Discrete Autoregressive Switching Processes in Sparse Graphical Modeling of Multivariate Time Series Data

Beniamino Hadj-Amar[*], Aaron M. Bornstein[†], Michele Guindani[‡]

and

Marina Vannucci[*]

June 6, 2024

Abstract

We propose a flexible Bayesian approach for sparse Gaussian graphical modeling of multivariate time series. We account for temporal correlation in the data by assuming that observations are characterized by an underlying and unobserved hidden discrete autoregressive process. We assume multivariate Gaussian emission distributions and capture spatial dependencies by modeling the state-specific precision matrices via graphical horseshoe priors. We characterize the mixing probabilities of the hidden process via a cumulative shrinkage prior that accommodates zero-inflated parameters for non-active components, and further incorporate a sparsity-inducing Dirichlet prior to estimate the effective number of states from the data. For posterior inference, we develop a sampling procedure that allows estimation of the number of discrete autoregressive lags and the number of states, and that cleverly avoids having to deal with the changing dimensions of the parameter space. We thoroughly investigate performance of our proposed methodology through several simulation studies. We further illustrate the use of our approach for the estimation of dynamic brain connectivity based on fMRI data collected on a subject performing a task-based experiment on latent learning.

---

[*]Department of Statistics, Rice University, Houston, TX
[†]Department of Cognitive Sciences, University of California, Irvine, CA
[‡]Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles,CA

1   Introduction

In this paper we consider the problem of estimating sparse Gaussian graphical models based on time series data. Time-changing dependencies and sparse structures are often encountered when investigating multi-dimensional physiological signals (Safikhani and Shojaie, 2022), environmental and sensor data (Lam and Yao, 2012), as well as macroeconomic and financial systems (Kastner and Huber, 2020). Among existing approaches, Song et al. (2009) introduced a time-varying dynamic Bayesian network for modeling the fluctuating network structures underlying non-stationary biological time series. Kolar et al. (2010) proposed a method for estimating time-varying networks based on temporally smoothed l_1-regularized logistic regression. Danaher et al. (2014) and Qiu et al. (2016) addressed the challenge of estimating multiple related Gaussian graphical models when observations belong to distinct classes, and Warnick et al. (2018) and Quinn et al. (2018) employed Hidden Markov Models (HMMs) for the estimation of recurrent brain connectivity networks during a neuroimaging experiment. Other procedures for modeling the temporal evolution of dynamic networks include change-point detection methods (Cribben et al., 2013; Xu and Lindquist, 2015) and time-varying parameter models (Lindquist et al., 2014; Zhang et al., 2021). Change-point techniques provide a data-driven approach for the temporal partitioning of the network structure into segments of adaptable length. However, these methods do not provide a system for identifying potentially recurring network patterns over time. Time-varying parametric methods offer a principled way of modeling dynamic correlations but are computationally intensive.

We propose a flexible Bayesian approach for sparse Gaussian graphical modeling of multivariate time series. In order to represent switching dynamics, we assume an unobserved hidden process, underlying the time series data, which at each time point exists in one of a finite number of states. We account for the temporal structure of this hidden process by assuming a Discrete Autoregressive (DAR) process of order P (Biswas and Song, 2009), which flexibly incorporates long-term dependencies by considering the P previous lags of the

process. Given the state of the latent process, we model the observations as conditionally independent of the observations and states at previous times and generated from state-specific multivariate Gaussian emission distributions. Under the multivariate Gaussian assumption, networks can be estimated by the graphical models induced by the state-specific inverse covariance matrices. We capture these spatial dependencies by modeling the state-specific precision matrices via graphical horseshoe priors.

The DAR hidden process construction we adopt is reminiscent of higher-order HMMs, where the present state depends not only on the immediately preceding state but also on prior states further back in time. First-order HMMs, which constrain the temporal dynamics of the hidden state sequence to be Markovian, have been successfully applied in many scientific fields, including neuroimaging (Warnick et al., 2018; Quinn et al., 2018), climate (Holsclaw et al., 2017) and animal behavior (DeRuiter et al., 2017), to cite a few. While higher-order HMMs have been suggested (Cappé et al., 2005), they require the estimation of transition probability matrices that grow exponentially in size as the order increases, making their estimation challenging (see, for a discussion, Sarkar and Dunson, 2019). In our proposed model, the state-switching behavior of the process is captured by the time-varying mixing probabilities of the DAR process. To model these probabilities, we propose a nonparametric zero-inducing cumulative shrinkage prior. Building upon the construction of the finite Dirichlet process (DP; see Ishwaran and James, 2001), the proposed prior accommodates zero-inflated parameters, to account for non-active components, and employs cumulative shrinkage (Legramanti et al., 2020) to handle increasing model complexity. This construction ensures that if a parameter in the DAR model is zero, then all subsequent lag parameters are also zero. This results in a flexible and computationally efficient framework for learning the time-varying mixing probabilities and the effective order of the process, as opposed to learning the entire transition matrix, as required in HMM modeling. Such reduction in the number of parameters leads to a substantial computational advantage. It also allows to learn the number of lags in a data-driven fashion. Related sparsity-inducing prior constructions have been developed by Heiner et al. (2019) for the simplex model, and by Tang and Chen (2019) for zero-inflated generalized Dirichlet multinomial regression models. These constructions are specific to those models and less flexible than our approach, which models the ordering of the lags as the process evolves in time while promoting lower-order

complexity. We complete our modeling framework with a sparsity-inducing Dirichlet prior that allows estimation of the effective number of hidden states in a data-driven manner. Drawing inspiration from the literature on overfitted finite mixture models (Rousseau and Mengersen, 2011; Malsiner-Walli et al., 2016), we consider more states than strictly necessary, while employing a prior that effectively constrains the model's complexity. This promotes sparsity while leading to more interpretable inferences.

For posterior inference, we take a fully Bayesian approach and develop a sampling procedure that accommodates the multiple model selection problems, namely the number of DAR lags and the number of states, while cleverly avoiding having to deal with the changing dimensions of the parameter space. Specifically, we implement a Gibbs sampler that alternates between updating the DAR parameters, the sparse emission parameters, and the latent state sequence. To update the DAR probabilities, we leverage the stick-breaking construction of the DP by augmenting the space with auxiliary indicator variables and design a joint sampling scheme that alternates between adding or removing the sticks of the zero-inducing DP formulation. Our zero-inducing cumulative shrinkage prior significantly accelerates the proposed sampler, particularly in regard to the forward-backward algorithm for updating the latent state sequence. Estimates of the number of hidden states and DAR order are determined based on the most frequently occurring number of active states and DAR order observed during MCMC sampling, respectively.

We thoroughly investigate performance of our proposed methodology through several simulation studies. We further illustrate the use of our proposed approach to estimate dynamic brain connectivity networks based on functional Magnetic Resonance Imaging (fMRI) data. Identifying the dynamic nature of brain connectivity is critical for understanding our current knowledge about human brain functioning. In our application, we consider data collected on a subject performing an experiment aimed at understanding neural representations that are formed during latent learning. Inferred networks by our method identify distinct regimes of functional connectivity, that can be mapped onto cognitive interpretation.

The rest of the paper is organized as follows. Section 2 introduces the proposed model, including the DAR process and the proposed prior structures, and the MCMC algorithm for posterior inference. Section 3 contains results from the simulation studies and Section 4 illustrates the application to fMRI data on latent learning. Section 5 provides concluding

remarks. Julia software is available on GitHub at XXX (to be released upon acceptance).

2 Sparse Modeling of Multivariate Time Series Data via Cumulative Shrinkage DAR

In this Section, we describe the proposed latent variable approach for modeling sparse multi-dimensional time series. Let $\boldsymbol{y} = \left\{\boldsymbol{y_t}\right\}_{t=1}^{T}$, $\boldsymbol{y_t} = (y_{t1}, \dots, y_{tD}) \in \mathbb{R}^D$, be the observed D-dimensional time series data, with T indicating the number of time points. We envision an unobserved, latent hidden process underlying the observations and assume that, at each time point, the process assumes one of a finite number of states, represented as $\boldsymbol{\gamma} = \left\{\gamma_t\right\}_{t=1}^{T}$, with $\gamma_t \in \{1, \dots, M\}$ and M denoting the (unknown) finite number of latent states. Given the value of γ_t, the observations $\boldsymbol{y_t}$ are assumed to be independent of both the observations and states at previous time points. We further assume that the state-specific emissions follow a D-variate Gaussian distribution

$$\boldsymbol{y_t} \,|\, \gamma_t, \,\boldsymbol{\mu}, \boldsymbol{\Omega} \sim \sum_{j=1}^{M} \mathbb{1}_{\{j\}}(\gamma_t) \, \mathcal{N}_D\left(\boldsymbol{y_t} \,|\, \boldsymbol{\mu}_j, \boldsymbol{\Omega}_j^{-1}\right), \tag{1}$$

with state-specific means $\boldsymbol{\mu}_j$ and precision matrices $\boldsymbol{\Omega}_j$, $j = 1, \dots, M$, $t = 1, \dots, T$. Conditional dependencies can be inferred from the off-diagonal entries of the precision matrices. Specifically, for a given state j, if the entry $\omega_{j,il}$ is zero, the corresponding variables y_{ti} and y_{tl} are conditionally independent given the other variables.

2.1 State Dynamics via Discrete Autoregressive Processes

In order to learn the dependence structure between time points, represented by the sequence $\boldsymbol{\gamma}$, we design an approach that employs a discrete autoregressive process, with a cumulative shrinkage prior that enables a computationally efficient estimation of the order of the process. More specifically, we assume that the evolution of the hidden state sequence γ_t follows a Discrete Autoregressive (DAR) process of order P (Biswas and Song, 2009), which allows the hidden sequence to incorporate long-term dependencies by considering the previous P lags. Formally, the conditional distribution of γ_t given the past values $\gamma_{t-1:t-P}$ is expressed

5

as

$$p(\gamma_t|\gamma_{t-1:t-P}, \boldsymbol{\phi}, \boldsymbol{\pi}) = \phi_1 \mathbb{1}_{\{\gamma_{t-1}\}}(\gamma_t) + \phi_2 \mathbb{1}_{\{\gamma_{t-2}\}}(\gamma_t) + \ldots + \phi_P \mathbb{1}_{\{\gamma_{t-P}\}}(\gamma_t) + \phi_0 \pi_{\gamma_t}, \qquad (2)$$

where $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_P)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$. We denote with $\{\phi_j\}_{j=0}^P$ the *autoregressive probabilities*, with $\phi_0 = 1 - \sum_{j=1}^P \phi_j$, while the *state innovation probabilities* $\{\pi_i\}_{i=1}^M$ are defined as $\pi_i := p(\gamma_t = i)$, for $i = 1, \ldots, M$, and allow the process to transition to any of the M states, including those not observed in the previous P lags. Here, $\mathbb{1}_{\{j\}}(i)$ is an indicator function equal to 1 if $i = j$ and 0 otherwise. According to (2), the value of γ_t is chosen as one of the latent states selected at previous time points, $t - 1$ to $t - P$, based on the autoregressive probabilities $\{\phi_j\}_{j=1}^P$. Alternatively, with probability ϕ_0, γ_t takes on any of the M possible states independently of the history of the latent sequence, according to the state innovation distribution $\boldsymbol{\pi}$.

The transition probabilities in the DAR process can be represented by a multi-dimensional array. The dimensions of this array are determined by the number of autoregressive lags, P, and the number of hidden states, M. As an illustration, when $P = 2$, the transition probabilities are described by an $[M \times M \times M]$ array, say $\boldsymbol{\eta}$. The individual components of this array, denoted as $\eta_{l,i,j}$, represent the probability $p(\gamma_t = j|\gamma_{t-1} = i, \gamma_{t-2} = l)$ for $l, i, j \in \{1, \ldots, M\}$, as defined in (2). However, as the number of lags P increases, the dimensionality of this array grows exponentially. Therefore, the DAR characterization simplifies inference by allowing us to focus only on making inferences on the ϕ parameters, as opposed to learning the entire transition matrix, which is the case with HMM models, for example. In fact, when dealing with higher order HMMs, the task involves estimating M^P parameters for the transition arrays. In contrast, our proposed method streamlines this process by estimating $(M + P)$ parameters, resulting in a substantial computational advantage.

2.1.1 Zero-inducing cumulative shrinkage prior for learning time dependence

The time-varying mixing probabilities of the DAR model, denoted as ϕ_j, characterize the state-switching behavior of the process. To model these probabilities, we propose a nonparametric zero-inducing cumulative shrinkage prior that accommodates zero-inflated parameters to account for non-active components, and that employs cumulative shrinkage (Legramanti

6

et al., 2020) to handle increasing model complexity. This prior modifies the stick-breaking construction to allow for an increasing probability of setting $\phi_j = 0$ as j increases. In addition, our formulation enforces that once ϕ_j becomes zero for a specific j, $j = 1, 2, \ldots, P$, subsequent lags obey the condition $p(\phi_{j+k} = 0 \,|\, \phi_j = 0) = 1$, $k = 1, \ldots, P - j$. To formally introduce our prior, we need to define a binary latent process, namely an "active order" latent indicator, denoted as $z_j \in \{0, 1\}$, $j = 1, \ldots, P$. If $z_j = 0$, then ϕ_j is almost surely non-zero. However, when the first j such that $z_j = 1$ occurs, then $\phi_j = 1 - \sum_{l=1}^{j-1} \phi_l$ and $z_l = 1$ almost surely for $l = j + 1, \ldots, P$. More formally, the mixing probabilities ϕ_j are generated via a modified stick-breaking construction,

$$\phi_j = v_j \prod_{l=0}^{j-1} (1 - v_l), \quad \text{for } j = 1, \ldots, P, \tag{3}$$

with $\phi_0 = v_0$, where the stick-breaking weights v_j are mixtures of a Beta distribution and a spike at one,

$$v_j \mid z_j \sim (1 - z_j)\, \text{Beta}\,(a_v, b_v) + z_j\, \delta_1, \tag{4}$$

with δ_x denoting a point mass at $\{x\}$, $j = 1, \ldots, P$, and by specifying $v_0 \sim \text{Beta}(a_0, b_0)$.

For $z_j = 0$, (3)–(4) define the stick-breaking construction typical of the Dirichlet process. If at some point $z_j = 1$ occurs, then $v_j = 1$, and $\phi_j = \prod_{l=0}^{j-1} (1 - v_l) = 1 - \sum_{l=1}^{j-1} \phi_l$. For all remaining lags, our construction ensures $\phi_l = 0$, $l = j+1, \ldots, P$. More specifically, to enforce the desired behavior and promote lower order model complexity, we leverage the increasing shrinkage prior construction of Legramanti et al. (2020) and assign increasing probability mass to selecting the spike component as the order of the DAR grows. In particular, we assume $z_j \mid \boldsymbol{v}_{0:j-1} \sim \text{Bern}\,(\xi_j)$ with probability $\xi_j = \sum_{i=0}^{j-1} \phi_i$ increasing with the lag j, where $z_1 | v_0 \sim \text{Bern}(v_0)$. See also Zhang et al. (2021), where an increasing shrinkage prior is used in a VAR model. Our construction ensures that $p\,(z_l = 1 | z_{l-1} = 1) = 1$ and $p\,(\phi_l = 0 \,|\, \phi_{l-1} = 0) = 1$. We define the effective order of the DAR process as the random element $\hat{P} = \inf_{j \in \{1, \ldots, P\}} \{z_j = 1\}$, that is the number of "active" lags of the DAR process. The proposition below demonstrates the aforementioned property.

Proposition 1. Let $\boldsymbol{\phi} = \{\phi_j \in \Delta_\phi : j = 0, \ldots, P\}$ with $\Delta_\phi = \{\phi_l : 0 \le \phi_l \le 1, \sum_{l=0}^{\infty} = 1\}$, be constructed according to (3), and $\boldsymbol{v} = \{v_i\}_{i=0}^{P}$ and $\boldsymbol{z} = \{z_i\}_{i=1}^{P}$ be specified

as in (4). Under these assumptions, the cumulative shrinkage DAR formulation implies that $p(z_{j+1} = 1|z_j = 1) = 1$, for $j = 1, \ldots, P$.

Proof. Recall that $z_j \,|\, \boldsymbol{v}_{0:\,j-1} \sim \text{Bern}\,(\xi_j)$ with probability $\xi_j = \sum_{i=0}^{j-1} \phi_i$, $j = 1, \ldots P$. Therefore, for $j = 1, \ldots, \hat{P}$, we can write

$$p(z_{j+1} = 1|z_j = 0, \boldsymbol{v}_{0:j}) = \sum_{i=0}^{j} \phi_i = v_0 + v_1(1 - v_0) + \cdots + v_j \prod_{l=0}^{j-1}(1 - v_l).$$

Thus, $p(z_{j+1} = 0|z_j = 0, \boldsymbol{v}_{0:j}) = 1 - p(z_{j+1} = 1|z_j = 0, \boldsymbol{v}_{0:j}) = \prod_{l=0}^{j}(1 - v_l)$. For $j = \hat{P}$, since $v_{\hat{P}} = 1$ a.s., we have $\sum_{j=0}^{\hat{P}} \phi_j = v_0 + v_1(1 - v_0) + \cdots + v_{\hat{P}} \prod_{l=0}^{\hat{P}-1}(1 - v_l) = 1$. Thus, for $j = \hat{P} + 1, \ldots, P - 1$, $p(z_{j+1} = 1|v_{0:\hat{P}}) = p(z_{j+1}|z_j = 1) = 1$. $\square$

Given the one-to-one relationship between the sequence z_j and $\hat{P}$, the process can be alternatively defined in terms of the random quantity $\hat{P}$, which is computationally convenient, as we explain in Section 2.3 below. We note that the previous characterization can also be extended to the case of $P = \infty$. However, for computational purposes, it is convenient to consider only a finite number of terms, say, P_{max}, and thus specify the autoregressive coefficients as $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{P_{max}})$, where $\phi_{P_{max}} = 1 - \sum_{l=0}^{P_{max}-1} \phi_j$. In implementations, this approach offers considerable versatility when P_{max} is set to a moderately high upper bound, and it is advisable to choose P_{max} such that it exceeds the expected number of lags.

2.1.2 Sparsity-inducing Dirichlet prior to infer state transitions and space size

As for the innovation probabilities $\boldsymbol{\pi}$, to facilitate a substantial reduction in the effective number of states compared to the maximum number, $M = M_{\max}$, we draw insights from recent literature on overfitted finite mixture models (Rousseau and Mengersen, 2011; Malsiner-Walli et al., 2016). Specifically, we assume a symmetric Dirichlet prior $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M) \sim \text{Dir}\,(\kappa_0, \ldots, \kappa_0)$, where the concentration parameter κ_0 is set at a very small value, so that the marginal densities of each π_j are spiked around the values zero and one, $j = 1, \ldots, M$. This approach results in estimating a reduced number of hidden states, denoted as $\hat{M}$, which is significantly less than M. Thus, unnecessary hidden states are effectively removed from the posterior distribution. The hyperparameter κ_0 plays a crucial role. Here, we set $\kappa_0 = 0.001$ following the recommendation by Malsiner-Walli et al. (2016). In Section 2.4, we propose to estimate the number of hidden states based on the most frequent number of active states

during MCMC sampling. By setting a large value for M, our approach provides a simple and automated framework for estimating the number of hidden states, without relying on computations of marginal likelihoods, post-MCMC model selection criteria, or reversible-jump MCMC.

2.2 Graphical Horseshoe Priors for the Precision Matrices

To induce prior sparsity in the state-specific precision matrices $\boldsymbol{\Omega}_j$'s, we employ the graphical horseshoe (GHS) prior proposed by Li et al. (2019). This prior utilizes normal scale mixtures with half-Cauchy hyperpriors for the off-diagonal entries of the precision matrix while using uninformative priors for its diagonal elements. Specifically,

$$
\begin{aligned}
\omega_{j,ii} &\propto 1, \\
\omega_{j,il\,:i<l} &\sim \mathcal{N}(0, \lambda_{j,il}^2 \tau_j^2), \\
\lambda_{j,il\,:i<l} &\sim C^+(0,1), \\
\tau_j &\sim C^+(0,1),
\end{aligned}
$$

for $i, l = 1, \ldots, D$, and $j = 1, \ldots, M$. The global shrinkage parameter τ_j plays a crucial role in promoting sparsity across the entire matrix $\boldsymbol{\Omega}_j$, by shrinking the estimates of all the off-diagonal values towards zero. On the other hand, the local shrinkage parameters $\lambda_{jil:i<l}$ allow to preserve the magnitudes of the nonzero off-diagonal elements, ensuring that the element-wise biases do not become too large. This combination of global and local shrinkage enables the GHS prior to induce sparsity in the precision matrices while capturing the relevant dependencies between the elements.

We complete the prior specification on the emission distributions by assuming Gaussian priors on the state-specific means, that is, $p(\boldsymbol{\mu}_j) \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{R}_0^{-1})$, for $j = 1, \ldots, M$.

2.3 Markov Chain Monte Carlo Algorithm

We now outline the MCMC algorithm we designed for posterior inference. For notational convenience, we collect all parameters except $\boldsymbol{\gamma}$ as the set $\boldsymbol{\theta} = \{\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\tau}, \boldsymbol{\Lambda}\}$ with $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_j\}_{j=1}^M$ and $\boldsymbol{\Lambda}_j = \{\lambda_{j,il}^2\}$ the matrices of local shrinkage parameters in the GHS prior,

$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_M)$ the global parameters, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M)$, and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_M)$. We then write the posterior distribution of $\boldsymbol{\theta}$ conditional upon the current value of $\boldsymbol{\gamma}$ as

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{y}, \boldsymbol{\gamma}) \propto \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\gamma}) \, p(\boldsymbol{v}, \boldsymbol{z}) \, p(\boldsymbol{\pi}) \, p(\boldsymbol{\mu}) \, p(\boldsymbol{\Omega}, \boldsymbol{\tau}, \boldsymbol{\Lambda}), \tag{5}$$

where the conditional likelihood is factorized as

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{\gamma}) = \prod_{t=P+1}^{T} p(\gamma_t \,|\, \gamma_{t-1:t-P}, \boldsymbol{v}, \boldsymbol{z}, \boldsymbol{\pi}) \, p(\boldsymbol{y}_t \,|\, \gamma_t, \boldsymbol{\mu}, \boldsymbol{\Omega}) \tag{6}$$

and where the joint prior $p(\boldsymbol{v}, \boldsymbol{z})$ of the indicator variables and the stick-breaking weights can be expressed as

$$p(\boldsymbol{v}, \boldsymbol{z}) = p(v_0) \prod_{j=1}^{\hat{P}-1} p(v_j | z_j) \prod_{j=0}^{\hat{P}} p(z_{j+1} | \boldsymbol{v}_{0:j}), \tag{7}$$

with

$$p(v_j | z_j) \propto \text{Beta}(a_v, b_v)^{z_j}, \qquad p(z_{j+1} | \boldsymbol{v}_{0:j}) \propto \text{Bern}(1)^{z_j} \, \text{Bern}(\xi_j)^{1-z_j}, \tag{8}$$

and the conditioning on $\boldsymbol{v}_{0:j-1}$ induced by the cumulative shrinkage parameter ξ_j.

Since the posterior distribution is not available in closed form, we develop a Gibbs sampler that alternates between: (i) drawing the stick-breaking weights $\boldsymbol{v}$ and auxiliary indicators $\boldsymbol{z}$. For this, we design a Metropolis-Hastings algorithm similar to Savitsky et al. (2011), that cleverly avoids having to deal with the changing dimensions of the parameter space via a joint update of the indicators and the weights; (ii) updating the innovation probabilities $\boldsymbol{\pi}$ related to the sparsity-inducing Dirichlet prior; (iii) sampling the multivariate sparse emission parameters, i.e. the mean vectors in $\boldsymbol{\mu}$, the precision matrices in $\boldsymbol{\Omega}$ and the global and local shrinkage parameters $\boldsymbol{\tau}$ and $\boldsymbol{\Lambda}$; (iv) updating the latent state sequence $\boldsymbol{\gamma}$, through a forward-backward algorithm, which is significantly accelerated by the proposed zero-inducing cumulative shrinkage prior formulation. We now describe these updates in full detail.

- **Update $\boldsymbol{z}$ and $\boldsymbol{v}$**: We perform a joint update of the indicators $\boldsymbol{z}$ and weights $\boldsymbol{v}$ by designing a Metropolis-Hastings sampler with *birth* and *death* moves, that increase or decrease the order of the DAR process by one. Formally, let us define the current number of active components $\hat{P}^{curr}$, stick-breaking weights $\boldsymbol{v}^{curr} = (v_0, v_1, \ldots, v_{\hat{P}^{curr}-1}, 1)$, and indicator variables $\boldsymbol{z}^{curr} = (0, 0, \ldots, 0, 1)$, of dimensions $\hat{P}^{curr} + 1$ and $\hat{P}^{curr}$, re-

spectively; note that $\boldsymbol{z}^{curr} = 1$ when $\hat{P}^{curr} = 1$. A new vector of indicators $\boldsymbol{z}$ is drawn by proposing at random one of the following two moves:

(i) *birth move:* Set $\hat{P}^{prop} = \hat{P}^{curr} + 1$ and construct $\boldsymbol{z}^{prop}$ from $\boldsymbol{z}^{curr}$ by adding a zero entry; for this move, the proposed vector of weights is constructed as $\boldsymbol{v}^{prop} = (v_0, v_1, \ldots, v_{\hat{P}^{curr}-1}, v_{\hat{P}^{prop}-1}, 1)$ with $v_{\hat{P}^{prop}-1}$ drawn from the prior, i.e. $v_{\hat{P}^{prop}-1} \sim \mathrm{Beta}(a_v, b_v)$, and $v_{\hat{P}^{prop}}$ set equal to one. This move is accepted or rejected with probability

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{v}^{prop}, \boldsymbol{z}^{prop} \,|\, \boldsymbol{\gamma}, \boldsymbol{y}, \cdot)}{p(\boldsymbol{v}^{curr}, \boldsymbol{z}^{curr} \,|\, \boldsymbol{\gamma}, \boldsymbol{y}, \cdot)} \frac{1}{\mathrm{Beta}(v_{\hat{P}^{prop}-1}|a_v, b_v)}\right\}, \qquad (9)$$

where the joint posterior distribution $p(\boldsymbol{z}, \boldsymbol{v}|\cdot)$ is easily available by appropriate conditioning of the relevant variables in Eq. (5) and (6), i.e.

$$p(\boldsymbol{v}, \boldsymbol{z} \,|\, \boldsymbol{\gamma}, \boldsymbol{y}, \cdot) \propto p(\boldsymbol{v}, \boldsymbol{z}) \prod_{t=P+1}^{T} p(\gamma_t \,|\, \gamma_{t-1:t-P}, \boldsymbol{v}, \boldsymbol{z}, \boldsymbol{\pi}), \qquad (10)$$

with the DAR probabilities $p(\gamma_t|\cdot)$ defined as in Eq. (2), recalling that $\boldsymbol{\phi}$ is a by-product of $\boldsymbol{v}$ and $\boldsymbol{z}$ using the formulation presented in Eq. (4).

(ii) *death move:* Set $\hat{P}^{prop} = \hat{P}^{curr} - 1$ and construct $\boldsymbol{z}^{prop}$ from $\boldsymbol{z}^{curr}$ by removing a zero entry; here, $\boldsymbol{v}^{prop}$ is obtained from $\boldsymbol{v}^{curr}$ by replacing the component $v_{\hat{P}^{curr}-1}$ with a one and setting $v_{\hat{P}^{curr}}$ equal to zero, namely $\boldsymbol{v}^{prop} = (v_0, v_1, \ldots, v_{\hat{P}^{curr}-2}, 1)$. This is move is accepted or rejected with probability the inverse of Eq (9) with the appropriate change of labeling.

After each death/birth move, to enhance the mixing efficiency of the MCMC algorithm, we further update each component of the weight vector $\boldsymbol{v}$ using a one-at-a-time slice sampler (Neal, 2003). Slice sampling is particularly advantageous for drawing samples from one-dimensional conditional distributions within a Gibbs sampling framework. Here, we focus on multivariate targets by iteratively sampling each variable. In particular, we obtain posterior samples from the target function $p(v_j \,|\, \boldsymbol{v}_{-j}, \cdot)$, for $j = 0, \ldots, \hat{P} - 1$, where $\boldsymbol{v}_{-j} = (v_0, \ldots, v_{j-1}, v_{j+1}, \ldots, v_{\hat{P}-1})$.

We remark here that the order P of the DAR process is not modeled as a random vari-

able, but rather inferred directly from $\boldsymbol{z}$ and $\boldsymbol{v}$, eliminating the necessity of employing a trans-dimensional MCMC sampler (Green, 1995).

- **Update $\boldsymbol{\pi}$:** We update the components of $\boldsymbol{\pi}$ with a one-at-a-time slice sampler, drawing samples from the target function $p(\pi_l \,|\, \boldsymbol{\pi}_{-l}, \cdot)$, for $j = 0, \ldots, M_{max} - 1$, where $\boldsymbol{\pi}_{-l} = (\pi_0, \ldots, \pi_{l-1}, \pi_{l+1}, \ldots, \pi_{M_{max}})$. Note that $\pi_{M_{max}}$ is automatically obtained from its simplex, i.e. $\pi_{M_{max}} = 1 - \sum_{l=0}^{M_{max}-1} \pi_l$.

- **Update $\boldsymbol{\Omega}_j$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\tau}$:** We use the augmented block Gibbs sampler method proposed by Li et al. (2019). We center the observations belonging to each state to its current value of the emission mean, $\boldsymbol{\mu}_j$, and consider a modified set of observations denoted as $\tilde{\boldsymbol{Y}}_j = \{\boldsymbol{y}_t - \boldsymbol{\mu}_j : \gamma_t = j\}$. By doing this, we can closely follow the scheme proposed by Li et al. (2019), which assumes zero-mean multivariate normal distributions. We apply the Gibbs sampler independently for each state j, from $j = 1$ to $M_{\max}$ and subsequently update the global shrinkage parameter τ_j and its corresponding augmented parameter ξ_j. We refer to the reader to Algorithm 1 of Li et al. (2019), for the details of the GHS sampler.

- **Update $\boldsymbol{\mu}_j$:** We sample the mean vectors $\boldsymbol{\mu}_j$ from the corresponding full conditional, as is typical in the context of Gaussian Bayesian regression settings (see e.g. Gelman et al. 1995). The posterior distribution is given by $\boldsymbol{\mu}_j|_{\boldsymbol{\Omega}_j, \boldsymbol{y}, \cdot} \sim \mathcal{N}(\boldsymbol{\mu}_j^\star, \boldsymbol{\Omega}_j^\star)$, where

$$\boldsymbol{\Omega}_j^{\star-1} = \boldsymbol{R_0} + N_j \boldsymbol{\Omega}_j, \quad \text{and} \quad \boldsymbol{\mu}_j^\star = \boldsymbol{\Omega}_j^\star (\boldsymbol{R_0}\boldsymbol{\mu}_0 + N_j \boldsymbol{\Omega}_j \boldsymbol{Y}_j), \tag{11}$$

and $\boldsymbol{Y}_j$ denotes the $(N_j \times D)$-dimensional matrix of observations assigned to state j, with N_j the corresponding number of observations belonging to that regime.

- **Update $\boldsymbol{\gamma}$:** We update the sequence of latent states $\boldsymbol{\gamma}$ with a block-wise approach that adapts the forward-backward procedure employed by Fox et al. (2011) and Hadj-Amar et al. (2021) to take into account temporal dynamics that extend beyond a simple Markovian structure. Conditional upon $\boldsymbol{\phi}$, $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, we harness the dependence structure of the DAR and develop an iterative sampling scheme based on the following

representation of the posterior distribution of the hidden states

$$p(\boldsymbol{\gamma}\,|\boldsymbol{y},\,\cdot) = p(\gamma_1|\boldsymbol{y},\cdot)\,p(\gamma_2|\gamma_1,\boldsymbol{y},\cdot)\ldots p(\gamma_{\hat{P}}|\gamma_{1:\hat{P}-1},\boldsymbol{y},\cdot)\prod_{t=\hat{P}+1}^{T}p(\gamma_t|\gamma_{t-1:t-\hat{P}},\boldsymbol{y},\cdot). \quad (12)$$

Under this factorization, we first sample $\gamma_1 \sim p(\gamma_1|\boldsymbol{y},\cdot)$, then, conditioning on the value of γ_1, we draw $\gamma_2 \sim p(\gamma_2|\gamma_1,\boldsymbol{y},\cdot)$, and so on, where we update $\gamma_t \sim p(\gamma_t|\gamma_{t-1:t-\hat{P}},\boldsymbol{y},\cdot)$, given the previous sampled states $\gamma_{t-1:t-\hat{P}}$. Assuming $M = M_{\max}$, the general form for the conditional posterior distribution of the states in Eq. (12) is given by

$$p(\gamma_t = j_0|\gamma_{t-1} = j_1,\ldots,\gamma_{t-\hat{P}} = j_{\hat{P}},\boldsymbol{y},\cdot) \propto \eta_{\{j_{\hat{P}},\ldots,j_1,j_0\}}\,p(\boldsymbol{y}_t|\gamma_t = j_0,\boldsymbol{\mu},\boldsymbol{\Omega})\,\beta_{t+1}(j_0),$$
$$(13)$$

for $t = \hat{P}+1,\ldots,T$, and $j_l \in \{1,\ldots,M\}$, $l = 1,\ldots\hat{P}$, where $\eta_{\{j_{\hat{P}},\ldots,j_1,j_0\}}$ are the DAR probabilities of selecting state j_0, given previous values $j_1,\ldots,j_{\hat{P}}$, as defined in Eq. (2), and $p(\boldsymbol{y}_t|\cdot)$ are the multivariate spiked Gaussian emission densities specified in Eq. (1). Here, we define the *backward messages* $\beta_t(j_1) = p(\boldsymbol{y}_{t:T}|\gamma_{t-1} = j_1,\cdot)$, as the probability of the partial observation sequence from time t to T given the state j_1 at time $t-1$, conditioned on all the other parameters. These messages can be recursively expressed as follows (see Proposition 2, Supplementary Material)

$$\beta_t(j_1) = \underbrace{\sum_{j_{\hat{P}}=1}^{M}\cdots\sum_{j_2=1}^{M}\sum_{j_0=1}^{M}}_{\hat{P}\text{ times}}\eta_{\{j_{\hat{P}},\ldots,j_2,j_1,j_0\}}p(\boldsymbol{y}_t|\gamma_t = j_0,\boldsymbol{\mu},\boldsymbol{\Omega})\beta_{t+1}(j_0),\quad t \leq T, \quad (14)$$

with $\beta_{T+1}(\cdot) = 1$. Our zero-inducing formulation for the DAR probabilities, described in Section 2.1.1, allows a significant speed-up of the proposed sampler, since in Eq. (14) we restrict summations to the active DAR terms only, rather than using the entire multi-dimensional array $\boldsymbol{\eta}$. Additionally, we specify the initial DAR probabilities $\eta_{\{\cdot\}}$ in Eq. (12) and (13) to be uniformly distributed.

Following similar practices as in Fox et al. (2011) and Hadj-Amar et al. (2021), we update only emission parameters for those states that have at least 1% of the assignments, while for those states that do not satisfy this condition we draw the corresponding emission parameters from their priors. For the GHS prior, we draw the diagonal entries of the precision matrix

using a diffuse prior $\omega_{ii} \sim U(0, 100)$.

We acknowledge that the proposed Bayesian procedure may be susceptible to the *label switching* problem (Jasra et al., 2005) due to the invariance of the likelihood (6) under permutations of the mixture components' labeling. To mitigate this issue, we adopt a post-processing approach using the Equivalence Classes Representatives (ECR) algorithm, initially introduced by Papastamoulis and Iliopoulos (2013) and later improved by Rodríguez and Walker (2014). The core idea of the ECR algorithm is to categorize analogous allocation vectors as mutually exclusive solutions to the label switching problem. In this context, two allocation vectors are considered analogous if one can be obtained from the other merely by permuting its labels. The ECR procedure divides the allocation vectors into analogous categories and identifies a representative from each category. Consequently, during post-processing, the ECR algorithm identifies the permutation corresponding to each MCMC iteration. This permutation is then applied to reorder the matching allocation with the aim of aligning it with the representative of its category.

2.4 Posterior Inference

After obtaining the (possibly relabeled) MCMC output, we first estimate the number of active DAR components by computing the posterior probabilities $p\,(\hat{P} = p\,|\cdot)$, $p = 1, \ldots, P_{max}$ and then identify the posterior mode as the value of $\hat{P}$ that maximizes such posterior probabilities. Similarly, to estimate the number of hidden states, we first calculate the posterior probabilities $p\,(M = m\,|\cdot)$ for $m = 1, \ldots, M_{\max}$ as

$$P(M = m|\cdot) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}(\hat{M}^{(s)} = m), \qquad \text{where} \quad \hat{M}^{(s)} = \sum_{j=1}^{M_{max}} \mathbb{1}\left(N_j^{(s)} \neq 0\right), \qquad (15)$$

with N_j the number of observations assigned to state j, and where the superscript (s) indicates the MCMC iteration for $s = 1, \ldots, S$. We then calculate the posterior mode to obtain the final estimate of the number of hidden states, $\hat{M}$. Next, conditional upon these estimates, we perform posterior inference on the model parameters $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\mu}}$, and $\hat{\boldsymbol{\Omega}}$ by averaging their sampled values across the MCMC iterations with number of hidden states $\hat{M}$ and DAR order $\hat{P}$.

As for inference on the sequence of latent states, we perform both global and local decoding. *Global decoding* refers to the determination of the most likely sequence of the entire vector of latent states $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_T)$. We obtain such a maximum a posteriori (MAP) estimate by using a variant of the scheme described in equation (12). Given the estimated parameters $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\mu}}$, and $\hat{\boldsymbol{\Omega}}$, we iteratively maximize the posterior distribution of the states, where at each time step t, we compute $\hat{\gamma}_t = \arg\max_{j=1,\ldots,\hat{M}} p(\gamma_t = j | \hat{\gamma}_{t-1:t-\hat{P}}, \boldsymbol{y}, \cdot)$. In contrast, *local decoding* of the hidden state at time t, $p(\gamma_t = j | \boldsymbol{y}, \cdot)$ refers to the determination of that state which is most likely at that time. This is achieved using

$$p(\gamma_t = j \,|\, \boldsymbol{y}, \cdot) \propto p(\gamma_t = j, \boldsymbol{y}_{1:t} | \cdot) p(\boldsymbol{y}_{t+1:T} | \gamma_t = j, \cdot) = \alpha_{t+1}(j)\beta_{t+1}(j), \tag{16}$$

where the backward messages are defined as $\beta_t(j) = p(\boldsymbol{y}_{t:T} | \gamma_{t-1} = j, \cdot)$ and the forward messages are expressed as $\alpha_t(j) = p(\boldsymbol{y}_{1:t-1}, \gamma_{t-1} = j | \cdot)$. In order to leverage the recursive nature of these messages, we also defined the DAR-forward messages $\alpha_t(j_1, \ldots, j_{\hat{P}}) = p(\boldsymbol{y}_{1:t-1}, \gamma_{t-1} = j_1, \ldots, \gamma_{t-\hat{P}} = j_{\hat{P}} | \cdot)$. Further details and the validity of these expressions are provided in the Supplementary Material.

For inference on the graphs, since the GHS approach is a shrinkage procedure, and thus it does not estimate the entries as exact zeros, we utilize posterior credible intervals to perform variable selection, as suggested by Li et al. (2019). Specifically, we use a 95% interval from the estimated posterior distribution of each off-diagonal element of the precision matrices, so that if the interval corresponding to an entry includes zero, that entry is assessed as non active. Note that Li et al. (2019) employed a 50% symmetric credible interval, arguing that such a procedure would have conservative properties, and would reduce false negatives while controlling for false positive. However, in our experiments, a 95% interval seemed to outperform the choice suggested by Li et al. (2019).

3   Simulation Studies

We investigate the performance of our proposed methodology using simulated data. We wish to assess the ability to recover the true generating DAR probabilities and the emissions parameters, as well as the numbers of autoregressive probabilities and hidden states.

3.1 Data Generation

We first consider a simulation framework where the data were generated with underlying time-varying means and structured precision matrices. We generated 30 distinct datasets from model (1)-(2), each consisting of $D = 15$-dimensional time series of length $T = 2,000$, and assumed $M = 5$ latent states and a DAR of order $P = 2$. The autoregressive probabilities were set to $\boldsymbol{\phi} = (0.1, 0.75, 0.15)$ and the innovations to $\boldsymbol{\pi} = (0.6, 0.1, 0.1, 0.1, 0.1)$. For each state j, the emission means $\boldsymbol{\mu}_j$ were independently simulated from a multivariate Gaussian distribution with mean vector $\boldsymbol{b}_0 = (-\frac{5}{D}, \ldots, -\frac{1}{D}, 0, \frac{1}{D}, \ldots, \frac{5}{D})$ and identity matrix as the covariance matrix, i.e. $\boldsymbol{\mu}_j \sim \mathcal{N}(\boldsymbol{b}_0, \boldsymbol{I}_D)$, and where the simulated components of these vectors were randomly shuffled. The state-specific precision matrices $\boldsymbol{\Omega}_j$ were assumed to be sparse with diagonal elements fixed to one and off diagonal elements constructed using the following five structures:

(i) Identity graph: this structure assumes that the components are independent, i.e. the off-diagonal elements are all set to zero.

(ii) Star graph: a configuration similar to the identity matrix, except for the first row and first column, whose elements are set to $\omega_{il} = -\frac{1}{D}$ if $i = 1$ or $l = 1$, and 0 otherwise.

(iii) Hub graph: this structure is organized into five blocks (hubs) of the same size. For any $l \neq i$ in the same block as i we specify $\omega_{il} = \omega_{li} = -\frac{2}{\sqrt{D}}$, and 0 otherwise.

(iv) AR(2) graph: in this structure the precision matrix displays an autoregressive pattern of order two over the main diagonal. The entries are specified as $\omega_{il} = \frac{1}{2}$ if $l = i - 1, i + 1$, $\omega_{il} = \frac{1}{4}$ if $l = i - 2, i + 2$, and 0 otherwise.

(v) Random sparse graph: for this setting, the precision matrix is generated by randomly selecting $\lfloor \frac{3}{2}D \rfloor$ off-diagonal entries, and drawing each ω_{jl} uniformly from the interval $[-1.0, -0.4] \cup [0.4, 1.0]$, while the diagonal elements are fixed at 1, and the other entries at 0. Each of the off-diagonal element is then divided by the sum of the off-diagonal elements in its row, and then the matrix is averaged with its transpose, to produce a symmetric, positive definite, matrix.

Partial correlation matrices corresponding to these five scenario are displayed in Figure 1 (top row). A single realization from the simulation setting described in this section is shown in Figure 2 (top panel), with vertical colored bands representing the true underlying state
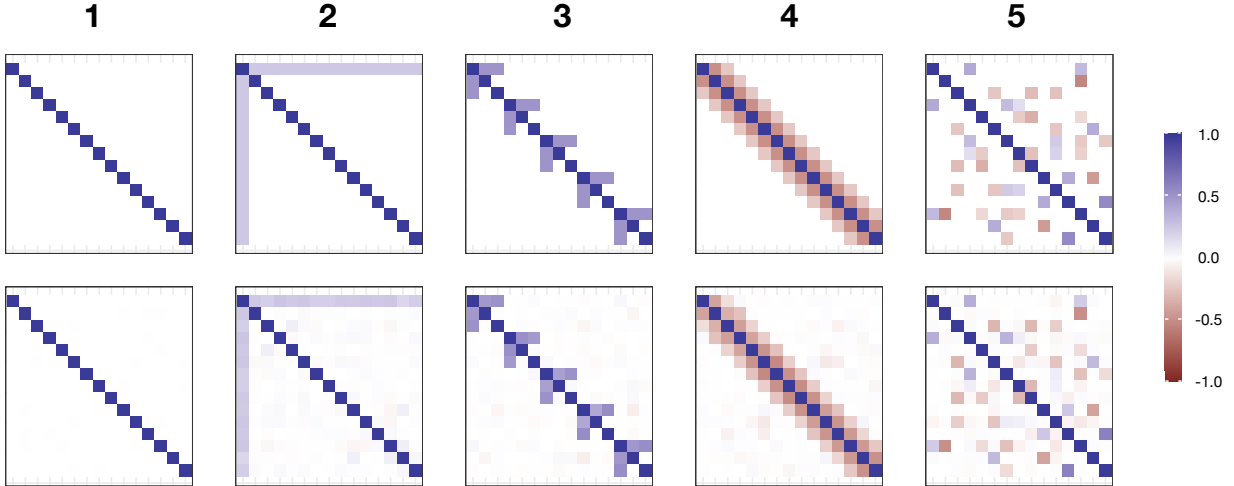
Figure 1: **Simulation study.** (top) true state-specific partial correlation matrices; (bottom) estimated state-specific partial correlation matrices. These results are conditioned upon the estimated modal number of states and autoregressive order.

sequence. Here, we have further scaled the time series realization, independently for each dimension $d = 1, \ldots, D$, in such a way that the corresponding standard deviation of those observations $\{y_{td}\}_{t=1}^{T}$ is equal to one. We note that the partial correlations are invariant under a change of scale and origin, allowing a meaningful comparison between true and estimated values of these matrices.

3.2   Parameter Settings

Results reported below were obtained by fixing the maximum number of states to $M_{max} = 10$ and the maximum DAR order to $P_{max} = 5$. The DAR hyperparameters were chosen as $a_0 = 1, b_0 = 10$, and $a_\nu = 10, b_\nu = 1$, so that the prior probabilities of innovation and autoregression were driven towards zero and one, respectively. The hyperparameters for the emission vector mean were specified as $\boldsymbol{\mu}_0 = \boldsymbol{0}$ and $\boldsymbol{R}_0 = (1/10)\,\boldsymbol{I}_D$, so that the mean components were a priori independent across different dimensions and with fairly large variance, hence reflecting weakly informative beliefs.

Initial values of the MCMC sampler were chosen as follows: the DAR parameters were sampled from the prior; the Gaussian emission means were fixed to the centers of a k-mean clustering and the covariance matrices were set to the identity. The GHS parameters were set to one. MCMC chains were run for 4,000 iterations, with 1,200 iterations discarded
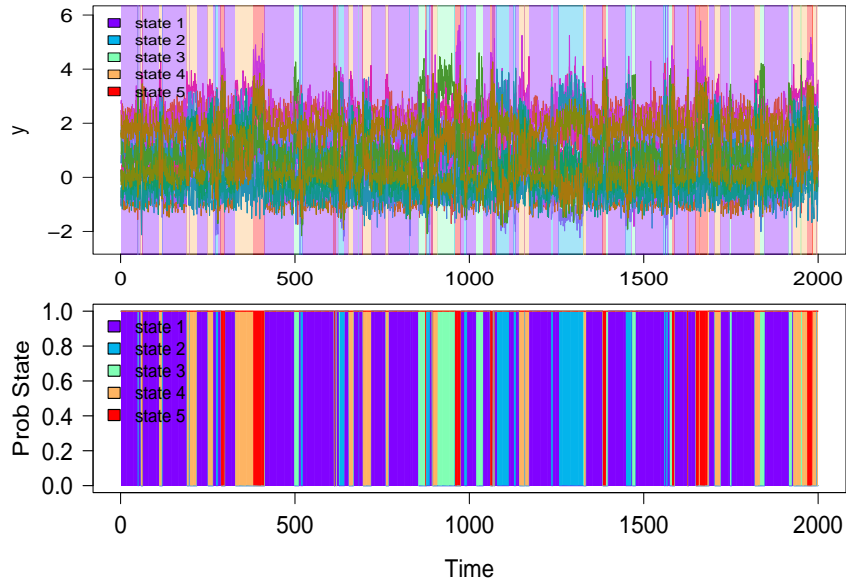
Figure 2: **Simulation Study.** (top) time series realization (lines), with each dimension represented by a different colored line; vertical colored bands represent the true underlying state sequence; (bottom) estimated time-varying probability plot.

as burn-in. The algorithm took approximately 10 minutes to run, for each simulated time series, with a program written in Julia 1.6 on an Intel® Core™ i5 2GHz Processor 16GB RAM. We verified convergence of the MCMC sampler by: (i) analyzing the trace plots of the parameters, e.g. the mean of the multivariate spiked Gaussian emissions, observing no pathological behavior; (ii) storing the values of the overall likelihood (Eq. 6) and plotting the corresponding trace, noting that it reached a stable regime; (iii) verifying the Heidelberger and Welch's convergence diagnostic (Heidelberger and Welch, 1981) for the likelihood trace. We report some of the results in the Supplementary Material.

3.3   Results

Our approach consistently estimated the correct number of states $\hat{M} = 5$ as the mode of the posterior distribution and the number of active DAR probabilities as $\hat{P} = 2$ with high posterior probability, on all simulated replicates. For a single replicate, in Figure 1 (bottom row) we show the estimates of the state-specific partial correlation matrices, conditioned upon the modal number of states and modal number of DAR parameters. Our approach successfully retrieves the distinct patterns of the true graphs. Figure 2 (bottom panel) displays a time-varying probability plot, namely the local decoding of the hidden state at
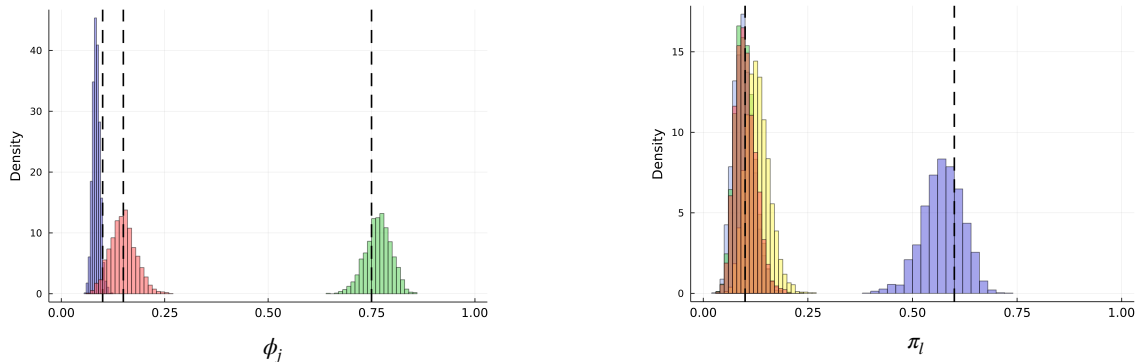
18

Figure 3: **Simulation Study.** Posterior histograms of the DAR parameters. (left) autoregressive probabilities ϕ_l, $l = 0, \ldots \hat{P}$; (right) innovation probabilities π_j, $j = 1, \ldots, \hat{M}$. Dotted vertical lines denote true parameters. These results are conditioned upon the modal number of states and autoregressive order. The figure is best viewed in colors.

time t, $p(\gamma_t = j \,|\, \boldsymbol{y}, \cdot)$, $j = 1, \ldots, \hat{M}$, as described in Section 2.4; these plots are constructed by plotting the local probabilities (which add to 1) cumulatively for each t, where each state is associated with a different color. The proposed approach appears to correctly retrieve the true latent state sequence. Additionally, Figure 3 displays the posterior histograms of autoregressive and innovation probabilities, with dotted vertical lines denoting the true generating values. Our proposed method appears to provide a good match between true and estimated values for the DAR parameters.

Next, we investigated the performance of our proposed approach over the 30 replicated datasets and performed comparisons with alternative methods. We focused on the recovery of the state-specific precision matrices and compare the proposed methodology, which will be referred to as `sggmDAR`, to two alternative approaches. For the first approach, which we call `mvHMM`, we fit a Bayesian multivariate HMM with Gaussian emissions, with a Normal inverse-Wishart prior on the state-specific emission parameters $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \sim NIW(\boldsymbol{\mu}_0, \boldsymbol{S}_0/\kappa_0; \nu_0, \boldsymbol{S}_0)$, where the hyper-parameters were specified in a weakly informative manner, i.e. $\boldsymbol{\mu}_0 = \boldsymbol{0}, \kappa_0 = 0.1, \nu_0 = D + 2, \boldsymbol{S}_0 = \boldsymbol{I}_D$. The number of states was set to five (i.e. the truth). The transition probabilities were assumed symmetric Dirichlet distributed, with concentration parameter equal to one. Since this HMM approach does not estimate precision entries as exact zeros, we once again used 95% posterior credible intervals to perform edge selection. In the second approach, named `glassoSlide`, we followed Allen et al. (2014) and employed a sliding window to compute time-varying sparse inverse precision matrices via graphical

lasso (Friedman et al., 2008) using the R package `glasso`. In order to obtain an estimate of the latent state sequence, the windowed estimates of the precision matrices were then utilized as feature vectors in the k-means clustering algorithm. Finally, sparse state-specific precision matrices were estimated by applying graphical lasso to the MLE estimates of the covariances relative to the set of observations corresponding to each distinct state. The number of states was set to five (i.e. the truth), while the size of the sliding window and the magnitude of the penalization parameter were selected in such a way to maximize model selection performances averaged over the different states.

		Identity	Star	Hub	AR(2)	Random
	sggmDAR	1.0 (0.0)	0.995 (0.007)	1.0 (0.002)	0.969 (0.021)	0.958 (0.027)
Acc	mvHMM	0.969 (0.032)	0.933(0.039)	0.933 (0.088)	0.822 (0.138)	0.883 (0.110)
	glassoSlide	0.993 (0.018)	0.835 (0.047)	0.805 (0.053)	0.690 (0.031)	0.714 (0.031)
	sggmDAR	1.0 (0.0)	0.995 (0.007)	1.0 (0.0)	0.995 (0.007)	0.997 (0.005)
Spec	mvHMM	0.969 (0.032)	0.942 (0.022)	0.933 (0.093)	0.840 (0.158)	0.924 (0.100)
	glassoSlide	0.993 (0.018)	0.937 (0.052)	0.879 (0.065)	0.899 (0.040)	0.869 (0.044)
	sggmDAR	-	0.979 (0.028)	0.998 (0.011)	0.919 (0.055)	0.868 (0.087)
MCC	mvHMM	-	0.734 (0.212)	0.765 (0.210)	0.591 (0.297)	0.630 (0.365)
	glassoSlide	-	0.140 (0.198)	-0.019 (0.040)	-0.017 (0.101)	-0.004 (0.073)
	sggmDAR	-	0.981 (0.025)	0.998 (0.011)	0.936 (0.045)	0.883 (0.083)
F1	mvHMM	-	0.761 (0.194)	0.747 (0.243)	0.683 (0.254)	0.683 (0.332)
	glassoSlide	-	0.193 (0.189)	0.081 (0.051)	0.125 (0.078)	0.152 (0.062)
	sggmDAR	-	0.994 (0.020)	0.996 (0.020)	0.895 (0.071)	0.807 (0.124)
Sens	mvHMM	-	0.874 (0.235)	0.927 (0.245)	0.772 (0.267)	0.732 (0.356)
	glassoSlide	-	0.171 (0.174)	0.103 (0.072)	0.089 (0.063)	0.126 (0.059)
	sggmDAR	0.002 (0.001)	0.034 (0.008)	0.021 (0.007)	0.049 (0.011)	0.042 (0.011)
RMSE	mvHMM	0.019 (0.012)	0.073 (0.013)	0.080 (0.035)	0.116 (0.057)	0.087 (0.041)
	glassoSlide	0.001 (0.004)	0.097 (0.003)	0.162 (0.002)	0.205 (0.002)	0.155 (0.004)

Table 1: **Simulation Study.** Accuracy, specificity, Matthew correlation coefficient (MCC), F1 score, sensitivity, and residual mean squared error (RMSE) of precision matrix estimates, for each state $j = 1, \ldots, \hat{M}$. Standard deviations over the 30 simulations are displayed in parentheses. Results are reported for `sggmDAR`, `mvHMM` and `glassoSlide`. Results for `sggmDAR` are conditioned upon the modal number of states and autoregressive order. A hyphen is used for those metrics that cannot be computed due to the structure of the underlying truth (e.g. TP+FN = 0).

To assess model selection performances we computed accuracy, sensitivity, specificity, $F1$-score and Matthew correlation coefficient (MCC), for each regime $j = 1, \ldots, \hat{M}$. In addition, to evaluate estimation accuracy, we calculated residual mean squared error (RMSE) of state-

specific off-diagonal entries of the precision matrices as $\text{RMSE}_j = \sqrt{\frac{1}{D} \sum_{i<l} \left(\omega_{jil} - \hat{\omega}_{jil} \right)^2}$.
Results from these measures are summarized in Table 1. Note that MCC, F1, and sensitivity for the Identity state are not presented since these metrics cannot be computed due to the structure of the underlying truth (e.g. TP+FN = 0). Overall, `sggmDAR` produced the best results both in estimation accuracy and model selection performances. Though accuracy and specificity of `glassoSlide` are somewhat high, this frequentist method is by far the worse, as illustrated by very low MCC scores. We remark that, while both `mvHMM` and `glassoSlide` need to specify the number of states in advance, our proposed approach produces an estimate of this parameter. In the Supplementary Material, we further investigate the performance of our proposed methodology for data-generating emissions with zero-mean, i.e. $\boldsymbol{\mu}_j = \mathbf{0}$, for $j = 1, \ldots, M$. The results confirm the superiority of our approach over both `mvHMM` and `glassoSlide` in terms of estimation and model selection accuracy. Additionally, the Supplementary Material contains a sensitivity analysis study that focuses on examining the impact of the hyperparameters a_v, b_v associated with the zero-inducing cumulative shrinkage prior (4). Results show that, for small and moderate T, different combinations of the hyperparameters may yield varying dynamics of the process. However, as T increases, such differences are not noticeable.

3.4 Simulations for Varying T and D

Next, we investigated performance for different values of the sample size T. For this, we generated 30 distinct time series for different sample sizes, $T = 100, 500, 1000, 5000, 10000$, assuming $M = 3$ states and DAR order $P = 2$, with autoregressive probabilities specified as $\boldsymbol{\phi} = (0.2, 0.5, 0.3)$ and innovations set to $\boldsymbol{\pi} = (0.5, 0.3, 0.2)$. The emission means were generated as in the main simulation above, whereas the precision matrices were constructed using patterns (i), (iii) and (v) from Section 3.1. Here, to perform Bayesian inference we fixed the maximum number of states to $M_{max} = 3$ and maximum DAR order to $P_{max} = 2$. The hyperparameters were specified as in Section 3.1. Figure 4 displays boxplots over the 30 simulations of the posterior distributions of $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$, for the different values of T, conditioned upon the the modal number of states and autoregressive order. As it was to be expected, estimates for both ϕ_j and π_j showed larger variability for small sample sizes. Conversely,
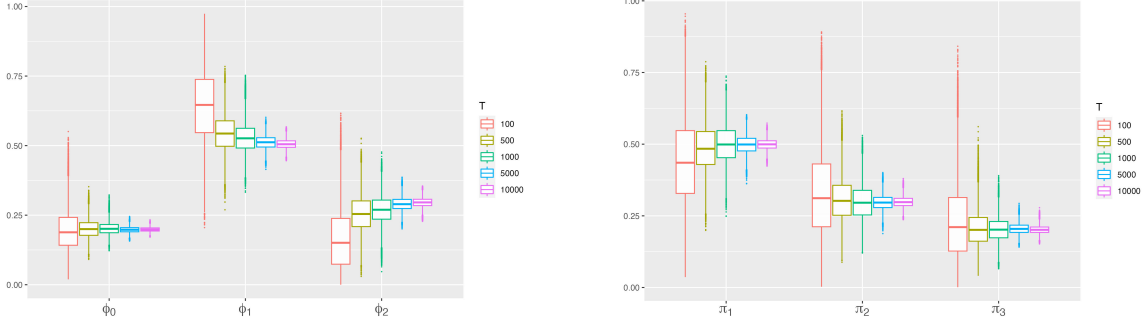
Figure 4: **Simulation Study.** Boxplots over 30 simulations of posterior distributions for (left) autoregressive probabilities ϕ_l, $l = 0, \ldots \hat{P}$, and (right) innovation probabilities π_j, $j = 1, \ldots, \hat{M}$, for different sample size $T = 100, 500, 1000, 5000, 10000$. These results are conditioned on the modal number of states and autoregressive order.

as T increases, the generating DAR dynamics became more apparent, and our inference procedure is indeed able to retrieve the correct parameters more accurately, for most cases.

Finally, we explored the performance of our approach in a scenario where the dimension D of the data is large. Here, we focused on assessing the ability of our proposed method in recovering the number of states, number of DAR parameters, and true sparse precision matrices. We simulated 30 time series, each consisting of $D = 100$-dimensional time series of length $T = 2000$, with $M = 3$ and $P = 2$, and with the emissions generated as in Section 3.1 and the precision matrices for the three states specified as Identity, Hub (with four blocks) and Random, respectively. The innovations were set to $\boldsymbol{\pi} = (0.6, 0.2, 0.2)$, while the rest of the data-generating parameters were set as in Section 3.1. Here we report results obtained by specifying the hyper-parameters as described in Section 3.1 and by running MCMC chains for 4,000 iterations, with 1,200 iterations discarded as burn-in.

As in the previous simulations, the correct number of states and DAR order were identified as those with the highest posterior probability for all replicated datasets. In the Supplementary Material, we report model selection and estimation accuracy performances for the off-diagonal component of the precision matrices, for sggmDAR, mvHMM, and glassoSlide. The MCC scores highlight the advantage of choosing our proposed method in large settings. Indeed, the number of parameters for each individual state is substantial, as there are 4950 distinct off-diagonal coefficients to be inferred for each precision matrix.

4   Application to fMRI Data

Identifying the dynamic nature of brain connectivity is critical for understanding and advancing our current knowledge about human brain functioning. Functional magnetic resonance imaging (fMRI), which measures brain activity by detecting changes associated with blood flow, has become a successful and effective instrument for studying how the brain functions. Here we consider the problem of estimating brain connectivity, i.e., statistical dependence between fMRI time series in distinct regions of the brain. Recent evidence has shown that the interactions among brain regions vary over the course of an fMRI experiment, suggesting that brain connectivity is a dynamic process (Cribben et al., 2013; Lindquist et al., 2014; Xu and Lindquist, 2015; Warnick et al., 2018). Flexible models that account for dynamic features, change-points and sparse structures are needed for the analysis of such data.

We apply the proposed model to fMRI data from a subject performing a task-based experiment where the interest is to identify the neural representations that are formed during latent learning of predictive sequences (Bornstein and Daw, 2012). In this experiment, participants carried out a task in which they observed a sequence of black-and-white natural scene images. They were instructed only to press a keyboard key ('d', 'f', 'j', 'k') that they had previously been trained to associate with each image. Throughout the trials, the series of pictures were generated according to a first-order Markov process, though the participants were not aware of this structure. This form of task has been used to examine the cognitive and neural architecture of latent learning and the use of learned representations in support of predictive lookahead, a core computational process supporting decision-making in humans and animals (Strange et al., 2005; Harrison et al., 2006; Bornstein and Daw, 2013; Morris et al., 2018; Rmus et al., 2022; Hunter et al., 2018; Yoo et al., 2023). Here, response times indicated the degree to which the participant implicitly expected the currently presented image, on the basis of the previously presented one. A consistent finding in these tasks is that participants implicitly learn the sequential structure, and that neural regions signal the degree to which they anticipate the upcoming image in the sequence. Several studies have identified more than one representation of sequential structure, each of which has an influence on behavior as estimated across the entire experiment. However, it is unclear how these multiple representations are arbitrated among to influence behavior – e.g., as a weighted

mixture at the single-trial level Wang et al. (2022); Khoudary et al. (2022); Nicholas et al. (2022), in alternation according to regimes of task statistics Daw et al. (2005); Poldrack et al. (2001); Lengyel and Dayan (2007); Yoo et al. (2023), or as a fixed proportion that varies according to individual differences such as in memory encoding precision Noh et al. (2023). Full details of the experiment are provided in Bornstein and Daw (2012).

The scanning session proceeded with four blocks consisting of 275 fMRI acquisitions. For the analyses of this paper we concatenated the four blocks and subtracted the mean of each block. $D = 18$ lateralized regions of interest (ROIs) were selected on the basis of prior findings using this task (Bornstein and Daw, 2012, 2013) as those most sensitive to one or more of the identified representations of sequential s tructure (dorsal and ventral striatum, hippocampus), to the degree of conflict between the representations (anterior cingulate cortex), or to the stimulus content (scene images; ventral visual stream regions). We scaled each dimension $d = 1, \ldots, D$, to have variance one. In our analysis, we seek to identify distinct regimes of functional connectivity that can be mapped onto cognitive interpretation – specifically, to identify the manner in which multiple representations combine to control behavior. Since we do not have prior information on the cognitive states that are manifested during the experiment, we assumed that the number of the states is unknown, as well as the latent learning structure driving the switching of regimes. We set the hyperparameters of the model as described in Section 3.1 and ran MCMC chains with 4000 iterations, 1200 of which were discarded as burn-in.

The time series data for the $D = 18$ selected ROIs are shown in the top panel of Figure 5, along with the estimated latent state sequence. Our model identified a mode at $\hat{K} = 2$ distinct states and a DAR order $\hat{P} = 2$, with estimated values of innovations and autoregressive parameters $\hat{\boldsymbol{\pi}} = [0.50, 0.50]$ and $\hat{\boldsymbol{\phi}} = [0.08, 0.87, 0.05]$, respectively. The bottom panel of Figure 5 displays the time-varying probability plot, namely the local decoding of the hidden state at time t, $p(\gamma_t = j \mid \boldsymbol{y},)$, for $t = 1, \ldots, T$, as described in Section 2.4. These probabilities are represented with a different color for each of the two inferred states, and they cumulatively add to one for each t. The state probability plot displays a clear transition from state 1 to state 2, approximately halfway through the task.

The top panel of Figure 6 shows the estimated state-specific partial correlation matrices, for the two estimated states, and the bottom panel the corresponding estimated connectiv-
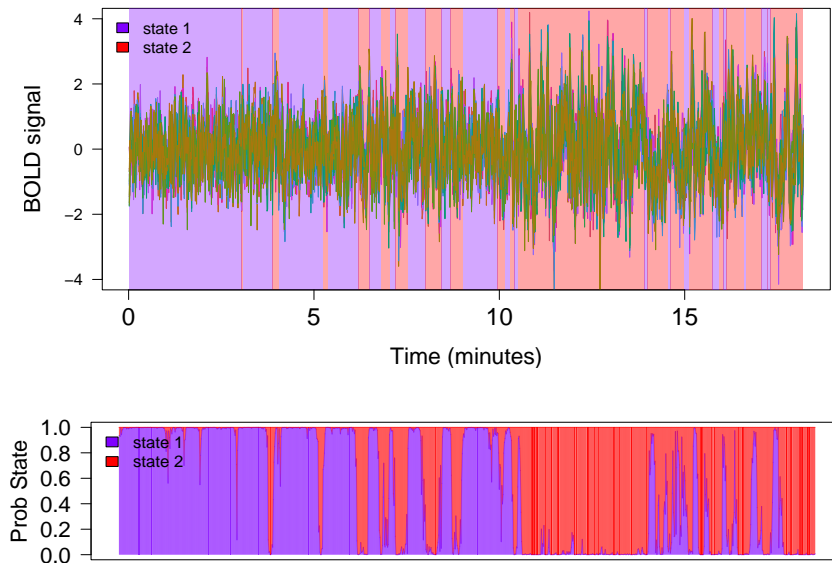
Figure 5: **Application to FMRI data**. (top) time series data from one participant, with each dimension represented by a different colored line; vertical colored bands represent the estimated state sequence. (bottom) estimated time-varying state probability plot.

ity graphs, with edges identified through the procedure described in Section 2.2. State 1 has relatively stronger connectivity between hippocampus (HC) and anterior cingulate cortex (ACC, ant_cing), with mean difference between states across ROI pairs equal to .028; whereas State 2 shows stronger connectivity between Caudate and ACC, with mean difference between states across ROI pairs equal to .048. Across all ROI pairs, the average difference in partial correlation values between states is .002.

These observations are consistent with findings in the literature that at least two distinct networks mediate expectations in this task: one centered on hippocampus and thought to encode stimulus-stimulus predictive relationships (e.g. "cognitive maps"), and the other centered on striatum and thought to encode response-response sequences (Bornstein and Daw, 2012, 2013). Each has different dynamics with regard to the predictiveness of the learned sequences: activity in the hippocampal network scales with increasing uncertainty about the next item in the sequence, consistent with its proposed role in "pre-fetching" upcoming states in support of decision-making (Johnson and Redish, 2007); separately, activity in the striatal network *decreases* with uncertainty about the next item in the sequence, consistent with observations that this structure is more strongly activated by highly predictive associ-
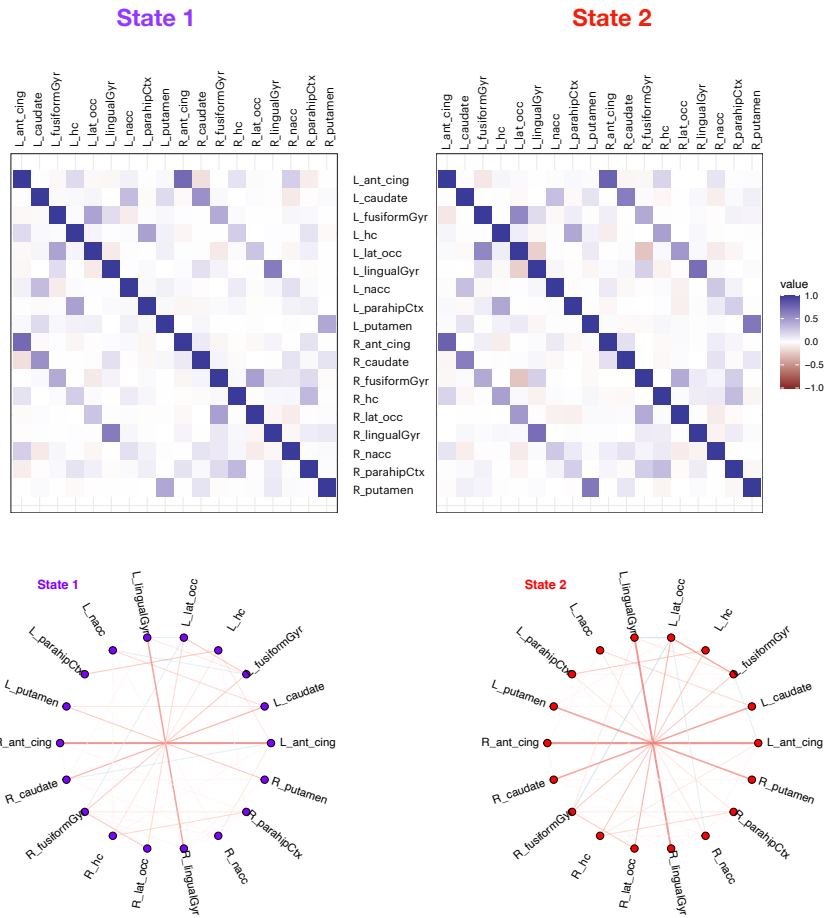
Figure 6: **Application to FMRI data**. (top) estimated partial correlation matrices, for each of the two inferred states. (bottom) estimated state-specific connectivity graphs.

ations (Smith and Graybiel, 2016). The observation that the network regime corresponds to shifts in its connectivity with anterior cingulate cortex is consistent with theoretical accounts of this region as signaling the "expected value of control", mediating the influence of internal representations on behavior (Shenhav et al., 2013). The transition between hippocampal and striatally-mediated regimes is consistent with extensive empirical findings that these regions "trade-off" in control of behavior across highly repeated tasks, with hippocampus driving responses early on and striatum taking over when sequences are more well-practiced (Poldrack et al., 2001; Lengyel and Dayan, 2007).

5   Concluding Remarks

We have presented a flexible Bayesian approach for estimating sparse Gaussian graphical models based on time series data. In order to represent switching dynamics of the time series data, we have assumed an unobserved hidden process underlying the data, with observations generated from state-specific multivariate Gaussian emission distributions. We have modeled the temporal structure of the hidden state sequence based on a DAR process, as a flexible approach to incorporate temporal dynamics that extend beyond simple Markovian structures. We have modeled the time-varying mixing probabilities capturing the state-switching behavior of the DAR process via a cumulative shrinkage non-parametric prior that accommodates zero-inflated parameters for non-active components. The proposed formulation ensures that if a parameter in the DAR model is zero, then all subsequent lag parameters are also zero, yielding a flexible and computationally efficient modeling framework for estimating the time-varying mixing probabilities as well as the effective order of the process. This considerably speeds up the posterior sampler, especially in regard to the forward-backward scheme for updating the latent state sequence. We have additionally integrated a sparsity-inducing Dirichlet prior to estimate the effective number of states in a data-driven manner. At the network level, we have assumed a graphical horseshoe prior to induce sparsity in the state-specific precision matrices. We have thoroughly investigated the performance of our methods through simulation studies and performed comparisons with competing approaches. We have further illustrated our proposed approach for the estimation of dynamic brain connectivity based on fMRI data collected on a subject performing a

task-based experiment on latent learning.

References

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., and Calhoun, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral Cortex*, 24(3):663–676.

Biswas, A. and Song, P. X.-K. (2009). Discrete-valued ARMA processes. *Statistics & Probability Letters*, 79(17):1884–1889.

Bornstein, A. M. and Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience*, 35(7):1011–1023.

Bornstein, A. M. and Daw, N. D. (2013). Cortical and hippocampal correlates of deliberation during model-based decisions for rewards in humans. *PLoS computational biology*, 9(12):e1003387.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag New York.

Cribben, I., Wager, T. D., and Lindquist, M. A. (2013). Detecting functional connectivity change points for single-subject fMRI data. *Frontiers in computational neuroscience*, 7:143.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711.

DeRuiter, S. L., Langrock, R., Skirbutas, T., Goldbogen, J. A., Calambokidis, J., Friedlaender, A. S., and Southall, B. L. (2017). A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *The Annals of Applied Statistics*, 11(1):362–392.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, pages 1020–1056.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics*, 9(3):432–441.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Hadj-Amar, B., Finkenstädt, B., Fiecas, M., and Huckstepp, R. (2021). Identifying the recurrence of sleep apnea using a harmonic hidden Markov model. *The Annals of Applied Statistics*, 15(3):1171.

Harrison, L. M., Duggins, A., and Friston, K. J. (2006). Encoding uncertainty in the hippocampus. *Neural Networks*, 19(5):535–546.

Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the Association for Computing Machinery*, 24(4):233–245.

Heiner, M., Kottas, A., and Munch, S. (2019). Structured priors for sparse probability vectors with application to model selection in Markov chains. *Statistics and Computing*, 29:1077–1093.

Holsclaw, T., Greene, A. M., Robertson, A. W., and Smyth, P. (2017). Bayesian nonhomogeneous Markov models via Pólya-Gamma data augmentation with applications to rainfall modeling. *The Annals of Applied Statistics*, 11(1):393 – 426.

Hunter, L. E., Bornstein, A. M., and Hartley, C. A. (2018). A common deliberative process underlies model-based planning and patient intertemporal choice. *bioRxiv*, page 499707.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Johnson, A. and Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, 27(45):12176–12189.

Kastner, G. and Huber, F. (2020). Sparse Bayesian vector autoregressions in huge dimensions. *Journal of Forecasting*, 39(7):1142–1165.

Khoudary, A., Peters, M. A., and Bornstein, A. M. (2022). Precision-weighted evidence integration predicts time-varying influence of memory on perceptual decisions. *Cognitive Computational Neuroscience*.

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010). Estimating time-varying networks. *The Annals of Applied Statistics*, pages 94–123.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, pages 694–726.

Legramanti, S., Durante, D., and Dunson, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107(3):745–752.

Lengyel, M. and Dayan, P. (2007). Hippocampal contributions to control: the third way. *Advances in neural information processing systems*, 20.

Li, Y., Craig, B. A., and Bhadra, A. (2019). The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics*, 28(3):747–757.

Lindquist, M. A., Xu, Y., Nebel, M. B., and Caffo, B. S. (2014). Evaluating dynamic bivariate correlations in resting-state fMRI: a comparison study and a new approach. *NeuroImage*, 101:531–546.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26(1):303–324.

Morris, R. W., Bornstein, A., and Shenhav, A. (2018). *Goal-directed decision making: computations and neural circuits*. Elsevier.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.

Nicholas, J., Daw, N. D., and Shohamy, D. (2022). Uncertainty alters the balance between incremental learning and episodic memory. *Elife*, 11:e81679.

Noh, S. M., Singla, U. K., Bennett, I. J., and Bornstein, A. M. (2023). Memory precision and age differentially predict the use of decision-making strategies across the lifespan. *Scientific Reports*, 13(1):17014.

Papastamoulis, P. and Iliopoulos, G. (2013). On the convergence rate of random permutation sampler and ECR algorithm in missing data models. *Methodology and Computing in Applied Probability*, 15(2):293–304.

Poldrack, R. A., Clark, J., Pare-Blagoev, E., Shohamy, D., Creso Moyano, J., Myers, C., and Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature*, 414(6863):546–550.

Qiu, H., Han, F., Liu, H., and Caffo, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(2):487–504.

Quinn, A. J., Vidaurre, D., Abeysuriya, R., Becker, R., Nobre, A. C., and Woolrich, M. W. (2018). Task-evoked dynamic network analysis through hidden Markov modeling. *Frontiers in neuroscience*, 12:603.

Rmus, M., Ritz, H., Hunter, L. E., Bornstein, A. M., and Shenhav, A. (2022). Humans can navigate complex graph structures acquired during latent learning. *Cognition*, 225:105103.

Rodríguez, C. E. and Walker, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Safikhani, A. and Shojaie, A. (2022). Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models. *Journal of the American Statistical Association*, 117(537):251–264.

Sarkar, A. and Dunson, D. B. (2019). Bayesian higher order hidden Markov models. *arXiv preprint arXiv:1805.12201*.

Savitsky, T., Vannucci, M., and Sha, N. (2011). Variable selection for nonparametric G aussian process priors: Models and computational strategies. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 26(1):130.

Shenhav, A., Botvinick, M. M., and Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240.

Smith, K. S. and Graybiel, A. M. (2016). Habit formation coincides with shifts in reinforcement representations in the sensorimotor striatum. *Journal of neurophysiology*, 115(3):1487–1498.

Song, L., Kolar, M., and Xing, E. (2009). Time-varying dynamic Bayesian networks. *Advances in neural information processing systems*, 22.

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., and Friston, K. J. (2005). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Networks*, 18(3):225–230.

Tang, Z.-Z. and Chen, G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713.

Wang, S., Feng, S. F., and Bornstein, A. M. (2022). Mixing memory and desire: How memory reactivation supports deliberative decision-making. *Wiley Interdisciplinary Reviews: Cognitive Science*, 13(2):e1581.

Warnick, R., Guindani, M., Erhardt, E., Allen, E., Calhoun, V., and Vannucci, M. (2018). A Bayesian Approach for Estimating Dynamic Functional Network Connectivity in fMRI Data. *Journal of the American Statistical Association*, 113(521):134–151.

Xu, Y. and Lindquist, M. A. (2015). Dynamic connectivity detection: an algorithm for determining functional connectivity change points in fMRI data. *Frontiers in neuroscience*, 9:285.

Yoo, J., Bornstein, A., and Chrastil, E. R. (2023). Cognitive graphs: Representational substrates for planning.

Zhang, W., Cribben, I., Petrone, S., and Guindani, M. (2021). Bayesian time-varying tensor vector autoregressive models for dynamic effective connectivity. *ArXiv:2106.14083*.

Supplement: We include further details about backward and forward messages for our sampling algorithm. We also report results from additional simulations, sensitivity analyses and convergence diagnostics of the MCMC.

Software: `sggmDAR` - a Julia software implementing the methodology outlined in the paper, accompanied by a comprehensive tutorial designed to guide users through replicating the findings detailed in the article.

A. Backward messages

Proposition 2. Let consider $\eta_{\{j_{\hat{P}},\ldots,j_1,j_0\}}$, i.e. the DAR probabilities of selecting state j_0, given previous values $j_1,\ldots,j_{\hat{P}}$, as defined in Eq. (2), and let $p(\boldsymbol{y}_t\,|\cdot)$ be the multivariate Gaussian emission densities specified in Eq. (1). Then, the backward messages $\beta_t(j_1) = p(\boldsymbol{y}_{t:T}|\gamma_{t-1}=j_1,\cdot)$ can be recursively expressed as in Eq. (14).

Proof: Let $M = M_{\max}$, and let $P = \hat{P}$ be the number of active DAR parameters. Then the proof proceeds as follows

$$
\begin{aligned}
\beta_t(j_1) &= p(\boldsymbol{y}_{t:T}|\gamma_{t-1}=j_1,\cdot)\\
&= \sum_{j_0=1}^{M} p(\boldsymbol{y}_{t:T},\gamma_t=j_0|\gamma_{t-1}=j_1,\cdot)\\
&= \sum_{j_P=1}^{M}\cdots\sum_{j_2=1}^{M}\sum_{j_0=1}^{M} p(\boldsymbol{y}_{t:T},\gamma_t=j_0|\gamma_{t-1}=j_1,\ldots,\gamma_{t-P}=j_P,\cdot)\\
&= \sum_{j_P=1}^{M}\cdots\sum_{j_2=1}^{M}\sum_{j_0=1}^{M} p(\gamma_{t-1}=j_1,\ldots,\gamma_{t-P}=j_P)\,p(\boldsymbol{y}_{t:T}|\gamma_t=j_0,\cdot)\\
&= \sum_{j_P=1}^{M}\cdots\sum_{j_2=1}^{M}\sum_{j_0=1}^{M} \eta_{\{j_P,\ldots,j_1,j_0\}}\,p(\boldsymbol{y}_t|\gamma_t=j_0,\boldsymbol{\mu},\boldsymbol{\Omega})\,\beta_{t+1}(j_0).
\end{aligned}
$$

B. Forward Messages for Local Decoding

The *forward messages* $\alpha_t(j_1) = p\left(\boldsymbol{y}_{1:t-1}, \gamma_{t-1} = j_1 \,|\, \cdot\right)$ utilized to conduct local decoding (Section 2.4) are defined as

$$\alpha_t(j_1) = \sum_{j_2=1}^{\hat{M}} \cdots \sum_{j_{\hat{P}}=1}^{\hat{M}} \alpha_t(j_1, j_2, \ldots, j_{\hat{P}}), \tag{S.1}$$

for $j_l \in \{1, \ldots, \hat{M}\}, l = 1, \ldots \hat{P}$, and the DAR-*forward messages* $\alpha_t(j_1, \ldots, j_{\hat{P}}) = p\left(\boldsymbol{y}_{1:t-1}, \gamma_{t-1} = j_1, \ldots, \gamma_{t-\hat{P}} = j_{\hat{P}} \,|\, \cdot\right)$ are described as the probability of the partial observations sequence $\boldsymbol{y}_{1:t-1}$, and states $\gamma_{t-1:t-\hat{P}}$ at time $t-1$, given all the other parameters. These messages can be recursively computed as follows

$$\alpha_t(j_1, \ldots, j_{\hat{P}}) = \sum_{j_{\hat{P}+1}=1}^{\hat{M}} \eta_{\{j_{\hat{P}+1}, \ldots, j_2, j_1\}}\, p(\boldsymbol{y}_{t-1}|\gamma_{t-1} = j_1, \boldsymbol{\mu}, \boldsymbol{\Omega})\, \alpha_{t-1}(j_2, \ldots, j_{\hat{P}+1}), \tag{S.2}$$

as shown in the following Proposition.

Proposition 3. Let $\alpha_t(j_1, \ldots, j_{\hat{P}}) = p\left(\boldsymbol{y}_{1:t-1}, \gamma_{t-1} = j_1, \ldots, \gamma_{t-\hat{P}} = j_{\hat{P}} \,|\, \cdot\right)$ be the DAR-*forward messages*, described as the probability of the partial observations sequence $\boldsymbol{y}_{1:t-1}$, and states $\gamma_{t-1:t-\hat{P}}$ at time $t-1$, given all the other parameters. Then, these messages can be recursively computed as in Eq. (S.2).

Proof: Let $M = \hat{M}$, and let $P = \hat{P}$ be the number of active DAR parameters. Then the proof proceeds as follows

$$
\begin{aligned}
\alpha_t(j_1, \ldots, j_P) &= p\left(\boldsymbol{y}_{1:t-1}, \gamma_{t-1} = j_1, \ldots \gamma_{t-\hat{P}} = j_P \,|\, \cdot\right) \\
&= \sum_{j_{P+1}=1}^{M} p\left(\boldsymbol{y}_{1:t-1}, \gamma_{t-1} = j_1, \ldots, \gamma_{t-(P+1)} = j_{P+1} \,|\, \cdot\right) \\
&= \sum_{j_{P+1}=1}^{M} p(\boldsymbol{y}_{1:t-2}, \gamma_{t-2} = j_2, \ldots, \gamma_{t-(P+1)} = j_{P+1})\, p(\gamma_{t-1} = j_1|\gamma_{t-2} = j_2, \ldots, \gamma_{t-(P+1)} = j_{P+1}) \\
&\qquad p(\boldsymbol{y}_{t-1}|\gamma_{t-1} = j, \boldsymbol{\mu}, \boldsymbol{\Omega}) \\
&= \sum_{j_{P+1}=1}^{M} \eta_{\{j_{P+1}, \ldots, j_1\}} p(\boldsymbol{y}_{t-1}|\gamma_{t-1} = j, \boldsymbol{\mu}, \boldsymbol{\Omega}) \alpha_{t-1}(j_2, \ldots, j_{P+1}).
\end{aligned}
$$

C. Simulated scenario with zero-mean observations

Here, we investigate the performance of our approach when the data-generating emissions are zero-mean, i.e. $\boldsymbol{\mu}_j = \mathbf{0}$, for $j = 1, \ldots, M$. We generated 30 distinct dataset, each consisting of $D = 15$-dimensional $T = 2000$, in a similar fashion as in Section 3.1. We assumed $M = 3$ states and DAR order $P = 2$, where the autoregressive probabilities and innovations were specified as $\boldsymbol{\phi} = (0.2, 0.5, 0.3)$ and $\boldsymbol{\pi} = (0.5, 0.3, 0.2)$. The precision matrices were constructed using patterns (i), (iv) and (v) from Section 3.1. A single realization from this simulation setting is shown in Figure 1 (top panel), where vertical colored bands represent the true underlying state sequence. We chose $M_{max} = 6$, and we set the rest of the hyperparameters as in Section 3.1.
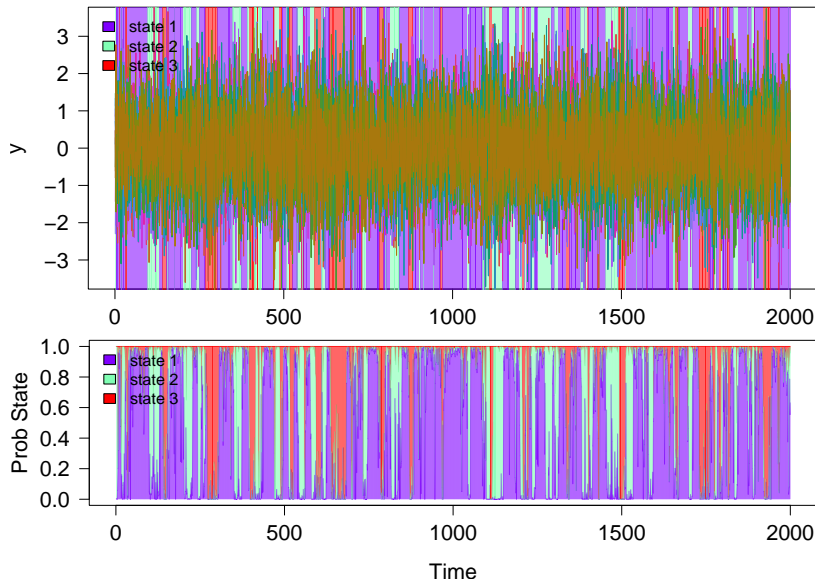

Figure 1: **Simulation with zero mean**. (top) time series realization (lines) where each dimension is represented by a different colored line; vertical colored bands represent the true underlying state sequence; (bottom) estimated time-varying probability plot.

Our approach consistently estimated the correct number of states $\hat{M} = 3$ as the mode of the posterior distribution and the number of active DAR probabilities $\hat{P} = 2$ with high posterior probability, on all simulated replicates. Figure 1 (bottom panel) displays a time-varying probability plot, namely the local decoding of the hidden state at time t, $p(\gamma_t = j \mid \boldsymbol{y}, \cdot)$, $j = 1, \ldots, \hat{M}$, as described in Section 2.4; these plots are constructed by plotting the local probabilities (which add to 1) cumulatively for each t, where each state is associated

with a different color. It is evident that our proposed approach correctly retrieves the true latent state sequence. We assessed the model selection performance of our approach in Table 2, showing accuracy, sensitivity, specificity, $F1$-score and Matthew correlation coefficient (MCC). To evaluate estimation accuracy we report RMSE of the state-specific off-diagonal entries of the precision matrices. As in Section 3, our proposed methodology is compared with `mvHMM` and `glassoSlide`. These results from our proposed approach are conditioned on the modal number of states and autoregressive order. As in the investigation carried out in Section 3.1, our approach seems to be superior to `mvHMM` and `glassoSlide`, for what concerns both estimation accuracy and model selection performances.

		Identity	AR(2)	Random
	sggmDAR	1.0 (0.0)	0.994 (0.007)	0.987 (0.011)
Acc	mvHMM	0.929 (0.072)	0.953 (0.026)	0.933 (0.057)
	glassoSlide	0.999 (0.004)	0.837 (0.049)	0.869 (0.069)
	sggmDAR	1.0 (0.0)	0.992 (0.01)	0.996 (0.006)
Spec	mvHMM	0.929 (0.072)	0.936 (0.035)	0.947 (0.039)
	glassoSlide	0.999 (0.004)	0.916 (0.062)	0.936 (0.047)
	sggmDAR	-	0.985 (0.018)	0.959 (0.034)
MCC	mvHMM	-	0.892 (0.055)	0.865 (0.076)
	glassoSlide	-	0.564 (0.126)	0.580 (0.267)
	sggmDAR	-	0.989 (0.013)	0.967 (0.029)
F1	mvHMM	-	0.918 (0.042)	0.802 (0.278)
	glassoSlide	-	0.652 (0.092)	0.646 (0.231)
	sggmDAR	-	0.999 (0.007)	0.949 (0.051)
Sens	mvHMM	-	1.0 (0.0)	0.877 (0.300)
	glassoSlide	-	0.607 (0.153)	0.617 (0.248)
	sggmDAR	0.003 (0.002)	0.028 (0.005)	0.030 (0.005)
RMSE	mvHMM	0.039 (0.012)	0.045 (0.006)	0.069 (0.036)
	glassoSlide	0.0 (0.001)	0.165 (0.019)	0.111 (0.019)

Table 2: **Simulation with zero-mean.** Accuracy, sensitivity, specificity, F1 score, Matthew correlation coefficient (MCC), and residual mean squared error (RMSE) of precision matrix estimates, for each regime $j = 1, \ldots, \hat{M}$. Standard deviation over the 30 simulations are displayed in brackets. Results are reported for our `sggmDAR` , `mvHMM` and `glassoSlide`. The results of our approach are conditioned on the modal number of states and autoregressive order. A hyphen is used for those metrics that cannot be computed due to the structure of the underlying truth (e.g. TP+FN = 0).

D. Large D setting

Here, we explore the performance of our approach in a scenario where the dimension D of the data is large, as discussed in Section 3.4. We focus on assessing the ability of our proposed method in recovering the number of states, number of DAR parameters, and true sparse precision matrices. Table 3 displays model selection and estimation accuracy performances for the off-diagonal component of the high-dimensional precision matrices, for `sggmDAR`, `mvHMM`, and `glassoSlide`. The MCC scores highlight the advantage of choosing our proposed method in high-dimensional settings. Indeed, the number of parameters for each individual state is substantial, as there are 4950 distinct off-diagonal coefficients to be inferred for each precision matrix.

We report model selection and estimation accuracy performances for the off-diagonal component of the high-dimensional precision matrices, for `sggmDAR`, `mvHMM`, and `glassoSlide`. The MCC scores highlight the advantage of choosing our proposed method in high-dimensional settings. Indeed, the number of parameters for each individual state is substantial, as there are 4950 distinct off-diagonal coefficients to be inferred for each precision matrix.

E. Sensitivity Analysis

We carried out a sensitivity analysis study by focusing on the impact of the hyperparameters of the zero-inducing cumulative shrinkage prior that characterizes the DAR process formulated in Section 2.1.1. In particular, we investigated the sensitivity of the hyperparameters of the Beta priors on the stick-breaking weights v_0 (i.e., a_0 and b_0), and v_j (i.e., a_j and b_j), recalling that the mixing probabilities $\{\phi_j\}_{j=0}^P$ are a by-product of the stick-breaking weights and that they determine the number of active DAR coefficients. We investigated four different scenarios: (i) $v_0 = v_j \sim \text{Beta}(0.5, 0.5)$, corresponding to a Jeffreys prior (Jeffreys, 1946); (ii) $v_0 = v_j \sim \text{Beta}(1, 1)$, namely a uniform prior; (iii) $v_0 \sim \text{Beta}(1, 10)$, $v_j \sim \text{Beta}(1, 10)$, so that the prior probability of autoregression is driven towards zero; (iv) $v_0 \sim \text{Beta}(1, 10)$, $v_j \sim \text{Beta}(10, 1)$, i.e. the hyperparameter setting chosen as in Section 3. We simulated 30 time series from the same simulation setting of Section 3, for different sample sizes $T \in \{100, 500\}$. Table 4 reports posterior probabilities of the number of DAR

		Identity	Hub	Random
Acc	`sggmDAR`	1.0 (0.0)	1.0 (0.0)	0.997 (0.001)
	`mvHMM`	0.923 (0.017)	0.862 (0.045)	0.861 (0.054)
	`glassoSlide`	0.997 (0.008)	0.963 (0.029)	0.932 (0.025)
Spec	`sggmDAR`	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
	`mvHMM`	0.923 (0.017)	0.861 (0.043)	0.861 (0.050)
	glassoSlide	0.997 (0.008)	0.981 (0.030)	0.959 (0.027)
MCC	`sggmDAR`	-	0.999 (0.002)	0.952 (0.02)
	`mvHMM`	-	0.301 (0.099)	0.352 (0.148)
	`glassoSlide`	-	0.006 (0.063)	0.015 (0.020)
F1	`sggmDAR`	-	0.999 (0.002)	0.953 (0.02)
	`mvHMM`	-	0.214 (0.064)	0.293 (0.104)
	`glassoSlide`	-	0.013 (0.033)	0.044 (0.022)
Sens	`sggmDAR`	-	1.0 (0.0)	0.911 (0.036)
	`mvHMM`	-	0.903 (0.209)	0.862 (0.256)
	`glassoSlide`	-	0.018 (0.061)	0.057 (0.043)
RMSE	`sggmDAR`	0.001 (0.0)	0.004 (0.001)	0.008 (0.001)
	`mvHMM`	0.031 (0.003)	0.063 (0.018)	0.067 (0.025)
	`glassoSlide`	0.001 (0.002)	0.028 (0.001)	0.059 (0.002)

Table 3: **Simulation Study.** Large D setting. Accuracy, sensitivity, specificity, F1 score, Matthew correlation coefficient (MCC), and residual mean squared error (RMSE) of precision matrix estimates, for each regime $j = 1, \ldots, \hat{M}$. Standard deviation over the 30 simulations are displayed in brackets. Results are reported for our `sggmDAR` , `mvHMM` and `glassoSlide`. The results of our approach are conditioned on the modal number of states and autoregressive order. A hyphen is used for those metrics that cannot be computed due to the structure of the underlying truth (e.g. TP+FN = 0).

parameters, $p(P = j|\cdot)$, over the 30 simulations, where we note that the true number of DAR parameters is $P_{true} = 2$. It appears that cases (1) and (2) behave similarly, by identifying a posterior mode at 2 for both $T \in \{100, 500\}$, noting that as T grows $p(P = 2|\cdot)$ increases considerably. As expected, case (3) seems to penalize the probability of autoregression at higher lags, since for both sample sizes, large posterior probability is located at 1. Case (4) identifies the right number of lags for both sample sizes and seems to slightly favor probability mass to larger numbers of lags, as the probability of identifying a DAR process of order 1 is 0.045 and 0.002, while the probabilities of selecting 3 lags are 0.103 and 0.140, for $T = 100$ and $T = 500$, respectively. In our investigations, we also noted that as $T > 1000$ the sensitivity is not very noticeable.

	$p(P = 1|\cdot)$	$p(P = 2|\cdot)$	$p(P = 3|\cdot)$	$p(P = 4|\cdot)$	$p(P = 5|\cdot)$
			$T = 100$		
case 1	0.094	0.557	0.247	0.089	0.014
case 2	0.104	0.401	0.268	0.146	0.081
case 3	0.754	0.107	0.067	0.012	0.061
case 4	0.045	0.850	0.103	0.002	0.000
			$T = 500$		
case 1	0.006	0.903	0.088	0.003	0.000
case 2	0.007	0.891	0.097	0.005	0.000
case 3	0.441	0.546	0.013	0.000	0.000
case 4	0.002	0.856	0.140	0.002	0.000

Table 4: **Simulation study.** Sensitivity analysis on the DAR parameters for the cases (1) $v_0 \sim \text{Beta}(0.5, 0.5)$, $v_j \sim \text{Beta}(0.5, 0.5)$; (2) $v_0 \sim \text{Beta}(1, 1)$, $v_j \sim \text{Beta}(1, 1)$; (3) $v_0 \sim \text{Beta}(1, 10)$, $v_j \sim \text{Beta}(1, 10)$; (4) $v_0 \sim \text{Beta}(1, 10)$, $v_j \sim \text{Beta}(10, 1)$. Note that $P_{true} = 2$.

We carried out further investigations on the impact of the hyperparameters of the zero-inducing cumulative shrinkage prior characterizing the DAR process. We studied the same four scenarios as above, by focusing on the posterior probability over the number of states. Table 5 reports, for each scenario, posterior probabilities of the number of states, $p(M = j|\cdot)$, over the 30 simulated replicates, where we note that the true number of states is $M_{true} = 5$. It appears that for the smaller sample size (e.g. $T = 100$) the sampler tends to estimate more states than necessary, for all combination of the hyperparameters. However, as the sample

size increases, the inference on the correct number of states increases considerably. In our investigations, we also noted that when $T > 1000$ the sensitivity is not very noticeable.

	$p(M=3\|\cdot)$	$p(M=4\|\cdot)$	$p(M=5\|\cdot)$	$p(M=6\|\cdot)$	$p(M=7\|\cdot)$	$p(M=8\|\cdot)$	$p(M=9\|\cdot)$
				$T=100$			
case 1	0.0	0.1	0.2	0.333	0.267	0.1	0.0
case 2	0.033	0.0	0.167	0.3	0.256	0.21	0.033
case 3	0.033	0.133	0.3	0.333	0.1	0.094	0.006
case 4	0.033	0.133	0.267	0.4	0.133	0.033	0.0
				$T=500$			
case 1	0.0	0.133	0.867	0.0	0.0	0.0	0.0
case 2	0.0	0.133	0.867	0.0	0.0	0.0	0.0
case 3	0.0	0.133	0.833	0.033	0.0	0.0	0.0
case 4	0.0	0.133	0.867	0.0	0.0	0.0	0.0

Table 5: **Simulation study.** Sensitivity analysis of the DAR parameters for the following scenarios: (1) $v_0 \sim \text{Beta}(0.5, 0.5)$, $v_j \sim \text{Beta}(0.5, 0.5)$; (2) $v_0 \sim \text{Beta}(1, 1)$, $v_j \sim \text{Beta}(1, 1)$; (3) $v_0 \sim \text{Beta}(1, 10)$, $v_j \sim \text{Beta}(1, 10)$; (4) $v_0 \sim \text{Beta}(1, 10)$, $v_j \sim \text{Beta}(10, 1)$. We report posterior probabilities of the number number of states, $p(M = j|\cdot)$, where we note that $M_{true} = 5$; results are shown for $T \in \{100, 500\}$.

F.   Convergence Diagnostics

We verified convergence of the MCMC sampler by: (i) analyzing the trace plots of the parameters, e.g. the mean of the multivariate spiked Gaussian emissions, observing no pathological behavior (see Figure 2 for representative examples of trace plots of the DAR parameters); (ii) storing the values of the overall likelihood of the system (Eq. (6)) and plotting the corresponding trace, noting that it reached a stable regime (see Figure 2, bottom plot, for a representative example of trace plot of the log likelihood); (iii) verifying the Heidelberger and Welch's convergence diagnostic test(Heidelberger and Welch, 1981): this diagnostic first tests the null hypothesis that the Markov Chain is in the stationary distribution and then tests whether the mean of a marginal posterior distribution can be estimated with sufficient precision, assuming that the Markov Chain is in the stationary distribution. In our experiments, this test was passed for every Markov chain we analyzed.
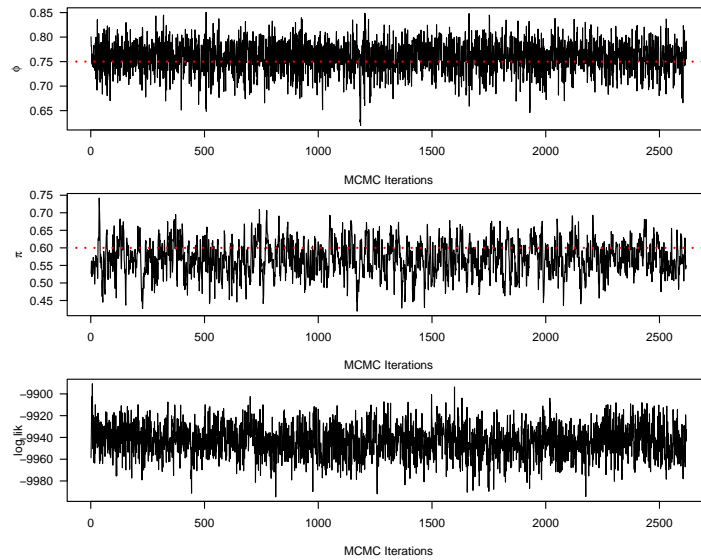
Figure 2: Examples of trace plots of the DAR parameters (ϕ_j and π_j) and of the likelihood. Dotted red lines corresponds to true generating value of the DAR parameters.