Statistics and Computing (2024) 34:91
https://doi.org/10.1007/s11222-024-10402-y

ORIGINAL PAPER

Bayesian variable selection for matrix autoregressive models

Alessandro Celani[1] · **Paolo Pagnottoni**[2] · **Galin Jones**[3]

Received: 10 August 2023 / Accepted: 5 February 2024 / Published online: 11 March 2024

Abstract
A Bayesian method is proposed for variable selection in high-dimensional matrix autoregressive models which reflects and exploits the original matrix structure of data to (a) reduce dimensionality and (b) foster interpretability of multidimensional relationship structures. A compact form of the model is derived which facilitates the estimation procedure and two computational methods for the estimation are proposed: a Markov chain Monte Carlo algorithm and a scalable Bayesian EM algorithm. Being based on the spike-and-slab framework for fast posterior mode identification, the latter enables Bayesian data analysis of matrix-valued time series at large scales. The theoretical properties, comparative performance, and computational efficiency of the proposed model is investigated through simulated examples and an application to a panel of country economic indicators.

Keywords Autoregressive models · Bayesian estimation · Matrix-valued time series · Maximum a posteriori probability · Stochastic search

1 Introduction

The emergence of high-dimensional time series observed in matrix form gives birth to new modeling challenges in economics, finance, and related fields. The existing approaches to dimension reduction in high-dimensional multivariate time series analysis can be organized in two major classes: (a) factor models (Bai and Ng 2002; Forni et al. 2005; Lam et al. 2011), and (b) modeling with frequentist regularization or Bayesian methods (Rothman et al. 2010; Song and Bickel 2011; Kock and Callot 2015; Park and Casella 2008; Bańbura et al. 2010; Gefang 2014; Ahelegbey et al. 2016; Korobilis 2021).

✉ Paolo Pagnottoni
paolo.pagnottoni@uninsubria.it

Alessandro Celani
a.celani@pm.univpm.it

Galin Jones
galin@umn.edu

[1] Department of Economics and Social Sciences, Marche Polytechnic University, Piazzale Martelli Raffaele 8, 60121 Ancona, AN, Italy

[2] Department of Economics, University of Insubria, Via Monte Generoso 71, 21100 Varese, VA, Italy

[3] School of Statistics, University of Minnesota Twin Cities, 224 Church Street S.E., Minneapolis, MN 55455, USA

While most of the extant modelling paradigms are designed to encourage parsimony by treating observations as time series vectors, in economics, finance and other fields, observations in matrix or tensor form are often generated over time. For instance, the collection of panel data forms matrix-valued time series observations, whose rows might represent indicators and columns countries—see Fig. 1. In this context, univariate time series analysis focuses on one element of the matrix at a time. Vector and panel time series analysis deal with the co-movement of one row in the matrix. Modeling each dimension separately annihilates the multidimensional structure of data, and can therefore lead to a significant loss of information or efficiency.

A strand of literature has therefore developed and studied multidimensional time series models, including matrix-valued ones (Hoff 2015, 2011; Chen and Yang 2021; Wang et al. 2019; Billio et al. 2022). However, when the matrix observation has large dimensions, the matrix autoregressive (MAR) model involves a large number of parameters, which requires further dimension reduction techniques to produce accurate estimation. So far, this has been primarily tackled through the introduction of factor autoregressive models for multidimensional time series (Wang et al. 2019; Chen and Fan 2021; Gao and Tsay 2021; Chen et al. 2022). Although factor approaches can cope well with the dimensionality problem, the resulting parameters lack a clear interpretation. As in the vector case, the factor approach in a matrix-valued

dataset involves replacing the original matrix of dependent variables with an estimated matrix of factors, featuring a sensibly lower number of rows and/or columns. Nevertheless, each coefficient refers to a factor, and not to an observed variable, thus hampering interpretation. In addition, factor models require choosing the number of factors, which is not a parameter that can be estimated. In light of this, dimensionality problems that preserve the original problem structure, such as the one performing variable selection and regularization, are preferred.

In parallel, research on high-dimensional statistics has widely studied sparse recovery in the context of normal linear regression models (George and McCulloch 1993; Barrett and Gray 1994; George and McCulloch 1997; Chen and Huang 2016; Wang et al. 2018; Samanta et al. 2022). In the Bayesian literature, the treatment of high-dimensional linear regression models mostly stems from the seminal papers on variable selection via Gibbs sampling (George and McCulloch 1993, 1997), hereafter identified as stochastic search variable selection (SSVS). More recently, a deterministic alternative to stochastic search was proposed based on an EM variable selection (EMVS) algorithm to find maximum a posteriori probability (MAP) estimates (Ročková and George 2014). As in SSVS, the EMVS method combines a spike-and-slab regularization procedure for the discovery of active predictor sets with subsequent evaluation of posterior model probabilities. However, EMVS provides an appealing alternative to SVSS in terms of computational efficiency.

Fully Bayesian variable selection has attracted attention in the multivariate time series context, starting from a stochastic search approach to selecting restrictions for vector autoregressive (VAR) models (George et al. 2008). However, the EMVS framework has yet to be exploited in multivariate time series. Moreover, variable selection has only been considered in vector-valued time series models to date, highlighting the need of identifying restrictions in multidimensional autoregressive models. The absence of restrictions on the regression coefficients results in a potentially large number of parameters relative to the available data and with a limited number of observations, over-parameterization can affect the precision of inference.

We propose a novel matrix autoregressive model where sparsity is induced naturally. By deriving a compact form for MAR models borrowed from the tensor linear regression framework, we design an estimation strategy using MCMC, which allows a full range of Bayesian inference, and a Bayesian EMVS procedure, which allows for fast posterior mode identification. That is, the latter substantially reduces the computational time required for the MCMC procedure, making it feasible when dealing with large-scale multidimensional time series.

The properties of the proposed model are demonstrated through simulations and examples in an application to macroeconomic data. The simulation experiments show: (a) the gain in small sample efficiency of the proposed estimators relative to maximum likelihood (ML) in high-dimensional sparse settings; and (b) that the proposed estimators perform generally better than standard VARs and several competing alternatives suited for longitudinal data, while limiting computational intensity in the EMVS formulation. The empirical application to macroeconomic data confirms that the model is able to: (a) handle high-dimensional longitudinal data; (b) outperform competing alternatives in high-dimensional settings; and (c) yield enhanced interpretability given by the autoregressive model in matrix form.

The proposed model can be readily extended to the tensor autoregressive (TAR) framework. In contrast to other work (e.g. Billio et al. 2022), it encompasses a Tucker structure in the matrix coefficient rather than a PARAFAC decomposition. The PARAFAC decomposition arises in the general context of tensor decomposition and is applicable to matrices in the bi-dimensional case. In particular, it achieves dimension reduction by factorizing a tensor in the outer (element-wise) product of vectors. The resulting tensor is of rank one, featuring a large reduction in the number of coefficients.[1] Thus, the PARAFAC decomposition has the advantage of being highly parsimonious but it can be restrictive for empirical purposes. This is magnified in economic and financial applications where the relationship among variables exhibits complex and highly interconnected behaviors. Conversely, the Tucker structure admits an arbitrary number of factor components in each mode. This allows modeling dimension asymmetric tensors and benefits from enhanced interpretability of mode-specific interrelationships.

The remainder proceeds as follows. Section 2 outlines the proposed sparse MAR model, while Sect. 3 discusses the shrinkage priors and Sect. 4 describes the MCMC and EMVS estimation procedures. Section 5 describes the dynamic analysis via Kronecker generalized forecast error variance decomposition (GFEVD). Section 6 evaluates the model performance and computational times, compared to some key competitors, through a simulation study. Section 7 considers an empirical application to macroeconomic data. Section 8 contains some concluding remarks.

2 Model and prior structure

2.1 The matrix autoregressive model

The MAR model takes advantage of the original structure of the data by modeling matrix autoregressive dynamics in a bilinear form. This is of paramount importance, for instance,

[1] A thorough exposition of tensor decomposition is available in Kolda and Bader (2009).

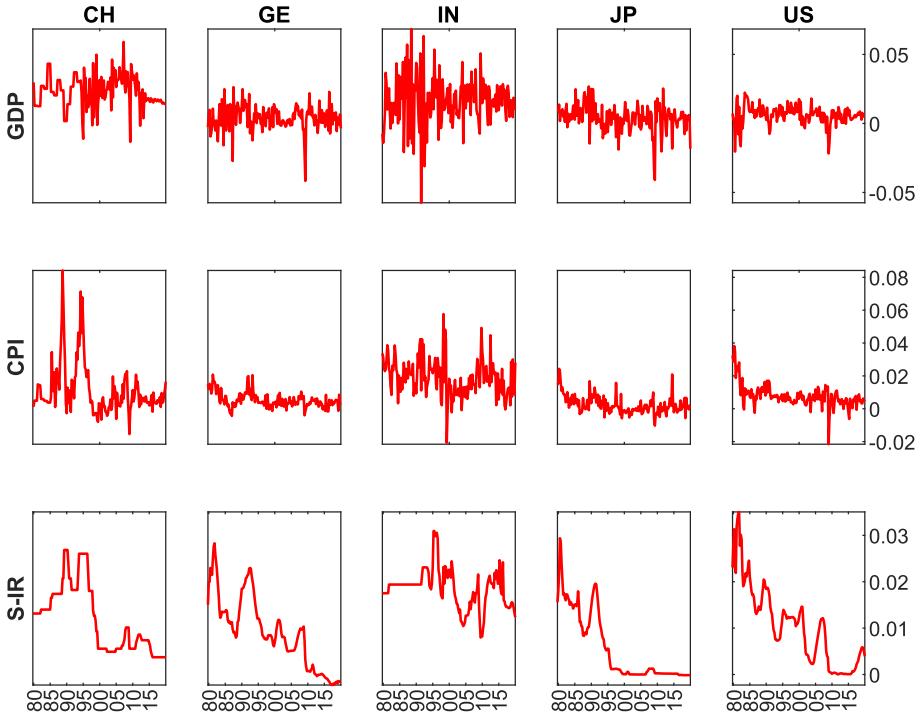


Fig. 1 Time series of three economic indicators: GDP growth, Consumer Price Index, and Short-term Interest Rate for five countries

in panel data applications where each observation at time t can be conceived as a matrix and whose rows represent indicators and columns represent countries.

We introduce a well-known formulation of the MAR model (Chen and Yang 2021) and then propose a parsimonious alternative model. The alternative can be useful in the estimation required for a model with several lag orders relative to the number of time series observations. Then, given the multidimensional structure, we cast the model into the tensor linear regression framework which will allow easier estimation. Finally, we consider the estimation of high-dimensional MAR models via variable selection in a Bayesian framework.

we observe $g = 1, ..., G$ indicators for a number of countries $g = 1, ..., G$, across $t = 1, ..., T$, $t = 1 \ldots, T$ time points. Let $\boldsymbol{y}_{n,t}$ be the, G dimensional vector that collects the domestic indicators for each country. $n = 1, \ldots, N$ Moreover, we define $\mathbf{Y}_t = [\mathbf{y}_{1,t}, ..., \mathbf{y}_{N,t}]$ to be the $[G \times N]$ matrix of endogenous variables, obtained by horizontally stacking all the country specific vectors. The conditional mean of the matrix observation is expressed as the product of P lagged observations, each one multiplied by two left and right autoregressive coefficient matrices $\mathbf{A}_1, ..., \mathbf{A}_P$ and $\mathbf{B}_1, ..., \mathbf{B}_P$, of dimension $[G \times G]$ and $[N \times N]$, repsectively. Specifically, t (Chen and Yang 2021, Section 2)

$$\begin{aligned} \mathbf{Y}_t &= \sum_{i=1}^{P} \mathbf{A}_i \mathbf{Y}_{t-i} \mathbf{B}_i^{'} + \mathbf{E}_t, \\ \mathbf{E}_t &\sim \mathcal{MN}(\mathbf{0}, \Sigma_1, \Sigma_2), \end{aligned} \tag{1}$$

where $\mathbf{E}_t \in \mathbb{R}^{G \times N}$ is a white noise matrix of the same dimension as $\mathbf{Y}_t$ with two symmetric positive definite covariance matrices Σ_1 and Σ_2 of dimensions $G \times G$ and $N \times N$, respectively. We use $\mathcal{MN}$ to denote the multilinear normal distribution (Gupta and N 1999; Ohlson et al. 2013).

Let vec$(\cdot)$ be the usual vectorization of a matrix by stacking its columns and $\otimes$ the usual Kronecker product. Let $\mathbf{y}_{t-i} = \text{vec}(\mathbf{Y}_{t-i})$ for $i = 0, \ldots, P$ and $\mathbf{e}_t = \text{vec}(\mathbf{E}_t)$. The vectorized form of the MAR model in Eq. (1) is

$$\begin{aligned} \mathbf{y}_t &= \sum_{i=1}^{P} (\mathbf{B}_i \otimes \mathbf{A}_i) \mathbf{y}_{t-i} + \mathbf{e}_t, \\ \mathbf{e}_t &\sim \mathcal{N}(\mathbf{0}, \Sigma_2 \otimes \Sigma_1). \end{aligned} \tag{2}$$

The representation in Eq. (2) shows that the MAR(P) model can be regarded as a special case of a VAR(P) model, with an autoregressive coefficient matrix given by a Kronecker product of the two mode-specific matrices. As a direct consequence of the equivalence, the MAR(P) in Eq. (1) is stationary if all the roots of $|\mathbf{I}_{GN} - (\mathbf{B}_1 \otimes \mathbf{A}_1)z + \cdots + (\mathbf{B}_P \otimes \mathbf{A}_P)z^P| = 0$ fall outside the unit circle. Under this assumption, the model in Eq. (1) can be rewritten as the infinite moving average representation

$$\mathbf{y}_t = \mathbf{e}_t + \boldsymbol{\Psi}_1 \mathbf{e}_{t-1} + \cdots + \boldsymbol{\Psi}_\infty \mathbf{e}_{t-\infty},$$

where the $GN \times GN$ coefficient matrices $\mathbf{\Psi}_k$ can be obtained via the following recursion

$$\mathbf{\Psi}_k = (\mathbf{B}_1 \otimes \mathbf{A}_1)\mathbf{\Psi}_{k-1} + \cdots + (\mathbf{B}_P \otimes \mathbf{A}_P)\mathbf{\Psi}_{k-P}, \ \ k = 1, 2, \ldots$$

the model is thenwhere $\mathbf{\Psi}_k = \mathbf{0}$ for $k < 0$ and $\mathbf{\Psi}_0 = \mathbf{I}_{GN}$.

2.2 Model interpretation

To ease the exposition, consider the case $P = 1$. In order to interpret the model parameters, it is helpful to highlight that, if $\mathbf{Y}_t$ follows a MAR(1), then for a specific pair $g = 1, \ldots, G$ and $n = 1, \ldots, N$, the conditional expected value of each entry $y_{gn,t}$ of $\mathbf{Y}_t$ is

$$E(y_{gn,t}|\mathbf{y}_{t-1}) = \sum_{i=1}^{G}\sum_{j=1}^{N} a_{g,i} b_{n,j} y_{ij,t-1},$$

where $a_{g,i} \in \mathbf{A}$ and $a_{n,j} \in \mathbf{B}$. Roughly speaking, $a_{g,i}$ embeds the impact of the rows of $\mathbf{Y}_{t-1}$ on the gth row of $\mathbf{Y}_t$. Analogously, $b_{n,j}$ describes the influence of the columns of $\mathbf{Y}_{t-1}$ on the nth column of $\mathbf{Y}_t$. This model could be referred to as a bilinear multiplicative model, given that the temporal effect on each observation is weighted by the product of two coefficients. Such restriction holds as long as the structure of the matrix dataset embeds two distinct dynamics, one for each dimension, which can be separately identified.

Consider the example shown in Fig. 1, where columns represent countries and rows macroeconomic indicators. Conceiving the model as a hierarchical structure, the conditional expectation of an economic indicator of one country is a linear combination of all the indicators of the other countries (globally adjusted combination):

$$\tilde{\mathbf{y}}_{n,t} = [b_{n,1} y_{ij,t-1}, \ldots, b_{n,N} y_{iN,t-1}]',$$

each one weighted for indicator-specific coefficients:

$$E(y_{gn,t}|\mathbf{y}_{t-1}) = [a_{g,1}, \ldots, a_{g,G}] \times \tilde{\mathbf{y}}_{n,t}.$$

This hierarchical structure, naturally induced by the matrix structure, is coherent with several restriction approaches proposed in the literature for Panel data in the vectorized framework (Pesaran et al. 2004; Canova and Ciccarelli 2009; Korobilis 2016; Camehl 2022). Analogous interpretations follow for the contemporaneous relationship captured by the left and right coefficient matrices Σ_1 and Σ_2. For a more detailed explanation of the model interpretation see Chen and Yang (2021).

The model can be made more parsimonious, if more restrictive, by assuming that the row and column effects matrices are time invariant ($\mathbf{A}_i = \mathbf{A}$, $\mathbf{B}_i = \mathbf{B}$), while embedding the effects of lagged observations in a row vector $\mathbf{c} \in \mathbb{R}^{1\times P}$. The conditional mean of $\mathbf{Y}_t$ is a linear combination of its P lagged values, pre and post multiplied by $\mathbf{A}$ and $\mathbf{B}$ as in a MAR(1). If $c_i \in \mathbf{c}$, the model is then

$$\mathbf{Y}_t = \mathbf{A}\left(\sum_{i=1}^{P} c_i \mathbf{Y}_{t-i}\right)\mathbf{B}' + \mathbf{E}_t \tag{3}$$

Although it can be expressed as a MAR, the model arises naturally as a special case of tensor autoregression (TAR), as will be shown in the following subsection. We therefore denote the model in Eq. (3) as MAR*(P).

It is clear that the MAR* formulation is quite restrictive and might not be a feasible specification for several types of applications, given that the dependence is expressed by the sole c_i coefficients, while $\mathbf{A}$ and $\mathbf{B}$ remain fixed. However, this formulation might be seen as a parsimonious alternative in particularly high-dimensional contexts. It is not uncommon to find empirical applications to macroeconomic data where the length of the time series requires the use of dimension reduction techniques to estimate model parameters. Despite being restrictive, the fixed-parameter formulation can be useful in modeling the impacts of a large number of lags relative to the available number of time series observations.

An unrestricted VAR(P) estimates $(GN)^2P \in \mathcal{O}(n^5)$ parameters, but due to the Kronecker structure imposed on its coefficient matrices, the MAR(P) model estimates only $(G^2 + N^2)P \in \mathcal{O}(n^3)$ parameters, where $\mathcal{O}(\cdot)$ denotes the order of parameter complexity. The further restriction imposed that yields the MAR*(P) results in $G^2 + N^2 + P \in \mathcal{O}(n^2)$ parameters to estimate. As a result, the number of parameters grows as a linear function of the lag order. This is clear from Fig. 2, in which a graphical comparison of the three models is depicted.

2.3 Compact form

Vector and matrix operations can be readily generalized to the tensor case, but notions of tensor algebra are necessary to proceed. See Appendix A.1 for some basics of tensor notation and calculus and Cichocki (2018) and Kolda and Bader (2009) for more details.

Although an estimation procedure for the MAR(1) model is readily available (Chen and Yang 2021), a compact form of the model is instrumental for developing a more general and coherent estimation procedure for the MAR(P). We therefore derive a comprehensive compact form for any general K-dimensional TAR(P), which admits the MAR(P) and MAR*(P) as special cases, by establishing a connection with the general Tensor Linear Regression (TLR) model. Indeed,

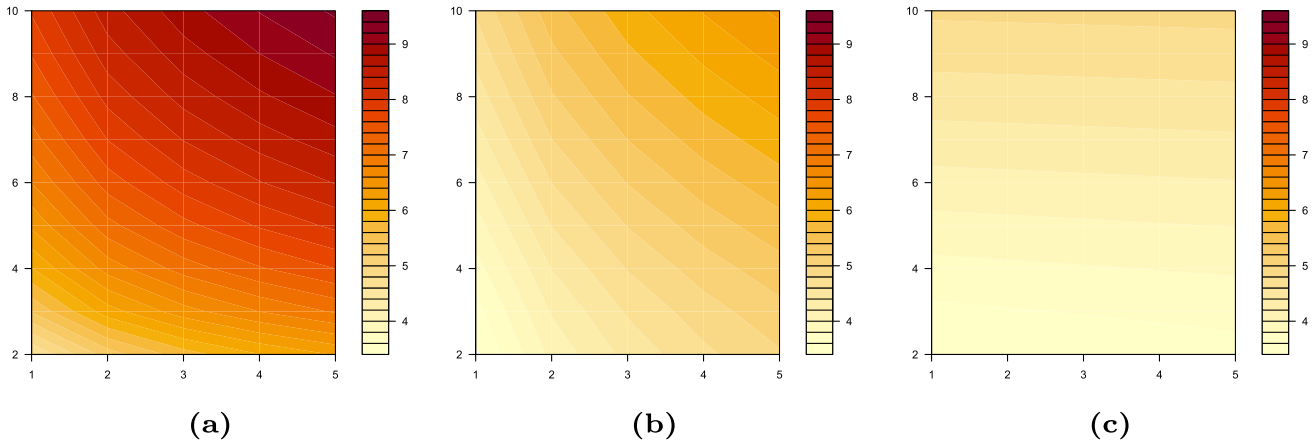


Fig. 2 Logarithm of the number of parameters in a Panel VAR (**a**), MAR (**b**) and MAR* (**c**) with $G = 5$, $N = 2, \ldots, 10$ (*y axis*) and lags $P = 1, \ldots, 5$ (*x axis*). The MAR is much less sensitive than the VAR to the increase in G, but it suffers from over-parameterization compared to the MAR*, for which this effect is negligible

as the VAR emerges as a multivariate linear regression (MLR) model in its compact form, the analogous counterpart for the MAR is a TLR model.

The compact form of the VAR can be easily derived via the matrix of dependent variables $\mathbf{Y}$ obtained by horizontally stacking each $\mathbf{y}_t$ for $t = 1, \ldots, T$. However, for each point in time, the variables in a MAR are already matrix shaped, so that its related compact form consists of a third order tensor.

Let us consider a K way (order) tensor $\mathcal{Y} \in \mathbb{R}^{J_1, \ldots, J_K}$ which is a K-dimensional array with entries $\mathcal{Y}_{j_1, \ldots, j_K}$ with $j_k = 1, \ldots, J_k$ for $k = 1, \ldots, K$. We define the response and explanatory tensor for the MAR in Eq. (1). Let $\mathcal{Y}$ be a $K = 3$ way response tensor of dimension $[G \times N \times T - P]$ with $\mathcal{Y}_{:,:,t} = \mathbf{Y}_t$. Then, define the explanatory tensor $\mathcal{X}$ to be of dimension $[GP \times NP \times T - P]$. When fixing the third dimension $j_3 = t = 1, \ldots, T - P$, we obtain tensor slices of dimensions $[GP \times NP]$, which we fill with the lagged values of $\mathbf{Y}_t$ and zeros otherwise:

$$\mathcal{X}_{:,:,t} = \begin{bmatrix} \mathbf{Y}_{t-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Y}_{t-2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Y}_{t-P} \end{bmatrix}.$$

In the case of the the MAR*(P) in (3), the response and explanatory objects of interest need to be slightly modified. We define $\mathcal{X}$ to be a $K = 4$ order tensor of dimensions $[G \times N \times P \times T - P]$ such that $\mathcal{X}_{:,:,i,t} = \mathbf{Y}_{t-i}$ and, for the sake of coherence, we modify $\mathcal{Y}$ to be a four-dimensional tensor $\mathcal{Y} \in \mathbb{R}^{G \times N \times 1 \times T - P}$ as well.

We may now write a unique compact form which encompasses both MAR(P) and MAR*(P). Let $\mathcal{B} = \{[\mathbf{A}_1, \ldots, \mathbf{A}_P], [\mathbf{B}_1, \ldots, \mathbf{B}_P], \mathbf{I}_{T-P}\}$ and $\mathbf{\Sigma} = \{\Sigma_1, \Sigma_2, \mathbf{I}_{T-P}\}$ in case of MAR(P), $\mathcal{B} = \{\mathbf{A}, \mathbf{B}, \mathbf{c}, \mathbf{I}_{T-P}\}$ and $\mathbf{\Sigma} = \{\Sigma_1, \Sigma_2, \Sigma_3, \mathbf{I}_{T-P}\}$[2] in case of MAR*$(P)$. Denoting $\bar{\times}$ as the Tucker product and $\mathcal{U}$ the tensor white noise, then we have

$$\begin{aligned} \mathcal{Y} &= \mathcal{X} \bar{\times} \mathcal{B} + \mathcal{E} \\ \mathcal{E} &= \mathcal{U} \bar{\times} \mathbf{\Sigma}^{1/2} \\ \mathcal{U} &\sim \mathcal{MN}(\mathbf{0}, \mathbf{I}_{J_1}, \ldots, \mathbf{I}_{J_K}). \end{aligned} \tag{4}$$

Notice that Eq. (4) can be seen as the multilinear generalization of the compact form of VAR models—see Lütkepohl (2005). We refer the reader to Appendix A.1 which describes the matricization and the Tucker product, the fundamental operators that are used to create and then manipulate $\mathcal{B}$ and Σ, and Kolda and Bader (2009); Bader and Kolda (2006) for further reference.[3]

It is convenient to derive the MLR representation of the models. To simplify the notation, let $\mathbf{\Phi}_1 = [\mathbf{A}_1, \ldots, \mathbf{A}_P]$ and $\mathbf{\Phi}_2 = [\mathbf{B}_1, \ldots, \mathbf{B}_P]$ for the MAR(P), $\mathbf{\Phi}_1 = \mathbf{A}$, $\mathbf{\Phi}_2 = \mathbf{B}$ and $\mathbf{\Phi}_3 = \mathbf{c}$ for the MAR*(P). Then, let $\tilde{\mathbf{Y}}^k = \text{mat}_k(\mathcal{Y})\Sigma_{-k}$, $\tilde{\mathbf{X}}^k = \text{mat}_k(\mathcal{X})\Sigma_{-k}$ and $\tilde{\mathbf{E}}^k = \text{mat}_k(\mathcal{E})\Sigma_{-k}$, where $\Sigma_{-k} = \Sigma_K \otimes \cdots \otimes \Sigma_{k+1} \otimes \Sigma_{k-1}, \ldots, \Sigma_1$, where $\text{mat}_k(\cdot)$ denotes the k-mode matricization operator as defined in Appendix A.1. By matricizing both sides of Eq. (4) for each k, one can easily derive a MLR model which highlights the kth way

[2] Σ_3 represents the third-way covariance matrix. Notice that, being the third dimension of $\mathcal{Y}$ of order 1, Σ_3 collapses to a scalar, i.e. $\Sigma_3 = \sigma_3^2$.

[3] The multilinear Normal distribution represents the generalization in higher dimension of the multivariate Normal, and the Matrix Normal. It has $K + 1$ parameters: a Kth order tensor of means, and K separate p.d. covariance matrices.

conditional mean and variance matrices $\{\mathbf{\Phi}_k, \Sigma_k\}$:

$$\begin{aligned}\tilde{\mathbf{Y}}^k &= \mathbf{\Phi}_k \tilde{\mathbf{X}}^k + \tilde{\mathbf{E}}^k \\ \tilde{\mathbf{E}}^k &\sim \mathcal{MN}(\mathbf{0}, \Sigma_k, \mathbf{I}_{J_{-k}}).\end{aligned} \tag{5}$$

The latter formulation allows the estimation to be carried out for each of the K modes separately. Notice that in the case of vector-valued time series, the compact form in Eq. (4) reduces to the formulation in Eq. (5).

3 Shrinkage priors

Consider the compact form in Eq. (4). Let $J_{-k} = \prod_{l \neq k} J_l$ so that the likelihood can be written as:

$$\begin{aligned}\mathcal{L}(\mathcal{Y}|\mathcal{B}, \mathbf{\Sigma}) \propto \prod_{k=1}^{K} |\Sigma_k|^{-\frac{J_{-k}}{2}} \\ \exp\left(-\frac{1}{2}||(\mathcal{Y} - \mathcal{X}\bar{\times}\mathcal{B})\bar{\times}\mathbf{\Sigma}^{-\frac{1}{2}}||_2^2\right).\end{aligned} \tag{6}$$

Also, notice that the likelihood for the MAR model in terms of $\{\mathbf{\Phi}_k, \Sigma_k\}$, given the other parameters, is proportional to the likelihood in Eq. (6):

$$\mathcal{L}(\mathcal{Y}|\mathbf{\Phi}_k, \Sigma_k) \propto \mathcal{L}(\mathcal{Y}|\mathbf{\Phi}, \Sigma).$$

This implies that both ML and Bayesian estimation can be carried for each dimension separately (see, e.g., Hoff 2015). As a consequence, sparsity can be induced in the two modes independently, easing the introduction of regularization methods for MAR models.

We will consider a spike-and-slab framework for the prior to induce sparsity in the two mode-specific coefficients of the MAR; a related approach is available for VAR models (George et al. 2008).

The proposed model is a multi-dimensional generalization of the uni-dimensional linear regression framework. Therefore, the variable selection for the MAR has much in common with the variable selection for the VAR in the bi-dimensional setting (George et al. 2008). The key feature of the variable selection approach is to construct a hierarchical prior for each lagged coefficient $\phi_{i,k} \in \boldsymbol{\phi}_k = \text{vec}(\mathbf{\Phi}_k)$. Each coefficient is endowed with a prior in the form of a mixture of two normal distributions, whose mixing coefficient is a latent variable $\gamma_{i,k} = \{0, 1\}$. The two normal distributions share the same mean, but have different variances. One has a large variance so that it mimics a uniform prior, whereas the other is tightly spiked around the shared mean. When $\gamma_{i,k} = 1$ (with probability $\theta_{i,k}$), the coefficient is said to be active, and governed by the normal distribution with large variance. When $\gamma_{i,k} = 0$ (with probability $\gamma_{i,k} = 1 - \theta_{i,k}$), the coefficient is said to be inactive, and thus drawn from the spiked normal. In this framework, variable selection can be performed through MCMC integration by evaluating the posterior of each $\gamma_{i,k}$. The indices i, k such that the posterior median of $\text{P}(\gamma_{i,k} = 1) > 0.5$ indicate the selected variables, that is, the relevant subset of the predictors.

For each coefficient $\phi_{i,k} \in \boldsymbol{\phi}_k = \text{vec}(\mathbf{\Phi}_k)$, we have a binary indicator $\gamma_{i,k} \in \{0, 1\}$, which encodes the state of $\phi_{i,k}$ (the "spike" inactive state for $\gamma_{i,k} = 0$ and the "slab" active state for $\gamma_{i,k} = 1$). Given $\gamma_{i,k}$, the conditional mixture prior for each $\phi_{i,k}$ can be expressed as

$$\phi_{i,k}|\gamma_{i,k} \sim (1 - \gamma_{i,k})\mathcal{N}(0, \tau_0) + \gamma_{i,k}\mathcal{N}(0, \tau_1),$$

which is controlled by the two hyperparameters τ_0 and τ_1. By selecting the two such that former approaches 0 whereas the latter is arbitrarily large, $\gamma_{i,k}$ is able to identify restrictions on $\phi_{i,k}$. The prior for the kth mode conditional mean parameters can be rewritten compactly as

$$\boldsymbol{\phi}_k|\boldsymbol{\gamma}_k \sim \mathcal{N}(\mathbf{0}, \mathcal{V}_k),$$

where $\mathcal{V}_k = \text{diag}(v_{1,k}, \ldots, v_{n_k,k})$ and $v_{i,k} = (1 - \gamma_{i,k})\tau_0 + \gamma_{i,k}\tau_1$, being n_k the cardinality of $\boldsymbol{\phi}_k$. We assume each $\gamma_{i,k}$ is independent Bernoulli, i.e.:

$$\gamma_{i,k}|\theta_k \sim \mathcal{B}\text{er}(\theta_k).$$

With a priori information on the level of sparsity in the coefficients, one can set θ_k accordingly. Therefore, we endow each indicator with a Beta-Bernoulli hierarchical prior:

$$\begin{aligned}\pi(\boldsymbol{\gamma}_k|\theta_k) &= \theta_k^{|\gamma_k|}(1 - \theta_k)^{n_k - |\gamma_k|} \\ \theta_k &\sim \mathcal{B}\text{eta}(\alpha_k, \beta_k),\end{aligned}$$

where $|\gamma_k| = \sum_i \gamma_{i,k}$.

Notice that the two covariance matrices of the MAR enter the likelihood in a multiplicative way, meaning their scales are not separately identifiable. Without imposing restrictions, such quantities would be determined completely by the prior covariance matrices. Further restrictions on the scales are therefore required without any additional a priori information.

As in Hoff (2011, 2015), we introduce dependence between the Inverse Wishart prior distribution of each Σ_k by adding a level of hierarchy through a hyperparameter ξ

$$\begin{aligned}\xi &\sim \mathcal{G}a(\eta_1, \eta_2) \\ \Sigma_k|\xi &\sim \mathcal{W}^{-1}(\xi\Omega_k, \nu_k),\end{aligned}$$

where $\mathcal{G}a(\eta_1, \eta_2)$ is the Gamma distribution with shape and scale parameters η_1, η_2 respectively. By setting $\Omega_k = \mathbf{I}_{J_k}/J_k$ and $\nu_k = J_k + 2$ the total variance is controlled only by a Kth power of ξ:

$$\mathrm{E}\Big[\prod_k \mathrm{tr}(\Sigma_k)\Big] = \xi^K.$$

Thus, if we let Δ_k be the collection of all the kth mode parameters except for $\boldsymbol{\gamma}_k$, the joint prior distribution takes the form

$$\pi(\Delta_1, \ldots, \Delta_K, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K) = \prod_k \pi(\boldsymbol{\phi}_k|\boldsymbol{\gamma}_k)\pi(\boldsymbol{\gamma}_k|\theta_k)\pi(\theta_k)\pi(\Sigma_k|\xi)\pi(\xi).$$

4 Bayesian estimation

We develop two computational methods for fitting the proposed Bayesian model (i) a Gibbs sampler and (ii) a maximum a posteriori (MAP) estimation procedure via EMVS. The Gibbs sampling algorithm will produce more accurate estimates and allow estimation of the full posterior. However, it is expected to be slower than the EMVS procedure, which aims only at identification of posterior modes by iteratively maximizing the conditional expectation of the log posterior.

An outline of the proposed Gibbs sampling procedure is given in Algorithm 1 in Appendix B along with the details on the full conditional posterior distributions. Notice that while we defer the description of how we obtain the MAP estimates, these are the ones we propose to use as starting values for the Gibbs sampler. As a matter of fact, posterior modes are often good starting values to use in MCMC simulation experiments (Geyer 2010; Jones and Qin 2022; Vats et al. 2021).

4.1 MAP estimation

A global optimization procedure to find the posterior mode can be set separately for each dimension K. However, given the mixture of prior for each $\boldsymbol{\phi}_k$, the direct optimization of the log conditional posterior $\log \pi(\Delta_k|\mathcal{Y})$ has no analytical solution. The presence of the sum prevents the logarithm from acting directly on the joint conditional posterior, which results in complicated expressions for the MAP solution.

We employ an Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin 1993), which indirectly maximizes $\log \pi(\Delta_k, \boldsymbol{\gamma}_k|\mathcal{Y})$ by iteratively maximizing its expected value under the posterior distribution of the latent variable. At iteration j, this expectation, denoted by $\mathcal{Q}(\Delta_k|\Delta_k^{[j-1]})$, is given by

$$\mathcal{Q}(\Delta_k^{[j]}|\Delta_k^{[j-1]}) = \mathrm{E}_{\boldsymbol{\gamma}_k|\Delta_k^{[j-1]}}\big[\log \pi(\Delta_k, \boldsymbol{\gamma}_k|\mathcal{Y})|\Delta_k^{[j-1]}\big] \quad (7)$$

which constitutes the E-step of the algorithm. In the M-step, we derive the revised estimate of all the other parameters by maximizing the function:

$$\Delta_k^{[j]} = \underset{\Delta_k}{\mathrm{argmax}} \ \mathcal{Q}(\Delta_k|\Delta_k^{[j-1]})$$

where $\mathcal{Q}(\Delta_k|\Delta_k^{[j-1]})$ can be decomposed as:

$$\begin{aligned}
\mathcal{Q}_{1,k}(\boldsymbol{\phi}_k^{[j]}|-) &= -\frac{1}{2}\Big[(\boldsymbol{\phi}_k - \hat{\boldsymbol{\phi}}_k)'(\tilde{\mathbf{X}}^{k'}\tilde{\mathbf{X}}^k \\
&\quad \otimes \Sigma_k^{-1})(\boldsymbol{\phi}_k - \hat{\boldsymbol{\phi}}_k) + \boldsymbol{\phi}_k' \mathcal{V}_k^{-1}\boldsymbol{\phi}_k'\Big] \\
\mathcal{Q}_{2,k}(\Sigma_k^{[j]}|-) &= -(J_k + J_{-k} + \nu_k + 1)\log|\Sigma_k| \\
&\quad - \mathrm{tr}\Big\{\Sigma_k^{-1}\Big[\xi\Omega_k + (\tilde{\mathbf{Y}}^k \\
&\quad - \boldsymbol{\Phi}_k\tilde{\mathbf{X}}^k)(\tilde{\mathbf{Y}}^k - \boldsymbol{\Phi}_k\tilde{\mathbf{X}}^k)'\Big]\Big\} \\
\mathcal{Q}_{3,k}(\theta_k^{[j]}|-) &= (|\boldsymbol{\gamma}_k| + \alpha_k - 1)\log(\theta_k) \\
&\quad + (n_k - |\boldsymbol{\gamma}_k| + \beta_k - 1)\log(1 - \theta_k) \\
\mathcal{Q}_4(\xi^{[j]}|-) &= \frac{1}{2}\sum_k J_k\nu_k \log(\xi) \\
&\quad - \frac{1}{2}\sum_k \mathrm{tr}\,(\Omega_k\Sigma_k^{-1})\xi + (\eta_1 - 1) \\
&\quad \log(\xi) - \eta_2\xi
\end{aligned}$$

Notice that there is an identifiability issue arising from the structure of the MAR. Given the properties of the Kronecker product, if $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_K$ are a solution of the problem, so are $f_1\boldsymbol{\Phi}_1, \ldots, f_K\boldsymbol{\Phi}_K$, with the condition $\prod_k f_k = 1$. To keep iterations of both the Gibbs and the EMVS stable, we choose f_k such that the magnitude between the various parameter matrices remains as stable as possible in the following way (see Hoff 2015):

$$f_k = \frac{\prod_{l\neq k}||\boldsymbol{\Phi}_l||^{\frac{1}{K}}}{||\boldsymbol{\Phi}_k||^{\frac{K-1}{K}}} \quad (8)$$

The same applies for the covariance matrices. We illustrate the complete EMVS estimation procedure in Algorithm 2 in Appendix C.

5 Kronecker GFEVD

Dynamic analysis via Generalized Impulse Response Functions (GIRF) and Generalized Forecast Error Variance Decomposition (GFEVD) can be carried out easily on a variable-by-variable basis. In this context, a modification of the GFEVD (Lanne and Nyberg 2016) obtained from a GIRF (Koop et al. 1996) is appealing, as it has the desired

property of unit row sum. Let $\Lambda = \text{diag}(\Sigma^{-1})$, where $\Sigma = \sigma_3 \otimes \Sigma_2 \otimes \Sigma_1$. The GFEVD for the vectorized MAR*(P) has the standard form

$$\theta_H^{ij} = \frac{\sum_{h=0}^{H} \mathbf{e}_i' \boldsymbol{\Psi}_h \Sigma \Lambda_{jj}^{1/2} \mathbf{e}_j}{\sum_{h=0}^{H} \mathbf{e}_j' \boldsymbol{\Psi}_h \Sigma \Lambda \Sigma \boldsymbol{\Psi}_h' \mathbf{e}_j}, \tag{9}$$

where $\mathbf{e}_i$ is a selection vector. The collection of $\theta_H^{ij} \in \boldsymbol{\Theta}_H$ is referred to the GFEVD matrix.

Although the MAR estimates interpretable lower dimensional coefficients via the Tucker product, dynamic analysis can be performed through the vectorized form of the model (Chen and Yang 2021). Nevertheless, this leads to a significant loss of information on mode-specific interactions that is masked via the computation of the Kronecker product of the objects of interest. Rather than studying solely variable-by-variable impacts, it is convenient to exploit the enhanced interpretability given by the bi-dimensional structure of the model and decompose the GFEVD into two lower-dimensional matrices representing the K-mode specific GFEVD. Among all possible decompositions of a GFEVD matrix into its mode-specific counterparts, the most reasonable approach given the form of the model is via a Kronecker decomposition

$$\boldsymbol{\Theta}_H^C \otimes \boldsymbol{\Theta}_H^I \approx \boldsymbol{\Theta}_H,$$

where $\boldsymbol{\Theta}_H^I \in \mathbb{R}^{G \times G}$ and $\boldsymbol{\Theta}_H^C \in \mathbb{R}^{N \times N}$ will be, respectively, the country and indicator GFEVD.

We now describe the Kronecker decomposition problem for the GFEVD derived from the MAR model. Recall that, for each forecast horizon, $\boldsymbol{\Theta}_H$ is a stochastic matrix, having $\boldsymbol{\Theta}_H \geq \mathbf{0}$ and $\boldsymbol{\Theta}_H \mathbf{1}_{GN} = \mathbf{1}_{GN}$, where $\mathbf{1}_J$ is a J-dimensional vector of ones. A simple approach to decompose $\boldsymbol{\Theta}_H$ into $\boldsymbol{\Theta}_H^C \otimes \boldsymbol{\Theta}_H^I$, so to reflect the indicator and country GFEVD structure is to project $\boldsymbol{\Theta}_H$ onto the space of Kronecker products under the squared Frobenius norm

$$\min_{\boldsymbol{\Theta}_H^V, \boldsymbol{\Theta}_H^C} \quad ||\boldsymbol{\Theta}_H - \boldsymbol{\Theta}_H^C \otimes \boldsymbol{\Theta}_H^I||_F^2,$$

which represents a Nearest Kronecker Product (NKP) problem in matrix computation (Van Loan and Pitsianis 1993; Van Loan 2000). This approach is also at the basis of the projection method for MAR estimation (Chen and Yang 2021), which can be used to find the starting values of the ML procedure given a VAR estimate.

However, the two resulting matrices minimizing this problem are not guaranteed to be stochastic as well, a necessary condition to constitute GFEVD. Thus, in order to get the best stochastic Kronecker Product (SKP) approximation $\boldsymbol{\Theta}_H^C$ and $\boldsymbol{\Theta}_H^I$, the following constrained nonlinear least squares must be solved:

$$\begin{aligned} \min_{\boldsymbol{\Theta}_H^I, \boldsymbol{\Theta}_H^C} \quad & ||\boldsymbol{\Theta}_H - \boldsymbol{\Theta}_H^C \otimes \boldsymbol{\Theta}_H^I||_F^2 \\ \text{s.t.} \quad & \boldsymbol{\Theta}_H^I \geq \mathbf{0}, \quad \boldsymbol{\Theta}_H^I \mathbf{1}_G = \mathbf{1}_G \\ & \boldsymbol{\Theta}_H^C \geq \mathbf{0}, \quad \boldsymbol{\Theta}_H^C \mathbf{1}_N = \mathbf{1}_N. \end{aligned} \tag{10}$$

Notice that all the entries in $\boldsymbol{\Theta}_H^C \otimes \boldsymbol{\Theta}_H^I$ are the same as all the entries in $\boldsymbol{\theta}_H^I \boldsymbol{\theta}_H^{C'}$, where $\boldsymbol{\theta}_H^I = \text{vec}(\boldsymbol{\Theta}_H^I)$ and $\boldsymbol{\theta}_H^C = \text{vec}(\boldsymbol{\Theta}_H^C)$, i.e. the matrices have the same set of elements, which only differ in their positions. By employing a rearrangement operator $\mathcal{G}(\cdot)$ such that $\mathcal{G}(\mathbf{X} \otimes \mathbf{Y}) = \text{vec}(\mathbf{Y})\text{vec}(\mathbf{X})'$, we can rewrite Eq. (10) as:

$$\begin{aligned} \min_{\boldsymbol{\theta}_H^I, \boldsymbol{\theta}_H^C} \quad & ||\mathcal{G}(\boldsymbol{\Theta}_H) - \boldsymbol{\theta}_H^I \boldsymbol{\theta}_H^{C'}||_F^2 \\ \text{s.t.} \quad & \boldsymbol{\theta}_H^I \geq \mathbf{0}, \quad \mathbf{R}_I \boldsymbol{\theta}_H^I = \mathbf{1}_G \\ & \boldsymbol{\theta}_H^C \geq \mathbf{0}, \quad \mathbf{R}_C \boldsymbol{\theta}_H^C = \mathbf{1}_N \end{aligned} \tag{11}$$

where $\mathbf{R}_I = [\mathbf{I}_G, \ldots, \mathbf{I}_G]$ and $\mathbf{R}_C = [\mathbf{I}_N, \ldots, \mathbf{I}_N]$ are linear equality constraint matrices of dimension $[G \times G^2]$ and $[N \times N^2]$. As it is expressed, the problem in Eq. (11) can be solved for $\boldsymbol{\theta}_H^I$ and $\boldsymbol{\theta}_H^C$ iteratively via standard constrained minimization routines.

6 Simulations

We design two simulation experiments. The first one is aimed at studying the small sample efficiency of the proposed method. The second one evaluates the estimation error, forecasting performance, and computational time of our proposed approach, relative to multiple existing estimation methods for longitudinal data. We further perform a comparative convergence analysis of the MCMC in Sect. 6.3.

We set our hyperparameters as follows:

- $\tau_0 = 0.01$, $\tau_1 = 4$. This choice is motivated by the fact that all the parameters we consider in our specifications are strictly lower than 1 in terms of magnitude. Thus, it is reasonable to consider a variance of 0.01 as "small", and a variance of 4 as "big".
- $\Omega_k = \mathbf{I}_{J_k}/J_k$, $\nu_k = J_k + 2$. This is standard in the VAR literature (Bańbura et al. 2010), being a combination that ensures minimal assumptions, but at the same time the existence of the expected value of the Inverse Wishart.
- $\alpha_k = 1, \beta_k = J_k - 1, \nu_1 = 1$. This choice has been recommended to obtain optimal posterior concentration rates in sparse settings (Ročková and George 2014; Castillo and van der Vaart 2012).

In the second experiment, we run all the Gibbs samplers for 1000 iterations. Given a confidence level of 0.05, this number is bigger than the minimum effective sample size (mESS) at a tolerance of 0.05, in the largest scenario, i.e. the one where $(G, N) = (8, 10)$ (Vats et al. 2019). Details on the convergence of the sampler of our proposed model can be found in Appendix D.4.

6.1 Small sample efficiency

We compare the small sample efficiency of ML and MAP estimators by letting the length of the time series T and the level of sparsity in the autoregressive coefficients vary. The main purpose of this experiment is to obtain qualitative understanding of the small sample covariances of the two estimators under different sparsity settings. We compare the MAP against the ML estimator for two main reasons, other than its computational convenience. Firstly, given that the MAP can be also seen as s frequentist regularized procedure, it is allegedly a fair competitor to the ML estimator. Secondly, in the comparison we focus on point estimates. In this setting, MAP and median or mean of the posterior distribution of the Gibbs sampler would give pretty similar results.

To this aim, we simulate our synthetic data as follows. We generate matrix-valued time series observations from a MAR(1) with dimensions $G, N = 8$, with different lengths of the time series $T = 50, \ldots, 1000$. We place true nonzero loadings in the left and right model coefficient matrices according to four different settings. For each setting $i = 1, \ldots, 4$, we place SP_i non-zero coefficients for each row of $\mathbf{A}$ and $\mathbf{B}$, respectively. In particular, we consider $SP_1 = 1$, $SP_2 = 2$, $SP_3 = 4$, and $SP_4 = 8$. Our data generating process (DGP) is such that the main diagonal blocks of $\mathbf{A}$ and $\mathbf{B}$ are $[SP_i \times SP_i]$ matrices whose elements are drawn from a $\mathcal{N}(0, 1)$, and zero otherwise. Notice that when $i = 1$ and $i = 4$ the two coefficient matrices are diagonal (1 out of 8 nonzero coefficient in each row) and full (8 out of 8 nonzero coefficient in each row), respectively. In the two intermediate cases, we have, respectively, 2 out of 8 (for $i = 2$) and 4 out of 8 (for $i = 3$) non-zero coefficient in each town of the coefficient matrices. The covariance matrices are set to $\Sigma_1 = \mathbf{I}_G$ and $\Sigma_2 = \mathbf{I}_N$. In setting our priors, we fix $\alpha_k = 1$, and choose $\beta_k = J_k - 1$ so to reflect a prior belief of a sparse DGP as the one in setting 1. Estimation errors are measured by the mean squared error (MSE):

$$MSE(\hat{\mathbf{\Phi}}) = \frac{\text{tr}\left[(\mathbf{\Phi} - \hat{\mathbf{\Phi}})'(\mathbf{\Phi} - \hat{\mathbf{\Phi}})\right]}{G \times N} \tag{12}$$

where for the MAR model $\mathbf{\Phi} = \mathbf{B} \otimes \mathbf{A}$ and $\hat{\mathbf{\Phi}} = \hat{\mathbf{B}} \otimes \hat{\mathbf{A}}$.

Figure 3 illustrates a comparison of the small sample efficiencies of the ML and MAP estimators, as measured by the average estimation errors over 50 repetitions of $MSE(\hat{\mathbf{\Phi}})$. The figure shows a decreasing trend of the average error of both estimators as T grows. As expected, given the imposed prior beliefs, the more sparse the DGP, the more the MAP estimator results in a higher efficiency in small samples. This is magnified in setting 1 (DGP with diagonal coefficient matrices), while the difference in efficiency between the two estimators gradually vanishes as the number of nonzero coefficients grow. Notice however that, even in setting 4 (DGP with full coefficient matrices) the efficiency of the two estimators is still comparable.

The efficiency of the MAP estimator depicted in Fig. 3 proves the capacity of our model to retrieve the true DGP across different sparsity settings. The gain is evident in small T and sparse scenarios, where the model outperforms the plain vanilla ML, bearing in mind that the latter is a special case of our model with an apriori fixed $\gamma_{i,k} = 1$. This, in turn, implies that the superior performance depends on the capacity of the model to shrink towards the true zero coefficients, as expected. Another way to look at this evidence is via θ_1 and θ_2, whose posterior distributions provide information on the level of sparsity predicted by the model, as a function of the prior and the data.

As for the efficiency factor, we now represent in Fig. 4 the posterior median of θ_1 and θ_2, under four different settings. Results are averaged over the 50 replications. The black lines represent the true proportion of nonzero coefficients in $\mathbf{A}$ and $\mathbf{B}$, whereas the blue and red ones are the median of the posterior distribution of θ_1 and θ_2, respectively. In all of the four cases, the two coefficients are increasing for the lowest values of T, and then tend to stabilize. This is in line with the fact that with small T, the noise is high, and thus the signal-to-noise ratio drops, inducing the model to consider as non-significant more coefficients. What emerges is that two coefficients lie above the true value in the first two settings. However, such deviation is relatively small in terms of magnitude, being lower than 5%. The third setting is the best captured by the model, which very well replicates the features of the DGP. In the last case, all the coefficients are nonzero, and the model slightly underestimates the real sparsity pattern. However, this is not a matter of concern.

A possible explanation is that the higher the number of nonzero parameters, the lower is their overall magnitude, imposed to ensure stationarity. Thus, it is reasonable that the model might underestimate the true sparsity pattern in this fully dense specification. It might tend to assign zeros where a real 0 is not present, but the overall magnitude of these coefficients is still minimal. This is confirmed by the forecast performance presented above, which is in line with the ML. Recall that the more rigid ML always implies $\theta_1 = \theta_2 = 1$, even in highly sparse settings. Our model scales pretty well with different levels of sparsity, thus emerging as a valid alternative model to large MAR.

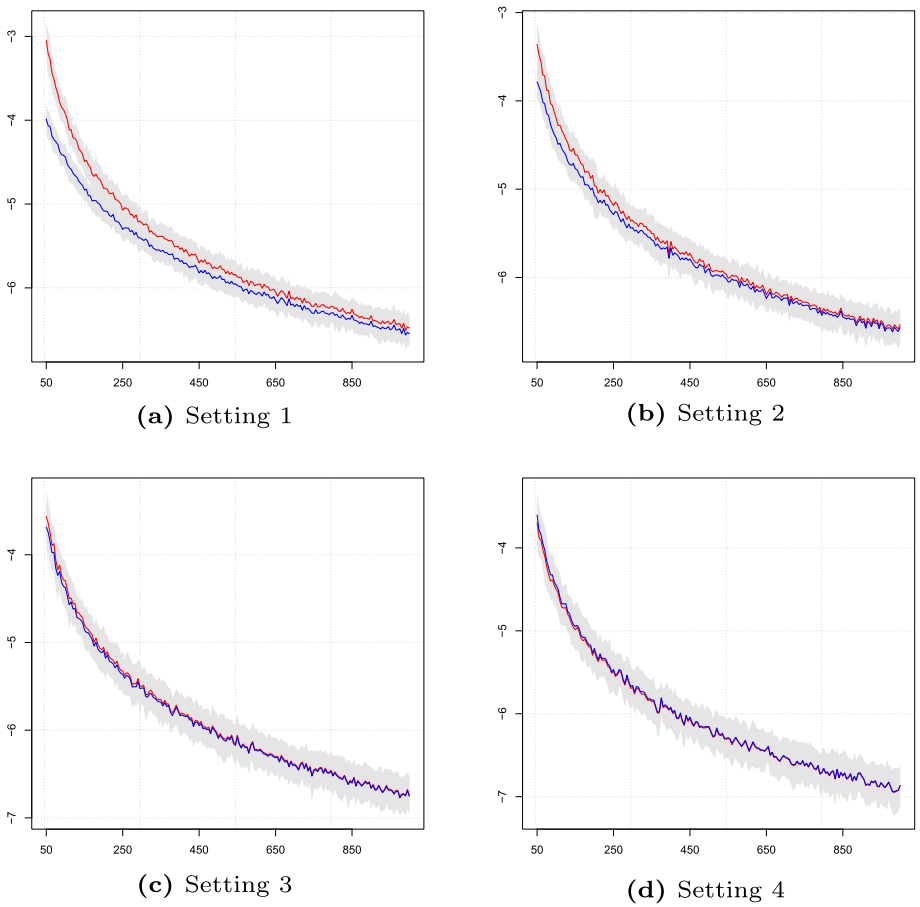


Fig. 3 Comparison of the efficiencies of ML (red) and MAP (blue) estimators over 50 repetitions for $T = 50, \ldots, 1000$ under four sparsity settings. The figure shows the logarithm of the $MSFE(\hat{\mathbf{\Phi}})$ (y axis) over different T (x axis). Gray shaded areas represent the 1 standard deviation confidence bounds

6.2 Comparative estimation, forecasting and computational performances

We aim to compare the estimation error, forecasting performance, and computation time of the illustrated estimators with other relevant competing alternatives for panel data. This is done for different choices of the matrix dimensions $(G, N) = (2, 3), (4, 6), (8, 10)$, so setup a "small", a "medium" and a "large" setting, relative to a reference sample size $T = 100$. We consider the three MAR estimators: ML, MCMC (Bayes), and MAP. The alternatives considered are the stacked VAR estimator (VAR), the country-block panel VAR (CB), the cross-sectional Shrinkage (CC) approach of Canova and Ciccarelli (2009, 2013), and the Stochastic Search Specification Selection (SSSS) of Koop and Korobilis (2016). In particular, the stacked VAR and the country-block PVAR are estimated with frequentist techniques, whereas the others with a Bayesian approach. We briefly describe the competing alternative models in Appendix D.1 along with the hyperparameter specification for the Bayesian ones in Appendix D.2.

We simulate a sparse VAR(1) able to reflect recurrent patterns in multi-country and multi-variable applications. Let Ξ be the $[GN \times GN]$ matrix of autoregressive coefficients and $\Xi_{i,j} \in \Xi$ be the $[G \times G]$ country j block of parameters in the country i equations, and consider its entries related to the indicators k, l:

$$\Xi_{i,j}^{k,l} = \begin{cases} \lambda \lambda_n^{|i-j|} \lambda_g^{|k-l|} & \text{if } |i-j| \leq r, |k-l| \leq r \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda = 0.95$ is an overall constant term, λ_n and λ_g are a country and an indicator penalty term, respectively, such that $\lambda_n = 0.5 + \mathcal{U}_{[-0.1,0.1]}$ and $\lambda_g = 0.5 + \mathcal{U}_{[-0.1,0.1]}$, and $r = 2$. The DGP is able to reflect heterogeneities in both dimensions such that coefficients are affected by a country penalty λ_n, an indicator penalty λ_g and both of them combined. Such penalties act when the row and column distances $|i-j|$ and $|k-l|$ of the elements of the matrix $\Xi_{i,j}^{k,l}$ do not exceed the threshold level r.

Estimation errors of the models are measured by the MSE as in Eq. (12), whereas the forecasting performance for any fixed forecast horizon H is assessed by means of the Mean Squared Forecast Error (MSFE):

$$MSFE(H) = \frac{1}{G \times N \times H}$$

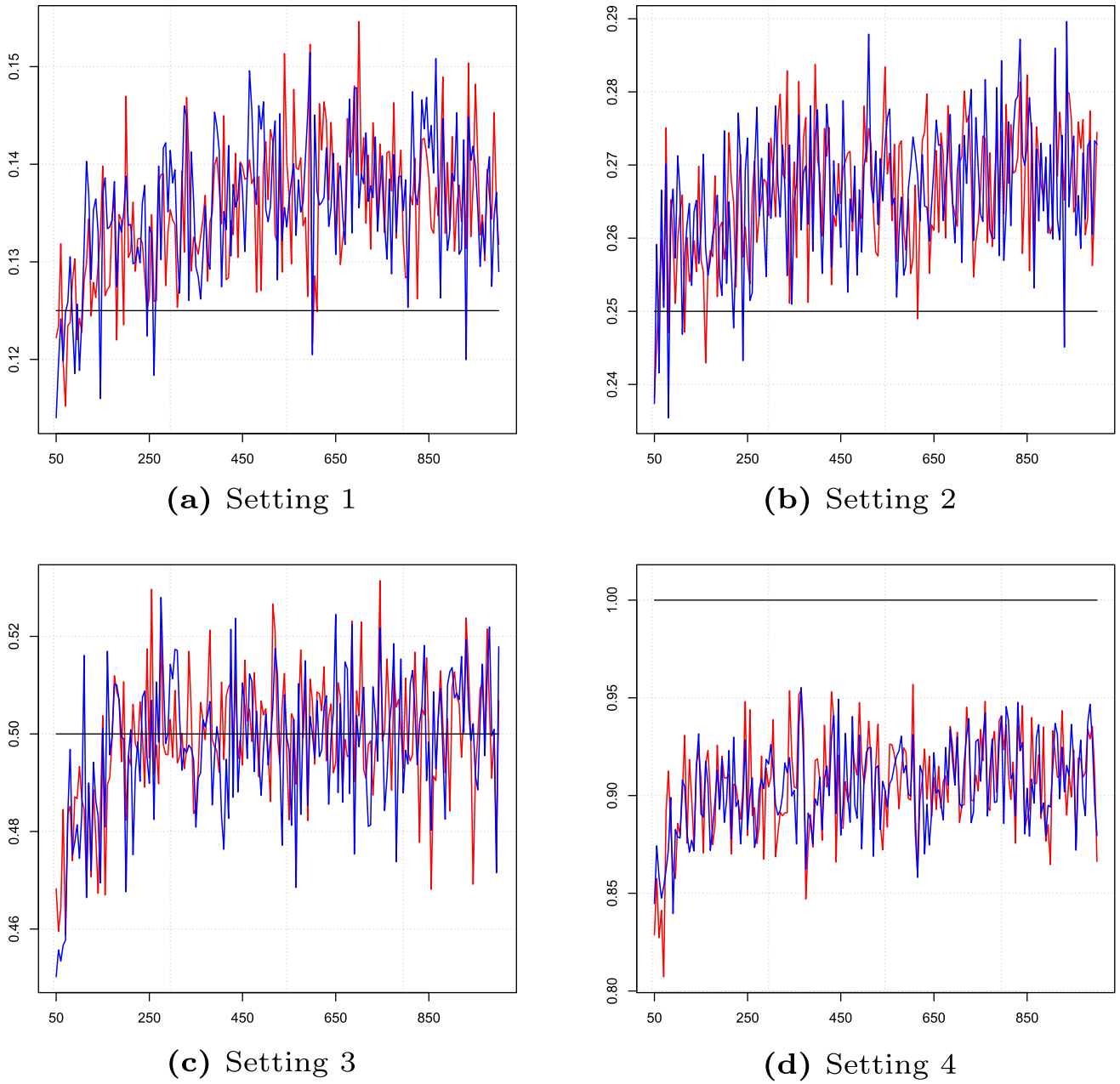


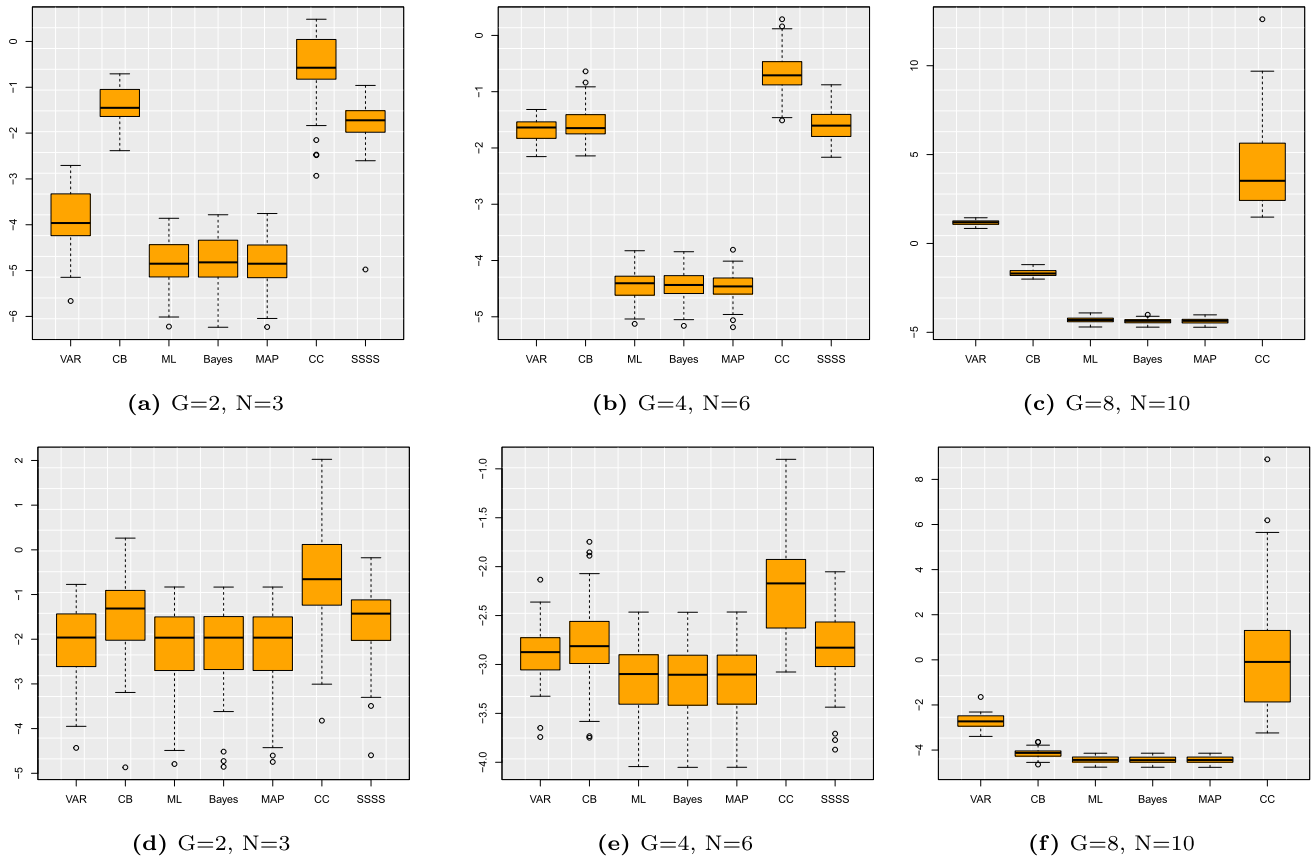
Fig. 4 The figure shows the median posterior of θ_1 (red lines), θ_2 (blue lines), and the true proportion of non-zero coefficients (black lines) over 50 repetitions for $T = 50, \ldots, 1000$ (*x axis*) under four sparsity settings

$$\sum_{h=1}^{H} (\hat{\mathbf{y}}_{T+h|T} - \mathbf{y}_{T+h})' (\hat{\mathbf{y}}_{T+h|T} - \mathbf{y}_{T+h}),$$

where $\hat{\mathbf{y}}_{T+h|T}$ is the H-step forecast obtained with information up to the last sample size T.

We illustrate in Fig. 5 numerical results on the model MSE and MSFE related to $H = 1$. When the estimation of some competitor model was infeasible, corresponding results are omitted. We report results related to repeated experiments with $T = 500$ and a sparsity setting generated by $r = 3$ in Appendix D.3.

The proposed estimators, along with ML, tend to overperform all competing alternatives, both in terms of estimation and forecasting performances. The difference in performances becomes more substantial when considering the "large" dimensional setting. Results with $T = 500$ and $r = 3$ are consistent with the expectations that as T grows and true coefficient matrices are less sparse, performance gains com-



(a) G=2, N=3 (b) G=4, N=6 (c) G=8, N=10

(d) G=2, N=3 (e) G=4, N=6 (f) G=8, N=10

Fig. 5 Average estimation error (MSE) (**a**–**c**) and 1 step ahead forecasting peformance (MSFE) (**d**–**f**) over 50 repetitions of each model for $(G, N) = (2, 3), (4, 6), (8, 10)$ with $T = 100$

pared to a standard VAR diminish in the "small" dimensional setting.

We now report the average computational time over 50 repetitions for the three different combination of G and N and $T = 100$ in Table 1. Being the estimators of VAR and CB closed form, their related execution times are naturally the lowest.

As expected, the computational times of MAP are much more favorable than those of the MCMC. The difference between the two lies in the fact that while the MCMC requires thousands of iterations to approximate the joint posterior distribution, the EMVS algorithm reaches convergence fairly quickly, usually in fewer than 100 iterations. Despite this difference, the two have exhibited comparable performances in estimating the autoregressive coefficient and forecasting performances.

In our settings, the MCMC is also generally slower than CC, which exploits factors to reduce dimensionality, but is much faster than SSSS in "medium" and "large" settings. Further, given the dimensionality reduction achieved by the MAR, computation times of the MCMC grow slower with increasing model parameters than those related to the CC and SSSS estimators. Notice that the computational times of MAP are distinctly lower, not only than those of its full Bayesian counterpart, but also than those related to the CC and SSSS estimators. Notwithstanding this, estimation and forecasting performances of MAP (and, in general, of the two proposed computational methods) are generally superior in high-dimensional settings.

6.3 MCMC output analysis

Here we assess and compare the competing models' MCMC algorithms based on effective sample size criteria (Vats et al. 2019; Gong and Flegal 2016). This criteria allows us to determine when should sampling stop in order to get reliable parameter estimates. Given a level of confidence α and tolerance ϵ, one simple strategy is to run the Markov Chain at least for a number of iterations larger than the minimum Effective Sample Size (mESS) (Vats et al. 2019). The multivariate mESS satisfies

$$mESS \geq \frac{2^{2/p}\pi}{[p\Gamma(p/2)]^{2/p}} \frac{\chi^2_{1-\alpha,p}}{\epsilon^2}, \tag{13}$$

Table 1 Average computational times (in seconds) over 50 repetitions for $(G, N) = (2, 3)$, $(4, 6)$ and $(8, 10)$, with $T = 100$

	G=2, N=3	G=4, N=6	G=8, N=10
VAR	8.7×10^{-5} (1.2×10^{-4})	1.1×10^{-4} (3.8×10^{-4})	4.7×10^{-4} (2.1×10^{-4})
CB	8.2×10^{-5} (1.5×10^{-4})	1.6×10^{-4} (7.3×10^{-4})	3.1×10^{-4} (2.1×10^{-4})
ML	4.6×10^{-4} (8.7×10^{-3})	0.01 (4.5×10^{-3})	0.01 (0.01)
Bayes	9.98 (0.51)	30.9 (10.1)	85.72 (23.24)
MAP	0.01 (7.2×10^{-3})	0.03 (0.02)	0.07 (0.02)
CC	0.26 (0.02)	1.82 (0.55)	25.67 (3.78)
SSSS	5.18 (0.24)	368.52 (76.58)	–

Standard deviations are shown in parentheses

where p is the number of parameters to be estimated. It should be noticed that $mESS$ is a function of p, α and ϵ and is therefore independent of the Markov chain or the underneath process. This paves the way for model comparison in terms of this quantity.

Notice that the three Gibbs samplers employed are characterized by a different number of parameters p. In particular, for each combination of $[G, N]$ of the simulation study, the Gibbs sampler for the MAR produces, respectively, two chains of dimensions $[G^2 \times MC]$ (MAR(B1)) and $[N^2 \times MC]$ (MAR(B2)), where MC is the number of clean post burn-in draws. The samplers for the CC and the SSSS procedures produce two chains of dimensions $[GNF \times MC]$ (with $F < GN$) and $[G^2N^2 \times MC]$. Moreover, the CC is a factor model, for which the number of factors can differ among datasets, and is sensitive to the statistical method employed to extract the relevant factors. As a consequence, it is not possible to choose a common number of simulations such that the tolerance ϵ is kept constant among all settings and models.

However, a possible strategy to overcome this issue is to use Eq. (13) in another way. Instead of choosing a different value of MC for each model and setting, one can set an arbitrarily large level so that it is guaranteed that in all settings at least a minimum pre-determined ϵ is reached, and then compare the samplers in terms of tolerance levels. We can determine a model specific estimated ESS, calculated after running the MCMC, and then get the corresponding tolerance level ϵ via Eq. (13). Given MC iterations in a Markov chain, the ESS measures the size of an i.i.d. sample with the same standard error. In a multivariate setting, the ESS is given by

$$ESS = MC\left(\frac{|\Lambda|}{|\Sigma|}\right)^{1/p},$$

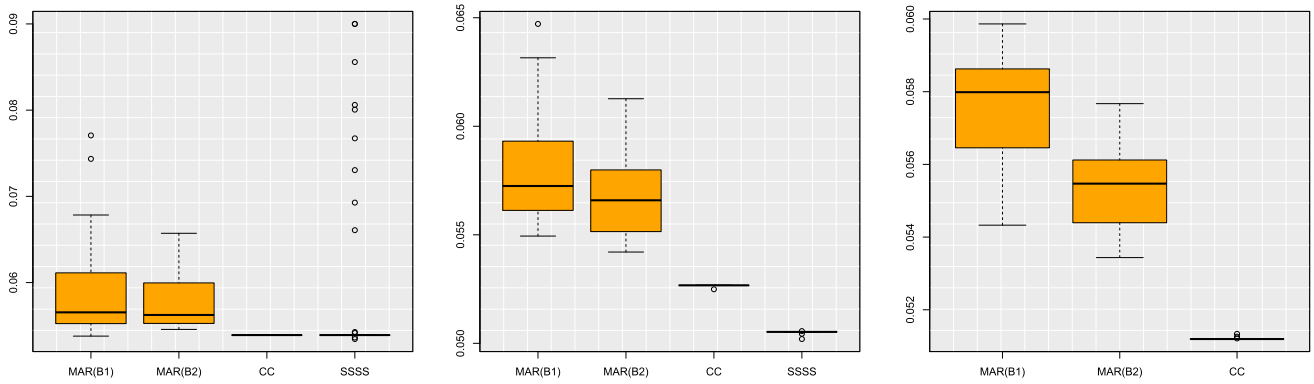
where Λ is the sample covariance matrix and Σ is an estimate of the variance of the asymptotic normal distribution. Replacing $mESS$ in Eq. (13) with ESS, we can express the former in terms of the tolerance ϵ, which can be viewed as a comparison in terms of convergence of the Gibbs sampler. The smaller the minimum effective samples, the larger the tolerance, and hence the smaller the number of simulations required.

A visual comparison of the tolerance level ϵ for the three models under the different setting is depicted in Fig. 6. The figure shows that our sampler achieves generally greater tolerance than the analyzed competing alternatives, particularly in high-dimensional settings.

7 Application: panel of country economic indicators

We now conduct an empirical application of the proposed model to a panel of $G = 9$ world countries, which currently represent approximately the 64% of the world total Gross Domestic Product (GDP): Canada (CA), China (CH), France (FR), Germany (GE), India (IN), Italy (IT), Japan (JP), United Kingdom (UK) and United States (US). We consider quarterly observations of $N = 3$ economic indicators: GDP, Consumer Price Index (CPI) and Short-term Interest Rates (S-IR), all expressed in log differences. The sample ranges from 1980Q1 to 2019Q4 ($T = 162$). The data is inherently multidimensional, as observations are generated in matrix form, where rows represent indicators and columns countries.

We estimate a fully Bayesian MAR*(3), which yields the full conditional posterior distribution of the parameters of interest. It is a standard procedure to fix an arbitrarily large lag length and then let the algorithm shrink the coefficients (George et al. 2008; Bańbura et al. 2010). By running the model with with several lag specifications (from 1 to 10 lags), we observed a progressively diminishing impact of lagged variables. In particular, three is the shortest number of lags such that the 95% credible interval for the lowest magnitude coefficient ($\hat{c}_3$ in our case) does not contain zero. Figure 7 displays the posterior distribution of the lag order coefficients $\hat{c}$ and shows a diminishing pattern of lag impacts over time, albeit still preserving a non negligible impact even for lag three. It is clear how values of the posterior distribution of $\hat{c}$ decrease with the lag order itself, as one could expect. In



(a) G=2, N=3 **(b)** G=4, N=6 **(c)** G=8, N=10

Fig. 6 Tolerance ϵ over 50 repetitions of each model for $(G, N) = (2, 3), (4, 6), (8, 10)$ with $T = 100$

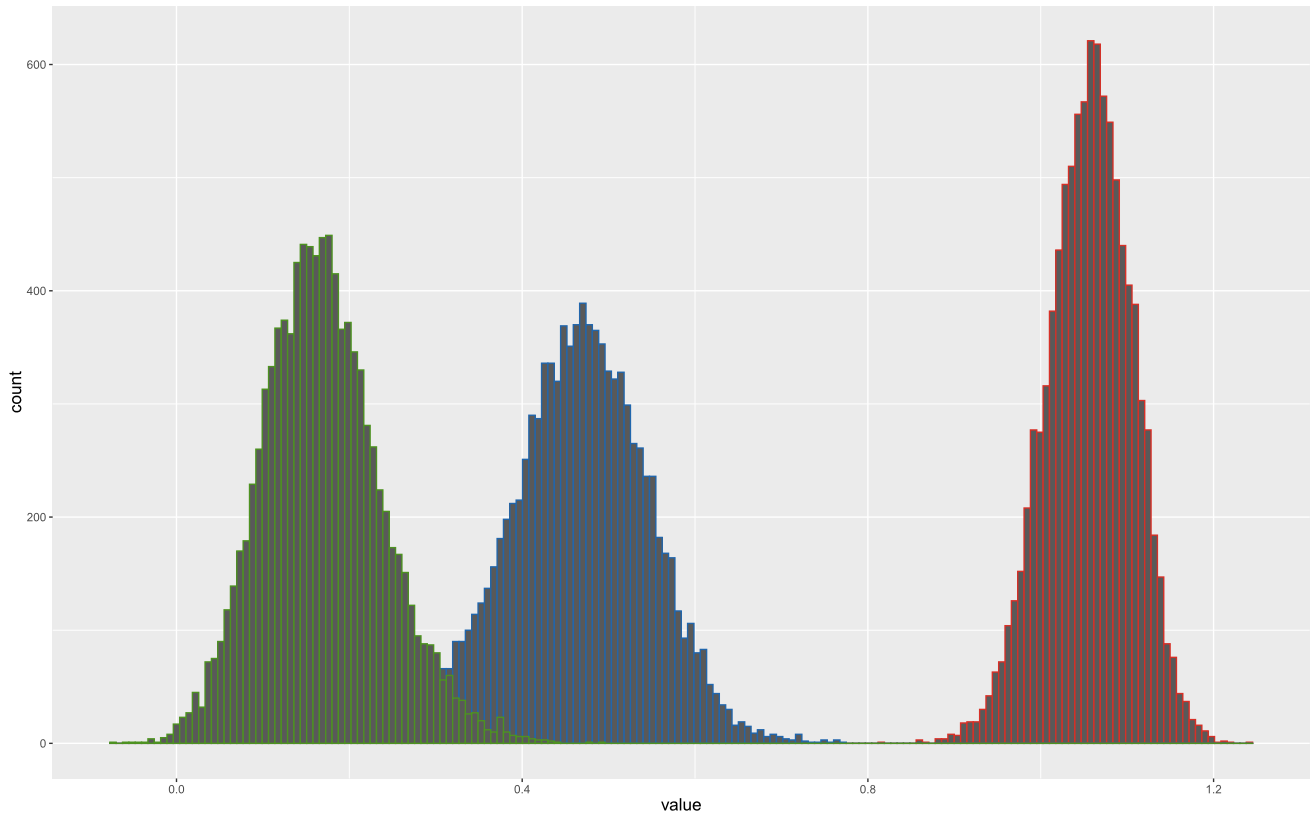
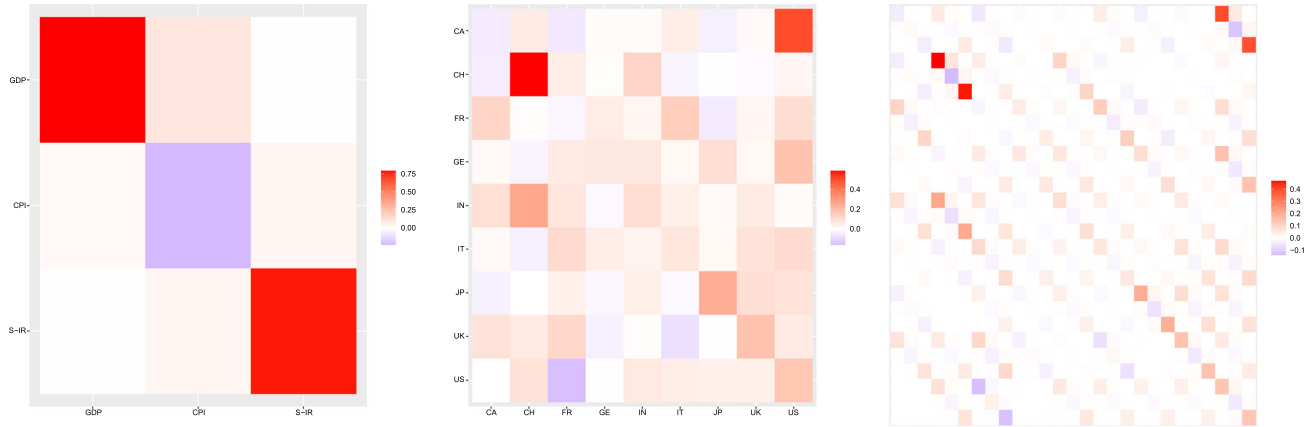


Fig. 7 Posterior distribution of $\hat{c}_1$ (red), $\hat{c}_2$ (blue), $\hat{c}_3$ (green)

Fig. 17 in Appendix D.4 we illustrate the posterior distribution of the mixing proportion parameters $\hat{\theta}$.

We illustrate in Fig. 8 the posterior median of the left ($\hat{\mathbf{A}}$) and right ($\hat{\mathbf{B}}$) coefficient matrices, along with the reconstructed coefficient matrix $\hat{\mathbf{B}} \otimes \hat{\mathbf{A}}$. Note that, given that only the Kronecker product $\hat{\mathbf{B}} \otimes \hat{\mathbf{A}}$ is uniquely identified, only magnitudes related to the left and right coefficient matrices can be meaningfully interpreted, rather than signs.

The figure shows that, as intuition suggests, the diagonal elements of the parameter matrices concur to a large portion of the system's autoregressive dynamics. This is more evident from the left coefficient matrix, indicating a strong autocorrelation in the variable dimension. From the first order right coefficient matrix estimate notice that the two largest impacts are those of China on itself and US on Canada. The former can be explained through the relatively low impact of other countries on the Chinese economy as a whole and the latter seems reasonable given the geographical proximity and large trade activity between the two.



(a) $\hat{\mathbf{A}}$ (b) $\hat{\mathbf{B}}$ (c) $\hat{\mathbf{B}} \otimes \hat{\mathbf{A}}$

Fig. 8 Median of the posterior entries of the first order left coefficient matrix $\hat{\mathbf{A}}$ (**a**), of the right one $\hat{\mathbf{B}}$ (**b**), and of $\hat{\mathbf{B}} \otimes \hat{\mathbf{A}}$ (**c**)

We evaluate out-of-sample rolling forecasting performances and computational times of the proposed methods against the same competing models analyzed in Sect. 6.2. In particular, starting from 1995Q1 (T = 60) to 2019Q2 (T = 120), we fit the corresponding models by means of all available data at time $t-1$ and compute the one step ahead MSFE. Results are summarized in Table 2. Further results related to the full set of estimations as well as the mixing of the Gibbs sampler of our proposed model are in Appendix D.4.

Notice that, on average, the MCMC outperforms all competing alternatives in the forecasting exercise, closely followed by the MAP and ML estimators. We then find the SSSS, CB and CC estimators which perform better than the stacked VAR. Despite its superior forecast accuracy, the MCMC method is more than 140 times slower than the EMVS. This might render the latter preferable in high-dimensional empirical applications, given the relatively little differences between the two in terms of performance.

A key advantage of the EMVS procedure relative to standard MCMC is that its lower computational intensiveness allows for dynamic posterior exploration (Ročková and George 2018). This consists of holding fixed at a high value the slab hyperparameter, while letting the spike hyperparameter gradually rise along a ladder of increasing values. In our case, dynamic posterior exploration can also be conducted on the "sparsity" parameters θ_k.

We therefore perform cross validation on a grid of ten values between 0.005 and 0.05 with a step of 0.005 for the spike parameter τ_0. Additionally, we let $\beta_k = J_k^{\zeta}$ vary on a grid of ten values of ζ from 0.1 to 1 with step 0.1. By performing dynamic posterior exploration, the average MSFE of the MAP estimator drops to 2.89×10^{-5} (std: 1.69×10^{-5}), while preserving a reasonable amount of computation time of 16.81 s (std: 5.86).

We now illustrate the resulting Kronecker decomposition of the estimated GFEVD in Fig. 9. We refer the reader to Appendix D.4 for a graphical representation of the spectrum of the companion matrix. For what concerns $\hat{\mathbf{\Theta}}_H^I$, a contribution of CPI and GDP to the GFEVD of S-IR is detected. Moreover, the figure shows that a portion of GFEVD in GDP is due to shocks in CPI and S-IR. Notice, however, that overall magnitudes of such cross-variance shares are small if compared to the country and full GFEVD matrices, meaning a weak dependence structure within the indicator dimension.

Conversely, we find stronger cross-variance shares in the country dimension, as reflected by $\hat{\mathbf{\Theta}}_H^C$. The largest pairwise contributions are those to Canada arising from shocks in the US economy, and vice versa, though with lower magnitude in the latter case. While both China and US are prone to transmit large portions of GFEVD to the rest of the countries, the former is generally more resilient to shocks in other countries. This with the exception of India, which is one of the largest exporters of China. Results also highlight noticeable cross variance shares across the EU countries, i.e. France, Germany and Italy. The full GFEVD matrix $\hat{\mathbf{\Theta}}_H$ reflects instead variable by variable interactions, which seem consistent with the ones obtained through the Kronecker decomposition problem.

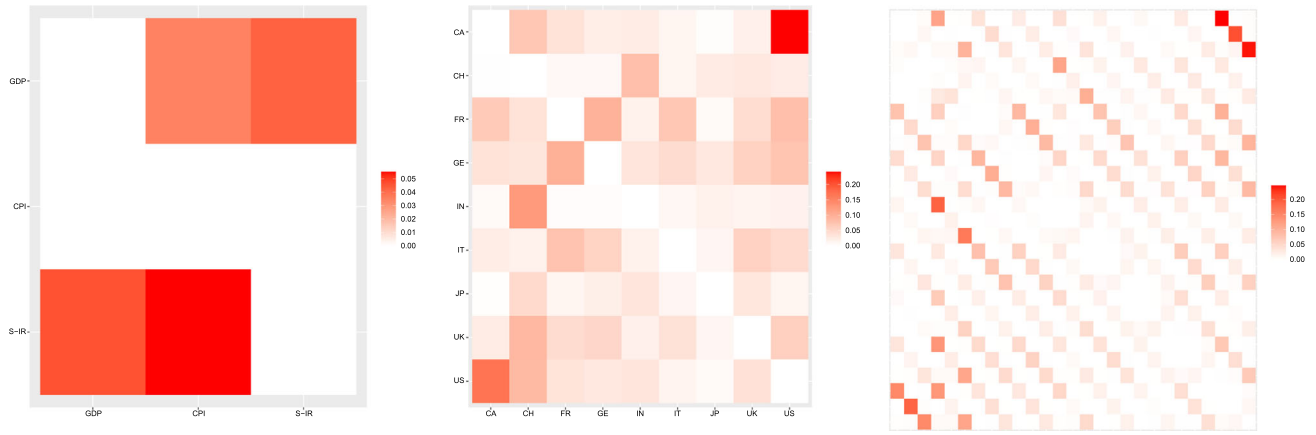
8 Concluding remarks

We developed a Bayesian method for variable selection in high-dimensional matrix autoregressive models which reflects and exploits the original matrix structure of data to reduce dimensionality and foster interpretability of multidimensional dependency structures. We firstly derived a compact form of the model stemming from the tensor linear regression framework, which facilitates the model estima-

Table 2 Average Logarithm of the MSFE and average computational time (in seconds) over 50 repeated estimations of three different MAR estimators (ML, Bayes, MAP) against competing alternatives

	Av. log MSFE	Av. computational time
VAR	4.41×10^{-5} (2.77×10^{-5})	1.91×10^{-4} (9.97×10^{-5})
CB	3.84×10^{-5} (1.96×10^{-5})	1.72×10^{-4} (7.44×10^{-5})
ML	3.21×10^{-5} (1.85×10^{-5})	1.83×10^{-2} (7.00×10^{-3})
Bayes	3.09×10^{-5} (1.79×10^{-5})	25.88 (5.10)
MAP	3.11×10^{-5} (1.79×10^{-5})	0.17 (0.07)
CC	4.40×10^{-5} (2.85×10^{-5})	1.9 (0.59)
SSSS	3.83×10^{-5} (1.75×10^{-5})	776.06 (51.82)

Standard deviations are shown in parentheses



(a) $\hat{\Theta}_H^I$ (b) $\hat{\Theta}_H^C$ (c) $\hat{\Theta}_H$

Fig. 9 Kronecker decomposition of the estimated GFEVD: indicator GFEVD $\hat{\Theta}_H^I$ (**a**), country GFEVD $\hat{\Theta}_H^C$ (**b**) and full GFEVD $\hat{\Theta}_H$ (**c**). Diagonal elements are omitted

tion. We then outlined two computational methods: a fully Bayesian MCMC algorithm and an EMVS estimation procedure, which foresees the forthcoming need of modeling matrix-valued time series at large scales, while allowing for fast dynamic posterior exploration.

We have numerically investigated the small sample efficiency of the proposed estimators, showing the gain with respect to ML in sparse, high-dimensional settings. We have also numerically explored the comparative estimation, forecasting and computational performances of the proposed estimators relative to key competing alternative models for longitudinal data. The experiment has shown that the estimation and forecasting performances of the Bayesian and MAP estimators are generally superior in sparse high-dimensional settings, with the latter drastically reducing computational intensiveness. The proposed methodology has been applied to a panel of nine world countries' economic indicators, for which we derived a method to decompose the GFEVD into its row and column dimensions, leading to country and indicator GFEVDs.

Our proposed method can be extended in several directions. The model can be easily generalized to a tensor autoregressive framework. Simultaneous sparsity both in the autoregressive coefficients and innovation covariance matrices can be introduced. Otherwise, the model can be equipped with different types of priors, e.g. those belonging to the class of global–local priors (Polson et al. 2011).

Furthermore, time variation can be embedded into the model. This also paves the way to the introduction of time varying parameter matrix autoregression with stochastic volatility (see Nakajima 2011), dynamic sparse factor matrix autoregressions (Rockova and McAlinn 2021), and dynamic covariance matrix estimation and prediction (Bucci et al. 2022).

Acknowledgements G.J. gratefully acknowledges partial support from NSF Grant DMS-2152746.

Funding Open access funding provided by Università degli Studi dell'Insubria within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare no competing interests.

Appendix A: Tensor operations and the Tucker product

A tensor is a multidimensional array, whose order expresses the number of dimensions, also known as ways or modes.[4] More formally, an Nth way tensor is an N dimensional array $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with entries $\mathcal{X}_{i_1 \ldots i_N}$ with $i_n = 1, \ldots, I_n$ and $n = 1, \ldots, N$.

Vectors are tensors of order one (denoted by boldface lowercase letters, e.g., $\mathbf{x}$) whereas matrices are tensors of order two (denoted by boldface capital letters, e.g., $\mathbf{X}$).

A.1 Tensor norm and inner product

The Frobenius norm of a tensor $\mathcal{X}$ is the square root of the squared sum of all its elements, ie.

$$||\mathcal{X}||_F = \sqrt{\sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} x_{i_1 \ldots i_N}^2} = \sqrt{\text{vec}(\mathcal{X})' \text{vec}(\mathcal{X})}$$

which is analogous to the Frobenious norm of a matrix.

The inner-product of two tensors of the same dimension $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{J_1, \ldots, J_N}$ is the sum of the product of their entries:

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{i_N} \cdots \sum_{i_N=1}^{i_N} x_{i_1 \ldots i_N} y_{i_1 \ldots i_N} = \text{vec}(\mathcal{X})' \text{vec}(\mathcal{Y})$$

A.2 Matricization and Tucker product

The process of reordering the elements of an N-way tensor into a matrix is called matricization. The nth way matricization of $\mathcal{X}$ is denoted by $\mathbf{X}^n = \text{mat}_n(\mathcal{X})$, and is obtained by reshaping the elements of the original tensor so that the resulting matrix is of dimension $[I_n \times \prod_{j \neq n} I_j]$. The special case of contemporaneous matricization along all the ways of a tensor is called vectorization, which is analogous to the vectorization of a matrix:

$$\mathbf{x} = \text{vec}(\mathcal{X}) = \text{mat}_{1,\ldots,N}(\mathcal{X})$$

Given the matrices $\mathbf{B}_1, \ldots, \mathbf{B}_N$ with $\mathbf{B}_n \in \mathbb{R}^{i_n \times j_n}$, a map from the space of $\mathcal{X}$ to the space generated by the rows of $\mathbf{B}_n$ ($\mathbb{R}^{i_1 \times \cdots \times i_N}$) is made by first obtaining $\mathbf{x}$, then computing:

$$\mathbf{m} = (\mathbf{B}_N \otimes \cdots \otimes \mathbf{B}_1)\mathbf{x}$$

and eventually forming an $[i_1 \times \cdots \times i_N]$ dimensional array $\mathcal{M}$ from $\mathbf{m}$. This transformation between the tensor $\mathcal{X}$ and the list $\mathbf{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_N\}$ is known as the Tucker product (Tucker 1966), and can be written as:

$$\mathcal{M} = \mathcal{X} \bar{\times} \{\mathbf{B}_1, \ldots \mathbf{B}_N\} \tag{A.1}$$

It is worth noting that the matricization operator connects the multidimensional Tucker product to the well known matrix multiplication, facilitating both understanding and computation of the former. In fact, by applying the nth way matricization to both sides of Eq. (A.1) we obtain the equivalent formulation:

$$\mathbf{M}^n = \mathbf{B}_n \mathbf{X}^n \mathbf{B}_{-n}^{'} \tag{A.2}$$

where $\mathbf{B}_{-n} = (\mathbf{B}_N \otimes \cdots \otimes \mathbf{B}_{n+1} \otimes \mathbf{B}_{n-1} \otimes \cdots \otimes \mathbf{B}_1)$. By repeating the operation for $n = 1, \ldots, N$, it emerges that the Tucker product can be expressed as a series of N matrix reshaping and multiplications. Matricization and vectorization applied to the Tucker product give raise to the following set of equivalences:

$$\begin{aligned} \mathcal{M} &= \mathcal{X} \bar{\times} \{\mathbf{B}_1, \ldots, \mathbf{B}_N\} \\ \mathbf{M}^n &= \mathbf{B}_n \mathbf{X}^n \mathbf{B}_{-n}^{'} \\ \mathbf{m} &= (\mathbf{B}_N \otimes \cdots \otimes \mathbf{B}_1)\mathbf{x} \end{aligned}$$

The VAR as well as the MAR equivalent form of a TAR can be easily derived with the abovementioned tools.

Appendix B: Conditional posterior distribution

The conditional posterior of each $\gamma_{i,k}$ is:

$$\pi(\gamma_{i,k}|-) \sim \mathcal{B}\text{er}(\bar{\theta}_{i,k}) \tag{B3}$$

where

$$\bar{\theta}_{i,k} = \frac{\theta_k \mathcal{N}(\phi_{i,k}, \tau_1)}{\theta_k \mathcal{N}(\phi_{i,k}, \tau_1) + (1 - \theta_k)\mathcal{N}(\phi_{i,k}, \tau_0)}$$

[4] To avoid confusion, we use the term way to express the dimension of a tensor, given that mode is already used to express the maximum value of a given distribution.

The conditional posterior of each $\boldsymbol{\phi}_k$ is:

$$\pi(\boldsymbol{\phi}_k|-) \sim \mathcal{N}(\bar{\mu}_k, \bar{\Omega}_k) \tag{B4}$$

where

$$\bar{\Omega}_k = (\tilde{\mathbf{X}}^k\tilde{\mathbf{X}}^{k'} \otimes \Sigma_k^{-1} + \mathcal{V}_k^{-1})^{-1}$$
$$\bar{\mu}_k = \bar{\Omega}_k(\tilde{\mathbf{X}}^k\tilde{\mathbf{X}}^{k'} \otimes \Sigma_k^{-1}\hat{\boldsymbol{\phi}}_k)$$

where $\hat{\boldsymbol{\Phi}}_k = (\tilde{\mathbf{Y}}^k\tilde{\mathbf{X}}^{k'})(\tilde{\mathbf{X}}^k\tilde{\mathbf{X}}^{k'})^{-1}$ and $\hat{\boldsymbol{\phi}}_k = \text{vec}(\hat{\boldsymbol{\Phi}}_k)$.

The conditional posterior of each Σ_k is:

$$\pi(\Sigma_k|-) \sim \mathcal{W}^{-1}(\bar{\Omega}_k, n_{2,k} + \nu_k) \tag{B5}$$

where $\bar{\Omega}_k = (\tilde{\mathbf{Y}}^k - \hat{\boldsymbol{\Phi}}_k\tilde{\mathbf{X}}^k)(\tilde{\mathbf{Y}}^k - \hat{\boldsymbol{\Phi}}_k\tilde{\mathbf{X}}^k)' + \xi\Omega_k$.

The conditional posterior of each θ_k is:

$$\pi(\theta_k|-) \sim \mathcal{B}\text{eta}(\bar{\alpha}_k, \bar{\beta}_k) \tag{B6}$$

where $\bar{\alpha}_k = |\boldsymbol{\gamma}_k| + \alpha_k$ and $\bar{\beta}_k = N_k + \beta_k + \alpha_k$.

The conditional posterior for ξ is:

$$\pi(\xi|-) \sim \mathcal{G}a(\bar{\nu}_1, \bar{\nu}_2) \tag{B7}$$

where $\bar{\nu}_1 = \frac{1}{2}\sum_k d_k\nu_k + \nu_1$ and $\bar{\nu}_2 = \frac{1}{2}\sum_k \Omega_k\Sigma_k^{-1} + \nu_2$.

We report in Algorithm 1 the proposed Gibbs sampling procedure.

Algorithm 1: MCMC

Starting values: MAP estimate $\hat{\boldsymbol{\Phi}}_k$, $\hat{\Sigma}_k$.
Hyperparameters: $\tau_0, \tau_1, \alpha_k, \beta_k, \Omega_k, \nu_k$.
Initialize : $\boldsymbol{\Phi}_k^{[0]} = \hat{\boldsymbol{\Phi}}_k$, $\Sigma_k^{[0]} = \hat{\Sigma}_k$, $\theta_k^{[it]} = \alpha_k/\beta_k$, n. of iterations MC, size of Burn in BU.
for $j = 1$ **to** $MC + BU$ **do**
  for $k = 1$ **to** K **do**
    for $i = 1$ **to** n_k **do**
      Draw $\gamma_{i,k}^{[j]}$ from the Bernoulli distribution (B3).
    Compute $\tilde{\mathbf{Y}}_k$ and $\tilde{\mathbf{X}}_k$ with $\boldsymbol{\Phi}_k^{[j-1]}$ and $\Sigma^{[j-1]}$ as in subsection 2.1.
    Draw $\boldsymbol{\phi}_k^{[j]}$ from the multivariate Normal distribution (B4).
    Draw $\Sigma_k^{[j]}$ from the Inverse Wishart distribution (B5).
    Draw $\theta_k^{[j]}$ from the Beta distribution (B6).
  Draw $\xi^{[j]}$ from the Gamma distribution (B7).
  Compute $\mathcal{B}^{[j]}$ and $\boldsymbol{\Sigma}^{[j]}$ with $\boldsymbol{\Phi}_1^{[j]}, \ldots, \boldsymbol{\Phi}_K^{[j]}$ and $\Sigma_1^{[j]}, \ldots, \Sigma_k^{[j]}$ and renormalize via (8).

Appendix C: E–M steps

For each k, the E-steps proceeds by computing the conditional expectations of $v_{i,k}^{-1} \in \mathcal{V}_k^{-1}$ in $\mathcal{Q}_{1,k}(\cdot)$ and of $\gamma_{i,k}$ for $|\boldsymbol{\gamma}_k|$ in $\mathcal{Q}_{3,k}(\cdot)$. Consider the latter first. At the jth step we have:

$$\begin{aligned}\text{E}_{\boldsymbol{\gamma}_k|\cdot}(\gamma_{i,k}) &= \text{P}(\gamma_{i,k} = 1|\boldsymbol{\phi}_k^{[j-1]}, \theta_k^{[j-1]}) \\ &= \frac{\theta_k^{[j-1]}\mathcal{N}(\phi_{i,k}|0, \tau_1)}{\theta_k^{[j-1]]}\mathcal{N}(\phi_{i,k}|0, \tau_1) + (1 - \theta_k^{[j-1]})\mathcal{N}(\phi_{i,k}|0, \tau_0)}\end{aligned} \tag{C8}$$

The E-step for the former is:

$$\begin{aligned}\text{E}_{\boldsymbol{\gamma}_k|\cdot}(v_{i,k}^{-1}) = \text{E}_{\boldsymbol{\gamma}_k|\cdot}[\tau_0(1 - \gamma_{i,k}) + \tau_1\gamma_{i,k}]^{-1} \\ = \frac{1 - \text{P}(\gamma_{i,k} = 1|\boldsymbol{\phi}_k^{[j-1]}, \theta_k^{[j-1]})}{\tau_0} \\ + \frac{\text{P}(\gamma_{i,k} = 1|\boldsymbol{\phi}_k^{[j-1]}, \theta_k^{[j-1]})}{\tau_1}\end{aligned} \tag{C9}$$

The maximization steps are:

$$\boldsymbol{\phi}_k = [\mathcal{V}^{-1} + (\tilde{\mathbf{X}}^{k'}\tilde{\mathbf{X}}^k \otimes \Sigma_k^{-1})]^{-1}[(\tilde{\mathbf{X}}^{k'}\tilde{\mathbf{X}}^k \otimes \Sigma^{-1})\hat{\boldsymbol{\phi}}_k] \tag{C10}$$

$$\Sigma_k = \frac{(\tilde{\mathbf{Y}}^k - \boldsymbol{\Phi}_k\tilde{\mathbf{X}}^k)(\tilde{\mathbf{Y}}^k - \boldsymbol{\Phi}_k\tilde{\mathbf{X}})' + \xi\Omega_k}{J_k + J_{-k} + \nu_k + 1} \tag{C11}$$

$$\theta_k = \frac{|\boldsymbol{\gamma}_k| + \alpha_k - 1}{n_k + \alpha_k + \beta_k + -2} \tag{C12}$$

$$\xi = \frac{\frac{1}{2}\sum_k(J_k\nu_k) + \eta_1 - 1}{\frac{1}{2}\sum_k \text{tr}(\Omega_k\Sigma_k^{-1}) + \eta_2} \tag{C13}$$

Algorithm 2: EMVS

Starting values: ML estimate $\hat{\boldsymbol{\Phi}}_k$, $\hat{\Sigma}_k$.
Hyperparameters: $\tau_0, \tau_1, \alpha_k, \beta_k, \Omega_k, \nu_k$.
Initialize : $\boldsymbol{\Phi}_k^{[0]} = \hat{\boldsymbol{\Phi}}_k$, $\Sigma_k^{[0]} = \hat{\Sigma}_k$, $\theta_k^{[it]} = \alpha_k/\beta_k$, $j = 0$, tolerance ϵ.
while $Tol > \epsilon$ **do**
  $j = j + 1$.
  for $k = 1$ **to** K **do**
    for $i = 1$ **to** n_k **do**
      Compute $\text{E}_{\boldsymbol{\gamma}_k|\cdot}(\gamma_{i,k})$ from Eq. (C8).
      Compute $\text{E}_{\boldsymbol{\gamma}_k|\cdot}(v_{i,k}^{-1})$ from Eq. (C9).
    Compute $\tilde{\mathbf{Y}}_k$ and $\tilde{\mathbf{X}}_k$ as in subsection 2.3.
    Update $\boldsymbol{\phi}_k^{[j]}$ from Eq. (C10).
    Update $\Sigma_k^{[j]}$ from Eq. (C11).
    Update $\theta_k^{[j]}$ from Eq. (C12).
  Update $\xi^{[j]}$ from Eq. (C13).
  Compute $\mathcal{B}^{[j]}$ and $\boldsymbol{\Sigma}^{[j]}$ with $\boldsymbol{\Phi}_1^{[j]}, \ldots, \boldsymbol{\Phi}_K^{[j]}$ and $\Sigma_1^{[j]}, \ldots, \Sigma_k^{[j]}$ and renormalize via (8).
  Compute $Tol = \max||\boldsymbol{\Phi}_k^{[j]} - \boldsymbol{\Phi}_k^{[j-1]}||_2^2$.

Appendix D: Additional simulation results

D.1 Competing models

In compact form, a PVAR mode can be rewritten as:

$$\mathbf{Y} = \mathbf{BX} + \mathbf{U},$$

or

$$\mathbf{y} = (\mathbf{X}^{'} \otimes \mathbf{I}_{GN})\boldsymbol{\beta} + \boldsymbol{u}, \tag{D14}$$

where $\mathbf{Y} = [\mathbf{y}_{P+1}, \ldots, \mathbf{y}_T]$ and the coefficient matrix $\mathbf{B} = [\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_P]$ is of dimension $GN \times GNP$. $\mathbf{X} = [\mathbf{X}_P, \ldots, \mathbf{X}_{T-1}]$, with $\mathbf{X}_t = [\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-P}]$. Equation (D14) is the vectorized form, where $\mathbf{y} = \text{vec}(\mathbf{Y})$, $\mathbf{u} = \text{vec}(\mathbf{U})$ and $\boldsymbol{\beta} = \text{vec}(\mathbf{B})$.

We consider the following competing models:

1. **CC**: Canova and Ciccarelli (2009, 2013) use a factorization approach of the parameters such that they can be divided into common, country-specific, and variable-specific factors. They specify the model in a hierarchical structure:

$$\begin{aligned}\boldsymbol{\beta}|\mathbf{F} &\sim \mathcal{N}(\Lambda\mathbf{F}, \Sigma \otimes \mathbf{I}_{GN}),\\ \mathbf{F} &\sim \mathcal{N}(0, \underline{c}\mathbf{F}),\end{aligned}$$

   where Λ is an $GN \times f$ matrix of loadings and $\mathbf{F}$ is an f-dimensional vector of factors where $f < GN$. The number of factors are, respectively, N for coefficients of each country and G for coefficients of each variable, and one common factor for all coefficients. There is only one hyperparameter to set is $\underline{c}$, related to the prior variance of the factors.
2. **SSVS**: George et al. (2008) specify a prior whereby each coefficient of $\mathbf{B}$ is drawn from a mixture of two normal distributions: the former with a small variance aiming at shrinking the coefficient towards 0 and the latter with a relatively large one. The higher the magnitude of $\mathbf{B}_{ij}$, the higher is the probability that it will be drawn from the second distribution, and viceversa.

$$\begin{aligned}\boldsymbol{\beta}_k|\gamma_k &\sim (1-\gamma_k)\mathcal{N}(0, \underline{\tau}_1^2) + \gamma_k\mathcal{N}(0, \underline{\tau}_2^2),\\ \gamma_k &\sim \mathcal{B}\mathbf{er}(\underline{\pi}_k),\end{aligned}$$

   with $k = 1, \ldots, G^2N^2P$.
3. **SSSS**: This algorithm designed by Koop and Korobilis (2016), who build on George et al. (2008) but taking in into account for panel restrictions. They specify two types of restrictions: dynamic interdependencies (DI) and cross-sectional homogeneity (CSH).

The DI works on off-diagonal blocks. Let $\mathbf{B}_{ij} \in \mathbf{B}$ be the $G \times G$ block embodying the parameters of country jth on country ith equations. The prior has the following form:

$$\begin{aligned}\text{vec}(\mathbf{B}_{ij})|\gamma_{ij}^{DI} &\sim (1-\gamma_{ij}^{DI})\mathcal{N}(0, \underline{\tau}_1^2\mathbf{I}_G{}^2)\\ &\quad + \gamma_{ij}^{DI}\mathcal{N}(0, \underline{\tau}_2^2\mathbf{I}_G{}^2),\\ \gamma_{ij}^{DI}|\pi_{ij}^{DI} &\sim \mathcal{B}\mathbf{er}(\pi_{ij}^{DI}), \forall j \neq i,\\ \pi_{ij}^{DI} &\sim \mathcal{B}\mathbf{eta}(1, \underline{\phi}),\end{aligned}$$

whereas the CSH prior works on the main block diagonal of $\mathbf{B}$. The prior reads as:

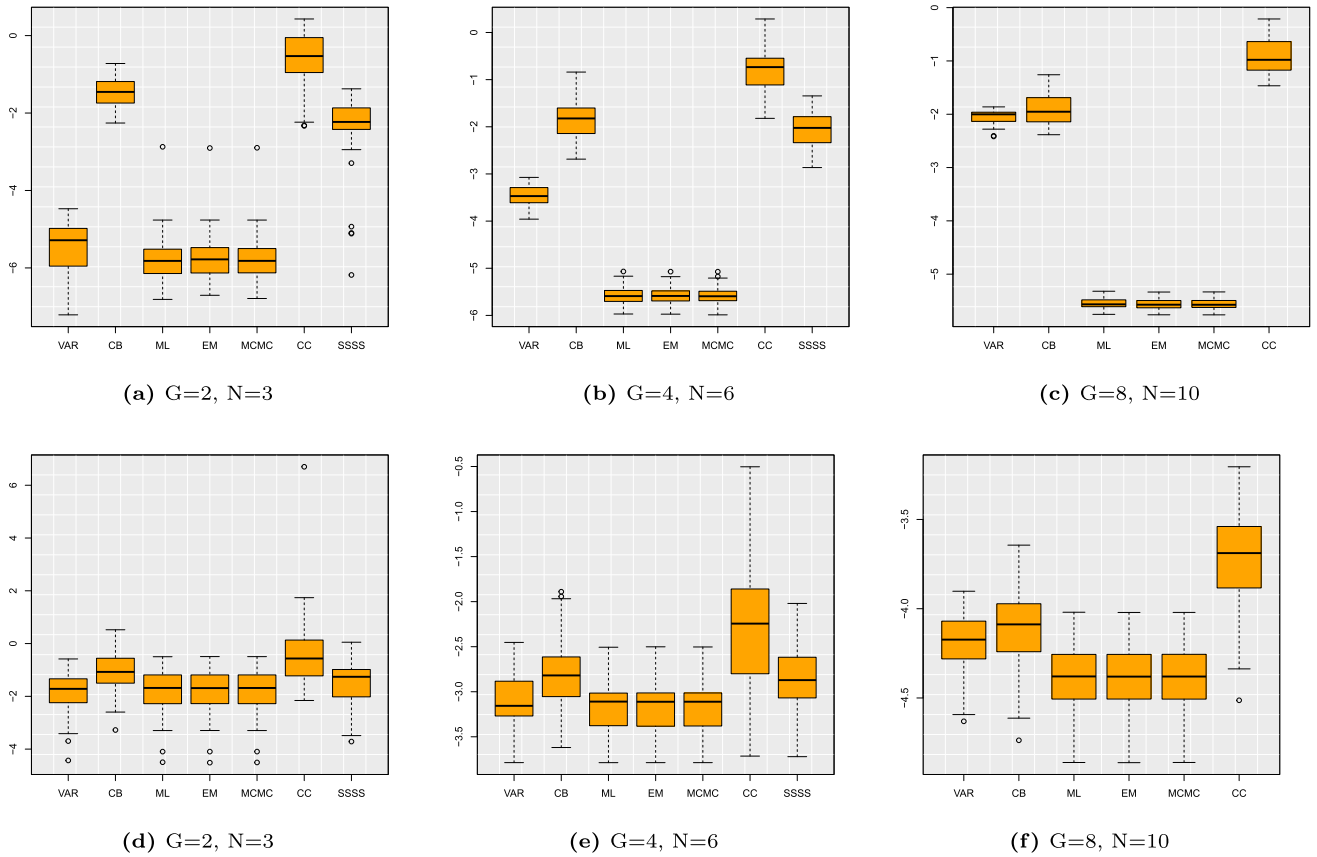
$$\begin{aligned}\text{vec}(\mathbf{B}_{ii})|\gamma_{ij}^{CSH} &\sim (1-\gamma_{ij}^{CSH})\mathcal{N}(\mathbf{B}_{jj}, \underline{\xi}_1^2\mathbf{I}_G^2)\\ &\quad + \gamma_{ij}^{CSH}\mathcal{N}(\mathbf{B}_{jj}, \underline{\xi}_2^2\mathbf{I}_G^2),\\ \gamma_{ij}^{CSH}|\pi_{ij}^{CSH} &\sim \mathcal{B}\mathbf{er}(\pi_{ij}^{CSH}), \quad \forall j \neq i,\\ \pi_{ij}^{CSH} &\sim \mathcal{B}\mathbf{eta}(1, \underline{\phi}).\end{aligned}$$

D.2 Hyper/regularization parameter tuning

- **CC**: We set $\underline{c} = 4$.
- **SSVS**: We set $\underline{\tau}_1^2 = 0.01$, $\underline{\tau}_2^2 = 4$ and $\underline{\pi}_k = 0.5$.
- **SSSS** We set $\underline{\tau}_1^2, \underline{\xi}_1^2 = 0.01$, $\underline{\tau}_2^2, \underline{\xi}_2^2 = 4$, $\pi_{ij}^{DI}, \pi_{ij}^{CSH} = 0.5$ and $\underline{\pi}_k = 1$.

D.3 Simulation results

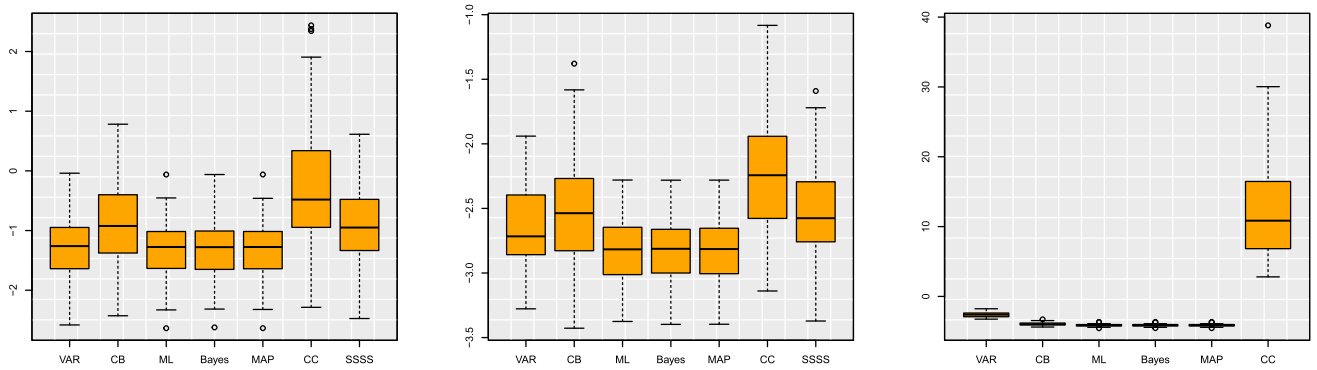
See Figs. 10, 11, 12, 13, 14, 15 and 16.



(a) G=2, N=3 (b) G=4, N=6 (c) G=8, N=10

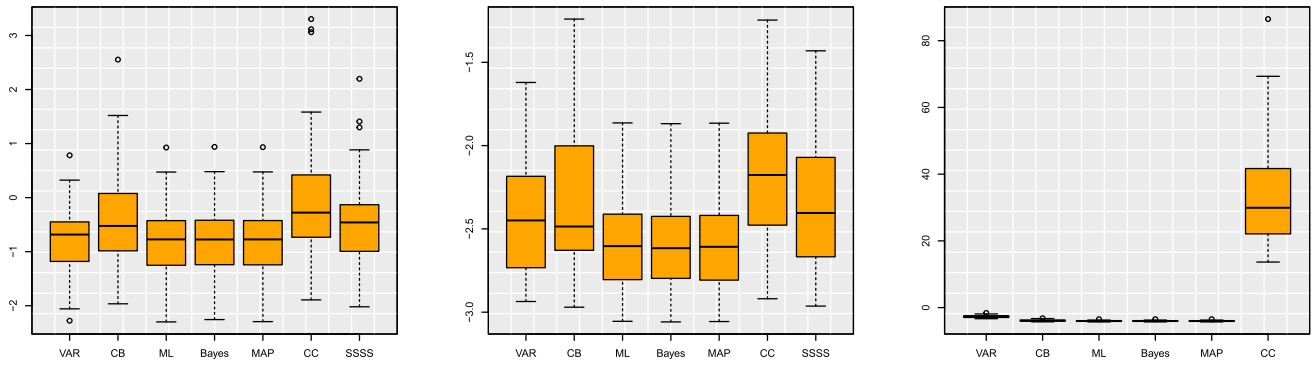
(d) G=2, N=3 (e) G=4, N=6 (f) G=8, N=10

Fig. 10 Average estimation error (MSE) (**a**–**c**) and 1 step ahead forecasting performance (MSFE) (**d**–**f**) over 50 repetitions of each model for $(G, N) = (2, 3), (4, 6), (8, 10)$ with $T = 500$



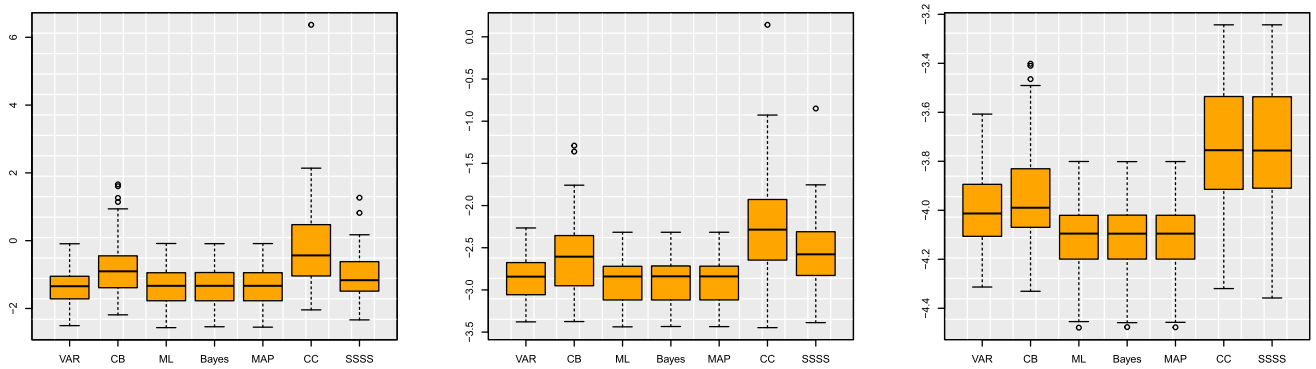
(a) G=2, N=3 (b) G=4, N=6 (c) G=8, N=10

Fig. 11 3 step ahead forecasting performance (MSFE) (**d**–**f**) over 50 repetitions of each model for $(G, N) = (2, 3), (4, 6), (8, 10)$ with $T = 100$



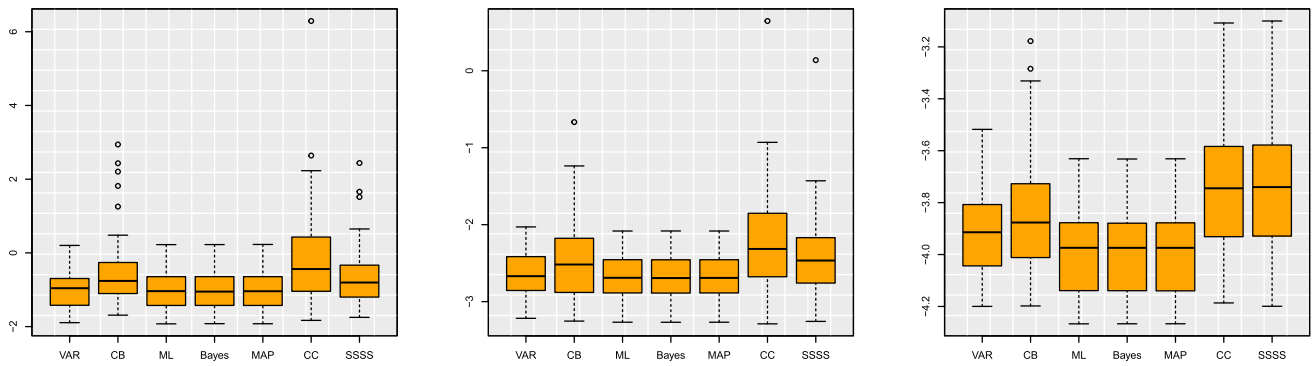
(a) G=2, N=3 (b) G=4, N=6 (c) G=8, N=10

Fig. 12 6 step ahead forecasting performance (MSFE) (**d**–**f**) over 50 repetitions of each model for $(G, N) = (2, 3), (4, 6), (8, 10)$ with $T = 100$



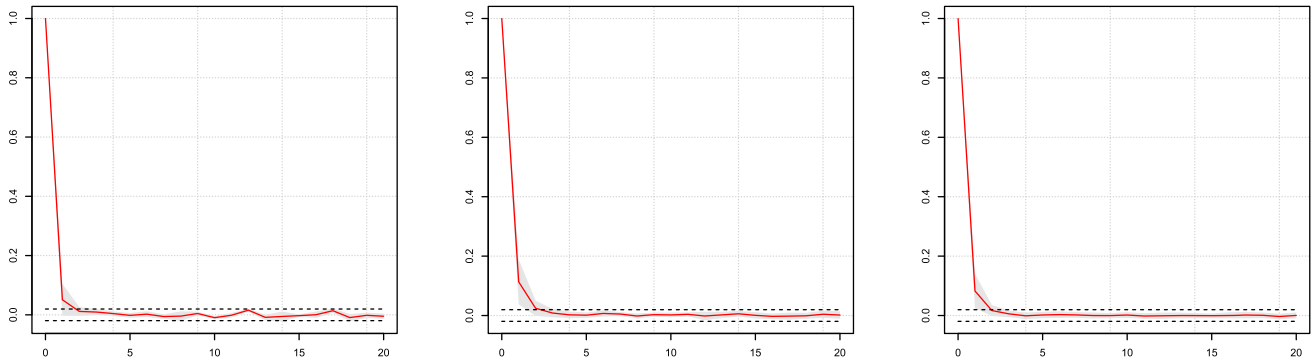
(a) G=2, N=3 (b) G=4, N=6 (c) G=8, N=10

Fig. 13 3 step ahead forecasting performance (MSFE) (**d**–**f**) over 50 repetitions of each model for $(G, N) = (2, 3), (4, 6), (8, 10)$ with $T = 500$



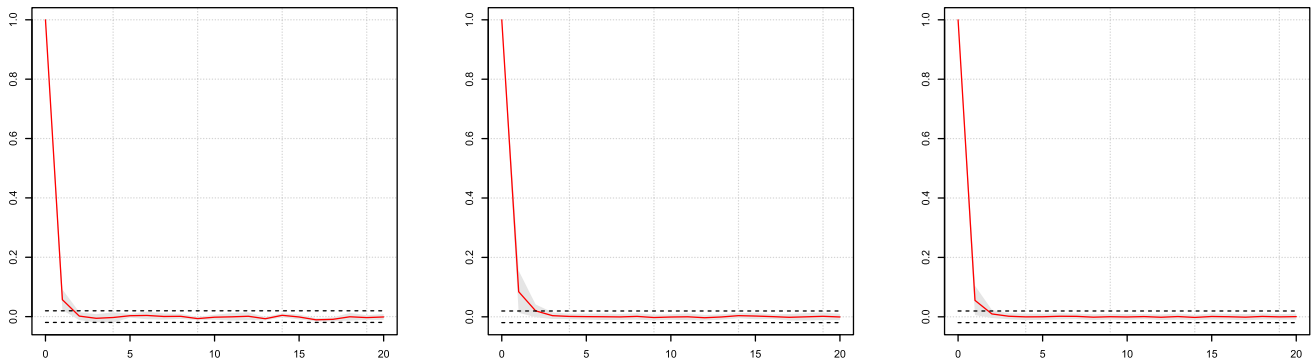
(a) G=2, N=3 (b) G=4, N=6 (c) G=8, N=10

Fig. 14 6 step ahead forecasting performance (MSFE) (**d**–**f**) over 50 repetitions of each model for $(G, N) = (2, 3), (4, 6), (8, 10)$ with $T = 500$



(a) G=2, N=3 (b) G=4, N=6 (c) G=8, N=10

Fig. 15 Overall autocorrelation function of the Markov Chain of our proposed model, averaged across the 50 repetitions, for all the coefficients of the left-hand matrix **A**



(a) G=2, N=3 (b) G=4, N=6 (c) G=8, N=10

Fig. 16 Overall autocorrelation function of the Markov Chain of our proposed model, averaged across the 50 repetitions, for all the coefficients of the right-hand matrix **B**

D.4 Empirical application

See Figs. 17, 18 and 19.

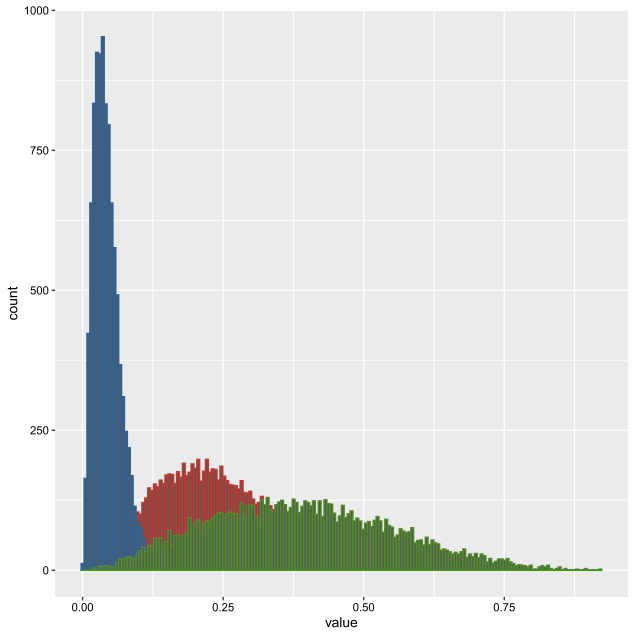
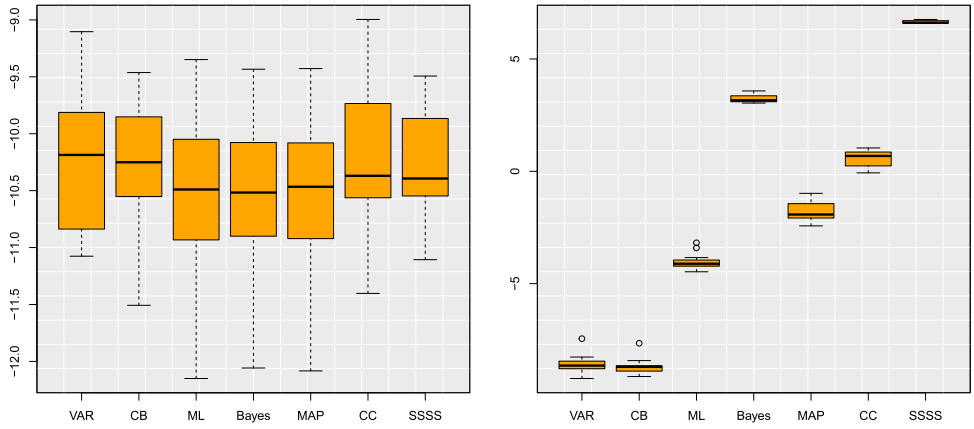


Fig. 17 Posterior distribution of $\hat{\theta}_1$ (red), $\hat{\theta}_2$ (blue) and $\hat{\theta}_3$ (green)



(a) MSFE **(b)** Computational time

Fig. 18 One step ahead MSFE (**a**) and computational time (**b**) of each model over different rolling windows spanning from 1995Q1 (T = 60) to 2019Q2 (T = 120)

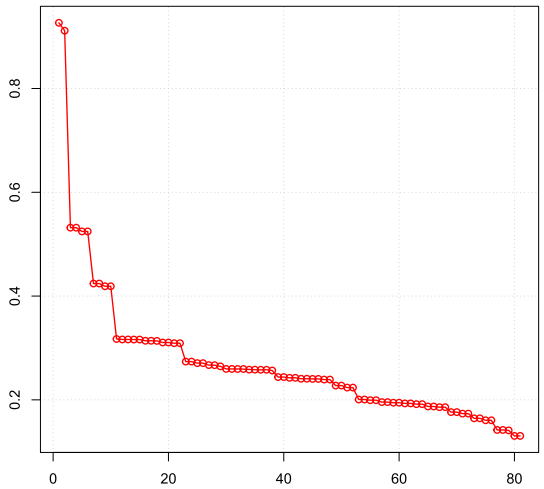


Fig. 19 Spectrum of the companion matrix

References

Ahelegbey, D., Billio, M., Casarin, R.: Bayesian graphical models for structural vector autoregressive processes. J. Appl. Econom. **31**, 357–386 (2016). (https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2443)

Bader, B., Kolda, T.: Algorithm 862: Matlab tensor classes for fast algorithm prototyping. ACM Trans. Math. Softw. **32**, 635–653 (2006). https://doi.org/10.1145/1186785.1186794

Bai, J., Ng, S.: Determining the number of factors in approximate factor models. Econometrica **70**, 191–221 (2002). https://doi.org/10.1111/1468-0262.00273

Bańbura, M., Giannone, D., Reichlin, L.: Large Bayesian vector auto regressions. J. Appl. Econom. **25**, 71–92 (2010). https://doi.org/10.1002/jae.1137

Barrett, B.E., Gray, J.B.: A computational framework for variable selection in multivariate regression. Stat. Comput. **4**, 203–212 (1994)

Billio, M., Casarin, R., Iacopini, M., Kaufmann, S.: Bayesian dynamic tensor regression. J. Bus. Econ. Stat. (2022). https://doi.org/10.1080/07350015.2022.2032721

Bucci, A., Ippoliti, L., Valentini, P.: Comparing unconstrained parametrization methods for return covariance matrix prediction. Stat. Comput. **32**, 90 (2022)

Camehl, A.: Penalized estimation of panel vector autoregressive models: a panel lasso approach. Int. J. Forecast. **6**, 66 (2022)

Canova, F., Ciccarelli, M.: Estimating multicountry VAR models. Int. Econ. Rev. **50**, 929–959 (2009). (https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-2354.2009.00554.x)

Canova, F., Ciccarelli, M.: Panel vector autoregressive models: a survey. Adv. Econom. **31**, 66 (2013)

Castillo, I., van der Vaart, A.: Needles and Straw in a Haystack: posterior concentration for possibly sparse sequences. Ann. Stat. **40**, 2069–2101 (2012). https://doi.org/10.1214/12-AOS1029

Chen, E.Y., Fan, J.: Statistical inference for high-dimensional matrix-variate factor models. J. Am. Stat. Assoc. **66**, 1–18 (2021)

Chen, L., Huang, J.Z.: Sparse reduced-rank regression with covariance estimation. Stat. Comput. **26**, 461–470 (2016)

Chen, R.H.X., Yang, D.: Autoregressive models for matrix-valued time series. J. Econom. **222**, 539–560 (2021)

Chen, R., Yang, D., Zhang, C.-H.: Factor models for high-dimensional tensor time series. J. Am. Stat. Assoc. **117**, 94–116 (2022)

Cichocki, A.: Fundamental tensor operations for large-scale data analysis using tensor network formats. Multidimens. Syst. Signal Process. **29**, 921–960 (2018). https://doi.org/10.1007/s11045-017-0481-0

Forni, M., Hallin, M., Lippi, M., Reichlin, L.: The generalized dynamic factor model. J. Am. Stat. Assoc. **100**, 830–840 (2005). https://doi.org/10.1198/016214504000002050

Gao, Z., Tsay, R.S.: A two-way transformed factor model for matrix-variate time series. Econom. Stat. **6**, 66 (2021)

Gefang, D.: Bayesian doubly adaptive elastic-net lasso for VAR shrinkage. Int. J. Forecast. **30**, 1–11 (2014). (https://www.sciencedirect.com/science/article/pii/S0169207013000770)

George, E.I., McCulloch, R.E.: Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**, 881–889 (1993)

George, E.I., McCulloch, R.E.: Approaches for Bayesian variable selection. Stat. Sin. **66**, 339–373 (1997)

George, E.I., Sun, D., Ni, S.: Bayesian stochastic search for VAR model restrictions. J. Econom. **142**, 553–580 (2008)

Geyer, C.J.: Computation for the Introduction to MCMC Chapter of Handbook of Markov chain Monte Carlo (2010)

Gong, L., Flegal, J.M.: A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. J. Comput. Graph. Stat. **25**, 684–700 (2016). https://doi.org/10.1080/10618600.2015.1044092

Gupta, A., Nagar, D.K.: Matrix Variate Distributions. Chapman & Hall/CRC, London (1999)

Hoff, P.D.: Separable covariance arrays via the Tucker product, with applications to multivariate relational data. Bayesian Anal. **6**, 179–196 (2011). https://doi.org/10.1214/11-BA606

Hoff, P.D.: Multilinear tensor regression for longitudinal relational data. Ann. Appl. Stat. **9**, 1169–1193 (2015). https://doi.org/10.1214/15-AOAS839

Jones, G.L., Qin, Q.: Markov chain Monte Carlo in practice. Annu. Rev. Stat. Appl. **9**, 557–578 (2022)

Kock, A., Callot, L.: Oracle inequalities for high dimensional vector autoregressions. J. Econom. **186**, 325–344 (2015)

Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**, 455–500 (2009)

Koop, G., Korobilis, D.: Model uncertainty in panel vector autoregressive models. Eur. Econ. Rev. **81**, 115–131 (2016)

Koop, G., Pesaran, M., Potter, S.M.: Impulse response analysis in nonlinear multivariate models. J. Econom. **74**, 119–147 (1996). (https://www.sciencedirect.com/science/article/pii/0304407695017534)

Korobilis, D.: Prior selection for panel vector autoregressions. Comput. Stat. Data Anal. **101**, 110–120 (2016). (https://www.sciencedirect.com/science/article/pii/S0167947316300275)

Korobilis, D.: High-dimensional macroeconomic forecasting using message passing algorithms. J. Bus. Econ. Stat. **39**, 493–504 (2021)

Lam, C., Yao, Q., Bathia, N.: Estimation of latent factors for high-dimensional time series. Biometrika **98**, 901–918 (2011). https://doi.org/10.1093/biomet/asr048

Lanne, M., Nyberg, H.: Generalized forecast error variance decomposition for linear and nonlinear multivariate models. Oxf. Bull. Econ. Stat. **78**, 595–603 (2016). (https://onlinelibrary.wiley.com/doi/abs/10.1111/obes.12125)

Lütkepohl, H.: New Introduction to Multiple Time Series Analysis. Springer, Berlin (2005)

Meng, X.-L., Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika **80**, 267–278 (1993). (http://www.jstor.org/stable/2337198)

Nakajima, J.: Time-varying parameter var model with stochastic volatility: an overview of methodology and empirical applications. Monet. Econ. Stud. **29**, 107–142 (2011)

Ohlson, M., Rauf Ahmad, M., von Rosen, D.: The multilinear normal distribution: introduction and some basic properties. J. Multivar. Anal. **113**, 37–47 (2013). (https://www.sciencedirect.com/science/article/pii/S0047259X11001047)

Park, T., Casella, G.: The Bayesian Lasso. J. Am. Stat. Assoc. **103**, 681–686 (2008). https://doi.org/10.1198/016214508000000337

Pesaran, M.H., Schuermann, T., Weiner, S.M.: Modeling regional interdependencies using a global error-correcting macroeconometric model. J. Bus. Econ. Stat. **22**, 129–162 (2004). https://doi.org/10.1198/073500104000000019

Polson, N., Scott, J., Clarke, B., Severinski, C.: Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction, Vol. 9780199694587 (Oxford University Press, Oxford, 2012)

Ročková, V., George, E.I.: Emvs: the EM approach to Bayesian variable selection. J. Am. Stat. Assoc. **109**, 828–846 (2014)

Ročková, V., George, E.I.: The spike-and-slab lasso. J. Am. Stat. Assoc. **113**, 431–444 (2018)

Rockova, V., McAlinn, K.: Dynamic variable selection with spike-and-slab process priors. Bayesian Anal. **16**, 233–269 (2021)

Rothman, A.J., Levina, E., Zhu, J.: Sparse multivariate regression with covariance estimation. J. Comput. Graph. Stat. **19**, 947–962 (2010). https://doi.org/10.1198/jcgs.2010.09188

Samanta, S., Khare, K., Michailidis, G.: A generalized likelihood-based Bayesian approach for scalable joint regression and covariance selection in high dimensions. Stat. Comput. **32**, 47 (2022)

Song, S., Bickel, P.: Large vector auto regressions. Papers, arxiv. org (2011)

Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **51**, 279–311 (1966). https://doi.org/10.1007/BF02289464

Van Loan, C.: The ubiquitous Kronecker product. J Comput. Appl. Math. **123**, 85–100 (2000)

Van Loan, C.F., Pitsianis, N.: Approximation with Kronecker Products, pp. 293–314 (Springer, Dordrecht, 1993). https://doi.org/10.1007/978-94-015-8196-7_17

Vats, D., Flegal, J.M., Jones, G.L.: Multivariate output analysis for Markov chain Monte Carlo. Biometrika **106**, 321–337 (2019). https://doi.org/10.1093/biomet/asz002

Vats, D., Flegal, J.M. & Jones, G.L.: Monte Carlo Simulation: Are we there yet?, pp. 1–15. Wiley, New York (2021). https://doi.org/10.1002/9781118445112.stat08283

Wang, T., Chen, M., Zhao, H., Zhu, L.: Estimating a sparse reduction for general regression in high dimensions. Stat. Comput. **28**, 33–46 (2018)

Wang, D., Liu, X., Chen, R.: Factor models for matrix-valued high-dimensional time series. J. Econom. **208**, 231–248 (2019). (https://www.sciencedirect.com/science/article/pii/S0304407618301787)