac{\mathbf{1}}{\boldsymbol{N}\times \#\text{ of type element}} \sum_{\boldsymbol{t}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\#\text{ of type element}} \frac{|\boldsymbol{Y}_{\boldsymbol{t}+\mathbf{1},\boldsymbol{i}} - \hat{\boldsymbol{Y}}_{\boldsymbol{t}+\mathbf{1},\boldsymbol{i}}|}{|\boldsymbol{Y}_{\boldsymbol{t}+\mathbf{1},\boldsymbol{i}}|}$,
- $\boldsymbol{NRMSE} = \frac{\frac{\mathbf{1}}{\boldsymbol{N}\times \#\text{ of type element}} \left(\sum_{\boldsymbol{t}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\#\text{ of type element}} (\boldsymbol{Y}_{\boldsymbol{t}+\mathbf{1},\boldsymbol{i}} - \hat{\boldsymbol{Y}}_{\boldsymbol{t}+\mathbf{1},\boldsymbol{i}})^2\right)}{\frac{\mathbf{1}}{\boldsymbol{N}\times \#\text{ of type element}} \sum_{\boldsymbol{t}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\#\text{ of type element}} \boldsymbol{Y}_{\boldsymbol{t}+\mathbf{1},\boldsymbol{i}}}$.

Table: The MAPEs and NRMSEs for the fitted and forecast results classified by type.

Type	Number	MAPE (%)	Range (%)	SD	NRMSE	Range	SD
In sample from 01/10/2013 to 31/03/2015							
Municipal	34	6.28	(3.20, 14.70)	1.93	0.09	(0.04, 0.36)	0.05
Industrial	11	11.69	(0.89, 34.22)	11.74	0.12	(0.01, 0.31)	0.10
Border	1	9.36	–	–	0.11	–	–
Others	5	13.05	(3.78, 36.30)	13.26	0.14	(0.05, 0.34)	0.12
Out of sample from 01/04/2015 to 30/09/2015							
Municipal	34	7.89	(4.66, 11.85)	1.60	0.10	(0.06, 0.15)	0.02
Industrial	11	15.98	(1.70, 54.57)	19.00	0.12	(0.02, 0.30)	0.10
Border	1	11.77	–	–	0.14	–	–
Others	5	8.65	(5.52, 13.47)	3.32	0.10	(0.07, 0.12)	0.02

Summary

- We focus on Bayesian analysis for NAR/VAR models, especially when there are many time series are taken into considered simultaneously.
- The simulation results support that the proposed VB approach not only identifies the proper active structures in NAR/VAR models but also significantly reduces the computational cost.
- Finally, German gas flow network data with 51 nodes are analyzed to illustrate the performance of the proposed method. The analytical results of the proposed VB method yield the trends of nodes, which may be useful for assisting operators with performing appropriate operations.

References

1 Carbonetto P, Stephens M. (2012), Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal 7: 73–108.

2 Christopher M. Bishop. (2006), Pattern Recognition and Machine Learning. Springer-Verlag New York.

3 Chi-Hsiang Chu, Mong-Na Lo Huang, Shih-Feng Huang, and Ray-Bing Chen. Bayesian structure selection for vector autoregression model. Journal of Forecasting, 38:422–439, 2019.

4 George, E. I. and McCulloch, R. E. (1993), Variable selection via Gibbs sampling. J. Am. Statist. Ass., 88, 881–889.

5 Mingxuan Cai, Mingwei Dai, Jingsi Ming, Heng Peng, Jin Liu, and Can Yang. Bivas: a scalable bayesian method for bi-level variable selection with applications. Journal of Computational and Graphical Statistics, 29(1):40–52, 2020.

6 Song, S. and Bickel, P. J. (2011), Large vector auto regressions. Preprint.

7 Titsias, M. K. and Lazaro-Gredilla, M. (2011), Spike and slab variational inference for multi-task and multiple kernel learning. In NIPS.