The Variational Bayesian Inference for Network Autoregression Models

Wei-Ting Lai

Department of Statistics, National Cheng Kung University, Tainan, Taiwan

Ray-Bing Chen

Department of Statistics, National Cheng Kung University, Tainan, Taiwan

Institute of Data Science, National Cheng Kung University, Tainan, Taiwan

Ying Chen

Department of Mathematics, National University of Singapore, Singapore

Risk Management Institute, National University of Singapore, Singapore

Institute of Data Science, National University of Singapore, Singapore

Thorsten Koch

Chair of Software and Algorithms for Discrete Optimization,

Technische Universität Berlin, Berlin, Germany

Department of Applied Algorithmic Intelligence Methods,

Zuse Institute Berlin, Berlin, Germany

Abstract: We develop a variational Bayesian (VB) approach for estimating large-scale dynamic network models in the network autoregression framework. The VB approach allows for the automatic identification of the dynamic structure of such a model and obtains a direct approximation of the posterior density. Compared to Markov Chain Monte Carlo (MCMC) based sampling approaches, the VB approach achieves enhanced computational efficiency without sacrificing estimation accuracy. In the simulation study conducted here, the proposed VB approach detects various types of proper active structures for dynamic network models. Compared to the alternative approach, the proposed method achieves similar or better accuracy, and its computational time is halved. In a real data analysis scenario of day-ahead natural gas flow prediction in the German gas transmission network with 51 nodes between October 2013 and September 2015, the VB approach delivers promising forecasting accuracy along with clearly detected structures in terms of dynamic dependence.

1 Introduction

Networks have emerged and become available in various fields, such as energy transmission, logistics and transportation, and financial systems. Networks are dynamic in terms of their temporal dependence, and they often have large scales. The understanding and inference of network dynamics have profound implications for operations and decision making in modern industries. Tremendous growth and heterogeneity in both nodes/edges and dependence over time are the key characteristics of such networks. However, conventional statistical methods either assume that networks are static or consider only low-dimensional temporal data. This creates a need for efficient computational approaches that are able to reveal the essential dependence structures in high dimensions and simultaneously deliver accurate inferences with a low computational cost.

Industrial networks contain series of temporal-spatial data collected over time. While the nodes/edges are often fixed or possess trivial changes, the lead-lag temporal dependence issue can no longer be ignored in network inference. Graph theory has been widely used for unraveling structural information in large-scale network analysis. For example, Fan et al. (2009) and Guo et al. (2011) proposed a sparse graphic network. Liu et al. (2012) proposed the semiparametric Gaussian copula graphical model. Despite their efficiency, these graphical models assume static networks, and the evolution of the network dependence is not considered in their estimation processes.

The temporal dependence of a network can be represented in the vector autoregressive (VAR) modeling framework. In the VAR framework, each node is considered as one time series, and the network dependence is measured as the lead-lag cross-correlations among multiple time series. Both the theoretical properties and empirical performance of VAR have been well studied with respect to multivariate data; see Lütkepohl (2007) and Bańbura et al. (2010). However, the application of VAR for large-scale network analysis is still challenging. Given a network with m models, which correspond to m time series, and supposing that the dynamics depend on the last p lags, to the result is that there are pm^2 unknown coefficients in the VAR model. When the number of nodes m becomes large or the temporal dependence p increases, VAR is overparameterized and this leads to low estimation accuracy or even

infeasibility with regard to model inference.

A unique feature in industrial networks is that their lead-lag temporal dependences, such as concurrent dependence among their nodes, are less dense than the networks themselves. Individual networks are also much sparser than other social networks. It is conceivable that large-scale industrial networks, are driven by a few essential and cohesive connections among their nodes to facilitate network evolution. This motivates the modeling of large-scale dynamic networks using sparse VAR. In particular, penalties are imposed on the parameter space of the VAR framework with various possible types of structural assumptions. For the purpose of both estimation and interpretability, structural sparsity can be enabled in elements, groups and lags. Lag sparsity investigates the effect of time-lagged information, while group sparsity highlights the impacts of certain nodes on others. In addition to the universal effect on a group of lag coefficients, a sparse element illustrates a single effect. Basu & Michailidis (2015) investigated the theoretical properties of ℓ_1-regularized estimates, where multiple time series were assumed to be stable Gaussian processes. Melnyk & Banerjee (2016) established bounds on the non-asymptotic estimation error of the Lasso-type estimator for structured VAR parameters. Nicholson et al. (2014) proposed several structures for VAR and Lasso, group Lasso and sparse group penalty functions to achieve sparsity in the elements and groups of a network; see also Hsu et al. (2008), Song & Bickel (2011), and Chen et al. (2020). Moreover, VAR models can be easily extended to the above three kinds of sparsity by building up a hierarchical lag structure in the autoregression model via the inclusion of high-dimensional exogenous variables.

The estimation of the structured of a VAR framework faces two challenges. First, the sparse structure needs to be specified to avoid the overfitting of the high-dimensional models. Second, the framework should be able to adapt to different kinds of stochastic behaviors, as empirical data are likely non-Gaussian. Bayesian methods are natural choices because they deliver stable performances without prespecified assumptions about the structure and distribution of the model. Stochastic search variable selection (SSVS), for example, is the most commonly used Bayesian variable selection approach. It introduces latent indicators embedded in the priors and stochastically searches subsets by generating posterior samples with the Markov Chain Monte Carlo (MCMC) algorithm; see George & McCulloch (1993).

Geweke (1996) proposed the component-wise Gibbs method for the purpose of improving computational efficiency. By using the spike-and-slab prior, it avoids computing inverse matrices and thus reduces the computational cost. However, the Gibbs sampler requests either random or systematic updating of the coefficients. Chen et al. (2011) proposed the stochastic matching pursuit (SMP) algorithm, which updates the coefficients of each step for obtaining the best fit based on the current residual vector. In terms of structural selection, Farcomeni (2010) introduced Bayesian-constrained variable selection. Chen et al. (2016) proposed the groupwise Gibbs sampler. As a proof of concept, Chu et al. (2019) implemented a Bayesian variable selection approach in the VAR framework and named it the VAGSA, the vector autoregression-based Gibbs sampler algorithm. For the high-dimensional VAR model, Kastner & Huber (2020) proposed a large Bayesian vector autoregression approach with a Dirichlet-Laplace prior and factor stochastic volatility (FSV), and they applied it on high-dimensional US economic data.

Nevertheless, these MCMC algorithms are known to be computationally expensive for sequentially generating posterior samples. Variational inference, as an alternative, shows great potential in terms of improving computational speed without sacrificing much accuracy. It obtains an approximation of the target posterior density using the Kullbak-Leibler divergence, based on which an EM-type algorithm is devised a reduced computational cost. Titsias & Lázaro-Gredilla (2011) and Carbonetto & Stephens (2012) introduced variational Bayesian approaches with spike-and-slab priors for dealing with variable selection problems in linear regression models. Cai et al. (2020) proposed a variational Bayesian method for sparse group selection in linear models and extended it to multiple response models.

In our study, we propose a variational Bayesian (VB) approach for estimating a large-scale dynamic network model. The serial dependence in a given network is represented in a vector autoregression framework with three possible types of structural assumptions and various nesting types. Here, we also call this model a network autoregression (NAR) model. We derive variational inferences and develop the corresponding algorithms. The VB approach allows for the automatic identification of the dynamic structure of data and obtains an approximation of the posterior density directly. Compared to MCMC-based sampling approaches, such as the VAGSA in Chu et al. (2019), the VB approach achieves enhanced

numerical performance with similar accuracy. A simulation study shows that compared with existing methods, the VB approach not only detects the proper active structures in various dynamic network models but also halves the computational time, with similar or better accuracies. In a real data analysis, we predict day-ahead natural gas flows for a German network with 51 nodes over 2 years from Oct 1, 2013, to Sep 30, 2015. Germany's gas transport system is essential to the European energy supply. The adequate, high-precision estimation of supply and demand is a crucial issue for efficient control and operations in gas transmission. The VB approach delivers a clear dynamic dependence structure, providing interpretability and insights for understanding and managing the gas transmission network. To the best of our knowledge, this is the first attempt to derive variational inference for a large-scale dynamic network analysis in a structured NAR/VAR framework.

This paper is organized as follows. Section 2 introduces the dynamic network model in the VAR framework. Several types of structural assumptions are also demonstrated. Section 3 presents the proposed variational Bayesian algorithms for large-scale dynamic network inference. Section 4 investigates the finite-sample performance of the proposed VB approach. Section 5 reports the network inference for day-ahead gas flow forecasting with 51 high-pressure nodes in the natural gas transmission network in Germany. Section 6 provides a brief conclusion of our work.

2   Model

Let $\boldsymbol{Y}_t \in \mathbb{R}^{1 \times m}$ denote a vectorized time series of networks with m nodes at time t over the time period $[1, T]$. Without loss of generality, $\boldsymbol{Y}_t$ is demeaned. We consider a dynamic network model in a vector autoregression framework $\boldsymbol{Y}_t$ which is assumed to depend on the past values of the network at lags 1 to p, i.e.,

$$\boldsymbol{Y}_t = \boldsymbol{Y}_{t-1} B_1 + \cdots + \boldsymbol{Y}_{t-p} B_p + \epsilon_t, \quad t = p+1, p+2, \ldots, T.$$

Here, we let $\boldsymbol{B}_\ell$ be an $m \times m$ coefficient matrix for lag-ℓ, $\ell = 1, 2, \cdots, p$, which is used to measure the lead-lag temporal dependence in the network. The number of coefficients in $\boldsymbol{B}_\ell$

grows quadratically with the number of nodes m. Given $m = 51$ in the gas transformation network, there are $51^2 = 2601$ unknown coefficients for each $\boldsymbol{B}_\ell$. Obviously, the diagonal elements of $\boldsymbol{B}_\ell$ indicate the serial dependence of each node on its own lag-ℓ value and the dependences of off-diagonal elements on other nodes' lag-ℓ values. The term $\{\epsilon_t\}_{t=p+1}^T$ is a sequence of serially uncorrelated $1 \times m$ random vectors, with a mean vector of zero and a covariance matrix Σ. Then, the dynamic network model can be represented in matrix form as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y} \in \mathbb{R}^{(T-p) \times m}$ is the response matrix, $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_\ell)$ is a $(T - p) \times m$ matrix with $\boldsymbol{X}_\ell = \left(\boldsymbol{Y}'_{p+1-\ell}, \boldsymbol{Y}'_{p+2-\ell}, \ldots, \boldsymbol{Y}'_{T-\ell} \right)'$, $\ell = 1, 2, \cdots, p$, $\boldsymbol{B} = \left(\boldsymbol{B}'_1, \boldsymbol{B}'_2, \ldots, \boldsymbol{B}'_p \right)'$, and $\boldsymbol{\epsilon} = \left(\boldsymbol{\epsilon}'_{p+1}, \boldsymbol{\epsilon}'_2, \cdots, \boldsymbol{\epsilon}'_T \right)'$.

In this paper, we consider a structured NAR/VAR framework, i.e., dynamic dependence is sparsely motivated by a large-scale industrial network, such as the German gas transformation network. Figure 1 displays the lag-1 and lag-3 cross-correlations of 11 nodes arbitrarily selected from the German natural gas transmission network. The 11 nodes belong to 4 different types: municipal (labeled with M), industrial (I), border (B), and others (O). The left-hand side of Figure 1 is the lag-1 cross-correlation matrix. The right-hand side of Figure 1 is the cross-correlation matrix for lag-3. This shows the coexistence of strong serial dependence and sparsity in elements, groups, and lags. According to Figure 1, due to their cross-correlation values, nodes sharing the same type may possess similar patterns. Thus, the dynamics of the network are not driven by each node individually, every group of nodes, or each lagged network in the past.

While element sparsity and lag sparsity are clear, there are different kinds of group sparsity. Following Song & Bickel (2011), we categorize the various structures into three types and discuss them as follows.

- UG Structure: The universal grouping (UG) structure in the coefficient matrix $\boldsymbol{B}_\ell$ means that the off-diagonal coefficients have the same sparsity pattern across the different columns. For example, municipal nodes M3, M4, M5, and M6 may have similar
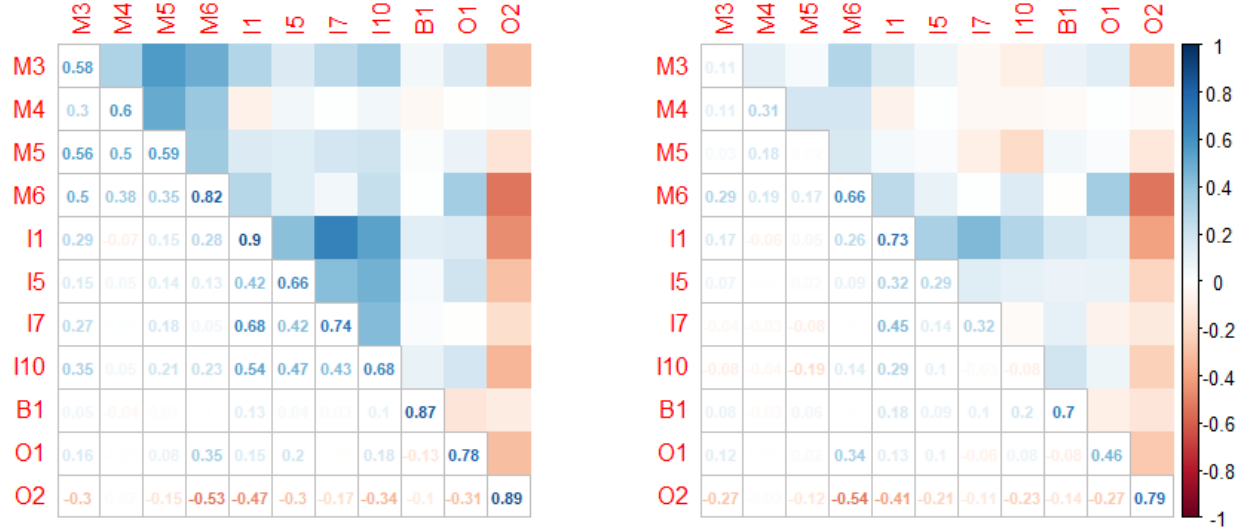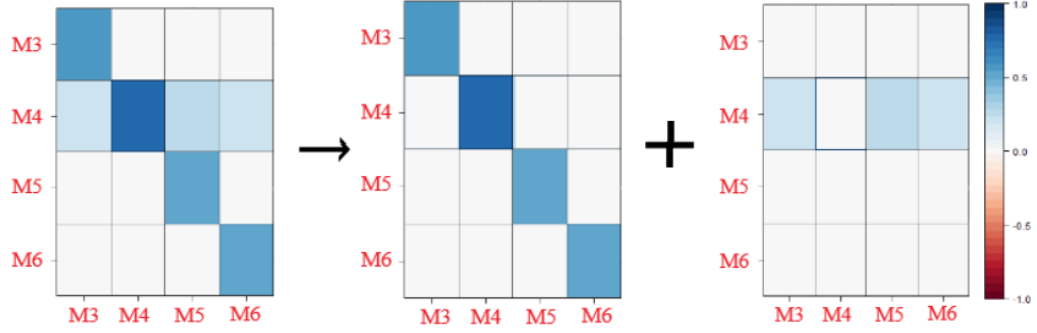
6

Figure 1: The cross-correlations of 11 nodes arbitrarily selected from the German natural gas transmission network.

patterns, and we can treat them as a group. Therefore, if one node affects another node, the others are all influenced by this node. As shown in Figure 2, node M4 affects other nodes, and nodes M3, M5 and M6 do not affect other nodes. As such, the coefficient matrix can be separated into a diagonal matrix reflecting the dynamic dependence of each node on its own, and a sparse matrix showing the temporal dependence of each node on others. In network analysis, the columns that have the same row sparsity are grouped together.

- SG Structure: For the segmentation grouping (SG) structure in B_ℓ, the nodes in the same segment interact with each other but are independent from the other nodes. This is associated with the empirical observation that there are different types of nodes in the network. Figure 3 illustrates the SG structure among municipal and industrial nodes. The network is divided into disjoint segments of the municipal and industrial types. The estimation process can be conducted segment by segment.

7

$$
\begin{pmatrix} \bullet & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & 0 \\ 0 & 0 & 0 & \bullet \end{pmatrix} \rightarrow \begin{pmatrix} \bullet & & & \\ & \bullet & & \\ & & \bullet & \\ & & & \bullet \end{pmatrix} + \begin{pmatrix} & 0 & 0 & 0 \\ \bullet & & \bullet & \bullet \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & \end{pmatrix}
$$
$$
\text{(Total)} \qquad\qquad \text{(Own)} \qquad\qquad \text{(Others)}
$$

Figure 2: An illustration of the coefficient matrix for the universal grouping structure.

- NG Structure: Consider the "no grouping" (NG) structure in B_ℓ. Each node has its own impact on the other nodes over time. This is a scenario where no regular pattern is identified among the nodes. Figure 4 shows an example with different types of nodes. In fact, there are no similar patterns among them. In this case, we separate them during the inference procedure.

It is easy to see that the first and third types of structures, UG and NG, are two special cases of the SG structure when the numbers of segments are 1 and m, respectively. While the three types of structures are defined for each coefficient matrix, they may appear in different time lags. The identification of a proper structure can help not only increase the estimation accuracy but also enhance computational time of the algorithm.
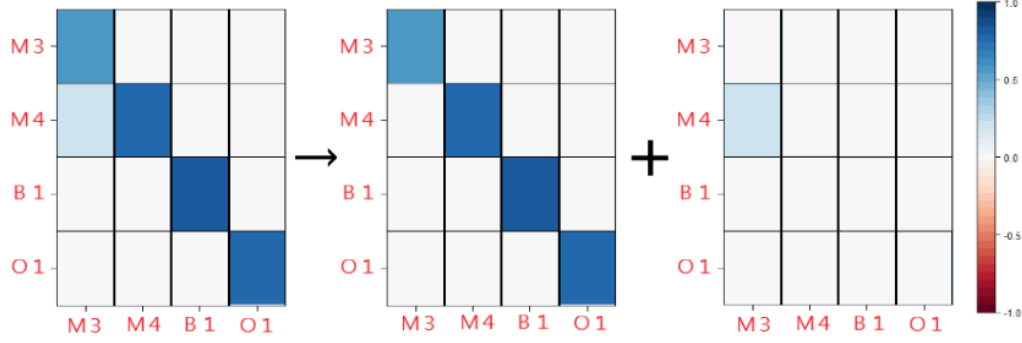
$$
\begin{pmatrix}
\bullet & 0 & 0 & 0 \\
\bullet & \bullet & 0 & 0 \\
0 & 0 & \bullet & \bullet \\
0 & 0 & 0 & \bullet
\end{pmatrix}
\rightarrow
\begin{pmatrix}
\bullet & & & \\
 & \bullet & & \\
 & & \bullet & \\
 & & & \bullet
\end{pmatrix}
+
\begin{pmatrix}
 & 0 & 0 & 0 \\
\bullet & & & 0 \\
0 & 0 & & \bullet \\
0 & 0 & 0 &
\end{pmatrix}
$$

(Total)   (Own)   (Others)

Figure 3: An illustration of the coefficient matrix for the segmentation grouping structure.

3   Variational Approximation Algorithm

We introduce the variational Bayesian approach for NAR/VAR structure inference in high-dimensional scenarios, and it is expected to automatically choose the proper structure and perform estimation at a low computational cost. Here, we derive only the VB method for the segmentation grouping structure because it nests the other two types of structures as special cases of segmentation grouping. Thus, it is a derivation under a general setup. To simplify the procedure, we assume that all coefficient matrices share the same segmentation grouping structure. Denote $S = \{s_1, s_2, \cdots, s_g\}$ as an overall index set for the columns in each B_ℓ, where s_k is the index set of the kth segment with size $0 < |s_k| < m$ and $\sum_{k=1}^{g} |s_k| = m$, and g is the total number of segments (groups). In this section, we start by introducing the

$$
\begin{pmatrix} \bullet & 0 & 0 & 0 \\ \bullet & \bullet & 0 & 0 \\ 0 & 0 & \bullet & 0 \\ 0 & 0 & 0 & \bullet \end{pmatrix} \rightarrow \begin{pmatrix} \bullet & & & \\ & \bullet & & \\ & & \bullet & \\ & & & \bullet \end{pmatrix} + \begin{pmatrix} & 0 & 0 & 0 \\ \bullet & & 0 & 0 \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & \end{pmatrix}
$$

$$
\text{(Total)} \qquad\qquad \text{(Own)} \qquad\qquad \text{(Others)}
$$

Figure 4: An illustration of the coefficient matrix for the "no grouping" structure.

Bayesian structure selection approach, and then we mention the details of the proposed VB method.

3.1   The Bayesian structure selection algorithm

For the Bayesian hierarchical model used to select segmentation structures, Chu et al. (2019) introduced two indicators $\gamma_{\ell,i}$ and $\eta_{\ell,i,k}$ to identify the active structures. These indicators denote the active nonzero coefficients of the kth segment in the ith row in B_ℓ with respect to the lag values of itself and others. For its own lags, $\gamma_{\ell,i} = 1$ denotes that the element of the ith row and the ith column in the coefficient matrix for lag-ℓ, $B_{\ell,i,i}$, is nonzero, and $\gamma_{\ell,i} = 0$ indicates a coefficient of zero, i.e., $B_{\ell,i,i} = 0$. In addition, $\eta_{\ell,i,k} = 1$ indicates that the ith row in B_ℓ and the columns of the kth segment for the off-diagonal elements in B_ℓ are all nonzero, i.e. $B_{l,i,\widetilde{s}_k} \neq \mathbf{0}$, and $\eta_{l,i,k} = 0$ otherwise. Consider the Bayesian structure

selection approach for determining the segmentation structure in the NAR/VAR model. Following the literature, the prior distributions of the indicators $\gamma_{\ell,i}$ and $\eta_{\ell,i,k}$ are chosen to be independent Bernoulli distributions with $P\left(\gamma_{\ell,i}=1\right)=\pi_1$ and $P\left(\eta_{\ell,i,k}=1\right)=\pi_2$, i.e., $Ber(\pi_1)$ and $Ber(\pi_2)$, respectively. The priors of the elements in matrix B_ℓ are dependent on $\boldsymbol{\gamma}_\ell$ and $\boldsymbol{\eta}_\ell$, which are the vectors of the indicators in lag-ℓ, as follows:

$$B_{\ell,i,i}|\gamma_{\ell,i},\sigma_B^2 \sim \gamma_{\ell,i}N(0,\sigma_B^2)+(1-\gamma_{\ell,i})\delta_0, \tag{2}$$

$$B_{\ell,i,\widetilde{s}_k}|\eta_{\ell,i,k},\sigma_B^2 \sim \eta_{\ell,i,k}MN_{1\times|\widetilde{s}_k|}(\mathbf{0},\boldsymbol{I},\sigma_B^2 I_{|\widetilde{s}_k|})+(1-\eta_{\ell,i,k})\delta_\mathbf{0}, \tag{3}$$

where $\widetilde{s}_k$ is the index set of the kth segment except i, i.e., $i\in s_k, \widetilde{s}_k = s_k\setminus\{i\}$, and $|\widetilde{s}_k|$ denotes the number of elements contained in $\widetilde{s}_k$. In addition, $MN_{1\times|\widetilde{s}_k|}(\mathbf{0},\boldsymbol{I},\sigma_B^2 I_{|\widetilde{s}_k|})$ is a $1\times|\widetilde{s}_k|$ multivariate normal distribution with a mean vector of $\mathbf{0}$ and a covariance matrix, $\sigma_B^2 I_{|\widetilde{s}_k|}$, and δ_0 and $\delta_\mathbf{0}$ are a point mass at 0 and a zero vector $\mathbf{0}$, respectively. Therefore, the coefficient prior is a mixture prior of the normal distribution and a point mass. Last, $\left(B_{\ell,i,i},\gamma_{\ell,i}\right)$ and $\left(B_{\ell,i,\widetilde{s}_k},\eta_{\ell,i,k}\right)$ for $k=1,2,\cdots,g,\ i=1,2,\ldots,m$ and $l=1,2,\cdots,p$ are assumed to be independent.

Based on these prior assumptions regarding the coefficients and the additional inverse-Wishart prior for the covariance matrix of the NAR/VAR model, coefficient inference can be performed by the vector autoregression Gibbs sampler (VAGSA, Chu et al., 2019). Basically, in the VAGSA, we need to iteratively generate the posterior samples of the two indicators and coefficients for the further inference. Similar to other MCMC algorithms, the VAGSA is computationally expensive, especially when the number of nodes increases. Instead of generating the posterior samples as an approximation of the posterior distribution, the variational Bayesian approach is adopted in our study to directly obtain the approximation of the posterior density function. Here, an EM-type algorithm is used to solve the corresponding optimization problem, and this is expected to significantly improve the computational efficiency of our approach.

3.2 The Variational inference procedure

Before introducing the variational Bayesian method, we first reparametrize the NAR/VAR model. Recall that in the VAGSA, the coefficient prior is a mixture distribution with a normal distribution and a delta function at point zero. When an indicator is equal to zero, we can simply set the corresponding coefficient to zero. To simplify the model structure, we reparametrize the coefficient as the product of the coefficient and the indicator. That is, $B_{\ell,i,i} = \gamma_{\ell,i} \widetilde{B}_{\ell,i,i}$ and $B_{\ell,i,\widetilde{s}_k} = \eta_{\ell,i,k} \widetilde{B}_{\ell,i,\widetilde{s}_k}$. Here, the priors of the indicators, $\gamma_{l,i}$ and $\eta_{l,i,k}$, are still independent Bernoulli distributions with probabilities π_1 and π_2, respectively. In addition, the prior of $\widetilde{B}_{\ell,i,i}$ is chosen as a normal distribution with mean zero and variance σ_β^2, and the prior of $\widetilde{B}_{\ell,i,\widetilde{s}_k}$ comes from the multivariate normal distribution with a mean zero vector and a covariance matrix $\sigma_\beta^2 I_{|\widetilde{s}_k|}$. Among them, the priors of $\widetilde{B}_{\ell,i,i}$ and $\widetilde{B}_{\ell,i,\widetilde{s}_k}$ are independent of $\gamma_{\ell,i}$ and $\eta_{\ell,i,k}$. Thus, $\gamma_{\ell,i} \widetilde{B}_{\ell,i,i}$ and $\eta_{\ell,i,k} \widetilde{B}_{\ell,i,\widetilde{s}_k}$ have the same effects as those shown in Eqs. (2) and (3).

Instead of generating the posterior samples directly, the variational Bayesian approach identifies the best approximate distribution of the true posterior for the further Bayesian inference. According to the ordinary variational Bayesian approach (Bishop, 2006), the Kullback-Leibler divergence (KL divergence) is used to measure the dissimilarity of the true posterior distribution from the approximated posterior distribution. Let $\theta = \{\pi_1, \pi_2, \Sigma, \sigma_B^2\}$ be the set of parameters and $\{\eta, \gamma, \widetilde{B}\}$ be the set of the indicator variables and coefficient matrices, where η is the vector of $\eta_{\ell,i,k}$, γ is the vector of $\gamma_{\ell,i}$, and $\widetilde{B} = \left(\widetilde{B}'_1, \widetilde{B}'_2, \ldots, \widetilde{B}'_p\right)'$. Given the prior assumptions, the posterior density function of $\eta, \gamma, \widetilde{B}$ is proportional to the the following joint density function:

$$
\begin{aligned}
P(\boldsymbol{Y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}} | \boldsymbol{X}, \theta) =& P(\boldsymbol{Y} | \boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}}, \boldsymbol{X}, \theta) P(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}} | \boldsymbol{X}, \theta) \\
=& MN_{T \times m}(\boldsymbol{XB}, \boldsymbol{I}, \Sigma) \prod_{\ell}^{p} \prod_{i}^{m} N(0, \sigma_B^2) \pi_1^{\gamma_{\ell,i}} (1 - \pi_1)^{(1 - \gamma_{\ell,i})} \\
& \prod_{k}^{g} MN_{1 \times |\widetilde{s}_k|}(\boldsymbol{0}, \boldsymbol{I}, \sigma_B^2 I_{|\widetilde{s}_k|}) \pi_2^{\eta_{\ell,i,k}} (1 - \pi_2)^{(1 - \eta_{\ell,i,k})}, \quad (4)
\end{aligned}
$$

Define $q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}})$ as an approximate posterior density function of $P(\widetilde{\boldsymbol{B}}, \boldsymbol{\eta}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{X})$. Since

$$
\begin{aligned}
KL(q||P) &= \int \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} q(\widetilde{\boldsymbol{B}}, \boldsymbol{\gamma}, \boldsymbol{\eta}) \log \left(\frac{q(\widetilde{\boldsymbol{B}}, \boldsymbol{\gamma}, \boldsymbol{\eta})}{P(\widetilde{\boldsymbol{B}}, \boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{Y}, \boldsymbol{X}; \theta)} \right) d\widetilde{\boldsymbol{B}} \\
&= \underbrace{\int \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} q(\widetilde{\boldsymbol{B}}, \boldsymbol{\gamma}, \boldsymbol{\eta}) \log \left(\frac{q(\widetilde{\boldsymbol{B}}, \boldsymbol{\gamma}, \boldsymbol{\eta})}{P(\widetilde{\boldsymbol{B}}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{Y}|\boldsymbol{X}; \theta)} \right) d\widetilde{\boldsymbol{B}}}_{-L(q)} \\
&+ \underbrace{\int \sum_{\boldsymbol{\gamma}} \sum_{\boldsymbol{\eta}} q(\widetilde{\boldsymbol{B}}, \boldsymbol{\gamma}, \boldsymbol{\eta}) \log P(\boldsymbol{Y}|\boldsymbol{X}; \theta) d\widetilde{\boldsymbol{B}}}_{\log P(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\theta})},
\end{aligned}
\tag{5}
$$

we have

$$
\log P(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\theta}) = L(q) + KL(q||P).
\tag{6}
$$

According to Eq. (5), the marginal likelihood $\log P(\boldsymbol{Y}|\boldsymbol{X}; \theta)$ is independent of q and can be treated as a fixed constant. Thus, $L(q)$ can be defined as the lower bound of the KL divergence between q and P. Minimizing the KL divergence with respect to q, is equivalent to maximizing the lower bound $L(q)$ of q. Here, one key is to specify the approximation density q via a factorization structure. Following Titsias & Lázaro-Gredilla (2011) and Cai et al. (2020), due to the independence assumption among the disjoint groups and their own lags, the following hierarchically factorized distribution is chosen as an approximate density function: $q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}})$, i.e.,

$$
q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}}) = \prod_{\ell}^{p} \prod_{i}^{m} \prod_{k}^{g} q_{\ell,i}(\widetilde{B}_{\ell,i,i}, \gamma_{\ell,i}) q_{\ell,i,k}(\widetilde{B}_{\ell,i,\widetilde{s}_k}, \eta_{\ell,i,k}),
$$

where $q_{\ell,i}(\widetilde{B}_{\ell,i,i}, \gamma_{\ell,i})$ and $q_{\ell,i,k}(\widetilde{B}_{\ell,i,\widetilde{s}_k}, \eta_{\ell,i,k})$ are chosen from the corresponding prior distributions. Then, a variational extension of the EM algorithm is adopted in the estimation process by maximizing the corresponding $L(q)$ with respect to the parameters. In the E-step, one would take the expectation of $L(q)$ with respect to $\boldsymbol{\eta}, \boldsymbol{\gamma}$ and $\widetilde{\boldsymbol{B}}$, and then in the M-step, one optimizes $L(q)$ with respect to θ. Iterate the two steps until the lower bound $L(q)$ converges.

In the E-step, $q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{B})$ is updated as follows:

$$
\begin{aligned}
q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{B}) \;=\; & \prod_{\ell}^{p} \prod_{i}^{m} \prod_{k}^{g} q_{\ell,i}(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i}) q_{\ell,i}(\gamma_{\ell,i}) q_{\ell,i,k}(\widetilde{B}_{\ell,i,\widetilde{s}_k}|\eta_{\ell,i,k}) q_{\ell,i,k}(\eta_{\ell,i,k}). \\
\;=\; & \prod_{\ell}^{p} \prod_{i}^{m} \prod_{k}^{g} \left(\phi_{1,\ell,i} N\left(\mu_{1,\ell,i,i}, \Sigma_{B_{\ell,i,i}}\right)\right)^{\gamma_{\ell,i}} \left(\left(1-\phi_{1,\ell,i}\right) N(0, \sigma_B^2)\right)^{(1-\gamma_{\ell,i})} \\
& \left(\phi_{2,\ell,i,k} MN_{1\times|\widetilde{s}_k|}(\boldsymbol{\mu}_{2,\ell,i,\widetilde{s}_k}, \boldsymbol{I}, \Sigma_{B_{\ell,i,\widetilde{s}_k}})\right)^{\eta_{\ell,i,k}} \left(\left(1-\phi_{2,\ell,i,k}\right) MN_{1\times|\widetilde{s}_k|}(\boldsymbol{0}, \boldsymbol{I}, \sigma_B^2 \boldsymbol{I}_{|\widetilde{s}_k|})\right)^{(1-\eta_{\ell,i,k})},
\end{aligned}
$$

where

$$
\begin{aligned}
\Sigma_{B_{\ell,i,i}} \;=\; & \left(\boldsymbol{X}_{\ell}^{(i)'} \boldsymbol{X}_{\ell}^{(i)}(\Sigma^{-1})_{i,i} + \sigma_B^2\right)^{-1}, \\
\mu_{1,\ell,i,i} \;=\; & \Sigma_{B_{\ell,i}} \left((\Sigma^{-1})_{i,i} \boldsymbol{X}_{\ell}^{(i)'} \left(\boldsymbol{Y}^{(i)} - \sum_{j\neq l}^{p} \boldsymbol{X}_j E(B_j^{(i)}) - \boldsymbol{X}_{\ell}^{(-i)} E(B_{\ell}^{(-i,i)})\right)\right. \\
& \left. + E\left(tr\left((\Sigma^{-1})_{-i,i} \boldsymbol{X}_{\ell}^{(i)'} \left(\boldsymbol{Y}^{(-i)} - \boldsymbol{X} \boldsymbol{B}^{(-i)}\right)\right)\right)\right), \\
\phi_{1,\ell,i} \;=\; & Inv-logit\left\{logit(\pi_1) - \frac{1}{2}\log(\sigma_B^2) + \frac{1}{2}\log(det(\Sigma_{B_{\ell,i,i}})) + \frac{(\Sigma_{B_{\ell,i,i}})^{-1}\mu_{1,\ell,i,i}^2}{2}\right\},
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_{B_{\ell,i,\widetilde{s}_k}} \;=\; & \left(\boldsymbol{X}_{\ell}^{(i)'} \boldsymbol{X}_{\ell}^{(i)}(\Sigma^{-1})_{\widetilde{s}_k,\widetilde{s}_k} + \sigma_B^2 \boldsymbol{I}_{|\widetilde{s}_k|}\right)^{-1}, \\
\boldsymbol{\mu}_{2,\ell,i,\widetilde{s}_k} \;=\; & \left(\boldsymbol{X}_{\ell}^{(i)'} \left(\boldsymbol{Y}^{(\widetilde{s}_k)} - \sum_{j\neq l}^{p} \boldsymbol{X}_j E(B_j^{(\widetilde{s}_k)}) - \boldsymbol{X}_{\ell}^{(-i)} E(B_{\ell}^{(-i,\widetilde{s}_k)})\right)(\Sigma^{-1})_{\widetilde{s}_k,\widetilde{s}_k}\right. \\
& \left. + E\left(tr\left(\boldsymbol{X}_{\ell}^{(i)'} \left(\boldsymbol{Y}^{(-\widetilde{s}_k)} - \boldsymbol{X} \boldsymbol{B}^{(-\widetilde{s}_k)}\right)(\Sigma^{-1})_{-\widetilde{s}_k,\widetilde{s}_k}\right)\right)\right) \Sigma_{B_{\ell,i,\widetilde{s}_k}}, \\
\phi_{2,\ell,i,k} \;=\; & Inv-logit\left\{logit(\pi_2) - \frac{1}{2}\log(det(\sigma_\beta^2 \boldsymbol{I}_{\widetilde{s}_k})) + \frac{1}{2}\log(det(\Sigma_{B_{\ell,i,\widetilde{s}_k}}))\right. \\
& \left. + \frac{1}{2}tr\left((\Sigma_{B_{\ell,i,\widetilde{s}_k}})^{-1}\boldsymbol{\mu}_{2,\ell,i,\widetilde{s}_k}'\boldsymbol{\mu}_{2,\ell,i,\widetilde{s}_k}\right)\right\}.
\end{aligned}
$$

Here, $\boldsymbol{X}_{\ell}^{(i)}$ and $\boldsymbol{X}_{\ell}^{(-i)}$ denote the ith column of $\boldsymbol{X}_{\ell}$ and matrix $\boldsymbol{X}_{\ell}$ excluding the ith column, respectively. The same definition structure is used for $\boldsymbol{Y}$ and $\boldsymbol{B}$. In addition, $B_{\ell}^{(-i,i)}$ and $(\Sigma)_{-i,i}^{-1}$ denote the ith column of B_{ℓ} and Σ^{-1} without ith element. $Inv-logit(\cdot)$ is an inverse logistic function.

In the M-step, we take the derivative of $L(q)$ with respect to θ and then set it as a zero

vector. Thus, the components in θ can be updated as the solutions of the normal equations, i.e.,

$$
\begin{aligned}
\pi_1 &= \frac{\sum_{\ell=1}^{p} \sum_{i=1}^{m} \phi_{1,\ell,i}}{mp}, \\
\pi_2 &= \frac{\sum_{\ell=1}^{p} \sum_{i=1}^{m} \sum_{k=1}^{g} \phi_{2,\ell,i,k}}{\sum_{k=1}^{g} |\widetilde{s}_k| mp}, \\
\Sigma &= \frac{E\left((\boldsymbol{Y} - \boldsymbol{XB})'(\boldsymbol{Y} - \boldsymbol{XB})\right)}{T}, \\
\sigma_B^2 &= \frac{\sum_{l}^{p} \sum_{i}^{m} \phi_{1,l,i}(\Sigma_{B,l,i,i} + \mu_{1,l,i}^2) + \sum_{l}^{p} \sum_{i}^{m} \sum_{k}^{g} \phi_{2,l,i,k} tr(\Sigma_{B,l,i,\widetilde{s}_k} + \boldsymbol{\mu}'_{2,l,i,\widetilde{s}_k} \boldsymbol{\mu}_{2,l,i,\widetilde{s}_k})}{\sum_{l}^{p} \sum_{i}^{m} \left(\phi_{1,l,i} + \sum_{k}^{g} |\widetilde{s}_k| \phi_{2,l,i,k}\right)}.
\end{aligned}
$$

More details regarding both steps are shown in the supplementary materials.

Since we iterate the E- and M-steps sequentially, a natural choice of a stopping criterion is that the difference between the values of $L(q)$ in two consecutive iterations is less than a certain threshold. Then one can approximate the posterior inclusion probabilities as follows:

$$
\begin{aligned}
P(\eta_{\ell,i,k} = 1 | \boldsymbol{Y}, \boldsymbol{X}, \hat{\theta}) &\approx q(\eta_{\ell,i,k} = 1 | \hat{\theta}) = \phi_{2,\ell,i,k}, \\
P(\gamma_{\ell,i,i} = 1 | \boldsymbol{Y}, \boldsymbol{X}, \hat{\theta}) &\approx q(\gamma_{\ell,i,i} = 1 | \hat{\theta}) = \phi_{1,\ell,i}.
\end{aligned}
$$

Thus, based on the median probability criterion (Barbieri & Berger, 2004), $B_{\ell,i,i}$ or $B_{\ell,i,\widetilde{s}_k}$ is identified as non zero if $\phi_{1,l,i}$ or $\phi_{2,l,i,k}$ is larger than or equal to $1/2$. Finally, the one-step head prediction $\hat{\boldsymbol{Y}}_{T+1}$ can be obtained by $\hat{\boldsymbol{Y}}_{T+1} = \boldsymbol{Y}_T \hat{B}_1 + \boldsymbol{Y}_{T-1} \hat{B}_2 + \ldots, \boldsymbol{Y}_{T+1-p} \hat{B}_p$, where the $\hat{B}_\ell$ are estimated based on the identified active structures.

The EM-type method is a locally optimal approach and may be sensitive to the initial status. Realistic initial values of θ may provide poor estimation results due to improper choices of the initial prior probabilities. In our study, the initial values of the prior probabilities, π_1 and π_2, are set to be small, e.g., $\pi_1 = \pi_2 = 0.01$. The initial values of the coefficient matrix, $\boldsymbol{B}$, are obtained via the least-squares estimation method, and the initial value of Σ is set as the sample variance of $\boldsymbol{Y}$ divided by 2. The threshold value of the stopping criterion is usually set to a prespecified value, which is less than or equal to 10^{-6}.

4 Simulation

This section investigates the finite-sample performance of the proposed VB approach compared with that of a known data generation process. We first revisit the simulation studies of VAGSA in Chu et al. (2019). Given dynamic networks, we care not only about the structure selection ability but also about the computational efficiency of our approach. For medium-sized examples, i.e., $m = 10$ and 20, we directly compare the performances of the VB method and the VAGSA. When the dimensionality increases to $m = 50$, the VAGSA requires an extremely long computational time, and thus, we illustrate only the results of the proposed VB method.

4.1 Medium-sized network examples

We follow the simulation setups in Chu et al. (2019) and generate medium-sized network series with $m = 10$ and 20 for a fair comparison. We assume that there are multiple lags p in the VAR framework. Three dependence structures are considered in the simulations.

m10UG and m20UG: The true models follow the universal grouping structure with $(m, p) = (10, 5)$ and $(20, 5)$, respectively. There are 72 and 145 nonzero coefficients in two cases.

m10SG and m20SG: The segmentation structure is considered in these two simulations with $m = 10$ and 20. Let $S_{m,g}$ denote a specific group structure with m time series for g disjoint groups. We set $S_{10,3} = \{(1, 2, 3), (4, 5, 6), (7, 8, 9, 10)\}$ for $(g, m, p) = (3, 10, 5)$ and $S_{20,4} = \{(1, 2, 3, 4, 5), (6, 7, 8, \ 9, 10), (11, 12, 13), (14, \ldots, 20)\}$ for the case where $(g, m, p) = (4, 20, 5)$. There are 40 and 109 nonzero coefficients, respectively.

m10NG and m20NG: In the "no grouping" cases, there are 18 and 27 nonzero coefficients for $(m, p) = (10, 5)$ and $(20, 5)$, respectively.

In addition, the error terms are generated from a multinormal distribution with zero mean. We consider two different covariance matrices when generating the stochastic noises, namely, an identity matrix $\boldsymbol{I}_m$ indicating that there are no concurrent correlations among the nodes and a symmetric matrix Σ defined as follows:

Σ_{10}: The diagonal of Σ_{10} is $(0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.8, 0.8, 0.8, 0.8)$, and the off-diagonal correlation coefficients of Σ_{10} are $0.4^{|i'-i|}$ for $i' \neq i$.

Σ_{20}: The diagonal of Σ_{20} is $(0.9, 0.9, 0.9, 0.8, 0.8, 0.8, 0.9, \cdots, 0.9)$, and the off-diagonal correlation coefficients of Σ_{20} are $0.4^{|i'-i|}$ for $i' \neq i$.

For each simulation, we generate data with $T = 301$, where the first 300 samples are used for model training and the last point Y_{301} is used to demonstrate the prediction ability of the proposed method. Each simulation is replicated $N = 100$ times, and the segmentation structure is assumed to be known. We simply fix the number of lags p to 10 and set the threshold value of the stopping criteria to 10^{-8} in all simulations.

As mentioned before, we also implement the VAGSA for the simulation cases for comparison purposes. According to Chu et al. (2019), first, we add the inverse Wishart prior for Σ with m degrees of freedom and a scale matrix I_m. For the other parameters, the prior probabilities, π_1 and π_2, are fixed at 0.5, and for the variance in the mixture prior distribution, σ_B is 0.5 for the cases of m10UG, m20UG, m10SG and m20SG and 15 for the other two cases, m10NG and m20NG. When we implement the VAGSA, there are 3000 sweeps in total, and we take the last 1000 samples for inference. For the details of implementing the VAGSA, please refer to Chu et al. (2019).

4.2 Large-scale network examples

To illustrate the feasibility of the proposed VB approach for high-dimensional scenarios, we conduct simulations with $m = 50$ nodes. The setups of the large-scale simulations are similar to those used in Section 4.1. Here, we also consider the three different grouping structures, and the details of three cases are shown as follows.

m50UG: There are 355 nonzero coefficients in B_1, B_3 and B_5 for $(m, p) = (50, 5)$.

m50SG: We set $S_{50,8} = \{(1, 2, 3, 4, 5), (6, 7, 8, 9, 10), (11, 12, 13), (14, \ldots, 20), (21, \ldots, 30), (31, \ldots, 35), (36, \ldots, 40), (41, \ldots, 50)\}$. There are 360 nonzero coefficients in B_1, B_3 and B_5 for $(m, p) = (50, 5)$.

m50NG: There are 128 nonzero coefficients in B_1, B_3 and B_6 for $(m, p) = (50, 5)$.

Figure 5 visualizes the true sparsity of the network dependence with the nonzero coefficients marked in bold. Note that only the active coefficient matrices are displayed, while the others, such as lag-2 and lag-4, are omitted as zero everywhere.

Similarly, two different noise covariance matrices are used for data generation, i.e., a 50×50 identity matrix $\boldsymbol{I}_{50}$ and a symmetric matrix $\boldsymbol{\Sigma}_{50}$, which is defined as follows:

$$\boldsymbol{\Sigma}_{50}: \text{The diagonal elements are} \left(0.9, 0.9, 0.9, 0.8, 0.8, 0.8, \underbrace{0.9, \cdots, 0.9}_{14}, \underbrace{0.8, \cdots, 0.8}_{20}, \underbrace{0.9, \cdots, 0.9}_{10} \right),$$

and the correlation coefficients of $\boldsymbol{\Sigma}_{50}$ are set as $0.4^{|i'-i|}$ for $i' \neq i$.

In this simulation example, we generate data with $T = 701$. The first 700 samples are used for model training, and the last sample Y_{701} is used to illustrate the prediction ability of the model.

4.3 Simulation results

Based on the simulation setups in Sections 4.1 and 4.2, we independently repeat each case 100 times. When we implement the proposed VB approach, the initial setups are basically the same as those use in Section 3, except for the case of m50NG with a covariance matrix of $\boldsymbol{\Sigma}_{50}$. For this case, we set $\pi_1 = \pi_2 = 0.5$, instead of 0.01. To illustrate the performances of the proposed method, we consider four measurements. The true positive rate (TPR) and the false positive rate (FPR) are designed to show the accuracy of structure identification. The average model size (AMS) measures how many active elements are identified. Compared to the TPR and the FPR, the AMS gives an overall indicator of identification accuracy. The last measurement is the average of the mean square prediction errors (MSPEs), and this is used to report the prediction ability of the model. The definitions of these four measurements

(a) m50UG   (b) m50SG   (c) m50NG

Figure 5: The true active coefficient matrices for all UG, SG and NG cases.

are shown as follows:

$$\text{TPR} = \frac{\text{\# correctly identified active variables}}{\text{\# true active variables}},$$

$$\text{FPR} = \frac{\text{\# incorrectly identified active variables}}{\text{\# true inactive variables}},$$

$$AMS = \frac{1}{N} \sum_{j}^{N} \text{\# identified active variables in the } j\text{th replicate},$$

where N is the number of replicates and

$$\text{MSPE} = \frac{1}{mN} \sum_{i=1}^{m} \sum_{j=1}^{N} \left(Y_{T+1,i,j} - \hat{Y}_{T+1,i,j} \right)^2.$$

Last, the average CPU time for a replication is also reported in seconds. Here, we run the R code for the proposed VB method, and we run the VAGSA based on its MATLAB code. All codes are implemented on a PC with an Intel(R) Core(TM) i7-4770 CPU @3.400 GHz and 16.0 GB of RAM.

The results of all cases are summarized in Table 1. For the medium-sized networks with $m = 10$ and $m = 20$, both the VB and VAGSA results are reported. For large-scale networks with $m = 50$, only the VB results are reported and the measurements for the VAGSA are denoted by "-". Consider the performances of the medium-sized networks. According to Table 1, both Bayesian methods achieve similar structural identifications and inference accuracies for medium-sized networks. In particular, the TPRs are perfect with values close to 100%, implying that all active elements are successfully identified. In terms of AMS, both methods share similar AMS values and are all close to the true model sizes. In addition, the VB method has better performance in terms of the FPR than the alternative VAGSA. For the UG and SG cases, the FDR values of the VB method are less than a quarter of those of the VAGSA. The VAGSA performs slightly better for the four NG cases; however, the differences between the two methods are minor. To check the selection results in detail, we find that the VB approach might not identify a variable with a small coefficient for the NG cases. Finally, consider the prediction ability. The VB method has slightly better accuracies than those of the VAGSA in most cases, as reflected in the MSPE. Most importantly, there is a dramatic

improvement in terms of the average CPU time when using the VB method. In general, it only needs approximately 1/7 of the computational cost required by the VAGSA. For some cases with $m = 10$ and I_{10}, the VB approach even saves 34/35 CPU time, without sacrificing accuracy.

The good performance of the VB method continues for the large-scale networks because it is computationally possible. Figures 6 and 7 show the average estimates of the coefficient matrices for two different covariance matrices. These coefficient matrices are all close to the true matrices. Use the UG case as an example. The structure identification results are perfect because the TPR = 100%, the FPR = 0% and the AMS is still close to the true value. In the NG case with a covariance matrix of, Σ_{50}, the corresponding TPR is 92%, which is slight lower than the TPRs in the other cases. This may be because the VB method is used to approximate the true posterior distribution, and thus, active variables with small coefficients might not be detected. Overall, the proposed VB method has good performance in identifying the active structures for large-scale network cases. In addition, the average CPU times show that the case with a higher level structure, namely, UG, requires the least computational time, followed by the SG and NG cases.

Table 1: Numerical Comparison between the VB Method and the VAGSA

VB/VAGSA	TPR(%)	FPR(%)	AMS	MSPE	Ave. CPU Time
m10UG-I_{10}	100/100	0.07/0.31	72.62/74.90	1.00/1.01	4/148
m10UG-Σ_{10}	100/100	0.06/0.28	72.51/74.59	1.00/0.83	5/138
m10SG-I_{10}	100/100	0.15/0.63	41.35/46.02	1.07/1.09	12/303
m10SG-Σ_{10}	100/100	0.13/0.26	41.17/46.02	0.86/1.09	11/303
m10NG-I_{10}	98/97	0.15/0.11	19.07/18.58	1.00/1.00	40/628
m10NG-Σ_{10}	99/99	0.11/0.08	18.86/18.55	0.89/0.88	36/653
m20UG-I_{20}	100/100	0.03/0.18	145.79/151.89	0.98/0.99	15/236
m20UG-Σ_{20}	100/100	0.02/0.14	145.56/150.54	0.95/0.95	180/237
m20SG-I_{20}	100/100	0.06/0.26	109.25/117.05	1.10/1.11	44/652
m20SG-Σ_{20}	100/100	0.06/0.21	109.22/115.20	0.90/0.90	51/672
m20NG-I_{20}	97/98	0.08/0.15	29.29/32.39	0.99/0.99	291/2183
m20NG-Σ_{20}	97/98	0.06/0.13	28.55/31.43	0.86/0.86	262/2247
m50UG-I_{50}	100/-	0.00/-	355.47/-	1.02/-	213/-
m50UG-Σ_{50}	100/-	0.00/-	355.57/-	0.87/-	239/-
m50SG-I_{50}	100/-	0.01/-	361.58/-	1.02/-	1082/-
m50SG-Σ_{50}	100/-	0.01/-	361.15/-	0.88/-	1767/-
m50NG-I_{50}	100/-	0.03/-	134.67/-	1.02/-	7860/-
m50NG-Σ_{50}	92/-	0.03/-	123.33/-	0.89/-	58194/-

5   Empirical Study

We apply the VB approach to estimate the temporal dependences of the German natural gas flow networks and perform day-ahead forecasting with the NAR/VAR framework. The data cover two years, from 1st October 2013 to 30th September 2015. The gas flows contain both inflows (supply) and outflows (demand) recorded 7 days a week at 51 distribution nodes belonging to 4 categories with different functions. There are 34 municipal nodes, that provide gas to local residential areas and small business districts. The 11 industry nodes are responsible for factory production. Moreover there is 1 border node, which is important in Germany, because these nodes serve as network transfer points for natural gas imported
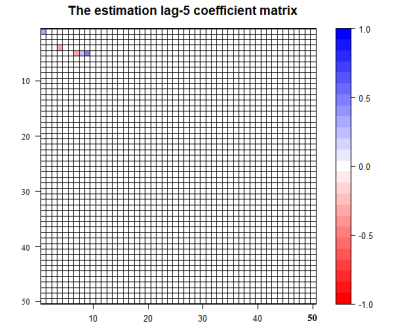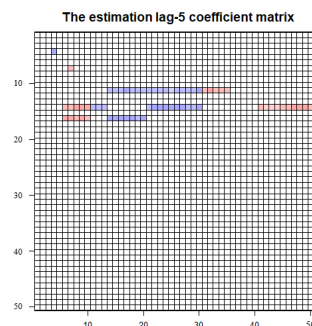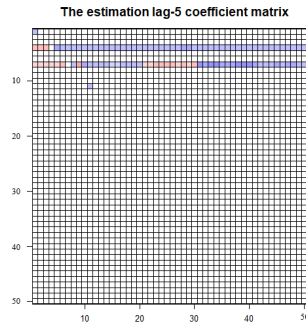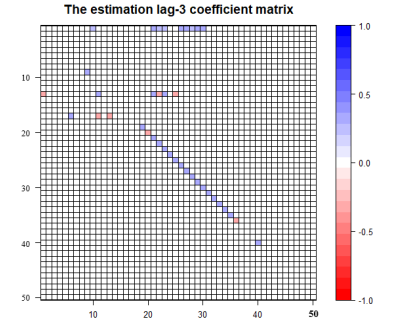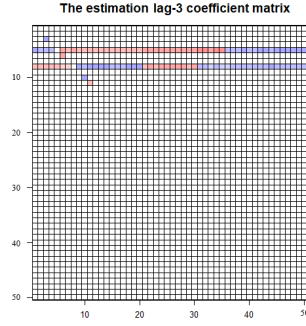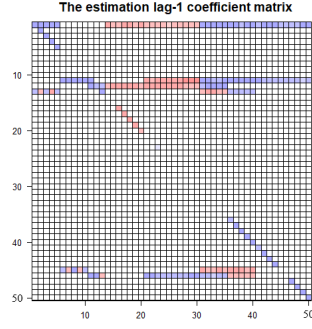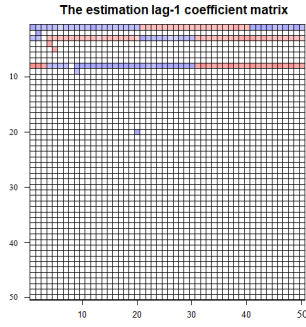
(a) m50UG　　　　(b) m50SG　　　　(c) m50NG

Figure 6: Estimated coefficient matrices with $\boldsymbol{\Sigma} = \boldsymbol{I}_{50}$ for all UG, SG and NG cases.
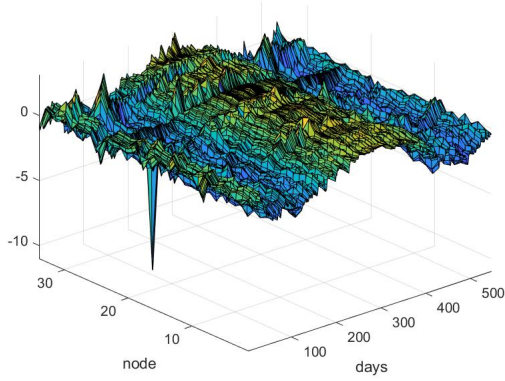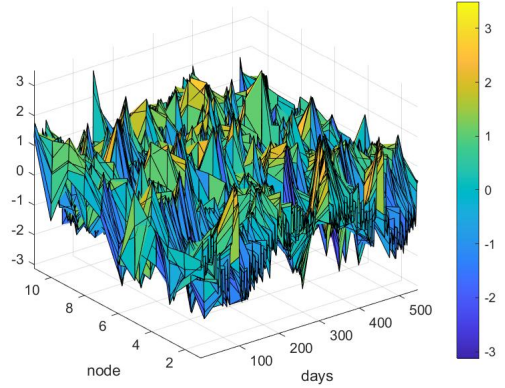
Figure 7: Estimated coefficient matrices with $\Sigma = \Sigma_{50}$ for all UG, SG and NG cases.
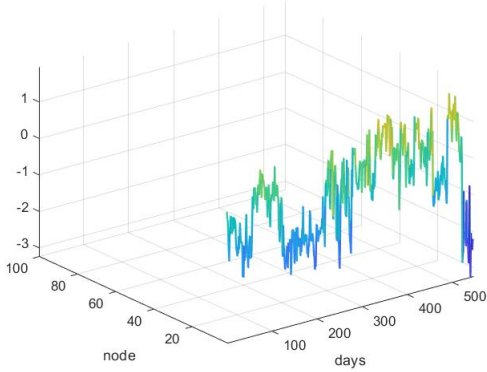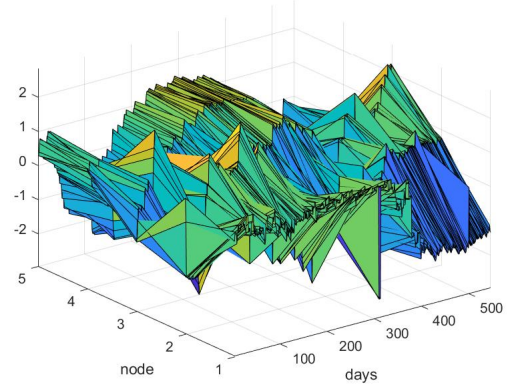
Municipal (34 nodes)

Industrial (11 nodes)

Border (1 node)

Others (5 nodes)

Figure 8: Normalize daily gas flows values in different types.

and exported via Germany. The rest 5 nodes are categorized to "others" that serve as switch nodes or perform other functions. Figure 8 displays the time series of the 51 nodes that show different dynamic patterns. For further analysis, we normalize the values of the flows. In addition, note that the data in municipal nodes have been seasonally adjusted via the daily temperature, and 3 nodes, O3, O4, and O5, in the "others" category have also been seasonally adjusted.

We adopt the NAR/VAR model with the segmentation structure to analyze the gas network data. Here, the segmented grouping structures are defined based on the types of nodes that are involved. That is, each type of node is treated as a group. In addition, we set the number of lags p as 14 to incorporate social dependence up to two weeks ahead.

We use the data from Oct 1st, 2013, to March 31, 2015, as the in-sample training data. The threshold value of the stopping criterion for the proposed VB method is set to be 10^{-6}, and the other initial settings are the same as those used we set in the simulation studies in Section 4. After learning the model via the VB method, we perform a one-step ahead forecast. That is, at each daily point, we shift one more day, use the expanded sample to retrain the model coefficients based on the active segmentation structures, and perform a one-day ahead forecast. Note that we repeat the iterative forecast for the remaining 0.5 years.

Overall, we obtain 183 forecasts. 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, and both measures are defined as follows:

$$MAPE = \frac{1}{183 \times 51} \sum_{t=547}^{729} \sum_{i=1}^{51} \frac{|Y_{t+1,i} - \hat{Y}_{t+1,i}|}{|Y_{t+1,i}|}, \tag{7}$$

$$NRMSE = \frac{1}{183 \times 51} \left(\sum_{t=547}^{729} \sum_{i=1}^{51} (Y_{t+1,i} - \hat{Y}_{t+1,i})^2 \right) / \frac{1}{183 \times 51} \sum_{t=547}^{729} \sum_{i=1}^{51} Y_{t+1,i}. \tag{8}$$

Figures 9 displays the estimated coefficient matrices for lag-1 to lag-3, and these three coefficient matrices are quite sparse. For lag-1, it can be observed that the 4th and 14th rows, corresponding to municipal nodes M4 and M14, affect the other municipal nodes. In the 35th row, an industry node, I1, influences the other industry nodes but not the other three types. The 51st row is the status of node O5; the others node only have relations with municipal nodes. On the other hand, the only border node is only related to itself and has no interactions with the others. For lag-2, only the 35th row I1 influences the nodes in the other category. For lag-3, some nodes have minor influences by themselves. As the lag increases, there are fewer and fewer nodes that may influence others. Furthermore, there is no more serial dependence in the network after lag-10.

Table 2 reports the overall model fitting and forecasting performances via the obtained MAPE and NRMSE values. The details of individual nodes are shown in the Appendix. Basically, the VB method delivers good performances for both the in-sample training period from 1st October 2014 to 31st March 2015 and the out-of-sample forecasting period from
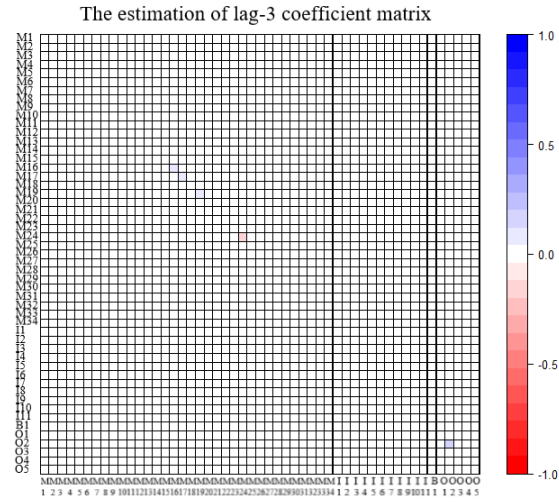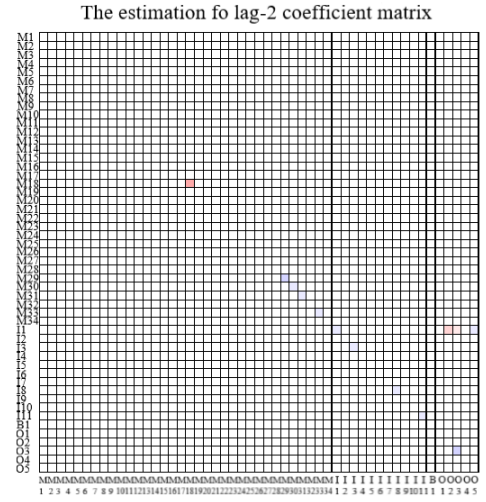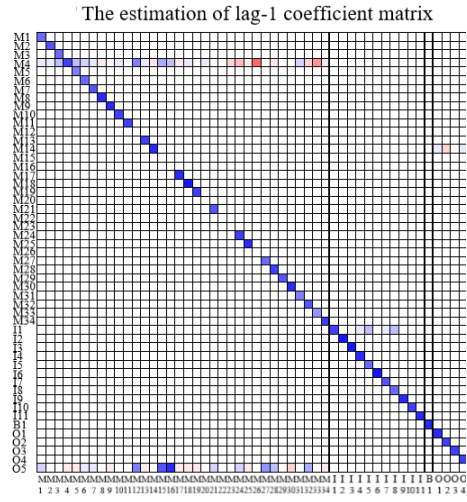
26

The estimation of lag-1 coefficient matrix

The estimation fo lag-2 coefficient matrix

The estimation of lag-3 coefficient matrix

Figure 9: The coefficient matrix for lag-1 to lag-3.

27

Table 2: The MAPEs and NRMSEs for the fitted and forecast results classified by type.

Type	Number	MAPE (%)	Range (%)	SD	NRMSE	Range	SD
In sample from 01/10/2014 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

1st April 2015 to 30th September 2015. According to Table 2, regardless of whether the in-sample training and out-of-sample forecasting results are examined, the average values of the MAPE are less than 15%, and the corresponding NRMSE values are all less than 0.15. Thus, we can claim that the NAR/VAR model with segmentation structures fits the in-sample training data well and provides good prediction capability for out-of-sample forecasting. According to Table A1 in the Appendix, most nodes have small mean values and standard deviations with repect to MAPE and NRMSE. For nodes I5, I8 and I10, both the MAPE and NRMSE values are large. A possible reason for this phenomenon may be the violation of the stationarity assumption in the NAR/VAR model because the corresponding trend of the daily data changes frequently. In addition, in the "others" type, due to some extreme values for a few special days, node O1 may not have good model fitting results. When we consider the out-of-sample forecasting performance, the corresponding MAPE and NRMSE values share similar patterns to those witnessed in the in-sample training results. Nodes I5, I8, and I10 still have large values for both measurements. However, node O1 does have good MAPE and NRMSE values because there are no extreme values shown in the prediction period. Overall, the out-of-sample forecasting performance is acceptable because the average MAPE and NRMSE values are 9.78% and 0.11, respectively.

6   Conclusion

In this paper, we focus on Bayesian analysis with NAR/VAR models, especially when there are many time series involved. First, to simplify the model complexity, the model structure

assumptions in Song & Bickel (2011) are adopted. Then, model inference can be performed via a Bayesian structure selection approach. Here, to denote active structures, indicators are added to the NAR/VAR model. Instead of generating the posterior samples of these indicators via an MCMC algorithm, we directly obtain the proper approximation of the posterior density function via a variational Bayesian approach. Based on a factorized approximation assumption for the latent structures, a variational EM-type algorithm is used to obtain the best approximation; then, according to the approximated posterior probabilities of the indicators, the median probability criterion is used to determine the active structures in the corresponding coefficient matrices. The simulation results support the notion that the proposed variational Bayesian 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.

There are several possible future research directions. The first concerns the parameter turning strategy for the proposed VB method. In our experience, the performance may be sensitive to the chosen initial parameters, especially for the prior probabilities π_1 and π_2. According to Ormerod et al. (2017) and Zhang et al. (2019), we suggest setting π_i as small probabilities, as this would tend to select a compact model. However, sometimes the model may not identify active variables with small coefficients. For example, in the simulated NG case with m = 50 and a covariance matrix of Σ_{50}, we set $\pi_1 = \pi_2 = 0.5$, instead of 0.01. Thus, choosing the proper initial setups should be an considered issue for the proposed VB method. The cross-validation approach should be examined as a possible strategy, and other possibilities should be related to proper information criteria. The second research direction is to take the group sparsity assumption into account. That is, the variables within an active segmentation still satisfy the elementwise sparsity assumption. Following Chen et al. (2016), we need to add new indicators for the variables within each segmentation. The idea of the variational Bayesian approach in Cai et al. (2020) could be modified for analyses using NAR models. In our real example, the segmentation structures are defined based on the types of nodes present. However, these may not be the best segmentation structures based on

these prespecified gas station types. Following Chu et al. (2019), we may apply a data-driven clustering approach to identify other segmentation structures for the corresponding NAR models. Thus, it would be an interesting research direction to integrate the clustering algorithm with the variational Bayesian approach. That is, we can identify the proper segmentation structures and determine the active structures simultaneously.

References

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*(1), 71–92.

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*, 870–897.

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

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Cai, M., Dai, M., Ming, J., Peng, H., Liu, J., & Yang, C. (2020). Bivas: a scalable bayesian method for bi-level variable selection with applications. *Journal of Computational and Graphical Statistics*, *29*(1), 40–52.

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

Chang, S.-M., Chen, R.-B., & Chi, Y. (2016). Bayesian variable selections for probit models with componentwise Gibbs samplers. *Communications in Statistics-Simulation and Computation*, *45*(8), 2752–2766.

Chen, R.-B., Chu, C.-H., Lai, T.-Y., & Wu, Y. N. (2011). Stochastic matching pursuit for Bayesian variable selection. *Statistics and Computing*, *21*, 247–259.

Chen, R.-B., Chu, C.-H., Yuan, S., & Wu, Y. N. (2016). Bayesian sparse group selection. *Journal of Computational and Graphical Statistics*, *25*, 665–683.

Chen, Y., Koch, T., Zakiyeva, N., & Zhu, B. (2020). Modeling and forecasting the dynamics of the natural gas transmission network in germany with the demand and supply balance constraint. *Applied Energy*, *278*, 115597.

Chu, C.-H., Lo Huang, M.-N., Huang, S.-F., & Chen, R.-B. (2019). Bayesian structure selection for vector autoregression model. *Journal of Forecasting*, *38*, 422–439.

Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, *3*(2), 521.

Farcomeni, A. (2010). Bayesian constrained variable selection. *Statistica Sinica*, *20*, 1043–1062.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics*, *5*, 609—620.

Guo, J., Levina, E., Michailidis, G., & Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, *98*(1), 1–15.

Hsu, N.-J., Hung, H.-L., & Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using LASSO. *Computational Statistics & Data Analysis*, *52*(7), 3645–3657.

Kastner, G., & Huber, F. (2020). Sparse bayesian vector autoregressions in huge dimensions. *Journal of Forecasting*.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, *40*(4), 2293–2326.

Lütkepohl, H. (2007). General-to-specific or specific-to-general modelling? An opinion on current econometric terminology. *Journal of Econometrics*, *136*(1), 319–324.

Melnyk, I., & Banerjee, A. (2016). Estimating structured vector autoregressive models. *International Conference on Machine Learning*, 830–839.

Nicholson, W. B., Matteson, D. S., & Bien, J. (2014). Structured regularization for large vector autoregressions. *Cornell University*.

Ormerod, J. T., You, C., Müller, S., et al. (2017). A variational Bayes approach to variable selection. *Electronic Journal of Statistics*, *11*(2), 3549–3594.

Song, S., & Bickel, P. J. (2011). Large vector auto regressions. *arXiv preprint arXiv:1106.3915*.

Titsias, M. K., & Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. *Advances in Neural Information Processing Systems*, *24*, 2339-2347.

Zhang, C.-X., Xu, S., & Zhang, J.-S. (2019). A novel variational Bayesian method for variable selection in logistic regression models. *Computational Statistics & Data Analysis*, *133*, 1–19.

A   Appendix

Table A1: The MAPEs and NRMSEs for the in-sample training period from 01/10/2014 to 31/03/2015 classified by type.

	Municipal:											
MAPE (%)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
mean	4.53	5.04	8.89	6.04	8.00	4.89	6.00	5.04	5.58	6.285	5.62	6.36
sd	3.59	3.58	7.30	4.66	6.58	3.98	4.81	4.04	4.80	5.30	4.39	5.07
NRMSE	0.06	0.06	0.11	0.08	0.10	0.06	0.08	0.07	0.07	0.08	0.07	0.08
MAPE (%)	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
mean	5.77	4.38	6.53	5.65	6.93	6.93	5.41	7.24	3.20	7.00	7.32	14.70
sd	4.59	3.46	4.89	4.23	5.53	5.70	4.74	5.31	2.44	5.31	5.37	23.15
NRMSE	0.07	0.06	0.09	0.07	0.09	0.09	0.07	0.09	0.04	0.09	0.10	0.36
MAPE (%)	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**		
mean	4.50	5.48	6.61	5.27	8.25	6.38	7.89	4.94	5.41	5.44		
sd	3.54	4.46	5.42	4.17	6.78	5.37	6.72	3.50	3.98	4.48		
NRMSE	0.06	0.07	0.09	0.07	0.10	0.8	0.10	0.06	0.07	0.07		
	Industry:											
MAPE (%)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
mean	11.04	8.30	2.80	10.94	24.35	1.68	4.71	27.73	1.96	34.22	0.89	
sd	10.20	9.05	2.34	11.57	33.19	1.76	4.51	39.21	1.63	41.65	0.10	
	0.12	0.10	0.04	0.14	0.22	0.02	0.06	0.22	0.03	0.31	0.01	
	Border:											
MAPE (%)	**1**											
mean	9.36											
sd	14.23											
NRMSE	0.11											
	Others:											
MAPE (%)	**1**	**2**	**3**	**4**	**5**							
mean	36.297	8.507	10.711	5.93	3.78							
sd	20.69	10.28	10.73	4.46	3.00							
NRMSE	0.34	0.07	0.13	0.08	0.05							

Table A2: The MAPEs and RMSEs for the out-of-sample forecasting period 01/01/2015 to 30/09/2015 classified by type.

Municipal:

MAPE (%)	1	2	3	4	5	6	6	8	9	10	11	12
mean	7.56	7.42	10.20	7.51	10.05	7.23	8.13	7.54	7.66	7.70	7.45	8.77
sd	5.66	5.69	7.63	5.74	7.23	5.60	6.34	6.10	5.85	6.15	5.91	6.78
NRMSE	0.10	0.10	0.13	0.10	0.13	0.10	0.11	0.10	0.10	0.10	0.10	0.11

MAPE (%)	13	14	15	16	17	18	19	20	21	22	23	24
mean	6.59	6.34	9.82	8.93	8.52	8.20	6.50	10.28	5.23	10.49	11.8	8.16
sd	4.91	5.11	7.75	7.37	6.45	6.71	8.35	8.75	4.25	8.88	9.30	6.31
NRMSE	0.09	0.08	0.12	0.12	0.11	0.10	0.09	0.13	0.07	0.13	0.15	0.12

MAPE (%)	25	261	27	28	29	30	31	32	33	34		
mean	4.65	7.91	8.79	7.67	8.85	5.81	8.06	6.60	5.92	5.90		
sd	3.79	5.69	6.44	5.90	7.54	4.31	6.74	4.95	4.61	4.31		
NRMSE	0.06	0.10	0.11	0.10	0.11	0.74	0.10	0.08	0.08	0.08		

Industry:

MAPE (%)	1	2	3	4	5	6	7	8	9	10	11
mean	9.32	8.52	2.51	9.93	32.62	2.98	6.30	54.57	1.70	45.35	1.94
sd	8.21	9.87	2.80	8.44	61.38	5.15	5.61	121.58	1.41	80.72	3.70
NRMSE	0.11	0.10	0.03	0.12	0.23	0.05	0.08	0.27	0.02	0.30	0.04

Border:

MAPE (%)	1
mean	11.77
sd	11.40
NRMSE	0.15

Others:

MAPE (%)	1	2	3	4	5
mean	5.76	9.73	13.47	9.03	5.53
sd	5.43	9.87	23.81	6.80	4.75
NRMSE	0.08	0.11	0.12	0.11	0.07

B Supplementary Document

B.1 Variational EM algorithm for th NAR model

In this extra supplementary document, we provide more details about the EM algorithm for the VB approach with respect to the NAR/VAR model.

B.1.1 E-Step

Let $\theta = \left\{\pi_1, \pi_2, \Sigma, \sigma_B^2\right\}$ be the collection of NAR model parameters and $\{\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{B}\}$ be the set of latent variables. The joint probability of $\boldsymbol{Y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}}$ and as follows:

$$
\begin{aligned}
P(\boldsymbol{Y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}}|\boldsymbol{X}, \theta) =& P(\boldsymbol{Y}|\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}}, \boldsymbol{X}, \theta)P(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}}|\boldsymbol{X}, \theta) \\
=& MN_{T\times m}(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}, \Sigma)\prod_l^p \prod_i^m N(0, \sigma_B^2)\pi_1{}^{\gamma_{\ell,i}}(1-\pi_1)^{(1-\gamma_{\ell,i})} \\
& \prod_k^g MN_{1\times|\widetilde{s}_k|}(\boldsymbol{0}, \boldsymbol{I}, \sigma_B^2 I_{|\widetilde{s}_k|})\pi_2{}^{\eta_{\ell,i,k}}(1-\pi_2)^{(1-\eta_{\ell,i,k})} .
\end{aligned}
$$

As mentioned before, in this study, we can maximize the lower bound $L(\boldsymbol{q})$ with respect to the approximate density function q. In the reparameterized spike-and-slab prior, each pair of variables $\left\{B_{\ell,i,i}, \gamma_{\ell,i}\right\}$ and $\left\{B_{\ell,i,\widetilde{s}_k}, \eta_{\ell,i,k}\right\}$ are strongly correlated since their product is the underlying variable that interacts with the data. Thus, a sensible approximation must treat each pair $\left\{B_{\ell,i,i}, \gamma_{\ell,i}\right\}$ and $\left\{B_{\ell,i,\widetilde{s}_k}, \eta_{\ell,i,k}\right\}$ as a unit so that $\left\{B_{\ell,i,i}, \gamma_{\ell,i}\right\}$ and $\left\{B_{\ell,i,\widetilde{s}_k}, \eta_{\ell,i,k}\right\}$ are placed in the same factor of the variational distribution. The simplest factorization that achieves this is:

$$
q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{\boldsymbol{B}}) = \prod_\ell^p \prod_i^m \prod_k^g q_{\ell,i}(\widetilde{B}_{\ell,i,i}, \gamma_{\ell,i})q_{\ell,i,k}(\widetilde{B}_{\ell,i,\widetilde{s}_k}, \eta_{\ell,i,k}),
$$

and we assume that q can have the following formulation:

$$q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{B}) = \prod_{\ell}^{p} \prod_{i}^{m} \prod_{k}^{g} q_{\ell,i}(\widetilde{B}_{\ell,i,i}, \gamma_{\ell,i}) q_{\ell,i,k}(\widetilde{B}_{\ell,i,\tilde{s}_k}, \eta_{\ell,i,k}), \tag{9}$$

$$= \prod_{l}^{p} \prod_{i}^{m} \prod_{k}^{g} q_{\ell,i}(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i}) q_{\ell,i}(\gamma_{\ell,i}) q_{\ell,i,k}(\widetilde{B}_{\ell,i,\tilde{s}_k}|\eta_{\ell,i,k}) q_{\ell,i,k}(\eta_{\ell,i,k}),$$

where $q_{\ell,i}(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i})$, $q_{\ell,i,k}(\widetilde{B}_{\ell,i,\tilde{s}_k}|\eta_{\ell,i,k})$, $q_{\ell,i}(\gamma_{\ell,i})$, and $q_{\ell,i,k}(\eta_{\ell,i,k})$ are the approximated posterior distribution of $\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i}$, $\widetilde{B}_{\ell,i,\tilde{s}_k}|\eta_{\ell,i,k}$, $\gamma_{\ell,i}$, and $\eta_{\ell,i,k}$, respectively, which were obtained from the prior distributions. We have assumed that segments are independent and that the elements with their own lags are also independent. With this assumption, we rewrite the lower bound as

$$L(q) = E_{q(\boldsymbol{\gamma}), q(\boldsymbol{\eta})} \left[E_{q(\widetilde{B}|\boldsymbol{\eta},\boldsymbol{\gamma})} \left[\log P\left(\boldsymbol{Y}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{B}|\boldsymbol{X}, \theta\right) - \log q\left(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{B}\right) \right] \right].$$

Hence, we have

$$L(q) = \sum_{\gamma} \prod_{\ell} \prod_{i} q(\gamma_{\ell,i}) \sum_{\eta} \prod_{\ell} \prod_{i} \prod_{k} q(\eta_{\ell,i,k}) \int_{\widetilde{B}} \left(\log P(Y, \gamma, \eta, \widetilde{B} | X, \theta) - \log q(\gamma, \eta, \widetilde{B}) \right)$$

$$\prod_{\ell}^{p} \prod_{i}^{m} \prod_{k}^{p} q(\widetilde{B}_{\ell,i,\widetilde{s}_k} | \eta_{\ell,i,k}) \prod_{\ell}^{p} \prod_{i}^{m} q(\widetilde{B}_{\ell,i,i} | \gamma_{\ell,i}) d\widetilde{B}$$

$$= \sum_{\gamma_{\ell,i}} q(\gamma_{\ell,i}) \sum_{\eta} \prod_{\ell} \prod_{i} \prod_{k} q(\eta_{\ell,i,k}) \int \int q(B_{\ell,i,i} | \gamma_{\ell,i}) \left[\int \log P(Y, \gamma, \eta, \widetilde{B} | X, \theta) \right.$$

$$\sum_{\gamma_{\ell',i}} \prod_{\ell' \neq \ell} \prod_{i} q(\gamma_{\ell',i}) q(\widetilde{B}_{\ell',i,i} | \gamma_{\ell',i,i}) d\widetilde{B}_{\ell',i,i} + \sum_{\gamma_{\ell,i'}} \prod_{\ell} \prod_{i' \neq i} q(\gamma_{\ell,i'}) q(\widetilde{B}_{\ell,i',i'} | \gamma_{\ell,i',i'}) d\widetilde{B}_{\ell,i',i'} \bigg]$$

$$d\widetilde{B}_{\ell,i,i} \prod_{\ell} \prod_{i} \prod_{k} q(\widetilde{B}_{\ell,i,\widetilde{s}_k} | \eta_{\ell,i,k}) d\widetilde{B}_{\ell,i,\widetilde{s}_k}$$

$$- \sum_{\gamma_{\ell,i}} q(\gamma_{\ell,i}) \sum_{\eta} \prod_{\ell} \prod_{i} \prod_{k} q(\eta_{\ell,i,k}) \int \int q(B_{\ell,i,i} | \gamma_{\ell,i}) \left[\log \left(q(B_{\ell,i,i} | \gamma_{\ell,i}) q(\gamma_{l,i}) \right) \right.$$

$$+ \log \left(q(\widetilde{B}_{\ell,i,\widetilde{s}_k} | \eta_{l,i,k}) q(\eta_{l,i,k}) \right) \bigg] d\widetilde{B}_{l,i,i} d\widetilde{B}_{l,i,\widetilde{s}_k} + \text{constant}$$

$$= E_{q(\eta),q(\widetilde{B}_\eta | \eta)} \left[E_{q(\widetilde{B}_{\ell,i,i} | \gamma_{\ell,i}=1)} \left[E_{\ell' \neq \ell,i} \left(\log \left(P(Y, \widetilde{B}, \gamma_{\ell',i}, \gamma_{\ell,i} = 1, \eta | X, \theta) \right) \right) \right. \right.$$

$$+ E_{\ell,i' \neq i} \left(\log \left(P(Y, \widetilde{B}, \gamma_{\ell,i'}, \gamma_{\ell,i} = 1, \eta | X, \theta) \right) \right) - \log \left(q(\widetilde{B}_{\ell,i,i} | \gamma_{\ell,i} = 1) \right) \bigg] \bigg] q(\gamma_{\ell,i} = 1)$$

$$+ E_{q(\eta),q(\widetilde{B}_\eta | \eta)} \left[E_{q(\widetilde{B}_{\ell,i,i} | \gamma_{\ell,i}=0)} \left[E_{\ell' \neq \ell,i} \left(\log \left(P(Y, \widetilde{B}, \gamma_{\ell',i}, \gamma_{\ell,i} = 0, \eta | X, \theta) \right) \right) \right. \right.$$

$$+ E_{\ell,i' \neq i} \left(\log \left(P(Y, \widetilde{B}, \gamma_{\ell,i'}, \gamma_{\ell,i} = 0, \eta | X, \theta) \right) \right) - \log \left(q(\widetilde{B}_{\ell,i,i} | \gamma_{\ell,i} = 0) \right) \bigg] \bigg] q(\gamma_{\ell,i} = 0)$$

$$+ \text{constant},$$

where $E_{\ell' \neq \ell,i}(\cdot)$ denotes that the expectation is taken with respect to all the ith column variables except lag-ℓ, and $E_{\ell,i' \neq i}(\cdot)$ denotes taking the expectation for all the lag-ℓ except the ith column. For a fixed $\gamma_{\ell,i} = 1$, $L(q)$ can be represented as

$$E_{q(\eta),q(\widetilde{B}_\eta | \eta)} \left[E_{q(\widetilde{B}_{\ell,i,i} | \gamma_{\ell,i}=1)} \left[E_{\ell' \neq l,i} \left(\log \left(P(Y, \widetilde{B}, \gamma_{\ell',i}, \gamma_{\ell,i} = 1, \eta | X, \theta) \right) \right) \right. \right.$$

$$+ E_{\ell,i' \neq i} \left(log \left(P(Y, \widetilde{B}, \gamma_{\ell,i'}, \gamma_{\ell,i} = 1, \eta | X, \theta) \right) \right) - log \left(q(\widetilde{B}_{\ell,i,i} | \gamma_{\ell,i} = 1) \right) \bigg] \bigg] .$$

This formulate can be treated as the negative KL divergence between $E_{\ell' \neq l,i} \left(\log \left(P(Y, \widetilde{B}, \gamma_{\ell',i}, \gamma_{\ell,i} = 1, \eta | X, \theta) \right) \right) + E_{\ell,i' \neq i} \left(\log \left(P(Y, \widetilde{B}, \gamma_{\ell,i'}, \gamma_{\ell,i} = 1, \eta | X, \theta) \right) \right)$ and $q(\widetilde{B}_{\ell,i,i} | \gamma_{\ell,i} = 1)$. Thus,

when

$$\log\left(q(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i}=1)\right) = E_{\ell'\neq\ell,i}\left(\log\left(P(Y,\widetilde{B},\gamma_{\ell',i},\gamma_{\ell,i}=1,\eta|X,\theta)\right)\right) \tag{10}$$
$$+ E_{\ell,i'\neq i}\left(\log\left(P(Y,\widetilde{B},\gamma_{\ell,i'},\gamma_{\ell,i}=1,\eta|X,\theta)\right)\right),$$

we can obtain the best approximation $q(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i}=1)$. Again, the cases for the given $\gamma_{\ell,i}=0$, $\eta_{\ell,i,k}=1$, and $\eta_{\ell,i,k}=0$ can be derived with the same procedure. Since both $\gamma_{\ell,i}$ and $\eta_{\ell,i,k}$ are from the independent Bernoulli distribution, with Eq. (10), we can add some variational parameters regarding $q(\gamma_{\ell,i})$ and $q(\eta_{\ell,i,k})$ and then derive the conditional distributiona of $\widetilde{B}_{\ell,i,i}$ given $\gamma_{\ell,i}$ and $\widetilde{B}_{\ell,i,\widetilde{s}_k}$ given $\eta_{\ell,i,k}$. Last, we optimize $L(q)$ to find the variational parameters. First, we derive $q(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i})$ and $q(\widetilde{B}_{\ell,i,\widetilde{s}_k}|\eta_{\ell,i,k})$ where this involves the joint probability function. To find the optimal form of Eq. (10), we rearrange the joint probability function to

retain only the terms involving ℓ, i, and $\widetilde{s}_k$ as follows:

$$\log\left(P(\boldsymbol{Y}, \widetilde{\boldsymbol{B}}, \boldsymbol{\gamma}, \boldsymbol{\eta}|\boldsymbol{X};\theta)\right) = \frac{-Tm}{2}\log(2\pi) - \frac{T}{2}\log(det(\Sigma)) \tag{11}$$
$$- \frac{1}{2}\left(-\frac{1}{2}tr\left(\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{XB})'(\boldsymbol{Y} - \boldsymbol{XB})\right)\right)$$
$$- \frac{pm}{2}\left(\log(2\pi) + \log(\sigma_B^2)\right) - \frac{\sum_\ell \sum_i \widetilde{B}_{\ell,i,i}^2}{2\sigma_B^2}$$
$$- \frac{pm(m-1)}{2}\left(\log(2\pi) + \log(\sigma_B^2)\right)$$
$$- \sum_\ell \sum_i \sum_{|\widetilde{s}_k|} \frac{1}{2}tr\left((\sigma_B^2 I_{|\widetilde{s}_k|})^{-1}\widetilde{B}'_{\ell,i,\widetilde{s}_k}\widetilde{B}_{\ell,i,\widetilde{s}_k}\right)$$
$$+ \sum_\ell \sum_i \gamma_{\ell,i}\log(\pi_1) + \sum_\ell \sum_i (1 - \gamma_{\ell,i})\log(1 - \pi_1)$$
$$+ \sum_\ell \sum_i \sum_k \eta_{\ell,i,k}\log(\pi_2) + \sum_\ell \sum_i \sum_k (1 - \eta_{\ell,i,k})\log(1 - \pi_2)$$
$$\propto -\frac{T}{2}\log(det(\Sigma)) - \frac{1}{2}tr\left(\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{XB})'(\boldsymbol{Y} - \boldsymbol{XB})\right)$$
$$- \sum_\ell \sum_i \frac{1}{2}\left((1 - \gamma_{\ell,i}) + \gamma_{\ell,i}\right)\log(\sigma_B^2) - \frac{\sum_\ell \sum_i \left((1 - \gamma_{\ell,i} + \gamma_{\ell,i})\right)\widetilde{B}_{\ell,i,i}^2}{2\sigma_B^2}$$
$$- \sum_\ell \sum_i \sum_k \left((1 - \eta_{\ell,i,k}) + \eta_{\ell,i,k}\right)\log(det(\sigma_B^2 I_{|\widetilde{s}_k|}))$$
$$- \frac{1}{2}tr\left((\sigma_B^2 I_{|\widetilde{s}_k|})^{-1}\left((1 - \eta_{\ell,i,k}) + \eta_{\ell,i,k}\right)\widetilde{B}'_{\ell,i,\widetilde{s}_k}\widetilde{B}_{\ell,i,\widetilde{s}_k}\right)$$
$$+ \sum_\ell \sum_i \gamma_{\ell,i}\log(\frac{\pi_1}{1 - \pi_1}) + \sum_\ell \sum_i \sum_k \eta_{\ell,i,k}\log(\frac{\pi_2}{1 - \pi_2})$$
$$+ pm\left(\log(1 - \pi_1)\right) + pmg\left(\log(1 - \pi_2)\right) + constant.$$

We can derive $\log(q(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i}))$ by taking the expectation in Eq. (10). When $\gamma_{\ell,i} = 1$, we have

$$\log\left(q(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i} = 1)\right) = -\frac{1}{2}\left[(\Sigma)_{i,i}^{-1}(\boldsymbol{X}_\ell^{(i)})'\boldsymbol{X}_\ell^{(i)} + \frac{1}{\sigma_B^2}\right]\widetilde{B}_{\ell,i,i}^2$$
$$- (\Sigma)_{i,i}^{-1}\widetilde{B}'_{\ell,i,i}\left[(\boldsymbol{X}_\ell^{(i)})'(\boldsymbol{Y}^{(i)} - \boldsymbol{X}_{-l}E(B_{-\ell}^{(i)}) - \boldsymbol{X}_\ell^{(-i)}E(B_\ell^{(-i,i)}))\right]$$
$$- E\left(tr((\Sigma)_{-i,i}^{-1}\widetilde{B}'_{\ell,i,i}((\boldsymbol{X}_\ell^{(i)})'(\boldsymbol{Y}^{(-i)} - (\boldsymbol{XB})^{(-i)})))\right) + constant.$$

We can find that this is a quadratic form of $\widetilde{B}_{\ell,i,i}$, the posterior of $q(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i} = 1)$ that follows a normal distribution in the form $N(\mu_{1,\ell,i,i}, \Sigma_{B_{\ell,i,i}})$, where

$$\Sigma_{B_{\ell,i,i}} = \left(\boldsymbol{X}_\ell^{(i)'}\boldsymbol{X}_\ell^{(i)}(\Sigma^{-1})_{i,i} + \sigma_B^2\right)^{-1},$$

$$\mu_{1,\ell,i,i} = \Sigma_{B_{\ell,i}}\left((\Sigma^{-1})_{i,i}\boldsymbol{X}_\ell^{(i)'}\left(\boldsymbol{Y}^{(i)} - \sum_{j\neq l}^{p}\boldsymbol{X}_j E(B_j^{(i)}) - \boldsymbol{X}_\ell^{(-i)}E(B_\ell^{(-i,i)})\right)\right.$$
$$\left. + E\left(tr\left((\Sigma^{-1})_{-i,i}\boldsymbol{X}_\ell^{(i)'}\left(\boldsymbol{Y}^{(-i)} - \boldsymbol{X}\boldsymbol{B}^{(-i)}\right)\right)\right)\right).$$

Similarly, $log(q(\widetilde{B}_{\ell,i,\widetilde{s}_k}|\eta_{\ell,i,k}))$ takes the expectation in Eq. (10); when $\eta_{\ell,i,k} = 1$, we have

$$\log\left(q(\widetilde{B}_{\ell,i,\widetilde{s}_k}|\eta_{\ell,i,k} = 1)\right) = -\frac{1}{2}tr\left((\Sigma)_{\widetilde{s}_k,\widetilde{s}_k}^{-1}(\boldsymbol{X}_\ell^{(i)})'\boldsymbol{X}_\ell^{(i)}\widetilde{B}'_{\ell,i,\widetilde{s}_k}\widetilde{B}_{\ell,i,\widetilde{s}_k}\right)$$
$$- \frac{1}{2}tr\left((\sigma_B I_{|\widetilde{s}_k|})^{-1}\widetilde{B}'_{\ell,i,\widetilde{s}_k}\widetilde{B}_{\ell,i,\widetilde{s}_k}\right)$$
$$- tr\left((\boldsymbol{X}_\ell^{(i)})'(\boldsymbol{Y}^{(\widetilde{s}_k)} - \boldsymbol{X}_{-l}E(B_{-\ell}^{(\widetilde{s}_k)}) - \boldsymbol{X}_\ell^{(-i)}E(B_\ell^{(-i,\widetilde{s}_k)}))(\Sigma^{-1})_{\widetilde{s}_k,\widetilde{s}_k}\right)$$
$$- E\left(tr\left((\boldsymbol{X}_\ell^{(i)})'(\boldsymbol{Y}^{(-\widetilde{s}_k)} - (\boldsymbol{X}\boldsymbol{B})^{(-\widetilde{s}_k)})(\Sigma^{-1})_{\widetilde{s}_k,i}\right)\right) + constant,$$

where the posterior $q(\widetilde{B}_{\ell,i,\widetilde{s}_k}|\eta_{\ell,i,\widetilde{s}_k} = 1)$ follows the $1 \times |\widetilde{s}_k|$ multivariate normal distribution $N_{1\times|\widetilde{s}_k|}(\mu_{2,\ell,i,\widetilde{s}_k}, \boldsymbol{I}, \Sigma_{B_{\ell,i,\widetilde{s}_k}})$, in which

$$\Sigma_{B_{\ell,i,\widetilde{s}_k}} = \left((\boldsymbol{X}_\ell^{(i)})'\boldsymbol{X}_\ell^{(i)}(\Sigma^{-1})_{\widetilde{s}_k,\widetilde{s}_k} + \sigma_B^2 I_{|\widetilde{s}_k|}\right)^{-1},$$

$$\mu_{2,\ell,i,\widetilde{s}_k} = \left((\boldsymbol{X}_\ell^{(i)})'\left(\boldsymbol{Y}^{(\widetilde{s}_k)} - \sum_{j\neq l}^{p}\boldsymbol{X}_j E(B_j^{(\widetilde{s}_k)}) - \boldsymbol{X}_\ell^{(-i)}E(B_\ell^{(-i,\widetilde{s}_k)})\right)(\Sigma^{-1})_{\widetilde{s}_k,\widetilde{s}_k}\right.$$
$$\left. + E\left(tr\left((\boldsymbol{X}_\ell^{(i)})'\left(\boldsymbol{Y}^{(-\widetilde{s}_k)} - \boldsymbol{X}\boldsymbol{B}^{(-\widetilde{s}_k)}\right)(\Sigma^{-1})_{-\widetilde{s}_k,\widetilde{s}_k}\right)\right)\right)\Sigma_{B_{\ell,i,\widetilde{s}_k}}.$$

According to the same process mentioned above, when $\gamma_{\ell,i} = 0$ and $\eta_{\ell,i,k} = 0$, we have

$$q(\widetilde{B}_{\ell,i,i}|\gamma_{\ell,i} = 0) \sim N(0, \sigma_B^2),$$
$$q(\widetilde{B}_{\ell,i,\widetilde{s}_k}|\eta_{\ell,i,\widetilde{s}_k} = 0) \sim N_{1\times|\widetilde{s}_k|}(\boldsymbol{0}, \boldsymbol{I}, \sigma_B^2 I_{|\widetilde{s}_k|}).$$

Therefore, $\phi_{1,\ell,i}$ and $\phi_{2,\ell,i,k}$ are the probabilities of $\gamma_{\ell,i} = 1$ and $\eta_{\ell,i,k} = 1$, respectively, and we have

$$q(\boldsymbol{\eta}, \boldsymbol{\gamma}, \widetilde{B}) = \prod_l^p \prod_i^m \left(\phi_{1,\ell,i} N \left(\mu_{1,\ell,i,i}, \Sigma_{B_{\ell,i,i}} \right) \right)^{\gamma_{\ell,i}} \left(\left(1 - \phi_{1,\ell,i} \right) N(0, \sigma_B^2) \right)^{(1-\gamma_{\ell,i})}$$

$$\prod_k^g \left(\phi_{2,\ell,i,k} N_{1\times|\widetilde{s}_k|}(\mu_{2,\ell,i,\widetilde{s}_k}, \boldsymbol{I}, \Sigma_{B_{\ell,i,\widetilde{s}_k}}) \right)^{\eta_{\ell,i,k}} \left(\left(1 - \phi_{2,\ell,i,k} \right) N_{1\times|\widetilde{s}_k|}(0, \boldsymbol{I}, \sigma_B^2 I_{|\widetilde{s}_k|}) \right)^{(1-\eta_{\ell,i,k})},$$

where $\phi_{1,\ell,i}$ and $\phi_{2,\ell,i,k}$ are

$$\phi_{1,\ell,i} = Inv - logit \left\{ logit(\pi_1) - \frac{1}{2} log(\sigma_B^2) + \frac{1}{2} log(det(\Sigma_{B_{\ell,i,i}})) + \frac{(\Sigma_{B_{\ell,i,i}})^{-1} \mu_{1,\ell,i,i}^2}{2} \right\} \text{ and}$$

$$\phi_{2,\ell,i,k} = Inv - logit \left\{ logit(\pi_2) - \frac{1}{2} log(det(\sigma_\beta^2 I_{|\widetilde{s}_k|})) + \frac{1}{2} log(det(\Sigma_{B_{\ell,i,k}})) \right.$$

$$\left. + \frac{1}{2} tr \left((\Sigma_{B_{\ell,i,\widetilde{s}_k}})^{-1} \mu'_{2,\ell,i,\widetilde{s}_k} \mu_{2,\ell,i,\widetilde{s}_k} \right) \right\}.$$

B.1.2 M-Step

During the M-step, we update the parameters $\theta = \{\pi_1, \pi_2, \Sigma, \sigma_B^2\}$ with $\frac{L(q)}{\partial\theta} = 0$. Considering π_1 and π_2, by setting $\frac{L(q)}{\partial\pi_1} = 0$ and $\frac{L(q)}{\partial\pi_2} = 0$, we obtain

$$\pi_1 = \frac{\sum_\ell^p \sum_i^m \phi_{1,\ell,i}}{pm},$$

$$\pi_2 = \frac{\sum_\ell^p \sum_i^m \sum_k^g \phi_{2,\ell,i,k}}{\sum_k^g |\widetilde{s}_k| pm}.$$

For Σ and σ_β^2, setting $\frac{L(q)}{\partial\Sigma} = 0$ and $\frac{L(q)}{\partial\sigma_\beta^2} = 0$, we obtain

$$\Sigma = \frac{E\left((\boldsymbol{Y} - \boldsymbol{XB})'(\boldsymbol{Y} - \boldsymbol{XB})\right)}{T},$$

$$\sigma_B^2 = \frac{\sum_\ell^p \sum_i^m \phi_{1,\ell,i}(\Sigma_{B,\ell,i,i} + \mu_{1,\ell,i}^2) + \sum_\ell^p \sum_i^m \sum_k^g \phi_{2,\ell,i,k} tr(\Sigma_{B,l,i,\widetilde{s}_k} + \mu'_{2,\ell,i,\widetilde{s}_k} \mu_{2,\ell,i,\widetilde{s}_k})}{\sum_\ell^p \sum_i^m \left(\phi_{1,\ell,i} + \sum_k^g |\widetilde{s}_k| \phi_{2,\ell,i,k} \right)}.$$