Penalized estimation of sparse Markov regime-switching vector auto-regressive models

Gilberto Chavez-Martinez * Ankush Agarwal † Abbas Khalili *
Syed Ejaz Ahmed ‡

April 3, 2023

Abstract

We consider sparse Markov regime-switching vector autoregressive (MSVAR) models in which the regimes are governed by a latent homogeneous Markov chain. In practice, even for moderate values of the number of Markovian regimes and data dimension, the associated MSVAR model has a large parameter dimension compared to a typical sample size. We provide a unified penalized conditional likelihood approach for estimating sparse MSVAR models. We show that our proposed estimators are consistent and recover the sparse structure of the model. We also show that, when the number of regimes is correctly or over-specified, our method provides consistent estimation of the predictive density. We develop an efficient implementation of the method based on a modified expectation-maximization (EM) algorithm. We discuss strategies for estimation of the number of regimes. We evaluate finite-sample performance of the method via simulations, and further demonstrate its utility by analyzing a real dataset. Supplementary materials for this paper are available online.

Keywords: EM algorithm, regularization methods, multivariate time series.

*Department of Mathematics and Statistics, McGill University, Montreal, Quebec H3A 0B9, Canada **e-mail**: gilberto.chavezmartinez@mail.mcgill.ca and abbas.khalili@mcgill.ca

†Adam Smith Business School, University of Glasgow, G12 8QQ Glasgow, United Kingdom **e-mail**: ankush.agarwal@glasgow.ac.uk

‡Faculty of Mathematics and Science, Brock University, St. Catharines, Ontario L2S 3A1, Canada **e-mail**: sahmed5@brocku.ca

1 Introduction

Markov regime-switching vector auto-regressive (MSVAR) models are ubiquitous in modelling heterogeneous and complex relationships between the variables of interest in multivariate time series analysis (Krolzig, 1997). The two stochastic components of an MSVAR are the observable time series vector $\mathbf{Y}_t \in \mathbb{R}^d$, and the latent Markov chain $S_t \in \{1, \ldots, M\}$, for some finite $M \in \mathbb{N}$. The process S_t determines the VAR regime under which the conditional distribution of $\mathbf{Y}_t|(S_t = s_t)$ evolves as a function of lagged $\mathbf{Y}_{t-j}, j = 1, \ldots, p \in \mathbb{N}$. In this work, we focus on Gaussian MSVARs where the noise component in each VAR regime follows a zero-mean, d-dimensional Gaussian distribution. These models are applicable in a wide range of disciplines. For example, in macroeconomic time series such as manufacturing activities, consumer price indices, and housing and asset prices ($\mathbf{Y}_t$), we typically see the effect of business cycles and the volatility clustering effect, which can be captured by S_t (Hamilton, 1988; Kim et al., 1998; Ang and Timmermann, 2012). In time series data related to natural phenomena such as wind power generation ($\mathbf{Y}_t$) across geographic regions, there are periods of high and low measurements that depend on wind speed (S_t) (Pinson and Madsen, 2012); similarly, there are periods of high and low temperature measurements that depend on annual meteorological cycles (S_t) corresponding to different seasons with warm and cold months (Monbet and Ailliot, 2017).

Maximum likelihood estimation (MLE) is the most common frequentist method of inference in MSVAR models. However, a limitation often encountered with MLE is the potentially large number of parameters to be estimated. In an MSVAR model with M regimes and autoregressive order p, the total number of parameters is $K = M\big(d + pd^2 +$

$d(d+1)/2\big) + M(M-1)$, which can be large even for moderate values of (d, p, M), compared to a typical sample size. For instance, in our case study each observation is 10-dimensional, and for an MSVAR model with AR-order $p = 1$ and the number of regimes M ranging from 1 to 5, there are $165 \leq K \leq 845$ parameters to estimate based on a sample of size 481. Therefore, besides an obfuscated model interpretation, it can also be difficult to perform stable MLE in large-dimensional parameter spaces. It thus becomes essential to consider strategies that enable more stable and interpretable parameter estimation. With this motivation, we perform parameter estimation using regularization techniques that have been successful in both high-dimensional VAR and covariance estimation problems (Basu and Matteson, 2021; Lam, 2020). These techniques arise from the assumption that many of the model parameters are null. In the context of MSVAR models, we assume that both VAR coefficient matrices and the noise covariance–or precision–matrices are sparse, that is, many of their entries are zero. This also results in more meaningful model interpretations.

There has been a huge surge in research on estimation in sparse VAR models (Basu and Michailidis, 2015; Nicholson et al., 2020, and references therein), and sparse covariance and precision matrix estimation (Friedman et al., 2008; Shojaie and Michailidis, 2010; Bien and Tibshirani, 2011; Lam, 2020, and references therein). This has been led by the development of penalized methods such as the LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006), SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). However, to the best of our knowledge, there are only a few works on regularized estimation for MSVARs. In the context of hidden Markov models (HMMs)–which are MSVARs with AR-order $p = 0$–Städler and Mukherjee (2013) studied the graphical LASSO algorithm for estimation of sparse precision matrices in

high-dimensions. Fiecas et al. (2017) proposed a shrinkage, rather than sparsity, approach to obtain stable covariance matrix estimates in high-dimensional HMMs. To simulate temperature data using a sparse MSVAR with a non-homogeneous Markov chain, Monbet and Ailliot (2017) performed simultaneous VAR coefficient and precision matrix estimation.

In this work, we focus on a general sparse estimation framework for MSVAR models, which is not available in the reviewed literature. We propose a penalized conditional likelihood approach that allows for sparse estimation of the regime-specific VAR coefficients and covariance/precision matrices. We condition on the initial state of the Markov chain S_t, which avoids the estimation of its initial distribution and thus simplifies the problem. We show that, irrespective of the initial state conditioning, our method consistently recovers the sparse MSVAR model. We also provide an estimator for the h-step-ahead predictive density, and show that as long as the true number of AR regimes is not under-specified, this estimator is consistent for the true predictive density. All of our theoretical results are provided in the Supplement, Section A2. We implement the proposed method by using a modified EM algorithm, combined with a generalized gradient descent method in the M-step. We investigate strategies for estimation of the number of AR regimes. We study finite-sample performance of the proposed methods via simulations, and further demonstrate their utility by analyzing a Canadian macroeconomic dataset.

Section 2 presents sparse MSVAR models. Section 3 describes the proposed estimation methods and their numerical implementation. We discuss the computation of predictive densities in Section 4. Section 5 presents a simulation study, and Section 6 contains a case study on a Canadian macroeconomic dataset to demonstrate the usage of our methodology.

2 Sparse Markov regime-switching vector auto-regressives

Let $\{S_t \in \{1, \ldots, M\}, t = 1, 2, \ldots\}$ be a latent homogeneous Markov chain for some finite $M \in \mathbb{N}$. We denote its $M \times M$ transition probability matrix by $\boldsymbol{P}$ with the (i, j)-th entry

$$\mathbb{P}(S_t = j | S_{t-1} = i, S_{t-2} = s_{t-2}, \ldots, S_1 = s_1) = \mathbb{P}(S_t = j | S_{t-1} = i) = \alpha_{ij},$$

which is the probability of entering state j at time t from state i at time $t - 1$, and $\sum_{j=1}^{M} \alpha_{ij} = 1$, $i = 1, \ldots, M$. Further, let $\{\mathbf{Y}_t \in \mathbb{R}^d, t = 1, 2, \ldots\}$ be an observable time series with observed values $\{\mathbf{y}_t, t = 1, 2, \ldots\}$. For any $t > p$, we assume that conditional on $\mathbf{Y}_{1:t-1} = \mathbf{y}_{1:t-1}$ and $S_{p:n} = s_{p:n}$, the distribution of $\mathbf{Y}_t$ only depends on the lagged $\mathbf{Y}_{t-p:t-1} = \mathbf{y}_{t-p:t-1}$ and $S_t = s_t$. In a Gaussian MSVAR with M regimes and AR-order p, the conditional distribution of $\mathbf{Y}_t$ given $S_t = m$ and $\mathbf{Y}_{t-j} = \mathbf{y}_{t-j}, j = 1, \ldots, p$, is a state-dependent multivariate Gaussian with covariance matrix $\boldsymbol{\Sigma}^{(m)} = (\sigma_{ij}^{(m)}) \in \mathbb{R}^{d \times d}$ and time-dependent mean vector

$$\boldsymbol{\mu}_t^{(m)} := \boldsymbol{\nu}^{(m)} + \boldsymbol{A}_1^{(m)} \mathbf{y}_{t-1} + \ldots + \boldsymbol{A}_p^{(m)} \mathbf{y}_{t-p}, \tag{1}$$

where $\boldsymbol{\nu}^{(m)} = (\nu_1^{(m)}, \ldots, \nu_d^{(m)})^\intercal \in \mathbb{R}^d$, and $\boldsymbol{A}_l^{(m)} = (a_{l,ij}^{(m)}) \in \mathbb{R}^{d \times d}, l = 1, \ldots, p$, are the AR coefficient matrices. In simple terms, the behaviour of $\mathbf{Y}_t$ within each regime–determined by the latent process S_t–over time is modeled by a Gaussian VAR with AR-order p.

The parameters of interest are the transition probability matrix $\boldsymbol{P}$, the AR intercepts $\boldsymbol{\nu}^{(1:M)} := \text{vec}([\boldsymbol{\nu}^{(1)}, \ldots, \boldsymbol{\nu}^{(M)}])$, the AR coefficients $\boldsymbol{A}^{(1:M)} := \text{vec}([\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(M)}])$, with $\boldsymbol{A}^{(m)} := \text{vec}([\boldsymbol{A}_1^{(m)}, \ldots, \boldsymbol{A}_p^{(m)}])$, $m = 1, \ldots, M$, the covariance matrices $\boldsymbol{\Sigma}^{(1:M)} := \text{vec}([\boldsymbol{\Sigma}^{(1)}, \ldots, \boldsymbol{\Sigma}^{(M)}])$, and the precision matrices $\boldsymbol{\Omega}^{(1:M)} := \text{vec}([\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(M)}])$, where $\boldsymbol{\Omega}^{(m)} = (\boldsymbol{\Sigma}^{(m)})^{-1} = (\omega_{ij}^{(m)})$. Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^K$ be the vector of all parameters with

$K = M\big(d + pd^2 + d(d+1)/2\big) + M(M-1)$. In our estimation procedure, we directly estimate either $\boldsymbol{\Sigma}^{(m)}$ or $\boldsymbol{\Omega}^{(m)}$, and the parameter $\boldsymbol{\theta}$ includes either $\boldsymbol{\Sigma}^{(1:M)}$ or $\boldsymbol{\Omega}^{(1:M)}$.

Conditional likelihood: Let $\mathbf{y}_{1:n} := \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be observations from a Gaussian MSVAR model with M regimes and AR-order p. We consider the joint conditional density of $\mathbf{Y}_{p+1:n}$ given $(\mathbf{y}_{1:p}, S_p = s_p)$. By the total probability rule, we have

$$f_{\boldsymbol{\theta}}(\mathbf{y}_{p+1:n}|\mathbf{y}_{1:p}, s_p) = \sum_{s_n=1}^{M} \cdots \sum_{s_{p+1}=1}^{M} f(\mathbf{y}_{p+1:n}|\mathbf{y}_{1:p}, S_{p:n} = s_{p:n}) \mathbb{P}(S_{p+1:n} = s_{p+1:n}|\mathbf{y}_{1:p}, s_p). \quad (2)$$

By the homogeneity of S_t, conditional independence of $\mathbf{Y}_t$'s, and the Gaussianity, we obtain

$$\mathbb{P}(S_{p+1:n} = s_{p+1:n}|\mathbf{y}_{1:p}, S_p = s_p) = \mathbb{P}(S_{p+1:n} = s_{p+1:n}|S_p = s_p) = \prod_{t=p+1}^{n} \alpha_{s_{t-1},s_t},$$

$$f(\mathbf{y}_{p+1:n}|\mathbf{y}_{1:p}, S_{p:n} = s_{p:n}) = \prod_{t=p+1}^{n} f(\mathbf{y}_t|\mathbf{y}_{1:t-1}, S_{p:n} = s_{p:n}) = \prod_{t=p+1}^{n} f(\mathbf{y}_t|\mathbf{y}_{t-p:t-1}, S_t = s_t)$$
$$= \prod_{t=p+1}^{n} \phi\big(\mathbf{y}_t; \boldsymbol{\mu}_t^{(s_t)}, \boldsymbol{\Sigma}^{(s_t)}\big),$$

where $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the d-dimensional Gaussian density. Therefore, the density (2) becomes

$$f_{\boldsymbol{\theta}}\big(\mathbf{y}_{p+1:n}|\mathbf{y}_{1:p}, s_p\big) = \sum_{s_n=1}^{M} \cdots \sum_{s_{p+1}=1}^{M} \Big(\prod_{t=p+1}^{n} \alpha_{s_{t-1},s_t}\Big)\Big(\prod_{t=p+1}^{n} \phi\big(\mathbf{y}_t; \boldsymbol{\mu}_t^{(s_t)}, \boldsymbol{\Sigma}^{(s_t)}\big)\Big),$$

where $\boldsymbol{\mu}_t^{(s_t)}$ is given in (1). The conditioning on $\mathbf{y}_{1:p}$ is standard in time series when fitting AR-type models. We also condition on $S_p = s_p$ to avoid specification of the initial distribution of S_t which is an intricate issue in itself. The effect of conditioning on S_p dissipates asymptotically (Douc et al., 2004). Thus, for any fixed s_p, the conditional log-likelihood is

$$l_n(\boldsymbol{\theta}; s_p) := \log\big(f_{\boldsymbol{\theta}}\big(\mathbf{y}_{p+1:n}|\mathbf{y}_{1:p}, s_p\big)\big). \quad (3)$$

One could obtain the (conditional) MLE of $\boldsymbol{\theta}$ by maximizing (3) with respect to $\boldsymbol{\theta}$. However, as discussed in the introduction, even for moderate values of (d, p, M), the parameter $\boldsymbol{\theta}$

has a large dimension which can result in an unstable estimation. Hence, to attain a more stable and meaningful estimation of $\boldsymbol{\theta}$, we assume that the data-generating model is sparse so that the number of parameters to estimate is much less than K. We develop an estimation framework based on penalization of the conditional likelihood (3) in Section 3.

Sparsity: Different sparsity structures on the AR coefficients $\{\boldsymbol{A}_l^{(m)}, m = 1, \ldots, M, l = 1, \ldots, p\}$ can be assumed for specific applications, for example, those in Nicholson et al. (2020). For generality, we do not assume any specific sparsity structure. Nonetheless, our method can be adapted to obtain structured sparse AR coefficients' estimates by an appropriate modification of the penalty function in Section 3.

Estimation of the covariance $\boldsymbol{\Sigma}^{(m)}$ or precision $\boldsymbol{\Omega}^{(m)}$ matrices when the dimension d is large is even more challenging, as their MLE estimates can be particularly unstable. To circumvent this, one could either assume a sparse structure on these matrices, or use a shrinkage-based approach to estimate them (Fiecas et al., 2017). We focus on the former approach, and assume that the aforementioned matrices have many off-diagonal zero entries. The sparsity assumption on covariance or precision matrices is application-dependent (Friedman et al., 2008; Bien and Tibshirani, 2011), and we study both scenarios separately. Figure 5 illustrates an example of a sparse 3-regime MSVAR model for macroeconomic variables and the estimated sparse MSVAR parameters and regimes (Section 6).

3 Maximum penalized likelihood estimation

We estimate $\boldsymbol{\theta}$ by penalizing the conditional log-likelihood in (3) using an appropriate penalty function. This results in more stable parameter estimates. Specifically, we obtain

the maximum penalized conditional likelihood estimate (MPLE) of $\boldsymbol{\theta}$ as

$$\widehat{\boldsymbol{\theta}}_n \in \arg\max\Big\{\mathcal{L}_n(\boldsymbol{\theta}; s_p) := \tfrac{1}{n-p} l_n(\boldsymbol{\theta}; s_p) - R\big(\boldsymbol{\theta}; \lambda_1, \lambda_2\big)\Big\}, \tag{4}$$

$$\text{with} \qquad R\big(\boldsymbol{\theta}; \lambda_1, \lambda_2\big) := \sum_{m=1}^{M}\sum_{l=1}^{p}\sum_{i,j=1}^{d} r_{\lambda_1}\big(a_{l,ij}^{(m)}\big) + \sum_{m=1}^{M}\sum_{\substack{i,j=1 \\ j\neq i}}^{d} r_{\lambda_2}\big(\gamma_{ij}^{(m)}\big),$$

where $a_{l,ij}^{(m)}$ are the AR coefficients, and $\gamma_{ij}^{(m)}$ correspond to either $\sigma_{ij}^{(m)}$ or $\omega_{ij}^{(m)}$. The r_λ is a penalty function with tuning parameter $\lambda > 0$ that determines the severity of the penalty and consequently the level of sparsity of $\widehat{\boldsymbol{\theta}}_n$. In our numerical studies, we use the well-known penalties LASSO, adaptive LASSO (ADALASSO), MCP and SCAD, which are given in the Supplement, Section A1.1. General assumptions on the choice of r_λ are given in the Supplement, Section A2. By design, many entries of $\widehat{\boldsymbol{\theta}}_n$ corresponding to $\widehat{\boldsymbol{A}}^{(1:M)}$ and $\widehat{\boldsymbol{\Sigma}}^{(1:M)}$ or $\widehat{\boldsymbol{\Omega}}^{(1:M)}$ are zero, resulting in an estimated sparse MSVAR. Theorem 1 (Supplement, Section A2) states that under the correct specification of M and appropriate choices of (r_λ, λ), there exists a local maximizer $\widehat{\boldsymbol{\theta}}_n$ of the penalized conditional likelihood $\mathcal{L}_n(\boldsymbol{\theta}; s_p)$ that is a consistent and sparse estimator of the true parameter $\boldsymbol{\theta}^\star$. As there is no closed-form solution to the optimization in (4), we provide a numerical algorithm to approximate $\widehat{\boldsymbol{\theta}}_n$.

3.1 Modified EM algorithm

The EM algorithm is commonly used to perform MLE in a wide range of latent variable models, including MSVAR (Krolzig, 1997). The main advantage of this algorithm compared to a direct gradient ascent method for maximizing $\mathcal{L}_n(\boldsymbol{\theta}; s_p)$ is its ease of implementation in the presence of the latent regime-governing S_t. Nevertheless, we use gradient ascent method in the maximization (M-) step of the EM algorithm for obtaining the estimates of AR coefficient and covariance/precision matrices. We adapt the EM algorithm to our

penalization method by first introducing a so-called penalized complete-data likelihood.

For each S_t, we introduce a vector $\boldsymbol{\xi}_t := (\xi_{t1}, \ldots, \xi_{tM})^\mathsf{T}$, where $\xi_{ti} = \mathbb{1}_{\{S_t=i\}}$ represents the membership of $\mathbf{y}_t$ to regime i, so that $\sum_{i=1}^M \xi_{ti} = 1$. We denote the complete data as $\{(\mathbf{y}_t, \boldsymbol{\xi}_t), t = p+1, \ldots, n\}$, and the penalized complete conditional log-likelihood is

$$\mathcal{L}_n^{\mathrm{c}}(\boldsymbol{\theta}; s_p) := \sum_{i,j=1}^{M} \sum_{t=p+1}^{n} \frac{\xi_{(t-1)i}\xi_{tj} \log \alpha_{ij}}{n-p} + \sum_{i=1}^{M} \sum_{t=p+1}^{n} \frac{\xi_{ti} \log \phi\big(\mathbf{y}_t; \boldsymbol{\mu}_t^{(i)}, \boldsymbol{\Sigma}^{(i)}\big)}{n-p} - R\big(\boldsymbol{\theta}; \lambda_1, \lambda_2\big).$$

Clearly, the function $\mathcal{L}_n^{\mathrm{c}}$ cannot be used directly to estimate $\boldsymbol{\theta}$ since $\boldsymbol{\xi}_t$ are non-observable. Instead, at the $(k+1)$-th iteration, the algorithm proceeds in two steps as follows.

E-step: It computes the conditional expectation of $\mathcal{L}_n^{\mathrm{c}}$ with respect to $\boldsymbol{\xi}_t$ given $\mathbf{y}_{1:n}$ and the current update $\boldsymbol{\theta}^{(k)}$. This results in

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, s_p) := \mathbb{E}(\mathcal{L}_n^{\mathrm{c}} | \mathbf{y}_{1:n}, \boldsymbol{\theta}^{(k)}, s_p) &= \frac{1}{n-p} \sum_{i,j=1}^{M} \sum_{t=p+1}^{n} \zeta_{t,ij}^{(k)} \log \alpha_{ij} \\ &- \frac{1}{n-p} \sum_{i=1}^{M} \sum_{t=p+1}^{n} \zeta_{ti}^{(k)} \Big(\log |\boldsymbol{\Sigma}^{(i)}| + \frac{1}{2} (\mathbf{y}_t - \boldsymbol{\mu}_t^{(i)})^\mathsf{T} \boldsymbol{\Omega}^{(i)} (\mathbf{y}_t - \boldsymbol{\mu}_t^{(i)}) \Big) - R\big(\boldsymbol{\theta}; \lambda_1, \lambda_2\big), \end{aligned}$$

where, for $t = p+1, \ldots, n$ and $i, j = 1 \ldots, M$, the weights

$$\zeta_{t,ij}^{(k)} = \mathbb{P}(\xi_{(t-1)i} = 1, \xi_{tj} = 1 | \mathbf{y}_{1:n}, \boldsymbol{\theta}^{(k)}, s_p), \quad \zeta_{ti}^{(k)} = \mathbb{P}(\xi_{ti} = 1 | \mathbf{y}_{1:n}, \boldsymbol{\theta}^{(k)}, s_p), \tag{5}$$

are computed using a recursive forward-backward procedure as follows.

We define the vectors $\boldsymbol{\eta}_t := \big(\phi\big(\mathbf{y}_t; \boldsymbol{\mu}_t^{(1)}; \boldsymbol{\Sigma}^{(1)}\big), \ldots, \phi\big(\mathbf{y}_t; \boldsymbol{\mu}_t^{(M)}; \boldsymbol{\Sigma}^{(M)}\big)\big)^\mathsf{T}$ and $\widehat{\boldsymbol{\xi}}_{t|\tau} := \mathbb{E}(\boldsymbol{\xi}_t | \mathbf{y}_{1:\tau}) = \big(\mathbb{P}(\xi_{t1} = 1 | \mathbf{y}_{1:\tau}), \ldots, \mathbb{P}(\xi_{tM} = 1 | \mathbf{y}_{1:\tau})\big)^\mathsf{T}$, for $p \le \tau \le t$ and $t \ge p+1$. We first compute $\widehat{\boldsymbol{\xi}}_{t|t}$ (forward recursion) and use them to obtain $\widehat{\boldsymbol{\xi}}_{t|n}$ (backward recursion). By Bayes' rule,

$$\mathbb{P}(\xi_{tm} = 1 | \mathbf{y}_{1:t}) = \mathbb{P}(\xi_{tm} = 1 | \mathbf{y}_t, \mathbf{y}_{1:t-1}) = \frac{\mathbb{P}(\mathbf{y}_t | \xi_{tm} = 1, \mathbf{y}_{1:t-1}) \mathbb{P}(\xi_{tm} = 1 | \mathbf{y}_{1:t-1})}{\mathbb{P}(\mathbf{y}_t | \mathbf{y}_{1:t-1})}. \tag{6}$$

The first term in the numerator of (6) is the m-th entry of $\boldsymbol{\eta}_t$, $m = 1, \ldots, M$. For the second term in the numerator, using the total probability rule and the homogeneity assumption,

$$\begin{aligned}\mathbb{P}(\xi_{tm} = 1|\mathbf{y}_{1:t-1}) &= \sum_{i=1}^{M} \mathbb{P}(\xi_{tm} = 1|\xi_{(t-1)i} = 1, \mathbf{y}_{1:t-1})\mathbb{P}(\xi_{(t-1)i} = 1|\mathbf{y}_{1:t-1}) \\ &= \sum_{i=1}^{M} \alpha_{im}\mathbb{P}(\xi_{(t-1)i} = 1|\mathbf{y}_{1:t-1}),\ m = 1, \ldots, M \end{aligned} \tag{7}$$

or, in its vector form, $\widehat{\boldsymbol{\xi}}_{t|t-1} = \boldsymbol{P}^{\mathsf{T}}\widehat{\boldsymbol{\xi}}_{t-1|t-1}$. For the denominator in (6), we thus have

$$\begin{aligned}\mathbb{P}(\mathbf{y}_t|\mathbf{y}_{1:t-1}) &= \sum_{m=1}^{M} \mathbb{P}(\mathbf{y}_t|\xi_{(t-1)m} = 1, \mathbf{y}_{1:t-1})\mathbb{P}(\xi_{(t-1)m} = 1|\mathbf{y}_{1:t-1}) \\ &= \boldsymbol{\eta}_t^{\mathsf{T}}\boldsymbol{P}^{\mathsf{T}}\widehat{\boldsymbol{\xi}}_{t-1|t-1} = \boldsymbol{\eta}_t^{\mathsf{T}}\widehat{\boldsymbol{\xi}}_{t|t-1}. \end{aligned} \tag{8}$$

Using the definition of $\boldsymbol{\eta}_t$, (7) and (8), we then compute (6). Next, we compute the elements of $\widehat{\boldsymbol{\xi}}_{t|n}$ as follows. By Bayes' rule and the model assumptions, for $m = 1, \ldots, M$,

$$\begin{aligned}\mathbb{P}(\xi_{tm} = 1|\mathbf{y}_{1:n}) &= \sum_{i=1}^{M} \mathbb{P}(\xi_{tm} = 1, \xi_{(t+1)i} = 1|\mathbf{y}_{1:n}) \\ &= \sum_{i=1}^{M} \frac{\alpha_{mi}\ \mathbb{P}(\xi_{tm} = 1|\mathbf{y}_{1:t})\mathbb{P}(\xi_{(t+1)i} = 1|\mathbf{y}_{1:n})}{\mathbb{P}(\xi_{(t+1)i} = 1|\mathbf{y}_{1:t})}, \end{aligned} \tag{9}$$

which, as shown in Section A1.2 of the Supplement, uses that

$$\mathbb{P}(\xi_{tm} = 1, \xi_{(t+1)i} = 1|\mathbf{y}_{1:n}) = \frac{\mathbb{P}(\xi_{(t+1)i} = 1|\xi_{tm} = 1)\mathbb{P}(\xi_{tm} = 1|\mathbf{y}_{1:t})\mathbb{P}(\xi_{(t+1)i} = 1|\mathbf{y}_{1:n})}{\mathbb{P}(\xi_{(t+1)i} = 1|\mathbf{y}_{1:t})}. \tag{10}$$

Letting $\odot$ and $\oslash$ be the element-wise vector product and division, respectively, we write (6) and (9) in their vector forms,

- Forward recursion (filtering): for $t = p + 1, \ldots, n$,

$$\widehat{\boldsymbol{\xi}}_{t|t} = \frac{\boldsymbol{\eta}_t \odot \widehat{\boldsymbol{\xi}}_{t|t-1}}{\mathbf{1}_M^{\mathsf{T}}\left(\boldsymbol{\eta}_t \odot \widehat{\boldsymbol{\xi}}_{t|t-1}\right)} = \frac{\boldsymbol{\eta}_t \odot \boldsymbol{P}^{(k)}\widehat{\boldsymbol{\xi}}_{t-1|t-1}}{\mathbf{1}_M^{\mathsf{T}}\left(\boldsymbol{\eta}_t \odot \boldsymbol{P}^{(k)}\widehat{\boldsymbol{\xi}}_{t-1|t-1}\right)}. \tag{11}$$

- Backward recursion (smoothing): for $t = n-1, \ldots, p+1$,

$$\widehat{\boldsymbol{\xi}}_{t|n} = \left[\boldsymbol{P}^{(k)\mathsf{T}}\left(\widehat{\boldsymbol{\xi}}_{t+1|n} \oslash \widehat{\boldsymbol{\xi}}_{t+1|t}\right)\right] \odot \widehat{\boldsymbol{\xi}}_{t|t}. \tag{12}$$

Finally, we set $\zeta_{ti}^{(k)} = \mathbb{P}(\xi_{ti} = 1|\mathbf{y}_{1:n}, \boldsymbol{\theta}^{(k)}, s_p)$ in (5) as the i-th entry of the vector $\widehat{\boldsymbol{\xi}}_{t|n}$, for $i = 1, \ldots, M$. The joint probabilities $\zeta_{t,ij}^{(k)} = \mathbb{P}(\xi_{(t-1)i} = 1, \xi_{tj} = 1|\mathbf{y}_{1:n}, \boldsymbol{\theta}^{(k)}, s_p)$ in (5), for $i, j = 1, \ldots, M$, are the $((j-1)M + i)$-th entries of the vector

$$\widehat{\boldsymbol{\xi}}_{t|n}^{\text{joint}} := (\zeta_{t,11}^{(k)}, \ldots, \zeta_{t,MM}^{(k)})^{\mathsf{T}} = \text{vec}(\boldsymbol{P}^{(k)}) \odot \left[\left(\widehat{\boldsymbol{\xi}}_{t|n} \oslash \widehat{\boldsymbol{\xi}}_{t|t-1}\right) \otimes \widehat{\boldsymbol{\xi}}_{t-1|t-1}\right], \tag{13}$$

which are computed using (10), for $t = p+1, \ldots, n$. Here, $\otimes$ denotes the Kronecker product. The details of all the above derivations are given in Section A1.2 of the Supplement.

M-step: It maximizes Q with respect to $\boldsymbol{\theta}$, resulting in the updates $\boldsymbol{\theta}^{(k+1)}$. Due to the structure of Q, we separately maximize it with respect to $\boldsymbol{P}$, $\{\boldsymbol{\nu}^{(m)}\}_{m=1}^M$, and $\{\boldsymbol{A}^{(m)}, \boldsymbol{\Sigma}^{(m)}\}_{m=1}^M$.

The updates for the transition probabilities α_{ij} and the intercepts $\boldsymbol{\nu}^{(m)}$ are given by

$$\alpha_{s_p,j}^{(k+1)} = \frac{\sum_{t=p+1}^n \zeta_{t,s_p,j}^{(k)}}{\sum_{l=1}^M \sum_{t=p+1}^n \zeta_{t,s_p,l}^{(k)}}, \quad \alpha_{ij}^{(k+1)} = \frac{\sum_{t=p+2}^n \zeta_{t,ij}^{(k)}}{\sum_{l=1}^M \sum_{t=p+2}^n \zeta_{t,il}^{(k)}}, \quad i, j = 1 \ldots, M,\ i \neq s_p, \tag{14}$$

$$\boldsymbol{\nu}^{(m,k+1)} = \frac{\sum_{t=p+1}^n \zeta_{tm}^{(k)}(\mathbf{y}_t - \sum_{i=1}^p \boldsymbol{A}_i^{(m,k)}\mathbf{y}_{t-i})}{\sum_{t=p+1}^n \zeta_{tm}^{(k)}}, \quad m = 1, \ldots, M. \tag{15}$$

Regarding the AR coefficients, we first update $\boldsymbol{A}^{(m)}$, and then $\boldsymbol{\Sigma}^{(m)}$ or $\boldsymbol{\Omega}^{(m)}$, for each regime $m = 1, \ldots, M$. The optimization problems for these updates are, respectively,

$$\min_{\{\boldsymbol{A}_l^{(m)}\}_{l=1}^p} \frac{1}{2(n-p)} \sum_{t=p+1}^n \zeta_{tm}^{(k)}\big(\mathbf{y}_t - \bar{\boldsymbol{\mu}}_t^{(m)}\big)^{\mathsf{T}} \widehat{\boldsymbol{\Omega}}^{(m,k)}\big(\mathbf{y}_t - \bar{\boldsymbol{\mu}}_t^{(m)}\big) + \sum_{l=1}^p \sum_{i,j=1}^d r_{\lambda_1}\big(a_{l,ij}^{(m)}\big), \tag{16}$$

$$\min_{\boldsymbol{\Sigma}^{(m)} \succ 0} \frac{1}{2(n-p)}\Big(\widehat{n}_m \log|\boldsymbol{\Sigma}^{(m)}| + \text{tr}\big(\boldsymbol{\Omega}^{(m)}\boldsymbol{S}^{(m)}\big)\Big) + \sum_{i \neq j=1}^d r_{\lambda_2}\big(\gamma_{ij}^{(m)}\big), \tag{17}$$

where $\bar{\boldsymbol{\mu}}_t^{(m)}$ is (1) with $\boldsymbol{\nu}^{(m)}$ replaced by $\boldsymbol{\nu}^{(m,k+1)}$. Further, we let $\widehat{n}_m := \sum_{t=p+1}^n \zeta_{tm}^{(k)}$, $\boldsymbol{S}^{(m)} := \boldsymbol{U}^{\mathsf{T}}\boldsymbol{\Xi}^{(m)}\boldsymbol{U}$, with $\boldsymbol{\Xi}^{(m)} = \text{diag}\{\zeta_{p+1,m}^{(k)}, \ldots, \zeta_{n,m}^{(k)}\}$, and $\boldsymbol{U} = \big(\mathbf{y}_{p+1} - \boldsymbol{\mu}_{p+1}^{(m,k+1)}, \ldots, \mathbf{y}_n - \boldsymbol{\mu}_n^{(m,k+1)}\big)^{\mathsf{T}}$. The term $\boldsymbol{\mu}_t^{(m,k+1)}$ is (1) with $(\boldsymbol{\nu}^{(m)}, \boldsymbol{A}^{(m)})$ replaced by $(\boldsymbol{\nu}^{(m,k+1)}, \boldsymbol{A}^{(m,k+1)})$.

We decouple the problem (16) coordinate-wise to alleviate the computation. Then both optimization problems in (16) and (17) can be written in a general form as $\arg\min_{\mathbf{x}}\{f(\mathbf{x})+R(\mathbf{x})\}$, where f is differentiable and R is a penalty function. Using a generalized gradient descent, the solution is obtained via the updates

$$\mathbf{x}^{(j+1)} \in \arg\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{x} - (\mathbf{z}^{(j)} - \tfrac{1}{u_{j+1}}\nabla f(\mathbf{z}^{(j)}))\|^2 + \tfrac{1}{u_{j+1}}R(\mathbf{x}),$$

where $\mathbf{z}^{(j)}$ interpolates between $\mathbf{x}^{(j)}$ and $\mathbf{x}^{(j-1)}$, and $1/u_{j+1}$ is the descent step size. Since the penalty function R in (4) is decomposable, the updates have an analytical form for all the penalty functions we consider. Full details are given in the Supplement, Sections A1.3-A1.5, including an initialization of the EM algorithm and a data-dependent selection of tuning parameters (λ_1, λ_2). Algorithm 1 summarizes the estimation steps.

Algorithm 1 Modified EM algorithm for maximum penalized likelihood estimation

1: **Initialize**: $k=0$, $\alpha_{ij}^{(k)}$, $\boldsymbol{\nu}^{(m,k)}$, $\{\boldsymbol{A}_l^{(m,k)}\}_{l=1}^p$, $\boldsymbol{\Sigma}^{(m,k)}$, for $i,j,m=1,\ldots,M$, $\lambda_1,\lambda_2 \geq 0$, $\varepsilon > 0$;

2: **E-step**: Compute ${\zeta_{ti}}^{(k)}$, $\zeta_{t,ij}^{(k)}$ in (5), for $i,j=1,\ldots,M$, using (11), (12), and (13);

M-step: For $i,j=1,\ldots,M$, update $\alpha_{ij}^{(k+1)}$ using (14);

For $m=1,\ldots,M$, update $\boldsymbol{\nu}^{(m,k+1)}$ using (15);

For $m=1,\ldots,M$, update $\{\boldsymbol{A}_l^{(m,k+1)}\}_{l=1}^p$ and $\boldsymbol{\Sigma}^{(m,k+1)}$ or $\boldsymbol{\Omega}^{(m,k+1)}$ solving (16)-(17);

3: If $\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k+1)}\|/\|\boldsymbol{\theta}^{(k)}\| < \varepsilon$, **return** $\boldsymbol{\theta}^{(k+1)}$, else $k = k+1$ and **go to** 2;

3.2 Selection of the number of regimes M

We use information-based criteria to estimate M when it is unknown. For a fitted MSVAR with M regimes and MPLE $\widehat{\boldsymbol{\theta}}_n$, we compute the regime-specific degrees of freedom as

$$D_m(\widehat{\boldsymbol{\theta}}_n) = \sum_{l=1}^{p}\sum_{i,j=1}^{d} \mathbb{1}_{\{(\widehat{a}_{l,ij}^{(m)})\neq 0\}} + \sum_{i,j=1}^{d} \mathbb{1}_{\{\widehat{\gamma}_{ij}^{(m)}\neq 0\}}, \quad m=1,\ldots,M,$$

and let $E_m(M) = M - 1 + D_m(\widehat{\boldsymbol{\theta}}_n)$. We denote the information criterion as

$$\mathcal{C}(M) = -2l_n(\widehat{\boldsymbol{\theta}}_n; s_p) + \sum_{m=1}^{M} c_m E_m(M), \tag{18}$$

for some $c_m > 0$. The common choices for c_m are 2 (AIC) and $\log(n-p)$ (BIC). Städler and Mukherjee (2013) introduced the mixture minimum description length (MDL) for HMMs by setting $c_m = \log(\widehat{n}_m)$, where $\widehat{n}_m = \sum_{i=p+1}^{n} \widehat{\zeta}_{im}$ is the estimated sample size for regime m and $\widehat{\zeta}_{im}$ are given in (5). We select the number of regimes over a set of candidate values $\mathcal{M} = \{1, 2, \ldots, M_{\max}\}$ as $\widehat{M} \in \underset{M \in \mathcal{M}}{\arg\min}\, \mathcal{C}(M)$, for some pre-specified upper bound $M_{\max} \in \mathbb{N}$. Selection of the AR-order p is discussed in the Supplement, Section A1.6.

4 Prediction

Prediction is a highly relevant task in time series analysis. In stationary VAR models, an optimal prediction at time $n+1$ given the observations $\mathbf{y}_{1:n}$, in terms of the minimum mean squared prediction error, is equal to the conditional expectation (Lütkepohl, 2007); h-step-ahead prediction is performed similarly. In this section, we compute the predictive density in MSVARs that can also be used to obtain the conditional expectations. For a fixed $h \in \mathbb{N}$, we compute the h-step-ahead predictive density $f_M(\mathbf{y}_{n+1:n+h}|\mathbf{y}_{1:n})$ as follows. The one-step-ahead predictive density ($h = 1$) is

$$f_M(\mathbf{y}_{n+1}|\mathbf{y}_{1:n}) = \sum_{m=1}^{M} \mathbb{P}(S_{n+1} = m|\mathbf{y}_{1:n}) \phi\big(\mathbf{y}_{n+1}; \boldsymbol{\mu}_{n+1}^{(m)}, \boldsymbol{\Sigma}^{(m)}\big), \tag{19}$$

where the mean $\boldsymbol{\mu}_{n+1}^{(m)}$ is given in (1), and the conditional probabilities $\mathbb{P}(S_{n+1} = m|\mathbf{y}_{1:n}) = \sum_{j=1}^{M} \alpha_{jm} \mathbb{P}(S_n = j|\mathbf{y}_{1:n})$ can be readily computed using the forward-backward procedure (Supplement, Section A1.2). An initial conditional distribution $\mathbb{P}(S_p = j|\mathbf{y}_{1:p})$ is also re-

quired to complete the computation. Since its effect on predictive density dissipates geometrically as n grows (Douc et al., 2004), and given the homogeneity of S_t, any non-informative distribution such as the discrete uniform $\mathbb{P}(S_p = j|\mathbf{y}_{1:p}) = 1/M, j = 1, \ldots, M$, suffices.

Similarly, for $h \geq 2$, by the model assumptions, the predictive density is

$$\begin{aligned} f_M(\mathbf{y}_{n+1:n+h}|\mathbf{y}_{1:n}) &= \sum_{m_{1:h}=1}^{M} \Big\{ \mathbb{P}(S_{n+1} = m_1, \ldots, S_{n+h} = m_h|\mathbf{y}_{1:n}) f(\mathbf{y}_{n+1:n+h}|\mathbf{y}_{1:n}, S_{n+1:n+h} = m_{1:h}) \Big\} \\ &= \sum_{m_{1:h}=1}^{M} \mathbb{P}(S_{n+1} = m_1|\mathbf{y}_{1:n}) \Big(\prod_{t=2}^{h} \alpha_{m_{t-1},m_t} \Big) \Big(\prod_{t=1}^{h} \phi\big(\mathbf{y}_{n+t}; \boldsymbol{\mu}_{n+t}^{(m_t)}, \boldsymbol{\Sigma}^{(m_t)}\big) \Big) \quad (20) \end{aligned}$$

Given the MPLE $\widehat{\boldsymbol{\theta}}_n$ for a model with M regimes, the estimated predictive density $\widehat{f}_M(\mathbf{y}_{n+1:n+h}|\mathbf{y}_{1:n})$ is computed for $h \geq 1$, using either (19) or (20). The optimal h-step ahead predictor $\widehat{\mathbf{y}}_{n+h}$ in terms of minimum mean squared error is then the conditional expected value of $\mathbf{Y}_{n+h}$, given $\mathbf{y}_{1:n}$. For $h = 1$, the predicted value is

$$\widehat{\mathbf{y}}_{n+1} = \sum_{m=1}^{M} \widehat{\mathbb{P}}(S_{n+1} = m|\mathbf{y}_{1:n}) \left(\widehat{\boldsymbol{\nu}}^{(m)} + \sum_{l=1}^{p} \widehat{\boldsymbol{A}}_l^{(m)} \mathbf{y}_{n+1-l} \right).$$

For $h \geq 2$, predictions are obtained recursively by computing the predictions for $\mathbf{y}_{n+j}$, $j = 1, \ldots, h-1$ (Lütkepohl, 2007; Krolzig, 1997). More specifically,

$$\widehat{\mathbf{y}}_{n+h} = \sum_{m=1}^{M} \widehat{\mathbb{P}}(S_{n+h} = m|\mathbf{y}_{1:n}) \left(\widehat{\boldsymbol{\nu}}^{(m)} + \sum_{l=1}^{p} \widehat{\boldsymbol{A}}_l^{(m)} \widetilde{\mathbf{y}}_{n+h-l} \right),$$

with $\widetilde{\mathbf{y}}_{n+h-l} = \widehat{\mathbf{y}}_{n+h-l}$, if $(h-l) > 0$ or $\mathbf{y}_{n+h-l}$, if $(h-l) \leq 0$.

Theorem 2 (Supplement, Section A2) states that, if M is not under-specified, the estimated predictive density of the over-fitted model based on MPLE consistently estimates the h-step ahead predictive density (20) of the true MSVAR model. As a consequence, in practice when the true number of regimes is unknown, a rather conservative choice of M considering the sample size n can guarantee a reliable estimate of (20), and of the optimal predictor in the sense of minimum mean squared prediction error.

5 Simulation study

Simulation design. We consider MSVAR models with $p = 1$, $M = 2, 3$, dimensions $d = 20, 40, 100$, and sample sizes n between 200 and 800. For these values of d, the parameter vector $\boldsymbol{\theta}^\star$ of the data-generating MSVAR models has 104, 202 and 504 nonzero entries, respectively, and is designed as follows. We set the number of nonzero entries per matrix to be $\lceil d/4 \rceil$, randomly scattered on each matrix, and thus not assuming a specific sparsity structure in the AR and covariance or precision matrices. We set the true nonzero AR coefficients for each regime $m = 1, \ldots, M$ as: $\nu_i^{(m)^\star} \sim \mathcal{N}(m-2, 0.01)$, $(a_l^{(m)^\star})_{ij} \sim \mathcal{N}((-1)^m, 0.01)$. We ensure that the resulting MSVAR is stationary. Note that a VAR process with AR coefficients $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ is stationary if the matrix

$$\begin{bmatrix} \boldsymbol{A}_1 \cdots \boldsymbol{A}_{p-1} & \boldsymbol{A}_p \\ \boldsymbol{I}_{d(p-1)\times d(p-1)} & \boldsymbol{0}_{d(p-1)\times d} \end{bmatrix} \tag{21}$$

has all its eigenvalues inside the unit circle (Lütkepohl, 2007, Section 9.1). In all of our simulations, to ensure stationarity of the VAR process for each Markov regime, we scale the AR matrices with entries $(a_l^{(m)^\star})_{ij}$ (generated as described above) to ensure that the modulus of each eigenvalue of the matrix in (21) is at most 0.75.

For the covariance matrices, we first sample $\gamma_{ij}^{(m)} \sim \mathcal{N}(\frac{(-1)^m}{2}, 0.01), i \neq j$, $\gamma_{ii}^{(m)} \sim \mathcal{N}(0, 0.01)$, and subsequently use one of the following formulas to obtain the nonzero elements, depending on the scenario, either covariance or precision matrices, under study: $\sigma_{ij}^{(m)^\star} = \frac{1}{2}(\gamma_{ij}^{(m)} + \gamma_{ji}^{(m)})$, $\omega_{ij}^{(m)^\star} = \frac{1}{2}(\gamma_{ij}^{(m)} + \gamma_{ji}^{(m)})$. To ensure positive-definiteness and numerical stability, if necessary, we shift the eigenvalues of matrices by adding a multiple of the identity matrix. This also aids in controlling the condition number (Rothman et al.,

2008). We denote scenarios **S1** and **S2** respectively corresponding to sparse covariance and precision matrices, along with sparse AR coefficients.

The transition probability matrix $\boldsymbol{P}^\star$ for $M = 2$ is set to: $\alpha_{11}^\star = 0.75, \alpha_{22}^\star = 0.60$; and for $M = 3$ to: $\alpha_{11}^\star = 0.5, \alpha_{12}^\star = 0.25, \alpha_{21}^\star = 0.2, \alpha_{22}^\star = 0.6, \alpha_{32}^\star = 0.25, \alpha_{33}^\star = 0.5$. These choices test the estimation performance for an unequal number of data points in each regime.

Performance measures. Recall the MPLE $\widehat{\boldsymbol{\theta}}_n \equiv \widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$, and let $\widetilde{\boldsymbol{\theta}}_n \equiv \widetilde{\boldsymbol{\theta}}$ be the maximum likelihood estimate of $\boldsymbol{\theta}^\star$ knowing its zero entries a priori. For a subvector $\mathbf{v}$ of $\boldsymbol{\theta}^\star$, let $\widehat{\mathbf{v}}$ and $\widetilde{\mathbf{v}}$ be the corresponding subvectors of $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$, respectively. Furthermore, let $\widehat{M}$ be the estimate of the true number of regimes M using either BIC or MDL introduced in Section 3.2. We consider the following performance measures:

- relative estimation error: $\mathrm{REE}(\widehat{\mathbf{v}}, \widetilde{\mathbf{v}}) = \mathrm{EE}(\widehat{\mathbf{v}})/\mathrm{EE}(\widetilde{\mathbf{v}}) = \frac{||\widehat{\mathbf{v}}-\mathbf{v}||}{||\mathbf{v}||} / \frac{||\widetilde{\mathbf{v}}-\mathbf{v}||}{||\mathbf{v}||} \in [0, \infty)$;
- true positive rate: $\mathrm{TPR}(\widehat{\mathbf{v}}) = \frac{\sum_i \mathbb{1}_{\{|\widehat{v}_i|>0\}} \mathbb{1}_{\{|v_i|>0\}}}{\sum_i \mathbb{1}_{\{|v_i|>0\}}} \in [0, 1]$;
- true negative rate: $\mathrm{TNR}(\widehat{\mathbf{v}}) = \frac{\sum_i \mathbb{1}_{\{|\widehat{v}_i|=0\}} \mathbb{1}_{\{|v_i|=0\}}}{\sum_i \mathbb{1}_{\{|v_i|=0\}}} \in [0, 1]$;
- correctly selected number of regimes: $\mathbb{1}_{\{\widehat{M}=M\}} \in \{0, 1\}$,

where $||\cdot||$ represents the Euclidean norm and $\mathbb{1}_{\{\cdot\}}$ is the indicator function. We compute EE, REE, TPR and TNR for different subvectors $\mathbf{v}$ corresponding to the AR coefficients and the covariance or precision matrices. We exclude the AR intercepts and the diagonals of the covariance or precision matrices when computing TPR and TNR.

Since in the computation of $\widetilde{\boldsymbol{\theta}}$ we use the knowledge of the true zero parameters, in general, we expect $\widetilde{\boldsymbol{\theta}}$ to outperform $\widehat{\boldsymbol{\theta}}$ which estimates the zero parameters through penalization. Thus, we focus on the REE to quantify the performance of our estimation methodology relative to $\widetilde{\boldsymbol{\theta}}$, and seek to obtain the empirical mean or median of all the performance measures, except EE, close to 1. However, in the case of estimation of precision

matrices, we observe REE < 1 which occurs due to the way we compute $\widetilde{\boldsymbol{\theta}}$ without directly using the knowledge of the true zero parameters in the precision matrices. More details are provided in the Supplement, Section A3.

Our results are based on 50 simulated samples from each MSVAR model with a fixed parameter setting as described in the **Simulation design**. We start recording the generated observations at $t = 401$, to allow for a simulation burn-in. The first state is always set as $s_{-399} = 1$. For the estimation stage, we condition on $s_p = 1$. We analyze the results for scenario **S1** (sparse covariance matrix) here; those for scenario **S2** (sparse precision matrix) are given in the Supplement, Section A3, along with additional simulation designs on sparsity structure. We use BIC for selection of the tuning parameters (λ_1, λ_2) using the method described in the Supplement, Section A1.5, as we find that AIC chooses over-parameterized models and therefore has a decreased performance.

The algorithms are implemented in `C` using the GNU Scientific Library (GSL) (Galassi et al., 2009) v2.7. The optimization procedure is implemented in parallel so that the estimation is done for all the regimes in an MSVAR model simultaneously.

5.1 Parameter estimation and sparsity

We fix $M = 2$, $p = 1$, and consider dimensions $d = 20$ with sample sizes $n \in \{200, 300, 400\}$, $d = 40$ with $n \in \{300, 400, 500\}$, and $d = 100$ with $n \in \{600, 700, 800\}$. For these values of d, the parameter vector $\boldsymbol{\theta}$ to be estimated by our method has dimensions $K = M\big(d + pd^2 + d(d+1)/2\big) + M(M-1) = 1262$, 4922 and 30302, respectively. Thus, even though the change from $d = 20$ to 100 may appear moderate, the corresponding number of parameters to be estimated is very large. The results for $d = 40, 100$ are discussed below, and those

for $d = 20$ are given in Section A3 of the Supplement; see also the end of this section.

Figure 1 shows the REE boxplots. We observe that for the AR coefficients, the ADALASSO and SCAD attain a performance close to $\widetilde{\boldsymbol{\theta}}$, followed by MCP and the LASSO. For covariance matrices, SCAD and MCP outperform the other penalties, and the medians reveal that they are comparable to $\widetilde{\boldsymbol{\theta}}$. Overall, SCAD and MCP attain a performance close to $\widetilde{\boldsymbol{\theta}}$, followed by the LASSO and ADALASSO. Estimation of covariance matrices is more difficult, as the medians depart from 1.0 more noticeably. In computing the overall REE, the estimation errors of the transition probabilities are included.

Figure 2 shows the TPR boxplots. We observe that the MPLE performs well under this measure. At least one penalty function attains an overall median TPR above 0.90 for the dimensions and sample sizes considered. The LASSO, SCAD and MCP perform similar to each other, while ADALASSO is less accurate. In terms of TNR (omitted), we find that for any setting, the median is at least 0.95 for both the AR and covariance matrices.

In summary, even though our theoretical results are proved in the setting of fixed dimension and large sample size, we conclude from the above experiments that our methodology has a good performance in finite-sample and moderate- to large-dimensions.

In Section A3 of the Supplement, we also provide the simulation results in terms of the EE and TPR for $d = 20$ and a wider range of sample sizes n to demonstrate the consistency and sparsity recovery property (Theorem 1) of the MPLE as n grows.

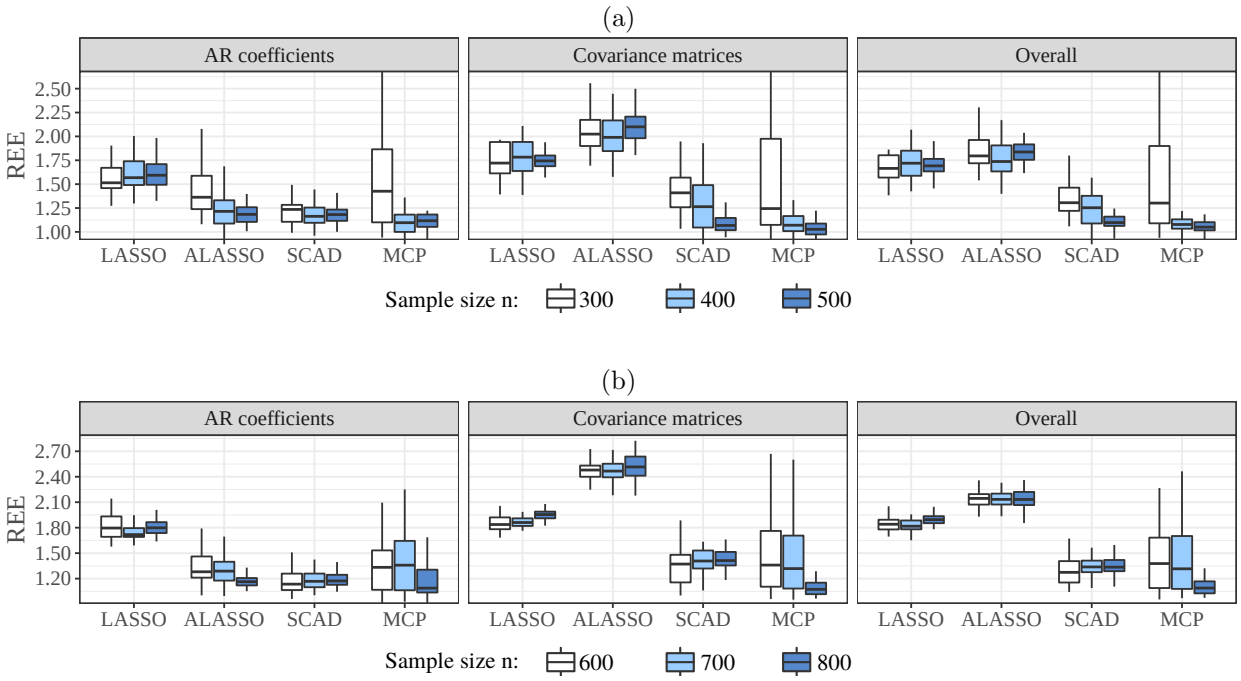

Figure 1: *Relative estimation error (REE) based on 50 samples: (a)* $d = 40, K = 4922$ *(b)* $d = 100, K = 30302$*, where* d *and* K *are resp. the data and parameter dimensions.*

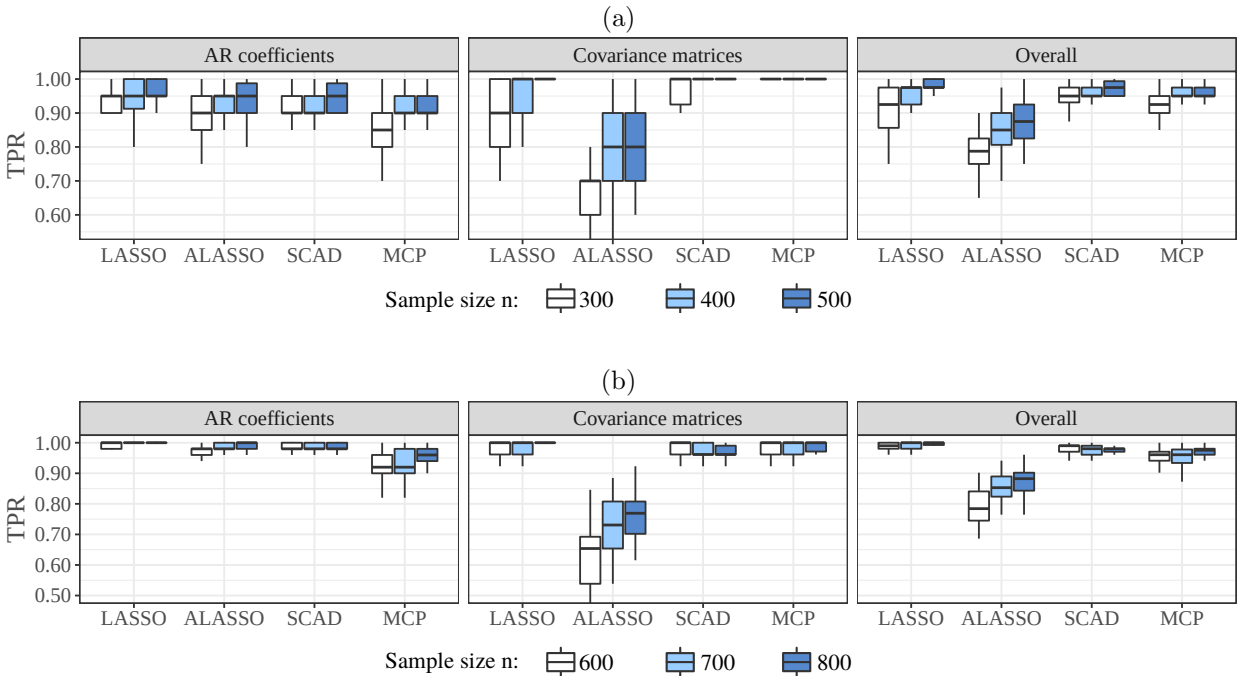

Figure 2: *True positive rate (TPR, nonzero parameter detection) based on 50 random samples: (a)* $d = 40, K = 4922$ *(b)* $d = 100, K = 30302$*, where* d *and* K *are resp. the data and parameter dimensions.*

5.2 Predictive density estimation and number of regime selection

We use models with $M \in \{2, 3\}$, $p = 1$, and generate samples of size $n+h$. We take $d = 20$, $n = 800$, and $h = 8$. We estimate the model parameters using n observations, setting the number of regimes M as either correctly or under/over-specified, ranging from 1 to 5. For the remaining h observations, we compute the estimated predictive densities in (20).

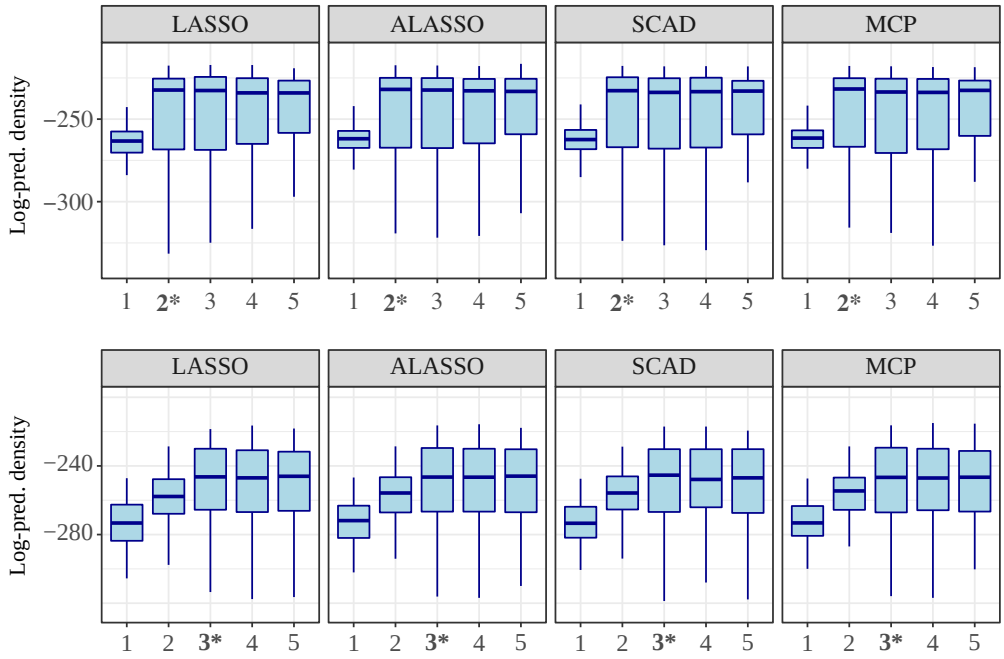

Figure 3: *h-step-ahead log-predictive density* (20) *based on 50 random samples, for the true number of regimes* $M = 2^*$ *(top),* 3^* *(bottom),* $d = 20$, $n = 800$, $h = 8$. *The x-axis shows* M *values used in estimation.*

Figure 3 shows boxplots of the estimated h-step ahead log-predictive density of $\mathbf{y}_{n+1:n+h}$, for different values of M. The results are based on 50 generated samples. We observe that the median values for the models with over-specified M fall within the inter-quantile range of the values for the model with correct M, which is expected as per the result of our Theorem 2 on the predictive density. The median for the under-specified M clearly deviates from the other medians, attaining a lower value. This is true for all the penalties.

Next, we examine the performance of the criterion (18) for estimating M when the AR-order p is fixed. The results for estimating p are given in the Supplement, Section A1.6. We consider $d = 20$ and $n \in \{400, 600, 800\}$. We generate the data using models with true number of regimes $M = 1, 2, 3$, and estimate M using the BIC over the candidate values ranging from 1 to 5 for all the penalties. The mean of the performance measure $\mathbb{1}_{\{\widehat{M}=M\}}$, with $\widehat{M}$ being the estimated number of regimes, is reported in Figure 4. The ADALASSO outperforms other penalties and estimates the true M with a mean rate of at least 80% for any sample size. The performance for $n = 600, 800$ is very good with rates above 95%. The LASSO also appears more stable than SCAD and MCP. The performance of the MDL is not as accurate and omitted, which can be attributed to the small sample size per regime.

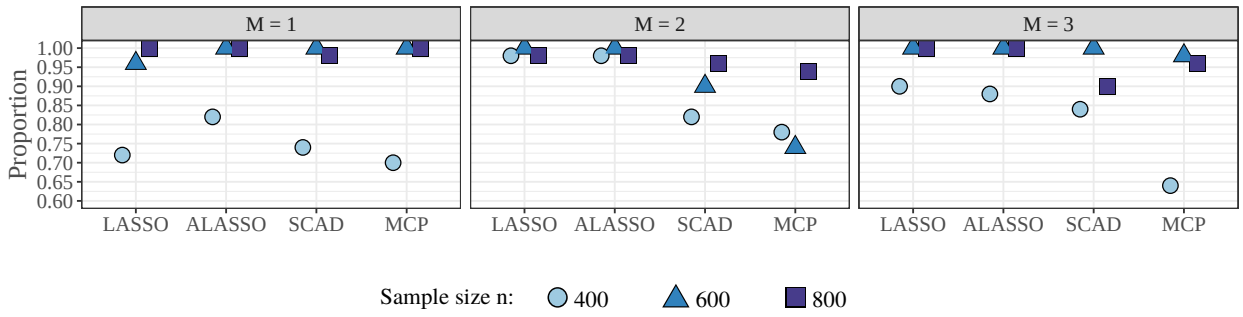

Figure 4: *Proportion of correctly selected M using BIC based on 50 random samples for $d = 20$.*

6 Case study: Identification of a recessionary regime

We consider a large-scale dataset of Canadian national and provincial macroeconomic indicators of observations from 1981 to 2022, containing 481 observations on 386 variables. The indicators belong to several categories, such as labour market, housing, construction and

manufacturing. Fortin-Gagnon et al. (2018) analyzed a part of this dataset using probit and logit LASSO models to identify the turning points of three economic recession events, and predict recession probabilities. As MSVARs have been used to successfully model business cycles (Hamilton, 1988), our aim is to fit a model to this data and identify the recessionary regimes. The effects of manufacturing decline and house boom on US non-employment levels have been studied before (Charles et al., 2019). With this motivation, we consider housing price index (HPI) and manufacturing index (MI) for the top 5 Canadian provinces in terms of population[1], hence the time series $\mathbf{Y}_t$ has dimension $d = 10$. We fit MSVARs with $M = 1, 2, 3, 4$, $p = 1, 2, 3$, each penalty, and scenarios of sparse covariance and precision matrices. We use BIC and MDL, and the AR-order selection method based on the hierarchical group-LASSO, to choose the final model based on the ADALASSO with $\widehat{M} = 3$, $\widehat{p} = 1$ and sparse covariance matrices; see Supplement, Section A4 for all the details.

Figure 5 displays the estimated AR coefficients and covariance matrices for each regime, with transition probabilities $(\widehat{\alpha}_{11}, \widehat{\alpha}_{12}) = (0.81, 0.13)$, $(\widehat{\alpha}_{21}, \widehat{\alpha}_{22}) = (0.23, 0.69)$, and $(\widehat{\alpha}_{31}, \widehat{\alpha}_{32}) = (0.40, 0.40)$. We also show the estimated regime for each observation at time t as $\arg\max_m \widehat{\mathbb{P}}(s_t = m|\mathbf{y}_{p+1:n})$. In the same figure, we also mark the recession periods of the Canadian economy, which took place over the periods: June 1981 to October 1982, March 1990 to April 1992, and October 2008 to May 2009.

The logit model used in Fortin-Gagnon et al. (2018) provided smooth estimates of the predictive recession probabilities, which overlapped with the actual recessionary periods. From Figure 5, our fitted MSVAR is also able to identify a regime whose occurrence prob-

[1]Statistics Canada, Table 17-10-0009-01; DOI: `https://doi.org/10.25318/1710000901-eng`

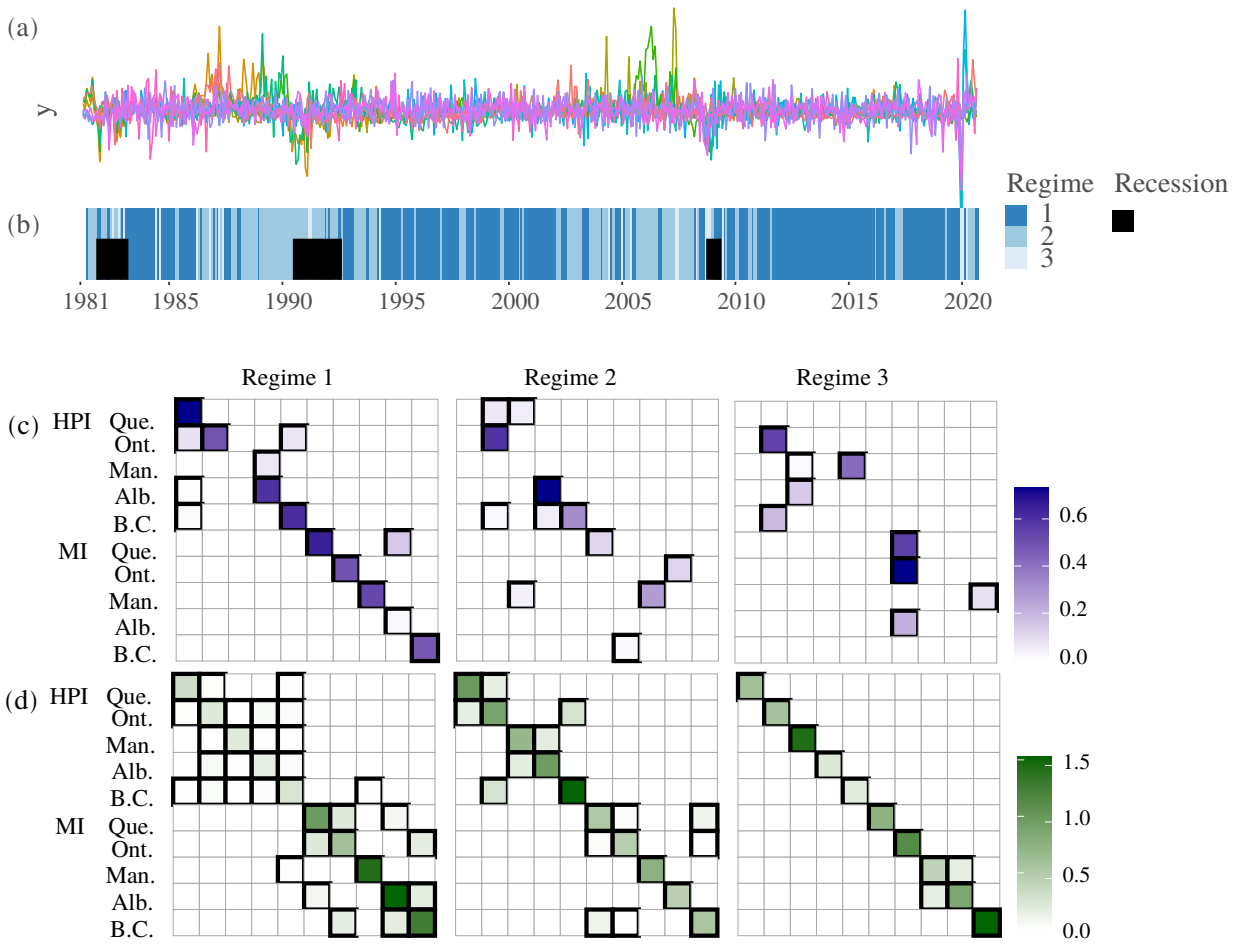

Figure 5: *(a) Time series of* $d = 10$ *Canadian province macroeconomic indicators: 5 housing price (HPI) and 5 manufacturing (MI) indices. (b) Estimated most probable membership of each observation to any of* $\widehat{M} = 3$ *regimes, and recession periods of Canadian economy. (c) Absolute value of estimated regime-AR coefficients, AR-order* $\widehat{p} = 1$*. (d) Absolute value of estimated regime-covariance matrix entries.*

ability is substantially high before and during the recession periods. This corresponds to regime 2, and agrees with the general definition of a recession as a persistent decline in an economy with a specified minimum duration. The plot also shows that regime 3 captures particularly extreme values of some variables (HPI-Ont., HPI-Man., HPI-B.C., MI-Que., MI-Ont., MI-Alb.) and seems to appear more often around the recessionary periods. A proportion of 5% (25) observations are classified to regime 3. Figure 5(b) suggests that our method is able to detect the recession periods based on 2 indicators from 5 provinces, compared to 368 national and provincial variables used in Fortin-Gagnon et al. (2018).

Prediction. We consider the 8-month recessionary period over October 2008 to May 2009, and the subsequent non-recessionary 8-month period over June 2009 to January 2010, for a total of 16 test observations. We fit a sparse MSVAR model ($\widehat{M} = 3, \widehat{p} = 1$) using all the data up to the month prior to a test point under consideration, and perform an out-of-sample one-step-ahead regime-membership prediction. For a test point, we determine that it belongs to a recession if its predicted maximum regime probability corresponds to regimes either 2 or 3.

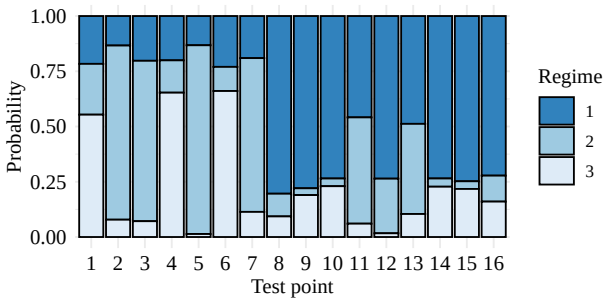

Figure 6: *Out-of-sample one-step-ahead predicted regime membership probabilities for the Canadian province macroeconomic test data. The first 8 points correspond to the third Canadian recession, the last 8 points are non-recessionary.*

Figure 6 shows the predicted probabilities for the 16 test observations. Note that even though we correctly identify the regimes for 14 test points, caution must be taken to predict recessions using the method since, as stated by Fortin-Gagnon et al. (2018), the presence of a recession is rather a rare event in the dataset.

7 Conclusion

We introduce a general penalized conditional likelihood method for fitting sparse MSVARs. It provides both sparse regime-specific AR coefficients and covariance or precision matrices, along with estimation of transition probabilities of the Markov chain. We prove consistency and sparsity properties of the proposed estimators. We also discuss estimation of the number of regimes and predictive density. We implement the method via an efficient modified EM algorithm. We examine the performance of the method via simulation and real data analysis. Simulations indicate that our method is reliable for fitting sparse MSVARs to moderate- and large-dimensional time series data. In practice, an increase in the number of regimes M and AR-order p has impact on the performance of the method, which is expected as the optimization problem (4) becomes more challenging. Our simulations also show that none of the penalties dominates the others in all the settings considered here. Thus, in practice, we suggest analyzing a dataset using all the penalties and choosing a sparse model that optimizes a selection criterion (e.g., BIC) as done in our real data analysis.

It can be shown that the BIC based on (18) does not under-estimate the true number of regimes M. A possible research direction is to investigate conditions under which BIC consistently estimates M. Another direction is the extension of our method to nonhomoge-

neous Markov chains. Shrinkage estimators can be incorporated by adjusting the penalty in (4) and also allowing the dimension d to increase with the sample size n as in the work of Fiecas et al. (2017), which require further investigation.

Supplementary materials

A Supplementary pdf document to this paper includes: (i) additional methodology details including the EM algorithm, (ii) theoretical results and proofs, (iii) complementary numerical results, and (iv) complementary case study results.

Code: "code_msvar.zip" contains the code to replicate Figure 1(a) and Figure 2(a).

Acknowledgments

The authors would like to thank the editor, an associate editor, and two referees for detailed comments and suggestions that led to an improved paper.

Funding

G. Chavez-Martinez is supported by the Mexican Council of Science and Technology (CONACyT) under the PhD scholarships program. A. Agarwal is supported by the University of Glasgow Early Career Mobility Scheme. A. Khalili and S.E. Ahmed are supported by the Natural Science and Engineering Research Council of Canada (NSERC RGPIN-2020-05011) and (NSERC RGPIN-2017-05228).

Disclosure statement: The authors report there are no competing interests to declare.

References

Ang, A. and Timmermann, A. (2012). Regime changes and financial markets. *Annual Review of Financial Economics*, 4(1):313–337.

Basu, S. and Matteson, D. S. (2021). A survey of estimation methods for sparse high-dimensional time series models. Unpublished manuscript.

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

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820.

Charles, K. K., Hurst, E., and Notowidigdo, M. J. (2019). Housing booms, manufacturing decline and labour market outcomes. *The Economic Journal*, 129(617):209–248.

Douc, R., Moulines, E., and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of Statistics*, 32(5):2254–2304.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fiecas, M., Franke, J., von Sachs, R., and Tadjuidje Kamgaing, J. (2017). Shrinkage estimation for multivariate hidden Markov models. *Journal of the American Statistical Association*, 112(517):424–435.

Fortin-Gagnon, O., Leroux, M., Stevanovic, D., and Surprenant, S. (2018). A large Canadian database for macroeconomic analysis. Technical report, CIRANO Working Paper 2018S-25.

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

Galassi, M., Davies, J., Theiler, J., Gough, B., and Jungman, G. (2009). *GNU Scientific Library Reference Manual*. Springer Science & Business Media. ISBN 0954612078.

Hamilton, J. D. (1988). Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates. *Journal of Economic Dynamics and Control*, 12(2-3):385–423.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393.

Krolzig, H.-M. (1997). *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*, volume 454. Springer Science & Business Media.

Lam, C. (2020). High-dimensional covariance matrix estimation. *WIREs Computational Statistics*, 12(2):e1485.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business.

Monbet, V. and Ailliot, P. (2017). Sparse vector Markov switching autoregressive models. Application to multivariate time series of temperature. *Computational Statistics & Data Analysis*, 108:40–51.

Nicholson, W. B., Wilms, I., Bien, J., and Matteson, D. S. (2020). High-dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, 21(166):1–52.

Pinson, P. and Madsen, H. (2012). Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models. *Journal of Forecasting*, 31(4):281–313.

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538.

Städler, N. and Mukherjee, S. (2013). Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. *The Annals of Applied Statistics*, 7(4):2157–2179.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1).

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.