# Supplement to "Penalized estimation of sparse Markov regime-switching vector auto-regressive models"

Gilberto Chavez-Martinez * Ankush Agarwal † Abbas Khalili *

Syed Ejaz Ahmed ‡

April 3, 2023

In this supplementary document, we provide complete details of the modified EM algorithm, proofs of the theoretical results on the asymptotic convergence and consistency of our penalized maximum likelihood estimates and predictive density estimates, and the results of additional numerical experiments not included in the main document. In the following, we denote the true model parameter vector as $\boldsymbol{\theta}^\star \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{\mathrm{K}}$, and denote a generic element of a parameter vector $\boldsymbol{\theta}$ as $\theta_j, j = 1, \ldots, \mathrm{K}$.

## A1 Methodology details

### A1.1 Penalty functions

The well-studied choices of the penalty function $r_\lambda$ in Section 3 of the main manuscript are:

- LASSO ($\mathrm{L}_1$ norm): $r_\lambda(\theta_j) = \lambda|\theta_j|$;
- Adaptive LASSO (ADALASSO, weighted $\mathrm{L}_1$ norm): $r_\lambda(\theta_j) = \lambda\widehat{w}_j|\theta_j|$, with $\widehat{w}_j = |\widehat{\theta}_j|^{-\gamma}$ and $\gamma > 0$, where $\widehat{\theta}_j$ is a $\sqrt{n}$–consistent estimator of the true parameter $\theta_j^\star$;
- SCAD: $r_\lambda(\theta_j)$ is such that $r'_\lambda(\theta) = \mathrm{sgn}(\theta)\left(\lambda\mathbb{1}_{\{|\theta|\leq\lambda\}} + \frac{(a\lambda - |\theta|)_+}{(a-1)}\mathbb{1}_{\{\theta>\lambda\}}\right)$, with parameter $a > 2$;
- MCP: $r_\lambda(\theta_j)$ is such that $r'_\lambda(\theta) = \mathrm{sgn}(\theta)\frac{(a\lambda - |\theta|)_+}{a}$, with parameter $a > 1$.

In Section A2 below we list general conditions on $r_\lambda$ required to prove our theoretical results.

### A1.2 Expectation (E-) step

Here we provide details of the forward-backward algorithm to compute the quantities

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

*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

at the $(k+1)$-th iteration of the EM algorithm. The *forward recursion* is also known as filtering, and the *backward recursion* as smoothing. We omit the terms $(\boldsymbol{\theta}^{(k)}, s_p)$ from the notation throughout.

Recall the variable

$$\boldsymbol{\xi}_t := \begin{bmatrix} \xi_{t1} \\ \vdots \\ \xi_{t\mathrm{M}} \end{bmatrix} = \begin{bmatrix} \mathbb{1}_{\{s_t=1\}} \\ \vdots \\ \mathbb{1}_{\{s_t=\mathrm{M}\}} \end{bmatrix}.$$

Let $\boldsymbol{\iota}_m$ denote the $m$-th column of the M-dimensional identity matrix. Using the model assumptions, we can write, for $t \geq p+1$,

$$\begin{aligned}
\mathbb{P}(\mathbf{y}_t|\boldsymbol{\xi}_{t-1} = \boldsymbol{\iota}_i, \mathbf{y}_{1:t-1}) &:= \mathbb{P}(\mathbf{y}_t|\xi_{(t-1)i} = 1, \mathbf{y}_{1:t-1}) \\
&= \sum_{m=1}^{\mathrm{M}} \mathbb{P}(\mathbf{y}_t|\xi_{(t-1)i} = 1, \xi_{tm} = 1, \mathbf{y}_{1:t-1}) \mathbb{P}(\xi_{tm} = 1|\xi_{(t-1)i} = 1, \mathbf{y}_{1:t-1}) \\
&= \sum_{m=1}^{\mathrm{M}} \mathbb{P}(\mathbf{y}_t|\xi_{tm} = 1, \mathbf{y}_{1:t-1}) \mathbb{P}(\xi_{tm} = 1|\xi_{(t-1)i} = 1) \\
&= \sum_{m=1}^{\mathrm{M}} \mathbb{P}(\mathbf{y}_t|\xi_{tm} = 1, \mathbf{y}_{1:t-1}) \alpha_{im} = \sum_{m=1}^{\mathrm{M}} \alpha_{im} \phi\big(\mathbf{y}_t; \boldsymbol{\mu}_t^{(m)}; \boldsymbol{\Sigma}^{(m)}\big)\big).
\end{aligned}$$

We collect the densities of $\mathbf{y}_t$ conditional on $\boldsymbol{\xi}_t$ and $\mathbf{y}_{1:t-1}$ in the vector

$$\boldsymbol{\eta}_t := \begin{bmatrix} \mathbb{P}(\mathbf{y}_t|\boldsymbol{\xi}_t = \boldsymbol{\iota}_1, \mathbf{y}_{1:t-1}) \\ \vdots \\ \mathbb{P}(\mathbf{y}_t|\boldsymbol{\xi}_t = \boldsymbol{\iota}_{\mathrm{M}}, \mathbf{y}_{1:t-1}) \end{bmatrix} = \begin{bmatrix} \phi\big(\mathbf{y}_t; \boldsymbol{\mu}_t^{(1)}; \boldsymbol{\Sigma}^{(1)}\big) \\ \vdots \\ \phi\big(\mathbf{y}_t; \boldsymbol{\mu}_t^{(\mathrm{M})}; \boldsymbol{\Sigma}^{(\mathrm{M})}\big) \end{bmatrix}.$$

Thus, we can rewrite the above conditional density as

$$\mathbb{P}(\mathbf{y}_t|\boldsymbol{\xi}_{t-1} = \boldsymbol{\iota}_i, \mathbf{y}_{1:t-1}) = \boldsymbol{\eta}_t^{\mathsf{T}} \mathbf{P}^{\mathsf{T}} \boldsymbol{\iota}_i. \tag{1}$$

As the regime-governing Markov chain is unobservable, the information at $t-1$ only consists of observations $\mathbf{y}_{1:t-1}$ and not the regime indicator vector $\boldsymbol{\xi}_{t-1}$. For the purpose of estimation, we replace the vector $\boldsymbol{\xi}_{t-1}$ by its conditional expected value, which in itself is estimated from the observed data as follows.

Denote the vectors

$$\widehat{\boldsymbol{\xi}}_{t|\tau} := \mathbb{E}(\boldsymbol{\xi}_t|\mathbf{y}_{1:\tau}) = \begin{bmatrix} \mathbb{P}(\boldsymbol{\xi}_t = \boldsymbol{\iota}_1|\mathbf{y}_{1:\tau}) \\ \vdots \\ \mathbb{P}(\boldsymbol{\xi}_t = \boldsymbol{\iota}_{\mathrm{M}}|\mathbf{y}_{1:\tau}) \end{bmatrix} = \begin{bmatrix} \mathbb{P}(\xi_{t1} = 1|\mathbf{y}_{1:\tau}) \\ \vdots \\ \mathbb{P}(\xi_{t\mathrm{M}} = 1|\mathbf{y}_{1:\tau}) \end{bmatrix}$$

for $p \leq \tau \leq t$ and $t \geq p+1$. Thus, using (1), the conditional probability density of $\mathbf{y}_t$ given $\mathbf{y}_{1:t-1}$ can be written as

$$\begin{aligned}
\mathbb{P}(\mathbf{y}_t|\mathbf{y}_{1:t-1}) &= \sum_{m=1}^{\mathrm{M}} \mathbb{P}(\mathbf{y}_t, \boldsymbol{\xi}_{t-1} = \boldsymbol{\iota}_m|\mathbf{y}_{1:t-1}) \\
&= \sum_{m=1}^{\mathrm{M}} \mathbb{P}(\mathbf{y}_t|\boldsymbol{\xi}_{t-1} = \boldsymbol{\iota}_m, \mathbf{y}_{1:t-1}) \mathbb{P}(\boldsymbol{\xi}_{t-1} = \boldsymbol{\iota}_m|\mathbf{y}_{1:t-1})
\end{aligned}$$

$$= \sum_{m=1}^{\mathrm{M}} \mathbb{P}(\mathbf{y}_t|\xi_{(t-1)m}=1,\mathbf{y}_{1:t-1})\mathbb{P}(\xi_{(t-1)m}=1|\mathbf{y}_{1:t-1})$$
$$= \boldsymbol{\eta}_t^{\mathsf{T}}\mathbf{P}^{\mathsf{T}} \sum_{m=1}^{\mathrm{M}} \boldsymbol{\iota}_m \, \mathbb{P}(\xi_{(t-1)m}=1|\mathbf{y}_{1:t-1}) = \boldsymbol{\eta}_t^{\mathsf{T}}\mathbf{P}^{\mathsf{T}}\widehat{\boldsymbol{\xi}}_{t-1|t-1}.$$

On the other hand, for $m = 1, \ldots, \mathrm{M}$, we have

$$\mathbb{P}(\xi_{tm}=1|\mathbf{y}_{1:t-1}) = \sum_{i=1}^{\mathrm{M}} \mathbb{P}(\xi_{tm}=1|\xi_{(t-1)i}=1,\mathbf{y}_{1:t-1})\mathbb{P}(\xi_{(t-1)i}=1|\mathbf{y}_{1:t-1})$$
$$= \sum_{i=1}^{\mathrm{M}} \alpha_{im}\mathbb{P}(\xi_{(t-1)i}=1|\mathbf{y}_{1:t-1}) = (\mathbf{P}^{\mathsf{T}}\widehat{\boldsymbol{\xi}}_{t-1|t-1})_m,$$

where $(\mathbf{v})_m$ refers to the $m$-th element of a vector $\mathbf{v}$. Thus, for all $t \geq p+1$, we have

$$\widehat{\boldsymbol{\xi}}_{t|t-1} = \mathbf{P}^{\mathsf{T}}\widehat{\boldsymbol{\xi}}_{t-1|t-1},$$

and hence we get

$$\mathbb{P}(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \boldsymbol{\eta}_t^{\mathsf{T}}\widehat{\boldsymbol{\xi}}_{t|t-1} = \mathbf{1}_{\mathrm{M}}^{\mathsf{T}}\left(\boldsymbol{\eta}_t \odot \widehat{\boldsymbol{\xi}}_{t|t-1}\right),$$

where $\odot$ denotes the element-wise matrix multiplication and $\mathbf{1}_{\mathrm{M}} = (1,\ldots,1)^{\mathsf{T}}$. Then

$$\widehat{\boldsymbol{\xi}}_{t|t} = \frac{\boldsymbol{\eta}_t \odot \widehat{\boldsymbol{\xi}}_{t|t-1}}{\mathbf{1}_{\mathrm{M}}^{\mathsf{T}}\left(\boldsymbol{\eta}_t \odot \widehat{\boldsymbol{\xi}}_{t|t-1}\right)}, \tag{2}$$

which follows since, by Bayes' rule,

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

Now we consider $\widehat{\boldsymbol{\xi}}_{t|n}$. First note that, by Bayes' rule and the model assumptions,

$$\begin{aligned}\mathbb{P}(\xi_{tm}=1|\xi_{(t+1)i}=1,\mathbf{y}_{1:n}) &= \mathbb{P}(\xi_{tm}=1|\xi_{(t+1)i}=1,\mathbf{y}_{1:t},\mathbf{y}_{t+1:n})\\ &= \frac{\mathbb{P}(\mathbf{y}_{t+1:n}|\xi_{tm}=1,\xi_{(t+1)i}=1,\mathbf{y}_{1:t})\mathbb{P}(\xi_{tm}=1|\xi_{(t+1)i}=1,\mathbf{y}_{1:t})}{\mathbb{P}(\mathbf{y}_{t+1:n}|\xi_{(t+1)i}=1,\mathbf{y}_{1:t})}\\ &= \frac{\mathbb{P}(\mathbf{y}_{t+1:n}|\xi_{(t+1)i}=1,\mathbf{y}_{1:t})\mathbb{P}(\xi_{tm}=1|\xi_{(t+1)i}=1,\mathbf{y}_{1:t})}{\mathbb{P}(\mathbf{y}_{t+1:n}|\xi_{(t+1)i}=1,\mathbf{y}_{1:t})}\\ &= \mathbb{P}(\xi_{tm}=1|\xi_{(t+1)i}=1,\mathbf{y}_{1:t}),\end{aligned}$$

which we use to write, once more using Bayes' rule,

$$\begin{aligned}\mathbb{P}(\xi_{tm}=1,\xi_{(t+1)i}=1|\mathbf{y}_{1:n}) &= \mathbb{P}(\xi_{tm}=1|\xi_{(t+1)i}=1,\mathbf{y}_{1:n})\mathbb{P}(\xi_{(t+1)i}=1|\mathbf{y}_{1:n})\\ &= \mathbb{P}(\xi_{tm}=1|\xi_{(t+1)i}=1,\mathbf{y}_{1:t})\mathbb{P}(\xi_{(t+1)i}=1|\mathbf{y}_{1:n})\\ &= \frac{\mathbb{P}(\xi_{(t+1)i}=1|\xi_{tm}=1,\mathbf{y}_{1:t})\mathbb{P}(\xi_{tm}=1|\mathbf{y}_{1:t})}{\mathbb{P}(\xi_{(t+1)i}=1|\mathbf{y}_{1:t})}\mathbb{P}(\xi_{(t+1)i}=1|\mathbf{y}_{1:n})\end{aligned}$$

$$= \frac{\mathbb{P}(\xi_{tm}=1|\mathbf{y}_{1:t})\mathbb{P}(\xi_{(t+1)i}=1|\xi_{tm}=1)}{\mathbb{P}(\xi_{(t+1)i}=1|\mathbf{y}_{1:t})}\mathbb{P}(\xi_{(t+1)i}=1|\mathbf{y}_{1:n}). \tag{3}$$

Using the above expression, for $m = 1, \ldots, \mathrm{M}$, we obtain

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

Collecting these into a vector, we get

$$\widehat{\boldsymbol{\xi}}_{t|n} = \begin{bmatrix} \mathbb{P}(\xi_{t1}=1|\mathbf{y}_{1:n}) \\ \vdots \\ \mathbb{P}(\xi_{t\mathrm{M}}=1|\mathbf{y}_{1:n}) \end{bmatrix} = \left[\mathbf{P}^{\mathsf{T}}\left(\widehat{\boldsymbol{\xi}}_{t+1|n} \oslash \widehat{\boldsymbol{\xi}}_{t+1|t}\right)\right] \odot \widehat{\boldsymbol{\xi}}_{t|t}, \tag{4}$$

denoting by $\oslash$ the element-wise division.

From (2) and (4), using the current estimate $\boldsymbol{\theta}^{(k)}$, we obtain the computations required in the E-step as follows:

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

$$\widehat{\boldsymbol{\xi}}_{t|t} = \frac{\boldsymbol{\eta}_t \odot \widehat{\boldsymbol{\xi}}_{t|t-1}}{\mathbf{1}_{\mathrm{M}}^{\mathsf{T}}\left(\boldsymbol{\eta}_t \odot \widehat{\boldsymbol{\xi}}_{t|t-1}\right)} = \frac{\boldsymbol{\eta}_t \odot \mathbf{P}^{(k)}\widehat{\boldsymbol{\xi}}_{t-1|t-1}}{\mathbf{1}_{\mathrm{M}}^{\mathsf{T}}\left(\boldsymbol{\eta}_t \odot \mathbf{P}^{(k)}\widehat{\boldsymbol{\xi}}_{t-1|t-1}\right)},$$

  where $\odot$ denotes the element-wise product.

- Backward recursion (smoothing): for $t = n-1, \ldots, p+1$, after plugging in the filtered probabilities, we compute

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

  denoting by $\oslash$ the element-wise division.

Finally, we set $\zeta_{ti}^{(k)} = \mathbb{P}(\xi_{ti}=1|\mathbf{y}_{1:n})$ as the $i$-th entry of the vector $\widehat{\boldsymbol{\xi}}_{t|n}$, for $i = 1, \ldots, \mathrm{M}$.

Regarding the joint probabilities $\zeta_{t,ij}^{(k)} = \mathbb{P}(\xi_{(t-1)i}=1, \xi_{tj}=1|\mathbf{y}_{1:n})$, as discussed in (Krolzig, 1997, Chapter 5), they can be expressed in terms of the smoothed probabilities $\widehat{\xi}_{t|n,j}$ ($j$-th entry of $\widehat{\boldsymbol{\xi}}_{t|n}$), the predicted probabilities $\widehat{\xi}_{t|t,j}$ ($j$-th entry of $\widehat{\boldsymbol{\xi}}_{t|t-1}$), the filtered probabilities $\widehat{\xi}_{t-1|t-1,i}$ ($i$-th entry of $\widehat{\boldsymbol{\xi}}_{t-1|t-1}$), and the transition probabilities estimates from the previous iteration. Using (3), the complete vector of joint probabilities, for $t = p+1, \ldots, n$, is estimated by

$$\widehat{\boldsymbol{\xi}}_{t|n}^{\text{joint}} := \operatorname{vec}(\mathbf{P}^{(k)}) \odot \left[\left(\widehat{\boldsymbol{\xi}}_{t|n} \oslash \widehat{\boldsymbol{\xi}}_{t|t-1}\right) \otimes \widehat{\boldsymbol{\xi}}_{t-1|t-1}\right],$$

where $\otimes$ denotes the Kronecker product, and whose $(j-1)\mathrm{M}+i$ entry corresponds to the term

$$\zeta_{t,ij}^{(k)} = \mathbb{P}(\xi_{(t-1)i}=1, \xi_{tj}=1|\mathbf{y}_{1:n}) = \frac{\alpha_{ij}^{(k)}(\widehat{\boldsymbol{\xi}}_{t-1|t-1})_i(\widehat{\boldsymbol{\xi}}_{t|n})_j}{(\widehat{\boldsymbol{\xi}}_{t|t-1})_j},$$

for $i, j = 1, \ldots, \mathrm{M}$. Recall $(\mathbf{v})_i$ refers to the $i$-th element of a vector $\mathbf{v}$.

### A1.3 Maximization (M-) step

#### A1.3.1 Transition probability matrix

The true transition probability matrix is estimated based on the joint distribution of $\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}$ given the full sample $\mathbf{y}_{1:n}$ as in (3). From (Krolzig, 1997, Equation 6.14), the estimate is given as

$$\operatorname{vec}(\mathbf{P}^{(k+1)}) = \left[\widehat{\boldsymbol{\xi}}^{\text{joint}}\right] \oslash \left[\mathbf{1}_{\text{M}} \otimes \left(\left(\mathbf{1}_{\text{M}}^{\mathsf{T}} \otimes \mathbf{I}_{\text{M}}\right) \widehat{\boldsymbol{\xi}}^{\text{joint}}\right)\right],$$

where $\widehat{\boldsymbol{\xi}}^{\text{joint}} := \sum_{t=p+1}^{n} \widehat{\boldsymbol{\xi}}_{t|n}^{\text{joint}}$, and $\mathbf{I}_{\text{M}}$ is the M-dimensional identity matrix.

#### A1.3.2 AR coefficient matrices

Recall the optimization problem with respect to the AR coefficients for regime $m$ is given as

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

To solve it, we utilize the block-wise coordinate descent algorithm for VAR models suggested by Basu and Michailidis (2015). Let us first define

$$\widehat{\Xi}_m := \begin{bmatrix} \zeta_{(p+1)m}^{(k)} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \zeta_{nm}^{(k)} \end{bmatrix}, \qquad \mathbf{y} := \begin{bmatrix} (\mathbf{y}_{p+1} - \boldsymbol{\nu}^{(m,k+1)})^{\mathsf{T}} \\ \vdots \\ (\mathbf{y}_n - \boldsymbol{\nu}^{(m,k+1)})^{\mathsf{T}} \end{bmatrix},$$

$$\mathbf{X} := \begin{bmatrix} \mathbf{X}_{p+1} \\ \vdots \\ \mathbf{X}_n \end{bmatrix}, \qquad \mathbf{X}_t := \begin{bmatrix} \mathbf{y}_{t-1}^{\mathsf{T}} & \mathbf{y}_{t-2}^{\mathsf{T}} & \cdots \mathbf{y}_{t-p}^{\mathsf{T}} \end{bmatrix}.$$

Let $\mathbf{y}_i^{(m)}$ and $\mathbf{b}_i^{(m)}$ denote the $i$-th column and the $i$-th row of $\widehat{\Xi}_m^{1/2}\mathbf{y}$ and $[\mathbf{A}_1^{(m)}, \ldots, \mathbf{A}_p^{(m)}]$, respectively. Also let $\mathbf{X}^{(m)} = \widehat{\Xi}_m^{1/2}\mathbf{X}$ and $\widehat{\boldsymbol{\Omega}}^{(m,k)} = (\omega_{ij})_{1 \le i,j \le d}$. The objective function can be rewritten as

$$\min_{\mathbf{b}_i^{(m)}, i=1,\ldots,d} \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \omega_{ij} \left(\mathbf{y}_i^{(m)} - \mathbf{X}^{(m)} \mathbf{b}_i^{(m)\mathsf{T}}\right)^{\mathsf{T}} \left(\mathbf{y}_j^{(m)} - \mathbf{X}^{(m)} \mathbf{b}_j^{(m)\mathsf{T}}\right) + \sum_{i=1}^{d} \sum_{j=1}^{dp} r_{\lambda_1}\left(b_{ij}^{(m)}\right),$$

where the $b_{ij}^{(m)}, j = 1, \ldots, dp$, are the elements of $\mathbf{b}_i^{(m)}, i = 1, \ldots, d$. As suggested in Basu and Michailidis (2015, Appendix C), we minimize the above objective function cyclically with respect to each $\mathbf{b}_i^{(m)}$ until convergence. We repeat the following procedure for $i = 1, \ldots, d$, until convergence:

1. Set $\mathbf{r}_i = \frac{1}{2\omega_{ii}} \sum_{j=1,\, j \neq i}^{d} \omega_{ij} \left(\mathbf{y}_j^{(m)} - \mathbf{X}^{(m)} \widehat{\mathbf{b}}_j^{\mathsf{T}}\right)$;
2. Update $\widehat{\mathbf{b}}_i = \arg\min_{\mathbf{b}_i} \frac{\omega_{ii}}{2} \|\mathbf{y}_i^{(m)} + \mathbf{r}_i - \mathbf{X}^{(m)} \mathbf{b}_i^{\mathsf{T}}\|_2^2 + \sum_{j=1}^{dp} r_{\lambda_1}\left(b_{ij}\right)$.

To solve the optimization problem in Step 2 above, we employ a generalized gradient descent. In particular, we use a modified fast iterative-shrinkage thresholding algorithm (FISTA, Beck and

Teboulle, 2009b), which was initially devised to solve optimization problems of the form $\min_{\mathbf{b}} g(\mathbf{b}) + r_\lambda(\mathbf{b})$. For a step size L, the iterative scheme is based on the update

$$\begin{aligned}\widehat{\mathbf{b}}_i^{(k+1)} &= \arg\min_{\mathbf{b}_i}\left[\frac{1}{2}\left\|\mathbf{b}_i - \left(\widehat{\mathbf{z}}_i^{(k)} - \mathrm{L}\omega_{ii}(\mathbf{y}_i^{(m)} + \mathbf{r}_i - \mathbf{X}^{(m)}(\widehat{\mathbf{z}}_i^{(k)})^\mathsf{T})^\mathsf{T}\mathbf{X}^{(m)}\right)\right\|^2 + \mathrm{L}\sum_{j=1}^{dp} r_{\lambda_1}(b_{ij})\right] \\ &=: h(\widehat{\mathbf{z}}_i^{(k)}, \mathrm{L}),\end{aligned} \tag{5}$$

where $\widehat{\mathbf{z}}_i^{(k)}$ is an interpolation between $\widehat{\mathbf{b}}_i^{(k)}$ and $\widehat{\mathbf{b}}_i^{(k-1)}$. Since the considered penalty function is decomposable with respect to individual elements of $\mathbf{b}_i$, the minimization problem above corresponds to a penalized least-squares problem with orthogonal predictors, and can be solved analytically for all the penalty functions we consider in this work. The exact formulas for this update for all the considered penalties are provided in Section A1.3.4.

Ideally, in an optimization algorithm the value of the objective function must not increase when computed over successive iterations. The original FISTA does not possess this property, making it vulnerable to divergence. This issue has been circumvented in another work by Beck and Teboulle (2009a), and we incorporate their approach in our implementation. Another enhancement we implement is known as FISTA with restart (Wen et al., 2017, and references therein), which resets the interpolation parameter every $\kappa$ iterations for a pre-specified $\kappa$, and gives faster convergence compared to the original version of FISTA. See Algorithm 2 below.

We compute a step size L that ensures convergence as follows. First note that a Lipschitz constant of the smooth term in the objective function is $\lambda_{\max}(\mathbf{X}^\mathsf{T}\widehat{\Xi}_m\mathbf{X})$, where $\lambda_{\max}(\mathbf{A})$ denotes the maximum eigenvalue of a real symmetric matrix $\mathbf{A}$. The knowledge of this constant provides a suitable step size (Beck and Teboulle, 2009b). For a given regime $m \in \{1, \ldots, \mathrm{M}\}$, we can approximate the corresponding Lipschitz constant by observing the relation

$$\mathbf{X}^\mathsf{T}\widehat{\Xi}_m\mathbf{X} \preceq \mathbf{X}^\mathsf{T}\mathbf{X},$$

which holds since each element of the diagonal matrix $\widehat{\Xi}_m$ lies in the interval $(0, 1)$. Hence, following Guo et al. (2016, Theorem 4.4), we set L as any value satisfying

$$\mathrm{L} < \begin{cases} \frac{2}{2\mathrm{C}+\lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X})} & \text{if } \mathrm{C} = 0, \\ \frac{2}{\mathrm{C}+\lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X})} & \text{if } 0 < \mathrm{C} \le \lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X})/2, \\ \frac{2}{2\mathrm{C}+\sqrt{\lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X})^2-\mathrm{C}^2}} & \text{if } \lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X})/2 < \mathrm{C} < \lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X}), \\ \frac{1}{\mathrm{C}} & \text{if } \lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X}) = \mathrm{C}, \end{cases} \tag{6}$$

where C is the weak-convexity constant of the penalty $r$, such that $\sum_{j=1}^{dp} r_{\lambda_1}(b_{ij}) + \frac{\mathrm{C}}{2}\|\mathbf{b}_i\|^2$ is convex. This constant C exists for all the penalties considered in this work (see Section A1.3.4 below). By using such L we ensure the convergence of the algorithm to a local minimum. Our procedure is summarized in Algorithm 2.

### A1.3.3 Covariance and precision matrices

Recall the optimization problem with respect to $\mathbf{\Sigma}^{(m)}$ or $\mathbf{\Omega}^{(m)}$ is

$$\min_{\mathbf{\Sigma}^{(m)} \succ 0} \frac{1}{2(n-p)}\left(\widehat{n}_m \log|\mathbf{\Sigma}^{(m)}| + \mathrm{tr}\left(\mathbf{\Omega}^{(m)}\mathbf{S}^{(m)}\right)\right) + \sum_{i\neq j=1}^{d} r_{\lambda_2}(\gamma_{ij}^{(m)}), \tag{7}$$

**Algorithm 2** Monotone FISTA with restart for sparse VAR parameter estimation

Initialization: $\widehat{\mathbf{z}}_i^1 = \widehat{\mathbf{b}}_i^0$, $t_1 = 1$, $\kappa \geq 1$, L satisfying (6), $\varepsilon > 0$

1: $\widehat{\mathbf{x}}_i^k = h(\widehat{\mathbf{z}}_i^k, \mathrm{L})$

2: $\widehat{\mathbf{b}}_i^k = \arg\min\{h(\mathbf{x}, \mathrm{L}) : \mathbf{x} \in \{\widehat{\mathbf{x}}_i^k, \widehat{\mathbf{b}}_i^{k-1}\}\}$

3: $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$

4: $\widehat{\mathbf{z}}_i^{k+1} = \widehat{\mathbf{b}}_i^k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\widehat{\mathbf{b}}_i^k - \widehat{\mathbf{b}}_i^{k-1}) + \left(\frac{t_k}{t_{k+1}}\right)(\widehat{\mathbf{x}}_i^k - \widehat{\mathbf{b}}_i^k)$

5: If $k \bmod \kappa = 0$ set $t_{k+1} = 1$

6: If $\|\widehat{\mathbf{b}}_i^k - \widehat{\mathbf{b}}_i^{k-1}\| / \|\widehat{\mathbf{b}}_i^{k-1}\| < \varepsilon$ return $\widehat{\mathbf{b}}_i^k$, else $k = k + 1$ and go to 1

with $\mathbf{\Gamma}^{(m)} = (\gamma_{ij}^{(m)})_{1 \leq i,j \leq d}$ being either $\mathbf{\Sigma}^{(m)}$ or $\mathbf{\Omega}^{(m)}$. To simplify the notation, we drop the index $m$ below. Loh and Wainwright (2015) proposed and theoretically studied an algorithm for regularized M-estimation for non-convex problems. Their algorithm performs a generalized gradient descent on the objective function, which is assumed to be the sum of a smooth function and a penalty function, similar to FISTA. We adopt their approach and implement it as follows. Let C be the weak-convexity constant of the penalty $r$, so that $\sum_{i,j=1}^d r_{\lambda_2}(\mathbf{\Gamma}) + \frac{\mathrm{C}}{2}\|\mathbf{\Gamma}\|_{\mathrm{F}}^2$ is convex (see Section A1.3.4). Next we provide the updates for the scenario where $\mathbf{\Sigma}$ is penalized, followed by the updates in the scenario where the penalty is instead on $\mathbf{\Omega}$.

**Penalization on the covariance matrix**

The update with step size $\mathrm{L}_k$ for the optimization problem for penalized covariance is given as

$$\mathbf{\Sigma}^{(k+1)} \in \underset{\mathbf{\Gamma} \succ 0}{\arg\min} \left\{ \frac{1}{2} \left\| \mathbf{\Gamma} - \left( \mathbf{\Sigma}^{(k)} - \mathrm{L}_k \left( \mathbf{\Omega}^{(k)} - \mathbf{\Omega}^{(k)} \mathbf{S} \mathbf{\Omega}^{(k)} - \mathrm{C}\mathbf{\Sigma}^{(k)} \right) \right) \right\|_{\mathrm{F}}^2 + \mathrm{L}_k \left( \sum_{i,j=1}^d r_{\lambda_2}(\gamma_{ij}) + \frac{\mathrm{C}}{2}\|\mathbf{\Gamma}\|_{\mathrm{F}}^2 \right) \right\} \tag{8}$$

$$=: h_{\mathbf{\Sigma}}(\mathbf{\Sigma}^{(k)}, \mathrm{L}_k),$$

which has a closed form for all the penalties we consider (see Section A1.3.4). Loh and Wainwright (2015) provided error bounds with respect to a global minimizer that hold with high probability, and stated that the iterates (8) quickly converge to a neighborhood of any global optimum under a set of smoothness and convexity conditions. Our objective function meets those requirements since its differentiable term has a Lipschitz-continuous gradient on a compact constraint set of the form $\{\Sigma : \Sigma \succeq \delta \mathrm{I}\}$ for some $\delta > 0$. This can be verified using ideas similar to those of Bien and Tibshirani (2011, Appendix 2). We compute the step size $\mathrm{L}_k$ using a backtracking line search; see, for example, Parikh and Boyd (2013, Section 4.2).

**Penalization on the precision matrix**

For the penalized precision matrix case, the difference arises on the differentiable part of the objective function and its gradient, in comparison with the covariance case. We perform each update as

$$\mathbf{\Omega}^{(k+1)} \in \underset{\mathbf{\Gamma} \succ 0}{\arg\min} \left\{ \frac{1}{2} \left\| \mathbf{\Gamma} - \left( \mathbf{\Omega}^{(k)} - \mathrm{L}_k \left( -\mathbf{\Sigma}^{(k)} + \mathbf{S} - \mathrm{C}\mathbf{\Omega}^{(k)} \right) \right) \right\|_{\mathrm{F}}^2 + \mathrm{L}_k \left( \sum_{i,j=1}^d r_{\lambda_2}(\gamma_{ij}) + \frac{\mathrm{C}}{2}\|\mathbf{\Gamma}\|_{\mathrm{F}}^2 \right) \right\} \tag{9}$$

$$=: h_{\mathbf{\Omega}}(\mathbf{\Omega}^{(k)}, \mathrm{L}_k).$$

The theoretical properties of the iterative procedure as investigated by Loh and Wainwright (2015) are also applicable. The required Lipschitz continuity is implied when imposing the constraint $\Sigma^{-1} \succeq \delta \mathrm{I}$ for some specific $\delta > 0$. Such a $\delta$ can be obtained with a derivation similar to Bien and Tibshirani (2011, Appendix 2). The final algorithm for solving (7) for a fixed regime $m$ is outlined in Algorithm 3.

**Algorithm 3** Generalized gradient descent for sparse VAR covariance or precision matrix estimation

Initialization: $\mathbf{\Sigma}^{(1)} = \mathbf{S}$, $\mathbf{\Omega}^{(1)} = \mathbf{S}^{-1}$, $\mathrm{L}_1 \leq 1$, $\varepsilon > 0$

1: Perform backtracking line search to obtain step size $\mathrm{L}_k$

2: (Penalized covariance) $\mathbf{\Sigma}^{(k+1)} = h_{\mathbf{\Sigma}}(\mathbf{\Sigma}^{(k)}, \mathrm{L}_k)$
(Penalized precision) $\mathbf{\Omega}^{(k+1)} = h_{\mathbf{\Omega}}(\mathbf{\Omega}^{(k)}, \mathrm{L}_k)$

3: If $\|\mathbf{\Sigma}^{(k+1)} - \mathbf{\Sigma}^{(k)}\|_{\mathrm{F}} / \|\mathbf{\Sigma}^{(k)}\|_{\mathrm{F}} < \varepsilon$ or $\|\mathbf{\Omega}^{(k+1)} - \mathbf{\Omega}^{(k)}\|_{\mathrm{F}} / \|\mathbf{\Omega}^{(k)}\|_{\mathrm{F}} < \varepsilon$ return $(\mathbf{\Sigma}^{(k+1)}, \mathbf{\Omega}^{(k+1)})$
else $k = k + 1$ and go to 1

### A1.3.4 The proximal operators

The updates in Algorithm 2 and 3 of Section A1.3.3 have a closed-form formula for the penalty functions we consider, which we provide here. The update steps in (5), (8), and (9) can be obtained by solving the following

$$\min_x \frac{1}{2}(x - z)^2 + \nu r_\lambda(x),$$

for $x \in \mathbb{R} \setminus \{0\}$, fixed $z \in \mathbb{R}$ and $\nu > 0$. For example, in Algorithm 2, each element of $\mathbf{b}_i$ as in (5), corresponds to $x$ here. The formulas can be obtained by equating the subgradient of the objective function above to zero and solving for $x$. Thus, for a particular choice of the penalty function $r_\lambda$, the corresponding formulas for the updates in Algorithm 2 and 3 are given as:

$$\widehat{x}_{\mathrm{L}_1} = \begin{cases} 0, & 0 \leq |z| \leq \nu\lambda, \\ z - \mathrm{sign}(z)\nu\lambda, & |z| \geq \nu\lambda; \end{cases} \tag{10}$$

$$\widehat{x}_{\mathrm{SCAD}} = \begin{cases} 0, & 0 \leq |z| \leq \nu\lambda, \\ z - \mathrm{sign}(z)\nu\lambda, & \nu\lambda < |z| \leq (1+\nu)\lambda, \\ \frac{(a-1)z - \mathrm{sign}(z)a\nu\lambda}{a-1-\nu}, & (1+\nu)\lambda < |z| \leq a\lambda, \\ z, & |z| \geq a\lambda; \end{cases}$$

$$\widehat{x}_{\mathrm{MCP}} = \begin{cases} 0, & 0 \leq |z| \leq \nu\lambda, \\ \frac{bz - \mathrm{sign}(z)b\nu\lambda}{b-\nu}, & \nu\lambda < |z| \leq b\lambda, \\ z, & |z| \geq b\lambda. \end{cases}$$

In the updates of Algorithm 2 and 3, we are required to set $\nu = (\mathcal{L} + \mathrm{C})^{-1}$, with C equal to 0, $(a-1)^{-1}$ or $b^{-1}$ for penalty functions $\mathrm{L}_1$-norm, SCAD or MCP, respectively; these are the weak convexity constants. For the adaptive LASSO, the update can be obtained by replacing $\lambda$ in (10) with the weighted version $\lambda\widehat{w}$, for a weight of the form $\widehat{w} = |\widehat{\beta}|^{-\gamma}$, where $\widehat{\beta}$ is a $\sqrt{n}$-consistent estimator of the parameter coordinate being estimated. We use the maximum-likelihood estimator with $\gamma = 1$. If the MLEs of the covariance or precision matrices are computationally singular, following the idea in ridge regression, we regularize them by adding a multiple of the identity matrix.

### A1.4 Model initialization and identification

The performance of the modified EM algorithm is heavily reliant on an appropriate initialization. Our proposal here is motivated by the desired behaviour from a good estimate, of having a moderately high likelihood value. Define

$$\mathbf{y} := \begin{bmatrix} \mathbf{y}_{p+1}^\mathsf{T} \\ \vdots \\ \mathbf{y}_n^\mathsf{T} \end{bmatrix}, \qquad \mathbf{X} := \begin{bmatrix} \mathbf{X}_{p+1} \\ \vdots \\ \mathbf{X}_n \end{bmatrix}, \qquad \mathbf{X}_t := \begin{bmatrix} 1 \ \mathbf{y}_{t-1}^\mathsf{T} \ \mathbf{y}_{t-2}^\mathsf{T} \ \dots \mathbf{y}_{t-p}^\mathsf{T} \end{bmatrix}.$$

For $m = 1, \dots, \mathrm{M}$, we set the first iterates as

$$\begin{aligned} [\boldsymbol{\nu}^{(m,1)}, \mathbf{A}^{(m,1)}]^\mathrm{T} &= [\mathbf{X}^\mathsf{T}\mathbf{X}]^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}, \\ \boldsymbol{\Sigma}^{(m,1)} &= \frac{1}{n-p}\left[\mathbf{y} - \mathbf{X}[\boldsymbol{\nu}^{(m,1)} \ \mathbf{A}^{(m,1)}]^\mathsf{T}\right]^\mathsf{T}\left[\mathbf{y} - \mathbf{X}[\boldsymbol{\nu}^{(m,1)} \ \mathbf{A}^{(m,1)}]^\mathsf{T}\right], \\ \mathbf{P}^{(1)} = {\alpha_{ij}}^{(1)} &= \begin{cases} \frac{1}{\mathrm{M}} + \delta, & i = j, \\ \frac{1}{\mathrm{M}-1}(1 - \alpha_{ii}^{(1)}), & i \neq j. \end{cases} \end{aligned}$$

In other words, we use the least-squares and MLE estimates for a VAR($p$) model (true $p$ being assumed as given) as initial coefficient and covariance matrices for all the regimes. We use a generalized inverse if necessary. The transition probability matrix initial estimate is just a symmetric matrix, whose diagonal is specified by the quantity $\delta \in (0, \mathrm{M}^{-1})$, and ensures that the estimates do not fall into a local maximum.

The interchange of regime labels in the MSVAR models during the estimation procedure can lead to a model identification problem. The model parameters belong to the same equivalence class if their regime labels are a permutation of any other label set in the class, and thus all the parameter combinations in an equivalence class lead to the same model. After convergence of the EM algorithm, we perform a permutation of the regime labels, as suggested by Krolzig (1997), according to a pre-specified ordering on the magnitude of the estimated VAR coefficients. For example, different regimes labels can be indexed based on the increasing order of mean of the intercept vector. This ensures that the labels of the final estimates always follow the pre-specified order which takes care of the model identifiability. Another requirement of model identification is the existence of a well-defined distribution function. This condition is trivially satisfied as we assume Gaussian noise in our MSVAR models.

### A1.5 Selection of the penalty parameters $\lambda_1$ and $\lambda_2$

Inspired by Zhang et al. (2010), we select the penalty parameters by optimizing a model selection criterion, which is composed of a loss function and a model complexity term which involves the number of degrees of freedom of a candidate model. Here, we assume a fixed number of regimes M and AR-order $p$.

Let $\widehat{\boldsymbol{\theta}}_n(\boldsymbol{\lambda})$ be the MPLE given a pair $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in \mathcal{I} \times \mathcal{I} \subset \mathbb{R}^2$, for a pre-specified set of values $\mathcal{I}$. We denote the regime-specific degrees of freedom of the corresponding fitted MSVAR model as

$$\mathrm{D}_m(\widehat{\boldsymbol{\theta}}_n(\boldsymbol{\lambda})) = \sum_{k=1}^{p}\sum_{i,j=1}^{d} \mathbb{1}_{\{(\widehat{a}_{k,ij}^{(m)}) \neq 0\}} + \sum_{i,j=1}^{d} \mathbb{1}_{\{\widehat{\gamma}_{ij}^{(m)} \neq 0\}}, \quad m = 1, \dots, \mathrm{M},$$

and define the selection criterion

$$\mathcal{C}_1(\boldsymbol{\lambda}) = -2l_n(\widehat{\boldsymbol{\theta}}_n(\boldsymbol{\lambda}); s_p) + c \sum_{m=1}^{\mathrm{M}} \mathrm{D}_m(\widehat{\boldsymbol{\theta}}(\boldsymbol{\lambda})),$$

where $l_n$ is the conditional log-likelihood and $c$ is a constant controlling the model complexity. A similar criterion, denoted by $\mathcal{C}$, is used in the main manuscript for the selection of the number of regimes M. In the above criterion, if $c = 2$ or $c = \log(n-p)$, we obtain the AIC or BIC, respectively. A data-dependent choice of the tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ is

$$\boldsymbol{\lambda} \in \underset{\boldsymbol{\lambda}_0 \in \mathcal{I} \times \mathcal{I}}{\arg\min}\, \mathcal{C}_1(\boldsymbol{\lambda}_0). \tag{11}$$

The usual practice is to take $\mathcal{I}$ to be a discretization of the set $[0, \widetilde{\lambda}]$ for some $\widetilde{\lambda} > 0$, and choose $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ as in (11). In our numerical studies, the pairs $(\lambda_1, \lambda_2)$ take values on a grid with increments of 0.25 on the logarithmic scale.

We propose a computationally efficient coordinate-descent approach which optimizes $\mathcal{C}_1$ by performing a cyclical search on either coordinate of the argument $\boldsymbol{\lambda}_0 \in \mathcal{I} \times \mathcal{I}$, while keeping the other fixed until reaching convergence. To increase the accuracy, we perform additional searches at different resolutions of $\mathcal{I}$. Thus, the coarsest grid corresponds to a global coordinate descent, whereas subsequent searches are performed at decreasing resolutions in a neighborhood of the optimum grid point found in the previous resolution. This refines the result further. The discrete nature of the problem enables this procedure to converge to a local optimum. For a step size $s > 0$, window radius $w \in \mathbb{N}$, and center $\lambda > 0$, let $\mathcal{I}_w(s, \lambda) := \{\max\{\lambda - ws, 0\}, \ldots, \lambda - 2s, \lambda - s, \lambda, \lambda + s, \lambda + 2s, \ldots, \min\{\lambda + ws, \widetilde{\lambda}\}\}$. Algorithm 4 summarizes the selection procedure of the penalty tuning parameters and number of regimes.

**Algorithm 4** Selection of Tuning Parameters and Number of Regimes

Input: $\mathrm{M}_{\max} \in \mathbb{N}$, $\widetilde{\lambda} > 0$, $\widehat{\lambda}_2^0 = 0$
For $\mathrm{M} = 1, \ldots, \mathrm{M}_{\max}$ do
  For $k = 1, 2, \ldots$ do

$$\widehat{\lambda}_1^k \in \underset{\lambda_1 \in \mathcal{I}_\infty(s, \widetilde{\lambda}/2)}{\arg\min} \mathcal{C}_1(\lambda_1, \widehat{\lambda}_2^{k-1}),$$
$$\widehat{\lambda}_2^k \in \underset{\lambda_2 \in \mathcal{I}_\infty(s, \widetilde{\lambda}/2)}{\arg\min} \mathcal{C}_1(\widehat{\lambda}_1^k, \lambda_2),$$

  until $\widehat{\lambda}_i^k = \widehat{\lambda}_i^{k-1}$, $i = 1, 2$
  Set $c_i = \widehat{\lambda}_i^k$, $i = 1, 2$
  For $j = 1, 2$ do
    $\widehat{\lambda}_2^0 = c_2$
    For $k = 1, 2, \ldots$ do

$$\widehat{\lambda}_1^k \in \underset{\lambda_1 \in \mathcal{I}_w(s/2^j, c_1)}{\arg\min} \mathcal{C}_1(\lambda_1, \widehat{\lambda}_2^{k-1}),$$
$$\widehat{\lambda}_2^k \in \underset{\lambda_2 \in \mathcal{I}_w(s/2^j, c_2)}{\arg\min} \mathcal{C}_1(\widehat{\lambda}_1^k, \lambda_2),$$

    until $\widehat{\lambda}_i^k = \widehat{\lambda}_i^{k-1}$, $i = 1, 2$
    Set $c_i = \widehat{\lambda}_i^k$, $i = 1, 2$
  Set $\boldsymbol{\lambda}_{\mathrm{M}} = (\widehat{\lambda}_1^k, \widehat{\lambda}_2^k)$
  Compute $\mathcal{C}(\mathrm{M})$ defined in the main manuscript using $\boldsymbol{\lambda}_{\mathrm{M}}$
Set $\widehat{\mathrm{M}} = \arg\min\{\mathcal{C}(1), \ldots, \mathcal{C}(\mathrm{M}_{\max})\}$
Return $(\boldsymbol{\lambda}_{\widehat{\mathrm{M}}}, \widehat{\mathrm{M}})$

### A1.6 AR-order selection

By adapting the approach for VARs by Nicholson et al. (2020), we devise a procedure for AR-order selection that has a superior performance for our models compared to a criterion-based approach. It is based on penalizing the AR coefficients using the group LASSO so that the spurious lag entries are shrunk to zero. This AR-order selection procedure, described next, is straightforward to incorporate into Algorithm 4.

First note that the AR-order selection can be done by relying on the consistency of the penalized estimator. We can consider a pre-specified AR-order $p_0 \geq p$, and express an MSVAR model with AR-order $p$ as an MSVAR model with AR-order $p_0$, with $\mathbf{A}_i^{(m)} \equiv 0_{d \times d}$ for $i > p$ and $m = 1, \ldots, \mathrm{M}$. However, a methodologically more suitable approach is to consider the group LASSO (Yuan and Lin, 2006) as done for VAR models by Song and Bickel (2011) and Nicholson et al. (2020).

Group LASSO shrinks all the elements in a given group simultaneously towards zero, that is, all the coefficients in a group are simultaneously estimated as either zero or nonzero. Nicholson et al. (2020) proposed a series of hierarchical-lag penalties derived from the group LASSO. The groups are designed in a nested manner so that if the group corresponding to lag $i$ is set to zero, then its nested groups, corresponding to a lag greater than $i$, remain zero. Among the various hierarchical penalties, the element-wise hierarchical penalty is the most suitable for our models as we do not assume any particular sparsity structure on the AR coefficient matrices. It is also straightforward

to optimize in our context. For regime $m$, we write

$$\min_{\{\mathbf{A}_l^{(m)}\}_{l=1}^{p}} \frac{1}{2(n-p)} \sum_{t=p+1}^{n} \zeta_{tm}^{(k)} \big(\mathbf{y}_t - \bar{\boldsymbol{\mu}}_t^{(m)}\big)^{\mathsf{T}} \widehat{\boldsymbol{\Omega}}^{(m,k)} \big(\mathbf{y}_t - \bar{\boldsymbol{\mu}}_t^{(m)}\big) + \mathrm{R}(\mathbf{A}_1^{(m)}, \ldots, \mathbf{A}_{p_0}^{(m)}; \lambda),$$

$$\mathrm{R}(\mathbf{A}_1^{(m)}, \ldots, \mathbf{A}_{p_0}^{(m)}; \lambda) = \lambda \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{i=1}^{p_0} \sqrt{p_0 - i + 1} \left( \sum_{\ell=i}^{p_0} (\mathbf{A}_\ell^{(m)})_{jk}^2 \right)^{1/2},$$

so that each group indexed by lag $i$ contains the element $(j,k)$ of all matrices of lag $\ell \geq i$. Furthermore, the adaptive version of the hierarchical penalty is straightforward to obtain as an extension of the adaptive group LASSO (Wang and Leng, 2008). After estimation we can set

$$\widehat{p} = \max\{k \in \{1, \ldots, p_0\} : \widehat{\mathbf{A}}_k^{(m)} \neq \mathbf{0}_{d \times d}\}. \tag{12}$$

An advantage of this AR-order selection method is that the penalty, and therefore the optimization problem, can be decomposed along coordinates, which enables us to apply the optimization framework we employ for the penalized AR coefficients estimation in Section A1.3.2.

We examine the performance of the above AR-order selection method via simulations. We set $\mathrm{M} = 2$, $d = 20$, and sample sizes $n \in \{400, 600, 800\}$. We consider the true AR-orders $p = 1, 2, 3$, and set the over-specified candidate $p_0 = 4$. We compute the mean of the performance measure $\mathbb{1}_{\{\widehat{p}=p\}}$, with $\widehat{p}$ obtained using (12). For both sparsity scenarios **S1** (sparse covariance matrices) and **S2** (sparse precision matrices), we obtain a mean performance measure of the true AR-order selection of at least 0.90. It is worth mentioning that using a AR-order selection procedure based on the BIC did not provide satisfactory results.

## A2 Theoretical results

In this section, we study the large sample properties of the MPLE $\widehat{\boldsymbol{\theta}}_n$ introduced in Section 3 of the main manuscript, when the true number of regimes M is correctly specified or over-estimated. We denote the true MSVAR model parameter vector by $\boldsymbol{\theta}^\star$, and partition it as $\boldsymbol{\theta}^\star = [(\boldsymbol{\theta}^{\star\mathcal{A}})^{\mathsf{T}}, (\boldsymbol{\theta}^{\star\mathcal{A}^c})^{\mathsf{T}}]^{\mathsf{T}}$, so that $\boldsymbol{\theta}^{\star\mathcal{A}^c} = \mathbf{0}$, and $\boldsymbol{\theta}^{\star\mathcal{A}}$ is a subvector of all the remaining non-zero parameters. Here, $\mathcal{A}$ also denotes the set of indices $k$ of all active (nonzero) parameters, and its complement $\mathcal{A}^c$, similarly contains the indices of all inactive (zero) parameters. Then, the cardinality of the index set of a parameter vector $\boldsymbol{\theta}$, denoted as $|\mathcal{A} \cup \mathcal{A}^c|$, is equal to $\mathrm{K} = \mathrm{M}(d + pd^2 + d(d+1)/2) + \mathrm{M}(\mathrm{M}-1)$. Furthermore, we let $\mathcal{E} \subset \mathcal{A} \cup \mathcal{A}^c$, denote the set of indices $k$ of the model parameters on which we perform the penalization. Recall, that we do not penalize the parameters corresponding to the transition probability matrix $\mathbf{P}$ of the hidden Markov chain, the AR intercepts $\boldsymbol{\nu}^{(m)}$, and the diagonal entries of the covariance matrices $\boldsymbol{\Sigma}^{(m)}$, $m = 1, \ldots, \mathrm{M}$. To establish our theoretical results, we require certain conditions on the process $\{\mathbf{Y}_t\}$, the penalty function $r_{\lambda_n}$ used in (4) of the main manuscript, and the tuning parameter $\lambda_n$. Let $r'_{\lambda_n}(\cdot)$ and $r''_{\lambda_n}(\cdot)$ denote the first and second derivatives, respectively, of the function $r_{\lambda_n}(\cdot)$ with respect to $\theta$.

**Assumption 1.** For all $\lambda \in \mathbb{R}$, $r_\lambda(0) = 0$. The function $r_\lambda(\theta)$ is continuous, symmetric, nonnegative, nondecreasing, continuously differentiable for all $|\theta| > 0$, and twice continuously differentiable for all $|\theta| > c\lambda$ for some constant $c > 0$.

**Assumption 2.** As $n \to \infty, \lambda_n = o(1)$, and for true value of model parameters $\theta_k^\star$,

$$a_n = \max_{k \in \mathcal{A} \cap \mathcal{E}} \{|r'_{\lambda_n}(\theta_k^\star)|\} = \mathrm{O}(n^{-1/2}), \qquad b_n = \max_{k \in \mathcal{A} \cap \mathcal{E}} \{|r''_{\lambda_n}(\theta_k^\star)|\} = o(1).$$

**Assumption 3.** $\liminf_{n\to\infty} \liminf_{|\theta|=\mathrm{O}(n^{-1/2})} \sqrt{n} r'_{\lambda_n}(\theta) = +\infty$.

Assumption 1 imposes a differentiability condition on the penalty function which allows to perform the Taylor expansion of the objective function around the true optimizer. Assumption 2 is essential to show the existence of a $\sqrt{n}$-consistent optimizer of the penalized conditional log-likelihood function, and Assumption 3 is required to prove the sparsity property for the estimator of the model parameter. For simplicity of exposition, we suppose that the penalty function $r_\lambda$ used for the covariance (or precision) matrix parameters is the same as for the VAR matrix coefficients, but this is not a requirement. In addition, in contrast to the related work of Monbet and Ailliot (2017), we employ a unique tuning parameter $\lambda_n$ for all the regimes; the computational burden is alleviated in this way and our experiments show that one tuning parameter for all the regimes also performs well. As we consider the situation that for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have $\max_{m\in\{1,\ldots,\mathrm{M}\}} \lambda_{\min}(\Sigma_m) \geq \delta > 0$, we do not penalize for a possibly singular $\Sigma_m$. Here, $\lambda_{\min}(\mathbf{A})$ denotes the smallest eigenvalue of a real symmetric matrix $\mathbf{A}$. We assume that the true number of regimes in the MSVAR model is already given.

We use the results on consistency and asymptotic normality of the maximum-likelihood estimators obtained by Douc et al. (2004) for the MSVAR models. All their assumptions follow in our setting of a homogeneous regime-governing Markov chain $\mathrm{S}_t$ and the Gaussian noise process.

**Proposition 6.1.** (Douc et al., 2004, Proposition 1-3) There exists a deterministic function $l(\cdot)$ such that

(i) for all $s_p \in \{1,\ldots,\mathrm{M}\}$, and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\lim_{n\to\infty} n^{-1} l_n(\boldsymbol{\theta}; s_p) = l(\boldsymbol{\theta}), \quad \mathbb{P}_{\boldsymbol{\theta}^\star}\text{-a.s. and in } \mathrm{L}^1(\mathbb{P}_{\boldsymbol{\theta}^\star});$$

(ii) $\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{s_p} |n^{-1} l_n(\boldsymbol{\theta}; s_p) - l(\boldsymbol{\theta})| = 0, \quad \mathbb{P}_{\boldsymbol{\theta}^\star}\text{-a.s.};$

(iii) $l(\boldsymbol{\theta}) \leq l(\boldsymbol{\theta}^\star)$ and $l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}^\star)$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$.

The above results ensure the strong consistency of the conditional MLE estimator. Next, we also have a central limit theorem for the score function and a law of large numbers for the observed Fisher information.

**Proposition 6.2.**

(i) (Douc et al., 2004, Theorem 2) (Central limit theorem) For any $s_p \in \{1,\ldots,\mathrm{M}\}$

$$n^{-1/2} \nabla_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}^\star; s_p) \xrightarrow{\mathbb{P}_{\boldsymbol{\theta}^\star}} \mathcal{N}\big(\mathbf{0}, \mathrm{I}(\boldsymbol{\theta}^\star)\big),$$

where $\mathrm{I}(\boldsymbol{\theta}^\star)$ is the asymptotic Fisher information matrix, defined as the covariance of asymptotic score function.

(ii) (Douc et al., 2004, Theorem 3) (Law of large numbers) For any $s_p \in \{1,\ldots,\mathrm{M}\}$ and a possibly stochastic sequence $\{\boldsymbol{\theta}_n^\star\} \in \boldsymbol{\Theta}$ converging to $\boldsymbol{\theta}^\star$ $\mathbb{P}_{\boldsymbol{\theta}^\star}$-almost surely, we have

$$-n^{-1} \nabla^2_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}_n^\star; s_p) \longrightarrow \mathrm{I}(\boldsymbol{\theta}^\star), \mathbb{P}_{\boldsymbol{\theta}^\star}\text{-a.s.}$$

As fixing the initial state of the regime-governing Markov chain does not affect our asymptotic convergence results, we simplify our notation for $l_n(\boldsymbol{\theta}; s_p)$ and $\boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta}; s_p)$ by rewriting them as $l_n(\boldsymbol{\theta})$ and $\boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta})$, respectively. The following results hold for any $s_p \in \{1, \ldots, \mathrm{M}\}$. We first show that under the correct specification of M and appropriate choices of $(r_{\lambda_n}, \lambda_n)$, there exists a local maximizer $\widehat{\boldsymbol{\theta}}_n$ of the penalized conditional likelihood function $\mathcal{L}_n(\boldsymbol{\theta})$ that is a consistent and sparse estimator of $\boldsymbol{\theta}^\star$.

**Theorem 1.** *Suppose that $\{\mathbf{Y}_t\}$ is generated according to a stationary and ergodic MSVAR process. Further assume that Assumption 1 - 3 on the tuning parameter $\lambda_n$ and penalty function $r_{\lambda_n}$ hold. Then, as $n \to \infty$,*
*(i) there exists a local maximizer $\widehat{\boldsymbol{\theta}}_n$ of the penalized log-likelihood $\mathcal{L}_n(\boldsymbol{\theta})$ such that $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\star\| = \mathrm{O_P}(n^{-1/2})$;*
*(ii) for any $\sqrt{n}$–consistent estimator $\widehat{\boldsymbol{\theta}}_n$ of the true sparse parameter $\boldsymbol{\theta}^\star$ with $\boldsymbol{\theta}^{\star\mathcal{A}^c} = \mathbf{0}$, we have $\mathbb{P}\big(\widehat{\boldsymbol{\theta}}_n^{\mathcal{A}^c} = \mathbf{0}\big) \to 1$.*

*Proof.* Let $\alpha_n = n^{-1/2} + a_n$. For the first claim, it suffices to show that for $n \geq n_0$ with $n_0$ large enough, for any $\varepsilon > 0$ there exists a large constant C such that

$$\mathbb{P}\Big( \sup_{\|\mathbf{u}\|=\mathrm{C}} \boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta}^\star + \alpha_n \mathbf{u}) < \boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta}^\star) \Big) \geq 1 - \varepsilon. \tag{13}$$

In other words, we need to show that for $\|\mathbf{u}\| = \mathrm{C}$, as $n \to \infty$, $\boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta}^\star + \alpha_n \mathbf{u}) - \boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta}^\star) < 0$ uniformly in $\mathbf{u}$ with probability approaching to one. As the penalty function is non-negative, for $k \in \mathcal{A}^c$, $r_{\lambda_n}\big(\theta_k^\star + \alpha_n u_k\big) \geq 0$. Also recall that $r_{\lambda_n}(0) = 0$. Then, we have the following inequality:

$$\begin{aligned}
\mathrm{V}_n(\mathbf{u}) &\equiv \big(\boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta}^\star + \alpha_n \mathbf{u}) - \boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta}^\star)\big) \\
&= \frac{1}{(n-p)} l_n(\boldsymbol{\theta}^\star + \alpha_n \mathbf{u}) - \sum_{k \in \mathcal{E}} r_{\lambda_n}\big(\theta_k^\star + \alpha_n u_k\big) - \frac{1}{(n-p)} l_n(\boldsymbol{\theta}^\star) + \sum_{k \in \mathcal{E}} r_{\lambda_n}\big(\theta_k^\star\big) \\
&\leq \frac{1}{(n-p)} \big(l_n(\boldsymbol{\theta}^\star + \alpha_n \mathbf{u}) - l_n(\boldsymbol{\theta}^\star)\big) - \sum_{k \in \mathcal{A} \cap \mathcal{E}} \Big(r_{\lambda_n}\big(\theta_k^\star + \alpha_n u_k\big) - r_{\lambda_n}\big(\theta_k^\star\big)\Big) \\
&=: \mathrm{D}_{1n}(\mathbf{u}) - \mathrm{D}_{2n}(\mathbf{u}, \lambda_n).
\end{aligned}$$

We use the Taylor expansion of the likelihood function $l_n$ to write:

$$\mathrm{D}_{1n}(\mathbf{u}) = \frac{\alpha_n}{(n-p)} \mathbf{u}^\mathsf{T} \nabla_\theta l_n(\boldsymbol{\theta}^\star) + \frac{1}{2(n-p)} \alpha_n^2 \mathbf{u}^\mathsf{T} \nabla_\theta^2 l_n(\widetilde{\boldsymbol{\theta}}_n) \mathbf{u}.$$

In the above $\widetilde{\boldsymbol{\theta}}_n$ is equal to $s\boldsymbol{\theta}^\star + (1-s)(\boldsymbol{\theta}^\star + \alpha_n \mathbf{u})$ for some $0 < s < 1$. From Proposition 6.2(i) we have for large enough $n$, $\nabla_\theta l_n(\boldsymbol{\theta}^\star) = \mathrm{O_P}(n^{1/2})$. This gives $\alpha_n \nabla_\theta l_n(\boldsymbol{\theta}^\star) = \mathrm{O_P}(\alpha_n n^{1/2}) = \mathrm{O_P}(\alpha_n^2 n)$, due to Assumption 2 on $a_n$. From Proposition 6.2(ii), for large enough $n$, we have $-\frac{1}{n}\nabla_\theta^2 l_n(\boldsymbol{\theta}) = \mathrm{I}(\boldsymbol{\theta}^\star)\big(1 + o_\mathrm{P}(1)\big)$. Then, we get

$$\mathrm{D}_{1n}(\mathbf{u}) = \frac{1}{(n-p)} \|\mathbf{u}\| \mathrm{O_P}(\alpha_n^2 n) - \frac{1}{2(n-p)} \mathbf{u}^\mathsf{T} \mathrm{I}(\boldsymbol{\theta}^\star) \mathbf{u} n \alpha_n^2 \big(1 + o_\mathrm{P}(1)\big). \tag{14}$$

In the above, it is clear that the second term dominates the first uniformly for $\|\mathbf{u}\| = \mathrm{C}$ with a sufficiently large C. Recall that for all $k \in \mathcal{A}, \theta_k^\star \neq 0$ and for large enough $n$, $|\theta_k^\star| > c\lambda_n$ for some constant $c > 0$, since $\lambda_n = o(1)$ due to Assumption 2. As $\alpha_n = o(1)$, we can perform a second-order Taylor expansion in $\mathrm{D}_{2n}(\mathbf{u}, \lambda_n)$ due to Assumption 1, and obtain the following inequality,

$$\mathrm{D}_{2n}(\mathbf{u}, \lambda_n) = \sum_{k \in \mathcal{A} \cap \mathcal{E}} \Big(\alpha_n r'_{\lambda_n}\big(\theta_k^\star\big) u_k + \frac{\alpha_n^2}{2} r''_{\lambda_n}\big(\theta_k^\star\big) u_k^2 \big(1 + o(1)\big)\Big)$$

$$\leq \sqrt{|\mathcal{A} \cup \mathcal{A}^c|}\alpha_n a_n \|\mathbf{u}\| + \frac{\alpha_n^2}{2} b_n \|\mathbf{u}\|^2. \tag{15}$$

From the assumption on the rates of decay of $a_n$ and $b_n$ in Assumption 2, the above upper bound can also be dominated by the second term of (14) by choosing a large enough C. Thus, from the results in (14) and (15), we can conclude that for sufficiently large $n$, $\mathrm{V}_n(\mathbf{u}) < 0$, and that the hypothesis in (13) holds.

To prove the result on the oracle property of the penalized estimator in Theorem 1(ii), we first show the following lemma.

**Lemma 6.3.** Suppose that $\boldsymbol{\theta}_n = [(\boldsymbol{\theta}_n{}^{\mathcal{A}})^\mathsf{T}, (\boldsymbol{\theta}_n{}^{\mathcal{A}^c})^\mathsf{T}]^\mathsf{T}$, is such that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star\| = \mathrm{O_P}(n^{-1/2})$ as $n \to \infty$. Then, for tuning parameter $\lambda_n$ and penalty function $r_{\lambda_n}$ satisfying Assumption 1–3, we have as $n \to \infty$, with probability tending to one that

$$\mathcal{L}_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^\mathsf{T}, (\boldsymbol{\theta}_n^{\mathcal{A}^c})^\mathsf{T}]^\mathsf{T}\big) < \mathcal{L}_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T}\big),$$

where $\boldsymbol{\theta}^\star = [(\boldsymbol{\theta}^{\star\mathcal{A}})^\mathsf{T}, (\boldsymbol{\theta}^{\star\mathcal{A}^c})^\mathsf{T}]^\mathsf{T}$, with $\mathcal{A}$ containing the indices of all nonzero elements of $\boldsymbol{\theta}^\star$.

*Proof.* Consider,

$$\begin{aligned}
&\Big(\mathcal{L}_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^\mathsf{T}, (\boldsymbol{\theta}_n^{\mathcal{A}^c})^\mathsf{T}]^\mathsf{T}\big) - \mathcal{L}_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T}\big)\Big) \\
&= \frac{1}{(n-p)}\Big(l_n(\boldsymbol{\theta}_n) - l_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T}\big)\Big) - \Big(\sum_{k\in\mathcal{E}} r_{\lambda_n}(\theta_{n,k}) - \sum_{k\in\mathcal{E}\cap\mathcal{A}} r_{\lambda_n}(\theta_{n,k})\Big).
\end{aligned}$$

By the second-order Taylor expansion, we have

$$\begin{aligned}
l_n(\boldsymbol{\theta}_n) - l_n(\boldsymbol{\theta}^\star) &= (\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star)^\mathsf{T}\nabla_\theta l_n(\boldsymbol{\theta}^\star) + \frac{1}{2}(\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star)^\mathsf{T}\nabla_\theta^2 l_n(\widetilde{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star), \\
l_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T}\big) - l_n(\boldsymbol{\theta}^\star) &= \big([(\boldsymbol{\theta}_n^{\mathsf{A}})^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T} - \boldsymbol{\theta}^\star\big)^\mathsf{T}\nabla_\theta l_n(\boldsymbol{\theta}^\star) \\
&\quad + \frac{1}{2}\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T} - \boldsymbol{\theta}^\star\big)^\mathcal{T}\nabla_\theta^2 l_n(\widetilde{\boldsymbol{\theta}}_n^{\mathcal{A}})\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T} - \boldsymbol{\theta}^\star\big).
\end{aligned}$$

In the above $\widetilde{\boldsymbol{\theta}}_n$ is $s\boldsymbol{\theta}^\star + (1-s)\boldsymbol{\theta}_n$ for some $0 < s < 1$ and $\widetilde{\boldsymbol{\theta}}_n^{\mathcal{A}}$ is $t\boldsymbol{\theta}^\star + (1-t)[(\boldsymbol{\theta}_n^{\mathcal{A}})^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T}$ for some $0 < t < 1$. We know that $\boldsymbol{\theta}^\star = [(\boldsymbol{\theta}^{\star\mathcal{A}})^\mathsf{T}, \mathbf{0}^\mathsf{T}]^\mathsf{T}$, which gives

$$(\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star)^\mathsf{T}\nabla_\theta l_n(\boldsymbol{\theta}^\star) = (\boldsymbol{\theta}_n^{\mathsf{A}} - \boldsymbol{\theta}^{\star\mathsf{A}})^\mathsf{T}\nabla_{\theta^\mathcal{A}} l_n(\boldsymbol{\theta}^\star) + (\boldsymbol{\theta}_n^{\mathsf{A}^c} - \boldsymbol{\theta}^{\star\mathsf{A}^c})^\mathsf{T}\nabla_{\theta^{\mathcal{A}^c}} l_n(\boldsymbol{\theta}^\star).$$

From Proposition 6.2(i) we have $\nabla_\theta l_n(\boldsymbol{\theta}^\star) = \mathrm{O_P}(n^{1/2})$. Then,

$$(\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star)^\mathsf{T}\nabla_\theta l_n(\boldsymbol{\theta}^\star) = (\boldsymbol{\theta}_n^{\mathsf{A}} - \boldsymbol{\theta}^{\star\mathsf{A}})^\mathsf{T}\nabla_{\theta^\mathcal{A}} l_n(\boldsymbol{\theta}^\star) + \mathrm{O_P}(n^{1/2})|\boldsymbol{\theta}_n^{\mathcal{A}^c}|.$$

As $n \to \infty$, $\widetilde{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^\star$ a.s., then from Proposition 6.2(ii) we have $-n^{-1}\nabla_\theta^2 l_n(\widetilde{\boldsymbol{\theta}}_n) = \mathrm{I}(\boldsymbol{\theta}^\star)(1 + o_\mathrm{P}(1))$. Furthermore, as $n \to \infty$, $\widetilde{\boldsymbol{\theta}}_n^{\mathcal{A}} \to \boldsymbol{\theta}^{\star\mathcal{A}}$ a.s., we have $-n^{-1}\nabla_\theta^2 l_n(\widetilde{\boldsymbol{\theta}}_n^{\mathcal{A}}) = \mathrm{I}_{11}(\boldsymbol{\theta}^\star)(1 + o_\mathrm{P}(1))$, where the Fisher information matrix corresponding to $\boldsymbol{\theta}^\star = [(\boldsymbol{\theta}^{\star\mathcal{A}})^\mathsf{T}, (\boldsymbol{\theta}^{\star\mathcal{A}^c})^\mathsf{T}]^\mathsf{T}$, is written as

$$\mathrm{I}(\boldsymbol{\theta}^\star) = \begin{bmatrix} \mathrm{I}_{11}(\boldsymbol{\theta}^\star) & \mathrm{I}_{12}(\boldsymbol{\theta}^\star) \\ \mathrm{I}_{12}(\boldsymbol{\theta}^\star) & \mathrm{I}_{22}(\boldsymbol{\theta}^\star) \end{bmatrix}.$$

Thus, for large enough $n$, we have

$$\begin{aligned} l_n(\boldsymbol{\theta}_n) - l_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}]^{\mathsf{T}}\big) &= \mathrm{O_P}(n^{1/2})|\boldsymbol{\theta}_n^{\mathsf{A}^c}| - \frac{1}{2} n\big(\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star\big)^{\mathsf{T}} \mathrm{I}(\boldsymbol{\theta}^\star)\big(\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star\big)(1 + o_{\mathrm{P}}(1)) \\ &\quad + \frac{1}{2} n\big(\boldsymbol{\theta}_n^{\mathcal{A}} - \boldsymbol{\theta}^{\star\mathcal{A}}\big)^{\intercal} \mathrm{I}_{11}(\boldsymbol{\theta}^\star)\big(\boldsymbol{\theta}_n^{\mathcal{A}} - \boldsymbol{\theta}^{\star\mathcal{A}}\big)(1 + o_{\mathrm{P}}(1)) \\ &= \mathrm{O_P}(n^{1/2})|\boldsymbol{\theta}_n^{\mathsf{A}^c}| - n\Big(\big(\boldsymbol{\theta}_n^{\mathsf{A}} - \boldsymbol{\theta}^{\star\mathsf{A}}\big)^{\mathsf{T}} \mathrm{I}_{12}(\boldsymbol{\theta}^\star)\boldsymbol{\theta}_n^{\mathcal{A}^c} \\ &\quad + \frac{1}{2}(\boldsymbol{\theta}_n^{\mathcal{A}^c})^{\mathsf{T}} \mathrm{I}_{22}(\boldsymbol{\theta}^\star)\boldsymbol{\theta}_n^{\mathcal{A}^c}\Big)(1 + o_{\mathrm{P}}(1)). \end{aligned}$$

Since $\|\boldsymbol{\theta}_n^{\mathcal{A}} - \boldsymbol{\theta}^{\star\mathcal{A}}\| = \mathrm{O_P}(n^{-1/2})$ and $\|\boldsymbol{\theta}_n^{\mathcal{A}^c}\| = \mathrm{O_P}(n^{-1/2})$, for large enough $n$ and some constants $\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3 > 0$, we get

$$l_n(\boldsymbol{\theta}_n) - l_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}]^{\mathsf{T}}\big) \leq \mathrm{C}_1 n^{1/2}|\boldsymbol{\theta}_n^{\mathcal{A}^c}| + n\mathrm{C}_2 n^{-1/2}|\boldsymbol{\theta}_n^{\mathcal{A}^c}|(1 + o_{\mathrm{P}}(1)) \leq n^{1/2}\mathrm{C}_3 \sum_{k \in \mathcal{A}^c} |\theta_{n,k}|. \quad (16)$$

Next, we notice that

$$\Big(\sum_{k \in \mathcal{E}} r_{\lambda_n}(\theta_{n,k}) - \sum_{k \in \mathcal{E} \cap \mathcal{A}} r_{\lambda_n}(\theta_{n,k})\Big) = \sum_{k \in \mathcal{E} \cap \mathcal{A}^c} r_{\lambda_n}(\theta_{n,k}).$$

From (16), for large enough $n$, we have

$$\boldsymbol{\mathcal{L}}_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^{\mathsf{T}}, (\boldsymbol{\theta}_n^{\mathcal{A}^c})^{\mathsf{T}}]^{\mathsf{T}}\big) - \boldsymbol{\mathcal{L}}_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}]^{\mathsf{T}}\big) \leq \frac{n^{1/2}}{(n-p)}\mathrm{C}_3 \sum_{k \in \mathcal{A}^c} |\theta_{n,k}| - \frac{1}{n} \sum_{k \in \mathcal{E} \cap \mathcal{A}^c} n r_{\lambda_n}(\theta_{n,k}). \quad (17)$$

As $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star\| = \mathrm{O_P}(n^{-1/2})$, for the coefficients belonging to the zero-elements set $\mathcal{A}^c$, we have $|\theta_{n,k}| = \mathrm{O_P}(n^{-1/2})$. Then, for each $k \in \mathcal{A}^c$, Assumption 3 is applicable on the penalty function value $r_{\lambda_n}(\theta_{n,k})$, which gives

$$\liminf_{n \to \infty} n r_{\lambda_n}(\theta_{n,k}) = \liminf_{n \to \infty} \sqrt{n}|\theta_{n,k}|\sqrt{n} r'_{\lambda_n}(\theta_{n,k}) = +\infty \text{ a.s.}.$$

Hence, as $n \to \infty$, from (17) we deduce that

$$\boldsymbol{\mathcal{L}}_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^{\mathsf{T}}, (\boldsymbol{\theta}_n^{\mathcal{A}^c})^{\mathsf{T}}]^{\mathsf{T}}\big) - \boldsymbol{\mathcal{L}}_n\big([(\boldsymbol{\theta}_n^{\mathcal{A}})^{\mathsf{T}}, \mathbf{0}^{\mathsf{T}}]^{\mathsf{T}}\big) < 0 \text{ a.s.}.$$

which proves the result. □

Now, we provide the remainder of the proof. From Lemma 6.3, we know that for any $\sqrt{n}$–consistent estimator $\widehat{\boldsymbol{\theta}}_n$, the parameter vector $\begin{bmatrix} \widehat{\boldsymbol{\theta}}_n^{\mathcal{A}} \\ 0 \end{bmatrix}$ maximizes $\boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta})$ with probability tending to one as $n \to \infty$, over any other choice $\boldsymbol{\theta}_n = \begin{bmatrix} \boldsymbol{\theta}_n^{\mathcal{A}} \\ \boldsymbol{\theta}_n^{\mathcal{A}^c} \end{bmatrix}$ such that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^\star\| = \mathrm{O_P}(n^{-1/2})$. From Theorem 1, we know that a $\sqrt{n}$-consistent maximizer $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\mathcal{L}}_n(\boldsymbol{\theta})$ exists when $\lambda_n$ and $r_{\lambda_n}$ satisfy Assumption 2. Then, it must be true that $\mathbb{P}(\widehat{\boldsymbol{\theta}}_n^{\mathcal{A}^c} = 0) \to 1$ as $n \to \infty$. □

The result in Theorem 1 shows that we have a $\sqrt{n}$–consistent estimator if $a_n = \mathrm{O}(n^{-1/2})$, with $\lambda_n \to 0$. For SCAD and MCP penalty functions, it is then enough that $\lambda_n \to 0$ as $n \to \infty$, since then $a_n = 0$. For the LASSO penalty, we must have at least that $\lambda_n = \mathrm{O}(n^{-1/2})$ since $a_n = \lambda_n$; and for

the ADALASSO, we require $\sqrt{n}\lambda_n = o(1)$. However, as is well-known, it is not possible to recover the sparsity structure using the LASSO penalty as we will then require that $\lim_{n\to\infty}\sqrt{n}\lambda_n = \infty$, to satisfy Assumption 3, which contradicts the earlier requirement on the tuning parameter decay rate for the LASSO penalty function. The sparsity structure can be recovered with the other three penalty functions, since we can choose $\lambda_n \sim n^{-1/2-\psi}$ for a $0 < \psi < \gamma/2$, for the adaptive LASSO, and $\lambda_n \sim n^{-1/2}\log n$ for the SCAD and MCP penalties.

If M is under-specified, the MPLE $\widehat{\boldsymbol{\theta}}_n$ converges to the minimizer of the Kullback-Leibler distance between the densities of the under-specified and true models, a property shared by the MLE under this misspecification (Douc and Moulines, 2012). On the other hand, in Theorem 2 we show that if M is over-specified, the density function of the over-fitted MSVAR model based on MPLE consistently estimates the density function of the true model, which is useful for prediction. In particular, we show that the estimated predictive density of the over-fitted model based on MPLE consistently estimates the $h$-step ahead predictive density of the true MSVAR model with M regimes, denoted by $f^\star(\mathbf{y}_{n+1:h}|\mathbf{y}_{1:n})$, and computed in Section 4 of the main manuscript. As a consequence of this result, in practice when the true number of regimes is unknown, a conservative choice of M considering the sample size $n$ can guarantee a reliable estimate of the $h$-step ahead predictive density, and the optimal predictor (see Section 4, main manuscript) in the sense of minimum mean squared prediction error.

**Theorem 2.** *Under Assumption 1–3, the estimated $h$-step ahead predictive density function $\widehat{f}_{\mathcal{M}}(\mathbf{y}_{n+1:h}|\mathbf{y}_{1:n})$ for number of regimes $\mathcal{M} \geq$ M, converges almost surely to the true $h$-step ahead predictive density $f^\star(\mathbf{y}_{n+1:h}|\mathbf{y}_{1:n})$ as $n \to \infty$ where* M *is the true number of regimes in the model.*

*Proof.* We provide the proof for the one-step estimated predictive density. The result for the $h$-step estimated predictive density can be shown similarly.

We denote the generic intercept vector for VAR model by $\nu$ and the vectorized VAR coefficients and covariance matrices as $a_i = \mathbf{vec}(\mathrm{A}_i), i = 1,\ldots,p$ and $\sigma = \mathbf{vec}(\Sigma)$. For number of regimes $\mathcal{M} >$ M and $n > p$, the one-step ahead predictive density function is given as

$$f_{\mathcal{M}}(\mathbf{y}_{n+1}|\mathbf{y}_{1:n}) = \sum_{m=1}^{\mathcal{M}} g\big(\mathbf{y}_{n+1},\mathbf{y}_{1:n};\nu^{(m)},\{a_i^{(m)}\}_{i=1}^p,\sigma^{(m)}\big)\mathbb{P}(s_{n+1}=m|\mathbf{y}_{1:n}),$$

where

$$\begin{aligned} g\big(\mathbf{y}_n+1,\mathbf{y}_{1:n};\nu,\{a_i\}_{i=1}^p,\sigma\big) &:= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\Bigg(-\frac{1}{2}(\mathbf{y}_{n+1}-\mu(\mathbf{y}_{1:n},\nu,\{a_i\}_{i=1}^p))^\mathsf{T}(\Sigma)^{-1} \\ &\qquad (\mathbf{y}_{n+1}-\mu(\mathbf{y}_{1:n},\nu,\{a_i\}_{i=1}^p))\Bigg), \\ \mu(\mathbf{y}_{1:n},\nu,\{a_i\}_{i=1}^p) &:= \nu + \mathrm{A}_1\mathbf{y}_n + \ldots + \mathrm{A}_p\mathbf{y}_{n-p+1}. \end{aligned}$$

The estimated one-step ahead predictive density for number of regimes $\mathcal{M}$ is given as

$$\widehat{f}_{\mathcal{M}}(\mathbf{y}_{n+1}|\mathbf{y}_{1:n}) = \sum_{j=1}^{\mathcal{M}} g\big(\mathbf{y}_{n+1},\mathbf{y}_{1:n};\widehat{\nu}^{(m)},\{\widehat{a}_i^{(m)}\}_{i=1}^p,\widehat{\sigma}^{(m)}\big)\widehat{\mathbb{P}}(s_{n+1}=m|\mathbf{y}_{1:n}),$$

with $\widehat{\mathbb{P}}(s_{n+1}=m|\mathbf{y}_{1:n})$ computed via forward and backward recursions and using the estimates of the transition probability matrix for $\mathcal{M}$ number of regimes. The estimates of VAR coefficients and covariance matrices are denoted by $(\widehat{\nu}^{(m)},\{\widehat{a}_i^{(m)}\}_{i=1}^p)$ and $\widehat{\sigma}^{(m)}$ respectively. We can alternatively write

$$\widehat{f}_{\mathcal{M}}(\mathbf{y}_{n+1}|\mathbf{y}_{1:n}) = \int g\big(\mathbf{y}_{n+1},\mathbf{y}_{1:n};\nu,\{a_i\}_{i=1}^p,\sigma\big)\mathrm{d}\widehat{\boldsymbol{\Phi}}_{\mathcal{M},n}(\nu,\{a_i\}_{i=1}^p,\sigma), \tag{18}$$

with

$$\widehat{\mathbf{\Phi}}_{\mathcal{M},n}(\nu,\{a_i\}_{i=1}^p,\sigma)=\sum_{m=1}^{\mathcal{M}}\widehat{\mathbb{P}}(s_{n+1}=m|\mathbf{y}_{1:n})\mathrm{H}(\nu\geq\widehat{\nu}^{(m)},\{a_i\}_{i=1}^p\geq\{\widehat{a}_i^{(m)}\}_{i=1}^p,\sigma\geq\widehat{\sigma}^{(m)}),$$

where $\mathrm{H}(\cdot)$ denotes the Heaviside step function. The true one-step ahead predictive density can be then written as

$$f^\star(\mathbf{y}_{n+1}|\mathbf{y}_{1:n})=\int g\big(\mathbf{y}_{n+1},\mathbf{y}_{1:n};\nu,\{a_i\}_{i=1}^p,\sigma\big)\mathrm{d}\mathbf{\Phi}_n^\star(\nu,\{a_i\}_{i=1}^p,\sigma),$$

with

$$\mathbf{\Phi}_n^\star(\nu,\{a_i\}_{i=1}^p,\sigma)=\sum_{m=1}^{\mathrm{M}}\mathbb{P}^\star(s_{n+1}=m|\mathbf{y}_{1:n})\mathrm{H}(\nu\geq\nu^{(m)\star},\{a_i\}_{i=1}^p\geq\{a^{(m)\star}_i\}_{i=1}^p,\sigma\geq\sigma^{(m)\star}).$$

The estimated and true filtered probabilities are computed by fixing the initial state $s_p$ and initializing with true conditional state probabilities $\mathbb{P}^\star(s_p=m|\mathbf{y}_{1:p})$ respectively. The effect of initial distribution dissipates geometrically fast in number of observations $n$ as shown in Douc et al. (2004, Corollary 1). Next, we define the following distance between the probability measures $\mathbf{\Phi}_n^\star(\nu,\{a_i\}_{i=1}^p,\sigma)$ and $\mathbf{\Phi}_{\mathcal{M},n}(\nu,\{a_i\}_{i=1}^p,\sigma)$ :

$$\begin{aligned}&\mathrm{D}(\mathbf{\Phi}_{\mathcal{M},n},\mathbf{\Phi}_n^\star)\\&=\int\int\big|\mathbf{\Phi}_{\mathcal{M},n}(\nu,\{a_i\}_{i=1}^p,\sigma)-\mathbf{\Phi}_n^\star(\nu,\{a_i\}_{i=1}^p,\sigma)\big|\mathrm{e}^{-(|\nu|+\sum_{i=1}^p|a_i|+|\sigma|)}\mathrm{d}\nu\,\mathrm{d}a_1\ldots\mathrm{d}a_p\,\mathrm{d}\sigma\,\mathrm{d}\mathbb{P}^\star(\mathbf{y}_{1:n}).\end{aligned}\tag{19}$$

The above distance metrizes the space of probability measures on the space of VAR model parameter $(\nu,\{a_i\}_{i=1}^p,\sigma)\subset\mathbf{\Theta}$. We can consider any other distance between probability measures on the space of VAR model parameter $(\nu,\{a_i\}_{i=1}^p,\sigma)$. For an arbitrary $\delta>0$, we consider a family

$$\mathcal{H}_n(\delta)=\{\mathbf{\Phi}_{\mathcal{M},n}:\mathrm{D}(\mathbf{\Phi}_{\mathcal{M},n},\mathbf{\Phi}_n^\star)>\delta\},$$

of probability measures such that each element in it is at least a $\delta$ distance away from $\mathbf{\Phi}_n^\star$. Then, clearly $\mathbf{\Phi}_n^\star\notin\mathcal{H}_n(\delta)$, and

$$\mathbb{E}^\star\left[\log\left(\frac{f_{\mathcal{M}}(\mathbf{y}_{n+1}|\mathbf{y}_{1:n})}{f^\star(\mathbf{y}_{n+1}|\mathbf{y}_{1:n})}\right)\right]<0.$$

Similarly, we also have for $t=p+1,\ldots,n$

$$\mathbb{E}^\star\left[\log\left(\frac{f_{\mathcal{M}}(\mathbf{y}_t|=\mathbf{y}_{1:t-1})}{f^\star(\mathbf{y}_t|\mathbf{y}_{1:t-1})}\right)\right]<0,$$

where the conditional densities $f_{\mathcal{M}}(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ and $f^\star(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ have their own representation similar to $f_{\mathcal{M}}(\mathbf{y}_{n+1}|\mathbf{y}_{1:n})$ and $f^\star(\mathbf{y}_{n+1}|\mathbf{y}_{1:n})$ but with their own probability measures $\mathbf{\Phi}_{\mathcal{M},t}\in\mathcal{H}_t(\delta)$ and $\mathbf{\Phi}_t^\star$, respectively. From the stationarity of the process and ergodic theorem,we can also conclude that

$$\frac{1}{n-p}\sum_{t=p+1}^{n}\log\left(\frac{f_{\mathcal{M}}(\mathbf{y}_t|\mathbf{y}_{1:t-1})}{f^\star(\mathbf{y}_t|\mathbf{y}_{1:t-1})}\right)=\frac{1}{n-p}\log\left(\frac{f_{\mathcal{M}}(\mathbf{y}_{p+1},\ldots,\mathbf{y}_n|\mathbf{y}_{1:p})}{f^\star(\mathbf{y}_{p+1},\ldots,\mathbf{y}_n|\mathbf{y}_{1:p})}\right)<-\epsilon(\delta)\text{ a.s.},\tag{20}$$

for some $\epsilon(\delta) > 0$ and $n$ large enough. Define $\boldsymbol{\gamma} := (\nu, \{a_i\}_{i=1}^p, \sigma)$ as the VAR model parameter vector. Then, the joint probability density under the assumption of $\mathcal{M} > \mathrm{M}$ regimes and true joint probability density can be expressed as follows:

$$f_{\mathcal{M}}(\mathbf{y}_{p+1}, \ldots, \mathbf{y}_n | \mathbf{y}_{1:p}) = \int \prod_{t=p+1}^{n} g\big(\mathbf{y}_t, \mathbf{y}_{1:t-1}; \boldsymbol{\gamma}_t\big) \mathrm{d}\boldsymbol{\Phi}_{\text{joint},\mathcal{M}}(\boldsymbol{\gamma}_{p+1:n}),$$

$$f^{\star}(\mathbf{y}_{p+1}, \ldots, \mathbf{y}_n | \mathbf{y}_{1:p}) = \int \prod_{t=p+1}^{n} g\big(\mathbf{y}_t, \mathbf{y}_{1:t-1}; \boldsymbol{\gamma}_t\big) \mathrm{d}\boldsymbol{\Phi}^{\star}_{\text{joint}}(\boldsymbol{\gamma}_{p+1:n}),$$

where $\boldsymbol{\gamma}_{p+1:n} = (\boldsymbol{\gamma}_{p+1}, \ldots, \boldsymbol{\gamma}_n)$, with $\boldsymbol{\gamma}_t, t = p+1, \ldots, n$, representing the choice of VAR model parameter at different instances. Furthermore,

$$g\big(\mathbf{y}_t, \mathbf{y}_{1:t-1}; \boldsymbol{\gamma}\big) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{y}_t - \mu(\mathbf{y}_{1:t-1}, \nu, \{a_i\}_{i=1}^p))^{\mathsf{T}} (\Sigma)^{-1} (\mathbf{y}_t - \mu(\mathbf{y}_{1:t-1}, \nu, \{a_i\}_{i=1}^p))\right),$$

$$\mu(\mathbf{y}_{1:t-1}, \nu, \{a_i\}_{i=1}^p) = \nu + \mathrm{A}_1 \mathbf{y}_{t-1} + \ldots + \mathrm{A}_p \mathbf{y}_{t-p},$$

and

$$\boldsymbol{\Phi}_{\text{joint},\mathcal{M}}(\boldsymbol{\gamma}_{p+1:n}) = \sum_{j_{p+1}=1}^{\mathcal{M}} \ldots \sum_{j_n=1}^{\mathcal{M}} \mathbb{P}(s_{p+1} = j_{p+1} | \mathbf{y}_{1:p}) \prod_{t=p+2}^{n} \alpha'_{j_{t-1}, j_t} \prod_{t=p+1}^{n} \mathrm{H}(\gamma_t \geq \gamma'_{j_t}),$$

$$\boldsymbol{\Phi}^{\star}_{\text{joint}}(\boldsymbol{\gamma}_{p+1:n}) = \sum_{j_{p+1}=1}^{\mathrm{M}} \ldots \sum_{j_n=1}^{\mathcal{M}} \mathbb{P}^{\star}(s_{p+1} = j_{p+1} | \mathbf{y}_{1:p}) \prod_{t=p+2}^{n} \alpha^{\star}_{j_{t-1}, j_t} \prod_{t=p+1}^{n} \mathrm{H}(\gamma_t \geq \gamma^{\star}_{j_t}).$$

For the joint distribution $\boldsymbol{\Phi}_{\text{joint},\mathcal{M}}(\boldsymbol{\gamma}_{p+1:n})$ we can similarly define a family

$$\mathcal{H}_{\text{joint}}(\delta) := \{\boldsymbol{\Phi}_{\text{joint},\mathcal{M}} : \mathrm{D}(\boldsymbol{\Phi}_{\text{joint},\mathcal{M}}, \boldsymbol{\Phi}^{\star}_{\text{joint}}) > \delta\},$$

based on its distance from the joint distribution $\boldsymbol{\Phi}^{\star}_{\text{joint}}(\boldsymbol{\gamma}_{p+1:n})$ with the distance being defined analogously as in (19) with inner integral being with respect to $\boldsymbol{\gamma}_{p+1:n}$ and the outside probability measure corresponding to $\mathbb{P}^{\star}(\mathbf{y}_{1:p})$.

From the result in (20) for any $\boldsymbol{\Phi}_{\text{joint},\mathcal{M}} \in \mathcal{H}_{\text{joint}}(\delta)$, and the fact that penalty terms are non-negative with order $o(1)$, we get

$$\sup_{\mathcal{H}_{\text{joint}}(\delta)} \left( \frac{1}{(n-p)} \log \left( \frac{f_{\mathcal{M}}(\mathbf{y}_{p+1}, \ldots, \mathbf{y}_n | \mathbf{y}_{1:p})}{f^{\star}(\mathbf{y}_{p+1}, \ldots, \mathbf{y}_n | \mathbf{y}_{1:p})} \right) - \Big( \sum_{k \in \mathcal{E}_{\mathcal{M}}} \mathrm{R}_{\lambda_n}(\theta_k^{\mathcal{T}}) - \sum_{k \in \mathcal{E}} \mathrm{R}_{\lambda_n}(\theta_k^{\star}) \Big) \right) < -\epsilon(\delta) \text{ a.s.}.$$

Hence, the joint probability distribution corresponding to the penalized maximum-likelihood estimate in coefficient space $\boldsymbol{\Theta}_{\mathcal{M}}$ cannot be an element of $\mathcal{H}_{\text{joint}}(\delta)$ almost surely as $n \to \infty$. This is true for any $\delta > 0$, thus we must have for the joint distribution corresponding to the penalized maximum-likelihood estimate in coefficient space $\boldsymbol{\Theta}_{\mathcal{M}}$ : $\widehat{\boldsymbol{\Phi}}_{\text{joint},\mathcal{M}}$ that $\mathrm{D}(\widehat{\boldsymbol{\Phi}}_{\text{joint},\mathcal{M}}, \boldsymbol{\Phi}^{\star}_{\text{joint}}) \to 0$ as $n \to \infty$. In other words, we have that $\widehat{\boldsymbol{\Phi}}_{\text{joint},\mathcal{M}}$ converges weakly to $\boldsymbol{\Phi}^{\star}_{\text{joint}}$ as $n \to \infty$. Then we must also have that $\widehat{\boldsymbol{\Phi}}_{\mathcal{M},n}$ converges weakly to the true distribution $\boldsymbol{\Phi}^{\star}_n$ as $n \to \infty$. As the function $g$ in (18) is bounded and continuous for the choice of parameter $\boldsymbol{\theta}$ in the compact parameter space $\boldsymbol{\Theta}$, we have that $\widehat{f}_{\mathcal{M}}(\mathbf{y}_{n+1} | \mathbf{y}_{1:n})$ converges almost surely to $f^{\star}(\mathbf{y}_{n+1} | \mathbf{y}_{1:n})$ as $n \to \infty$. □

## A3 Complementary numerical results

Recall the scenarios **S1** and **S2** for sparse covariance and precision matrices, respectively, along with sparse AR matrices. We use MLE* to refer to the MLE obtained by incorporating the knowledge of the true zero parameters of an MSVAR model, which we denote it by $\widetilde{\boldsymbol{\theta}}$ in the main paper. We set $\mathrm{M} = 2$, $p = 1$, 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 $(d, \mathrm{M}, p)$, the parameter vector $\boldsymbol{\theta}$ has dimensions $\mathrm{K} = \mathrm{M}\big(d + pd^2 + d(d+1)/2\big) + \mathrm{M}(\mathrm{M}-1) = 1262$, 4922 and 30302, respectively. The corresponding dimensions of the parameter vector of the true data-generating MSVAR models are 104, 202 and 504, respectively. The simulation result for each $d$ should be analyzed on its own, since the parameter configurations of the underlying models corresponding to $d = 20, 40, 100$, are different; see Section 5 of the main manuscript on "Simulation design".

**Sparsity scenario S1.** Figure 1 shows the relative estimation error (REE) and true positive rate (TPR) for $d = 20$. The results for dimensions $d = 40, 100$, are presented in the main manuscript. In Figure 1, we observe that in terms of the overall REE for the smallest sample size, SCAD and MCP perform better compared to the other two penalty functions. On the other hand, ADALASSO and MCP outperform when $n = 300, 400$, while also being comparable to the MLE* (median REE at most 1.2). In terms of TPR, the performance of the method based on all four penalties is reasonable.

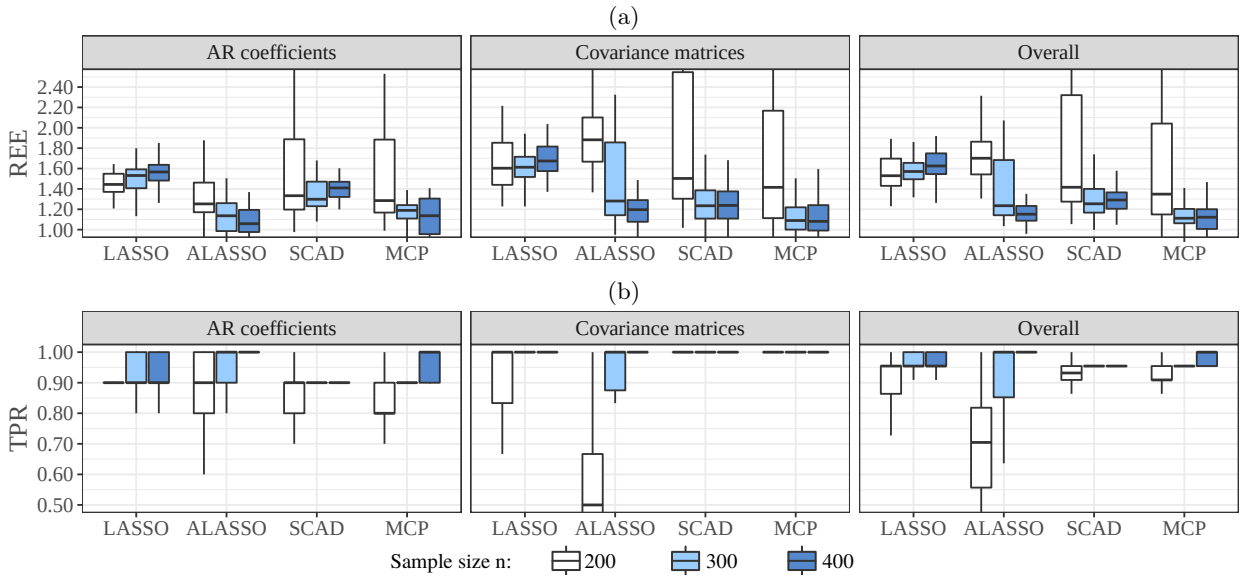


Figure 1: *(a) Relative estimation error (REE) and (b) true positive rate (TPR, nonzero parameter detection, bottom) based on 50 random samples for data dimension* $d = 20$*, parameter dimension* $\mathrm{K} = 1262$*, and sparsity scenario* ***S1****.*

For $d = 20$, we also investigate the performance of the MPLE for a wider range of sample sizes $n \in \{120, 200, 400, 800, 2000, 5000, 10000\}$. Figure 2 shows the estimation error (EE) and the TPR values. For the sample sizes $n = 120, 200$, the results show that the EE are relatively large and the TPR are relatively low. This is expected as these sample sizes are close to the number of parameters in the true data-generating MSVAR model, which is 104. On the other hand, the results show that as the sample size increases to 400 and beyond the standard deviation of the estimates decreases, which is expected as per our result in Theorem 1-(i) on the consistency of the MPLE. In addition, the TPR reaches the value 1.0 as $n$ increases, which confirms the sparsity recovery property of the

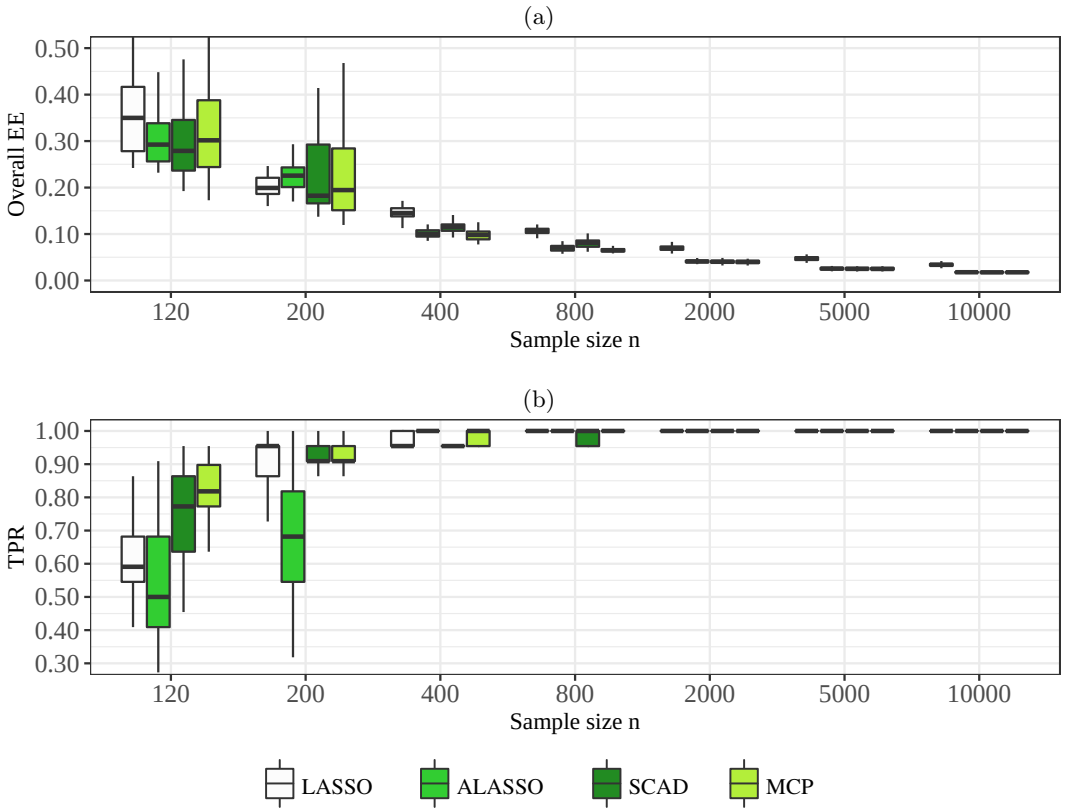


Figure 2: *(a) Overall estimation error (EE) and (b) true positive rate (TPR) for data dimension* $d = 20$*, parameter dimension* $\mathrm{K} = 1262$*, and sparsity scenario* ***S1****.*

MPLE in Theorem 1-(ii). We also performed simulations for M = 2, $p = 1$ and $d = 40$ (K = 4922) with a minimum sample size $n = 220$ (the number of parameters in the true data-generating MSVAR model is 202), and we observed the same behaviour of the MPLE as for $d = 20$.

**Sparsity scenario S2.** Relative estimation errors (REE) with respect to the MLE* for this scenario are presented in Figure 3 for $d = 20, 40, 100$. Different from scenario **S1**, here we observe that the MPLE attains a lower estimation error compared to the MLE* $\widetilde{\boldsymbol{\theta}}$, the reason being as follows. To obtain $\widetilde{\boldsymbol{\theta}}$ for the case of sparse precision matrices $\boldsymbol{\Omega}^{(m)}$, we first compute the regime-specific MLE of $\boldsymbol{\Sigma}^{(m)}$ and we then set the entries of the MLE of $(\boldsymbol{\Sigma}^{(m)})^{-1}$ to zero, corresponding to the zero entries of the true precision matrices. This estimation procedure does not directly use the knowledge of the true zero parameters in the precision matrices resulting in REE $< 1$, that is, a higher estimation error (EE) for $\widetilde{\boldsymbol{\theta}}$ compared to the proposed penalized estimators. From the results we can see that regarding the AR matrices, the ADALASSO outperforms the other penalties, followed by SCAD, MCP and LASSO. About the precision matrices and also the overall REE, the four penalties perform similar to each other.

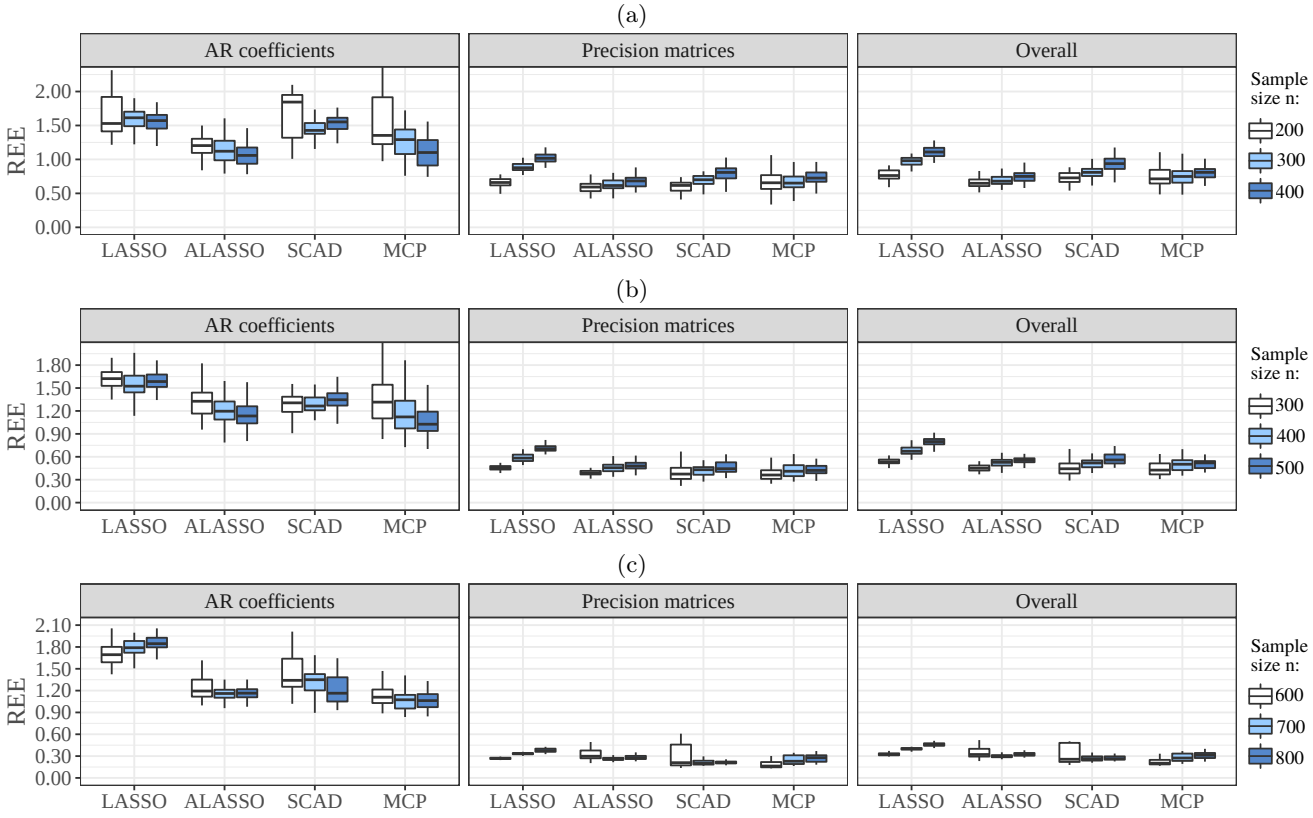


Figure 3: *Relative estimation error (REE) based on 50 random samples for sparsity scenario **S2**: (a)* $d = 20, \mathrm{K} = 1262$ *(b)* $d = 40, \mathrm{K} = 4922$ *(c)* $d = 100, \mathrm{K} = 30302$*, where* $d$ *and* K *are the data and parameter dimensions, respectively.*

Figure 4 shows the boxplots of the TPR. We can see that the method has a reasonable performance, above 0.85 overall. As in scenario **S1**, we observe mean true negative rate (TNR) above 0.90 for all cases (omitted).

Similar to scenario **S1**, for $d = 20$, we also investigate the performance of the MPLE for a wider range of sample sizes $n \in \{120, 200, 400, 800, 2000, 5000, 10000\}$, and for $d = 40$ with $n \in \{220, 300, 400, 800, 2000, 5000, 10000\}$. The performance of the method in terms of the overall EE

and TPR improves as the sample size increases (the results omitted), which is expected as per the result of Theorem 1.

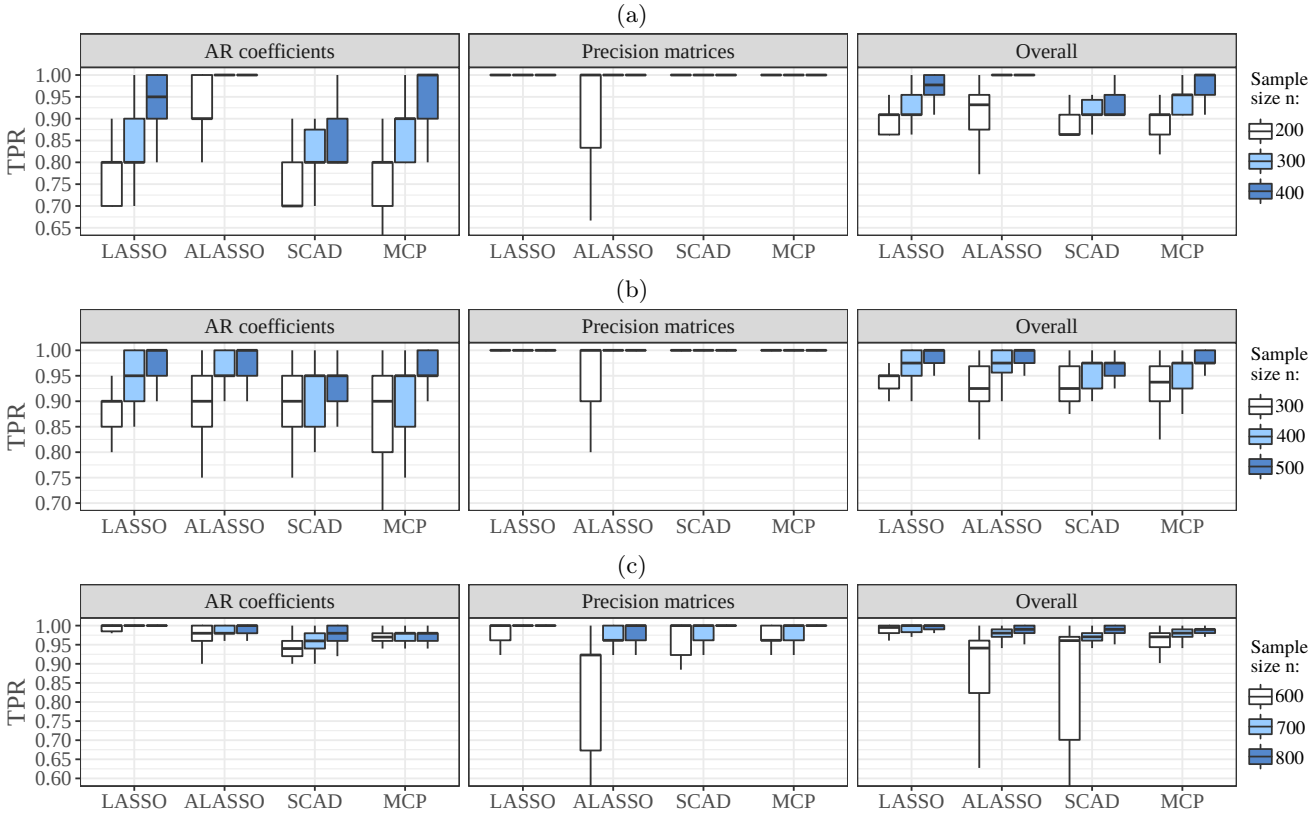


Figure 4: *True positive rate (TPR, nonzero parameter detection) based on 50 random samples for scenario* ***S2****: (a)* $d = 20, \mathrm{K} = 1262$ *(b)* $d = 40, \mathrm{K} = 4922$ *(c)* $d = 100, \mathrm{K} = 30302$*, where* $d$ *and* $\mathrm{K}$ *are the data and parameter dimensions, respectively.*

**Diagonal covariance matrices.** In our simulation experiments considered in the main manuscript, we considered non-structured sparse covariance matrices. Here, we present the results for an experiment for $d = 20$, M $= 2$, $p = 1$ and diagonal true covariance matrices $\mathbf{\Sigma}^{(m)}$. If we incorporate the knowledge of the true covariance matrices being diagonal, the number of parameters (dimension of $\boldsymbol{\theta}$) to be estimated by the MPLE reduces from K $= 1262$ to K $= 882$.

In Figure 5, corresponding to K $= 1262$, we observe that the median REE across all the four penalties is less than 1.6, with the ADALASSO and LASSO performing closer to the MLE* compared to the other two penalties, as also shown by the overall EE. In terms of TPR and TNR, we observe that the MPLE is able to recover both the the true nonzero and zero (off-diagonal) entries of the covariance matrices, with median rates above 0.80 and 0.90, respectively.

In Figure 6, corresponding to K $= 882$, we observe that, by incorporating the knowledge of the covariance matrices being diagonal, the medians of all the performance measures remain roughly the same as in Figure 5. The main differences are that (i) the standard deviations of the REE and EE are lower in Figure 6 for the smallest sample size, and (ii) the TNR corresponding to the covariance matrices is now 1.0. This shows that the performance of our method is at par with the performance of the estimator that uses the knowledge of the true sparsity structure. In simulation designs with covariance matrices being multiples of the identity matrix, as well as for $d = 40$, we observe the same phenomenon (results omitted).

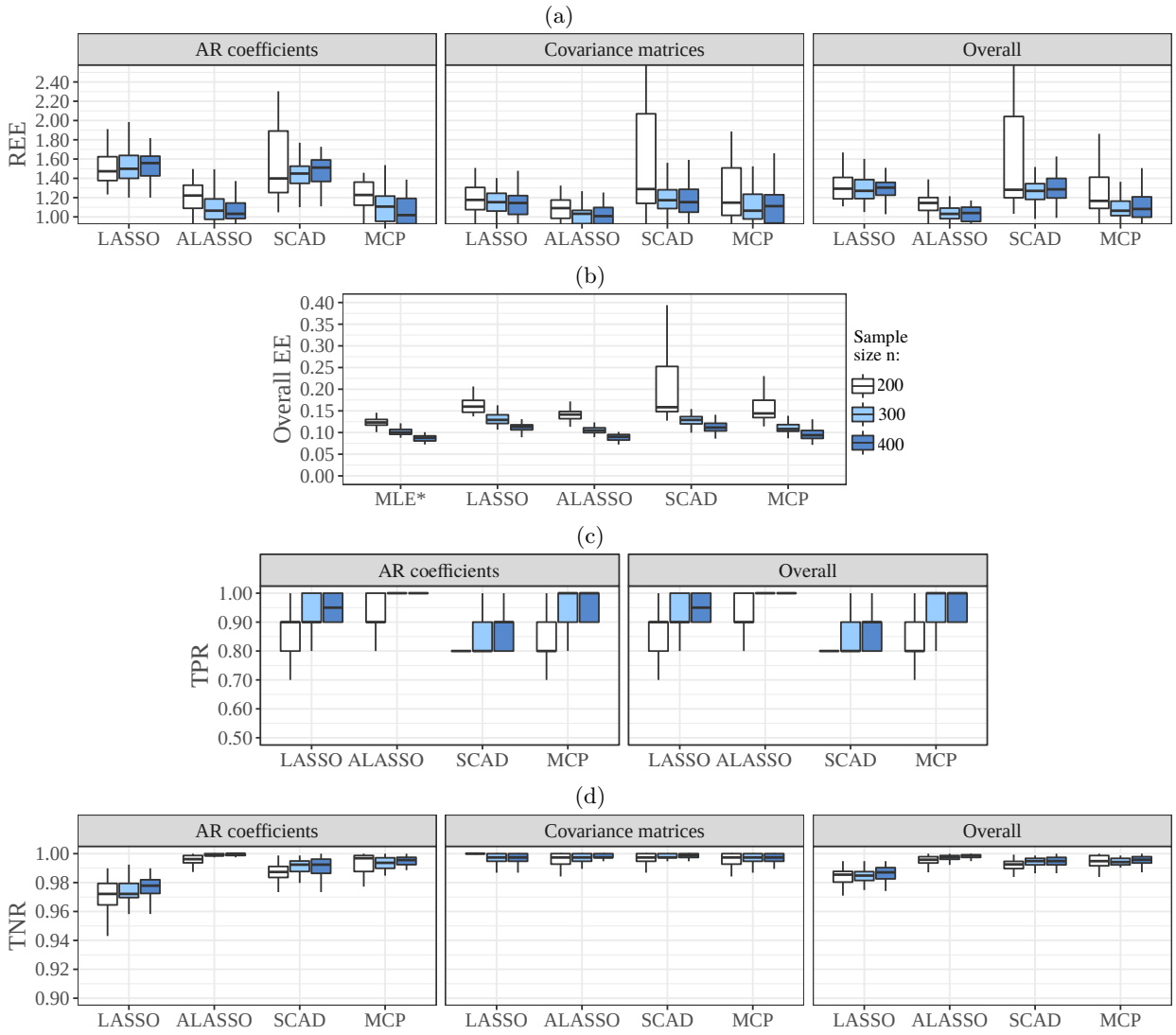


Figure 5: *(a) Relative estimation error (REE), (b) overall estimation error (EE), (c) true positive rate (TPR) and (d) true negative rate (TNR) for model with diagonal covariance matrices and data dimension* $d = 20$ *(parameter dimension* $\mathrm{K} = 1262$*). MLE* represents maximum likelihood estimation knowing the location of the zero parameters.*

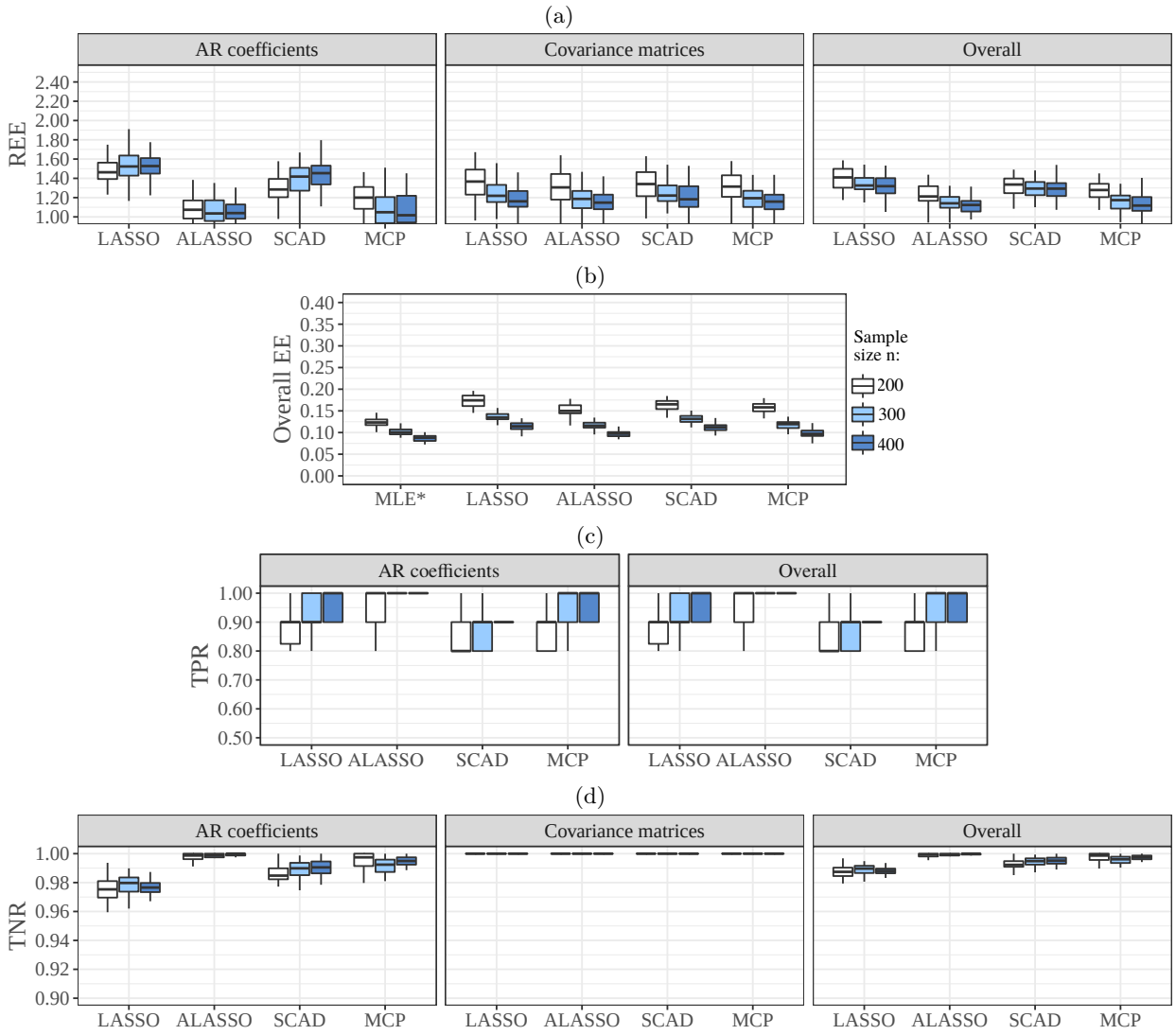


Figure 6: *(a) Relative estimation error (REE), (b) overall estimation error (EE), (c) true positive rate (TPR) and (d) true negative rate (TNR) for model with diagonal covariance matrices and data dimension* $d = 20$*, using the knowledge of the true covariance matrices being diagonal (parameter dimension* $\mathrm{K} = 882$*). MLE* represents maximum likelihood estimation knowing the location of the zero parameters.*

## A4 Model fitting results for the macroeconomic dataset case study

In the case study based on the Canadian macroeconomic dataset as presented in Section 6 of the main manuscript, we fit MSVAR models with $\mathrm{M} = 1, 2, 3, 4$, and $p = 1, 2, 3$, and for each penalty and sparsity scenario **S1** (sparse covariance matrices) and scenario **S2** (sparse precision matrices). We compute BIC and MDL to choose the final model. From Table 1 below, we conclude that: (i) scenario **S1** yields the lowest values of BIC in all cases, and of MDL in almost all cases, compared to **S2**; (ii) under **S1** and for any penalty, the number of regimes $\mathrm{M} = 3$ is consistently selected based on BIC and MDL; (iii) under **S1** and for $\mathrm{M} = 3$, the ADALASSO and MCP provide the lowest values of both BIC and MDL. The change in BIC and MDL when varying M from 3 to 4 is not significant. Moreover, the optimization problems associated to $\mathrm{M} \geq 4$ are obviously more sensitive to their initialization, and more prone to bad local optima.

We employ our proposed AR-order selection method based on the hierarchical group-LASSO penalized estimation (Section A1.6), and obtain $\widehat{p} = 1$ for the estimated $\widehat{\mathrm{M}} = 3$. Finally, regarding the choice of the penalty function, since we are also interested in the predictive performance of the model, we choose between ADALASSO and MCP by comparing the value of their predictive densities obtained with $\widehat{\mathrm{M}} = 3$ and $\widehat{p} = 1$, evaluated on out-of-sample observations from the last 6 months in the dataset. The ADALASSO penalty yields the highest value of the log-predictive density (-162.4) compared to LASSO (-180.2), SCAD (-218.4) and MCP (-220.5). Thus, using ADALASSO , we obtain the final model with $\widehat{\mathrm{M}} = 3$, $\widehat{p} = 1$.

| LASSO | | M = 1 | | M = 2 | | M = 3 | | M = 4 | |
|---|---|---|---|---|---|---|---|---|---|
| | | BIC | MDL | BIC | MDL | BIC | MDL | BIC | MDL |
| **S1** | $p = 1$ | 12224.3 | – | 11319.8 | 11219.4 | **11273.4** | **11096.9** | 11356.2 | 11170.5 |
| | $p = 2$ | 12095.8 | – | 11246.8 | 11158.9 | **11103.9** | **10959.7** | 11174.1 | 10960.8 |
| | $p = 3$ | 12049.2 | – | 11341.5 | 11257.3 | **11182.6** | **10650.2** | 11218.1 | 11034.4 |
| **S2** | $p = 1$ | 12285.7 | – | **11938.6** | 11824.6 | 12123.4 | **11744.9** | 12071.9 | 11796.1 |
| | $p = 2$ | 12161.2 | – | **11822.4** | **11697.3** | 12507.9 | 11968.1 | 12149.1 | 11766.7 |
| | $p = 3$ | 12119.9 | – | 12061.6 | 11992.6 | **11778.0** | **11632.3** | 11958.6 | 11691.6 |

| ALASSO | | M = 1 | | M = 2 | | M = 3 | | M = 4 | |
|---|---|---|---|---|---|---|---|---|---|
| | | BIC | MDL | BIC | MDL | BIC | MDL | BIC | MDL |
| **S1** | $p = 1$ | 12094.5 | – | 11051.7 | 10974.2 | **11009.9** | 10861.7 | 11095.1 | **10769.3** |
| | $p = 2$ | 11931.5 | – | 10912.0 | 10833.0 | **10893.1** | **10758.6** | 10957.2 | 10778.9 |
| | $p = 3$ | 11872.9 | – | **10937.1** | 10848.1 | 10976.1 | 10828.9 | 10945.6 | **10739.6** |
| **S2** | $p = 1$ | 12246.8 | – | 12617.7 | 12024.7 | 12125.2 | **11577.5** | **12090.3** | 11580.3 |
| | $p = 2$ | 12111.0 | – | **11970.6** | 11826.1 | 12311.5 | 11941.4 | 12618.7 | **11765.5** |
| | $p = 3$ | 12045.9 | – | 11622.8 | 11507.8 | **11551.3** | **11377.3** | 11787.7 | 11484.0 |

| SCAD | | M = 1 | | M = 2 | | M = 3 | | M = 4 | |
|---|---|---|---|---|---|---|---|---|---|
| | | BIC | MDL | BIC | MDL | BIC | MDL | BIC | MDL |
| **S1** | $p = 1$ | 12187.4 | – | 11137.5 | 11059.7 | **11105.5** | 10917.0 | 11115.4 | **10854.5** |
| | $p = 2$ | 12053.8 | – | 11043.6 | 10968.5 | **10943.6** | 10788.9 | 11014.4 | **10783.0** |
| | $p = 3$ | 12019.9 | – | 11186.0 | 11070.8 | **11051.9** | 10914.7 | 11110.8 | **10896.7** |
| **S2** | $p = 1$ | 12322.3 | – | 12676.2 | 12305.7 | **11761.1** | **11539.9** | 12598.5 | 11710.1 |
| | $p = 2$ | 12182.4 | – | **11948.9** | 11817.0 | 12169.0 | **11731.6** | 11997.9 | 11845.1 |
| | $p = 3$ | **12118.2** | – | 12671.9 | 12306.7 | 12394.5 | **10049.0** | 12862.4 | 12069.9 |

| MCP | | M = 1 | | M = 2 | | M = 3 | | M = 4 | |
|---|---|---|---|---|---|---|---|---|---|
| | | BIC | MDL | BIC | MDL | BIC | MDL | BIC | MDL |
| **S1** | $p = 1$ | 12074.3 | – | 11016.8 | 10953.2 | **10987.6** | **10754.9** | 11387.9 | 10976.8 |
| | $p = 2$ | 11897.1 | – | 10885.6 | 10809.5 | **10840.4** | 10644.4 | 10917.7 | **10551.2** |
| | $p = 3$ | 11863.4 | – | 10905.2 | 10841.1 | **10757.4** | **10600.7** | 10982.2 | 10630.4 |
| **S2** | $p = 1$ | 12239.9 | – | 12166.7 | 11903.8 | **11864.8** | **11486.9** | 12409.5 | 11759.5 |
| | $p = 2$ | 12063.3 | – | **11927.1** | **11691.2** | 12205.6 | 11746.1 | 12603.9 | 12006.7 |
| | $p = 3$ | **12036.0** | – | 12572.1 | **10385.3** | 12412.1 | 11786.2 | 13314.0 | 11959.3 |

Table 1: *BIC and MDL values based on the macroeconomic dataset of Section 6 of the main manuscript, for* $M = 1, 2, 3, 4$*,* $p = 1, 2, 3$*, sparsity scenarios* ***S1*** *and* ***S2****, and all the penalties considered. Minima for fixed p and criterion are highlighted in boldface.*

# References


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

Beck, A. and Teboulle, M. (2009a). Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434.

Beck, A. and Teboulle, M. (2009b). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

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

Douc, R. and Moulines, E. (2012). Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *The Annals of Statistics*, 40(5):2697–2732.

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.

Guo, K., Yuan, X., and Zeng, S. (2016). Convergence analysis of ISTA and FISTA for "strongly + semi" convex programming. Unpublished manuscript.

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

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616.

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.

Parikh, N. and Boyd, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3).

Song, S. and Bickel, P. J. (2011). Large vector auto regressions. *arXiv:1106.3915 [stat.ML]*.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12):5277–5286.

Wen, B., Chen, X., and Pong, T. K. (2017). Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM Journal on Optimization*, 27(1):124–145.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323.