Supplementary Material for: Sparseness, consistency and model selection for Markov regime-switching Gaussian autoregressive models

Abbas Khalili and David A. Stephens
Department of Mathematics and Statistics
McGill University, Montreal, Canada

September 14, 2020

1 PREAMBLES

For a stationary time series Y_t, its population lag-l partial autocorrelation function (PACT) is defined as

$$\pi_Y(l) = \text{Corr}\left(Y_t, Y_{t+l}\middle| Y_{t+1:t+l-1} = y_{t+1:t+l-1}\right) \quad , \quad l \geq 2$$

such that $\pi_Y(0) = 1$ and $\pi_Y(1) = \text{Corr}(Y_t, Y_{t+1})$. We now state a result that shows properties of the partial autocorrelation function (PACT) of a stationary MSAR model.

Proposition 1 *Suppose the true* MSAR *model of Y_t is stationary, and its maximal* AR*-order is $q^* = \max_{1\leq j\leq K} q_j$. Then, $\pi_Y(l) = 0$, for any $l \geq q^* + 1$.*

Proof. For lag $l = q^* + 1$, we have that

$$\text{Cov}(Y_t, Y_{t+q^*+1}|y_{t+1:t+q^*}) = \sum_{j=1}^{K} \text{Cov}(Y_t, Y_{t+q^*+1}|S_{t+q^*+1} = j, y_{t+1:t+q^*})P(S_{t+q^*+1} = j|y_{t+1:t+q^*}).$$

By the MSAR model assumptions, the conditional distribution of $(Y_{t+q^*+1}|S_{t+q^*+1} = j, y_{t+1:t+q^*})$ is Gaussian with variance σ_j^2, and mean $\mu_{t+q^*+1,j} = \theta_{j0}+\theta_{j1}y_{t+q^*}+\ldots+\theta_{jq^*}y_{t+1}$ which clearly does not depend on y_t. The latter is because the maximal AR-order across the AR-regimes in the MSAR model is q^*. This implies that conditioning on $(S_{t+q^*+1}, Y_{t+1:t+q^*})$, the variables Y_t and Y_{t+q^*+1} are independent. Thus, all the conditional covariances in the above sum are zero, which results in $\text{Cov}(Y_t, Y_{t+q^*+1}|y_{t+1:t+q^*}) = 0$, or equivalently $\pi_Y(q^*+1) = 0$. Similarly, we have that $\pi_Y(l) = 0$, for any $l \geq q^*+2$. This completes the proof. ♠

The rest of this section is a preparation for the proof of our main results. Recall from Section 2 of the manuscript, the conditional log-likelihood of a MSAR model with K regimes and a maximum AR-order q is given by

$$\ell_n(\mathbf{\Phi}_K; s_q) = \log\{f_3(y_{q+1:n}|y_{1:q}, s_q, \mathbf{\Phi}_K)\}, \tag{A.1}$$

where $\mathbf{\Phi}_K = (\nu_1, \ldots, \nu_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \mathbb{P} = \{\alpha_{ij}\})$, and $\boldsymbol{\theta}_j = (\theta_{j0}, \theta_{j1}, \ldots, \theta_{jq})^\top$ is the coefficient vector of the k-th AR-process. Since in Theorems 1 and 2 we regard K as known, for simplicity in notation we drop the subscript K from the vector of parameters. In Theorem 3, we show that the RBIC does not under-estimate K when is unknown, and we also show that the conditional h-step ahead predictive density can be estimated consistently when the number of regimes is estimated by the RBIC.

Before we embark on the proofs, we state Lemma 1 which shows the large sample properties of the (conditional) score and observed information matrix, respectively,

$$\ell'_n(\mathbf{\Phi}; s_q) = \frac{\partial \ell_n(\mathbf{\Phi}; s_q)}{\partial \mathbf{\Phi}} \quad , \quad \ell''_n(\mathbf{\Phi}; s_q) = \frac{\partial^2 \ell_n(\mathbf{\Phi}; s_q)}{\partial \mathbf{\Phi} \partial \mathbf{\Phi}^\top}$$

for all $s_q \in \{1, \ldots, K\}$. The result is due to Douc et al. (2004) and we omit its proof.

Lemma 1 *Suppose the true* MSAR *model of* Y_t *is strictly stationary, ergodic and* $E|Y_t|^{(4+2\delta)} < \Delta < \infty$, *for some* $\delta > 0$. *Then, for all* $s_q \in \{1, \ldots, K\}$,

(i) *For any* $\mathbf{\Phi} \in \mathbf{\Theta}$,

$$n^{-1}E\left\{[\ell'_n(\mathbf{\Phi}; s_q)][\ell'_n(\mathbf{\Phi}; s_q)]^\top \middle| Y_{1:q}, s_q\right\} = -n^{-1}E\left\{\ell''_n(\mathbf{\Phi}; s_q)\middle| Y_{1:q}, s_q\right\},$$

and, at $\mathbf{\Phi} = \mathbf{\Phi}^*$, both sides converge to the Fisher information matrix $\mathbf{I}(\mathbf{\Phi}^*)$, in probability, as $n \to \infty$.

(ii) *For any possibly random sequence $\widetilde{\mathbf{\Phi}}_n$ in $\mathbf{\Theta}$ such that $\widetilde{\mathbf{\Phi}}_n \xrightarrow{p} \mathbf{\Phi}^*$, as $n \to \infty$,*

$$-\frac{1}{n}\ell''_n(\widetilde{\mathbf{\Phi}}_n; s_q) \xrightarrow{p} \mathbf{I}(\mathbf{\Phi}^*).$$

(iii) *Asymptotic normality: $\frac{1}{\sqrt{n}}\ell'_n(\mathbf{\Phi}^*; s_q) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}(\mathbf{\Phi}^*))$, as $n \to \infty$.*

Regularity conditions on the penalty r_n and tuning parameter λ_n:

We now state the regularity conditions on the tuning parameter λ_n and the penalty function r_n. For each $1 \leq j \leq K$, denote

$$\Im_j^* = \{1 \leq l \leq q : \theta_{jl}^* \neq 0\}$$

as the set of indices corresponding to the true non-zero AR-coefficients of the j-th regime of the MSAR model. The regularity conditions are as follow.

$\mathbf{C}_1$. For all n and λ, the penalty function $r_n(\theta; \lambda)$ is symmetric, nonnegative, nondecreasing and it has first derivative for all $\theta \in (0, \infty)$. The function is also continuously twice differentiable for all $\theta \in (c\lambda, \infty)$, and some constant $c > 0$. In addition, $r_n(0; \lambda) = 0$.

$\mathbf{C}_2$. As $n \to \infty$, we have $\lambda_n = o(1)$ such that $\min_{l \in \Im_k^*} |\theta_{jl}^*|/\lambda_n \longrightarrow \infty$,, and we have $a_n = o(n^{1/2})$, $b_n = o(1)$, where

$$a_n = \max_{j,l}\{|r'_n(\theta_{jl}^*; \lambda_n)|/\sqrt{n};\ l \in \Im_j^*\} \quad , \quad b_n = \max_{j,l}\{|r''_n(\theta_{jl}^*; \lambda_n)|/n;\ l \in \Im_j^*\}.$$

The $r'_n(\theta_{jl}^*; \lambda)$ and $r''_n(\theta_{jl}^*; \lambda)$ are the first and second derivatives of $r_n(\theta; \lambda)$ at $\theta_{jl}^* \neq 0$.

$\mathbf{C}_3$. $\lim_{n\to\infty} \inf\{r'_n(\theta; \lambda_n) : 0 < \theta \leq n^{-1/2}\log n)\}/\sqrt{n} = +\infty$.

$\mathbf{C}_1$ is a smoothness condition on the penalty which facilitates obtaining estimators by differentiating the objective function and for studying the asymptotic properties of the estimators of the true non-zero AR-coefficients. $\mathbf{C}_2$ is required to obtain $\sqrt{n}$-consistent estimators of the true non-zero coefficients, while $\mathbf{C}_3$ is required for consistency in AR-order selection or sparsity.

Examples of r_n: Let $\omega_{jl} > 0$ be some pre-specified (possibly random) weights.

– LASSO: $r_n(\theta_{jl}; \lambda) = (n-q)\lambda|\theta_{jl}|$.

– adaptive LASSO (ADALASSO): $r_n(\theta_{jl}; \lambda) = (n-q)\lambda\, \omega_{jl}|\theta_{jl}|$.

– SCAD: it is most often characterized by its first derivative,

$$r_n'(\theta_{jl}; \lambda) = (n-q)\left\{\lambda I(|\theta_{jl}| \le \lambda) + \frac{(a\lambda - |\theta_{jl}|)_+}{(a-1)}\, I(|\theta_{jl}| > \lambda)\right\} \times \operatorname{sgn}(\theta_{jl})$$

for some constant $a > 2$, where $I(\cdot)$ and $\operatorname{sgn}(\cdot)$ are the indicator and sign functions, and $(\cdot)_+$ is the positive part of the input, respectively.

The common choice of the weights in ADALASSO are $\omega_{jl} = |\tilde{\theta}_{jl}|^{-\gamma}$, for some $\gamma > 0$, where $\tilde{\theta}_{jl}$ is the (conditional) MLE of θ_{jl}. Park and Sakaori (2013) proposed the weights $\omega_{jl}(\alpha) = |\tilde{\theta}_{jl}\ \alpha(1-\alpha)^l|^{-\gamma}$, for some $0 < \alpha < 1$, where the lag effect l is also taken into consideration: the larger the lag l, the heavier the penalty on the θ_{jl}; these weights were designed for estimation and AR-order selection in AR models, under the assumptions of no seasonality effects and that generally the $|\theta_{jl}|$ decline with increasing lag length l.

2 PROOFS

Proof of Theorem 1:

Let $\eta_n = n^{-1/2}(1 + a_n) = o(1)$. For any $s_q \in \{1, \ldots, K\}$, it suffices to show that for any given $\epsilon > 0$, there exists a constant $\mathcal{C}$ such that

$$\lim_{n\to\infty} P\{\sup_{\|\boldsymbol{u}\|_2 = \mathcal{C}} \mathcal{L}_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{u}; s_q, \lambda_n) < \mathcal{L}_n(\boldsymbol{\Phi}^*; s_q, \lambda_n)\} \ge 1 - \epsilon. \tag{A.2}$$

To prove (A.2), we verify that $D_n(\boldsymbol{u}; s_q, \lambda_n) = \mathcal{L}_n(\boldsymbol{\Phi}^* + \eta_n\boldsymbol{u}; s_q, \lambda_n) - \mathcal{L}_n(\boldsymbol{\Phi}^*; s_q, \lambda_n) < 0$ uniformly in $\boldsymbol{u}$ with probability approaching one, as $n \to \infty$. Recall the partitioning

$\mathbf{\Phi}^* = (\mathbf{\Phi}_1^*, \mathbf{0})$ given in Section 6 of the manuscript. Since $r_n(0; \lambda) = 0$, we write

$$\begin{aligned} D_n(\boldsymbol{u}; s_q, \lambda_n) &= \{\ell_n(\mathbf{\Phi}^* + \eta_n \boldsymbol{u}; s_q) - \ell_n(\mathbf{\Phi}^*; s_q)\} - \sum_{j=1}^{K} \{p_n(\tilde{\nu}_j^*) - p_n(\nu_j^*)\} \\ &- \{\mathcal{R}_n(\mathbf{\Phi}^* + \eta_n \boldsymbol{u}; \lambda_n) - \mathcal{R}_n(\mathbf{\Phi}^*; \lambda_n)\} \\ &\leq \{\ell_n(\mathbf{\Phi}^* + \eta_n \boldsymbol{u}; s_q) - \ell_n(\mathbf{\Phi}^*; s_q)\} - \sum_{j=1}^{K} \{p_n(\tilde{\nu}_j^*) - p_n(\nu_j^*)\} \\ &- \{\mathcal{R}_n(\mathbf{\Phi}_1^* + \eta_n \boldsymbol{u}_1; \lambda_n) - \mathcal{R}_n(\mathbf{\Phi}_1^*; \lambda_n)\} \\ &= D_{1n}(\boldsymbol{u}; s_q) - D_{2n}(\boldsymbol{u}) - D_{3n}(\boldsymbol{u}; \lambda_n), \end{aligned}$$

where $\boldsymbol{u}_1$ is a sub-vector of $\boldsymbol{u}$, and $p_n(\tilde{\nu}_j^*)$ is used for the penalty at a point shifted from ν_j^*. Its exact value does not matter. We now proceed to evaluate the orders of three differences $D_{1n}(\boldsymbol{u}; s_q)$, $D_{2n}(\boldsymbol{u})$ and $D_{3n}(\boldsymbol{u}; \lambda_n)$. By Taylor's expansion, we have

$$D_{1n}(\boldsymbol{u}; s_q) = \eta_n \ \boldsymbol{u}^\top \ell_n'(\mathbf{\Phi}^*; s_q) + (\eta_n^2/2) \boldsymbol{u}^\top \ell_n''(\mathbf{\Phi}_n^*; s_q) \ \boldsymbol{u} = E_{1n} + E_{2n}$$

where $\mathbf{\Phi}_n^*$ lies between $\mathbf{\Phi}^*$ and $\mathbf{\Phi}^* + \eta_n \boldsymbol{u}$. By Lemma 1, $\ell_n'(\mathbf{\Phi}^*; s_q) = O_p(\sqrt{n})$. Thus, for large n, we have that uniformly in $\boldsymbol{u}$,

$$E_{1n} = O_p(\sqrt{n}\eta_n).$$

On the other hand, from $\eta_n = o(1)$, we get that $\mathbf{\Phi}_n^* \xrightarrow{p} \mathbf{\Phi}^*$, and by Lemma 1, uniformly in $\boldsymbol{u}$, $n^{-1}\ell_n''(\mathbf{\Phi}_n^*; s_q) - \mathbf{I}(\mathbf{\Phi}^*) = o_p(1)$ which implies that, for large n,

$$E_{2n} = -(n\eta_n^2/2) \ \boldsymbol{u}^\top \mathbf{I}(\mathbf{\Phi}^*) \boldsymbol{u} \ \{1 + o_p(1)\}.$$

Combining the two assessments, for all $s_q \in \{1, \ldots, K\}$ and for large n we have

$$D_{1n}(\boldsymbol{u}; s_q) = O_p(\sqrt{n}\eta_n) - (n\eta_n^2/2) \ \boldsymbol{u}^\top \mathbf{I}(\mathbf{\Phi}^*) \boldsymbol{u} \ \{1 + o_p(1)\}.$$

Next, we check the order of $D_{2n}(\boldsymbol{u})$. Under the assumption of stationary, ergodic and $E|Y_t|^{(4+2\delta)} < \Delta < \infty$, for some $\delta > 0$, it is clear that $\mathcal{V}_n^2 \xrightarrow{p} \mathrm{Var}(Y_t)$, as $n \to \infty$. Thus, in a $o_p(1)$ neighbourhood of ν_j^*, we have $p_n(\nu) = O_p(n^{-1/2})$ uniformly in $\boldsymbol{u}$; see the definition of this penalty in equation (3.2) of the manuscript. This implies that $D_{2n}(\boldsymbol{u}) = o_p(n^{-1/2})$. In other words, this part of the penalty has minimal effect when $\mathbf{\Phi}$ is near $\mathbf{\Phi}^*$.

We now focus on $D_{3n}(\boldsymbol{u};\lambda_n)$. First, we have

$$|D_{3n}(\boldsymbol{u};\lambda_n)| = |\mathcal{R}_n(\boldsymbol{\Phi}_1^*+\eta_n\boldsymbol{u}_1;\lambda_n)-\mathcal{R}_n(\boldsymbol{\Phi}_1^*;\lambda_n)| \leq \sum_{j=1}^{K}\sum_{l\in\Im_j^*}|r_n(\theta_{jl}^*+\eta_n u_{jl};\lambda_n)-r_n(\theta_{jl}^*;\lambda_n)|$$

Since $\lambda_n = o(1)$ and $\theta_{jl}^* \neq 0$ for all $l \in \Im_j^*$, we have that for large n, $\theta_{jl}^* \in (c\lambda_n, \infty)$, for some $c > 0$. Also, we have $\eta_n = o(1)$. Thus, by conditions $\mathbf{C}_1$ and $\mathbf{C}_2$, the penalty $r_n(\theta;\lambda_n)$ accepts its second-order continuous derivatives at $\theta = \theta_{jl}^*$ and any point between θ_{jl}^* and $\theta_{jl}^* + \eta_n u_{jl}$. Thus, by a second order Taylor's expansion of $r_n(\theta;\lambda_n)$,

$$\begin{aligned}
|D_{3n}(\boldsymbol{u};\lambda_n)| &\leq \sum_{j=1}^{K}\sum_{l\in\Im_j^*}|[\eta_n r_n'(\theta_{jl}^*;\lambda_n)u_{jl} + \frac{\eta_n^2}{2}r_n''(\theta_{jl}^*;\lambda_n)u_{jl}^2(1+o(1))]| \\
&\leq \tilde{q}\,\sqrt{n}\eta_n a_n\|\boldsymbol{u}\|_2 + n\eta_n^2\,\frac{b_n}{2}\|\boldsymbol{u}\|_2^2
\end{aligned}$$

where $\tilde{q} = \max_{1\leq j\leq K}|\Im_j^*|$ is the maximum cardinality of the index sets $\Im_j^*$. Since $b_n = o(1)$, by the order assessment of $D_n(\boldsymbol{u};s_q,\lambda_n) = D_{1n}(\boldsymbol{u};s_q) - D_{2n}(\boldsymbol{u}) - D_{3n}(\boldsymbol{u};\lambda_n)$, we can see that, for large n, the negative quantity

$$-\frac{1}{2}n\eta_n^2\;\boldsymbol{u}^\top\mathbf{I}(\boldsymbol{\Phi}^*)\boldsymbol{u}\;[1+o_p(1)] = -\frac{1}{2}(1+a_n)^2\;\boldsymbol{u}^\top\mathbf{I}(\boldsymbol{\Phi}^*)\boldsymbol{u}\;[1+o_p(1)] < 0$$

is the dominant term in $D_n(\boldsymbol{u};s_q,\lambda_n)$, for large n. Therefore, (A.2) holds and this completes the proof. ♠

Proof of Theorem 2:

The result of Lemma 2 below will be used to prove the AR-order estimation consistency property of the regularization method, claimed in Part (i) of Theorem 2. As described in the beginning of Section 6 of the manuscript, recall the partitioning of an arbitrary parameter vector $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) \in \boldsymbol{\Theta}$, where $\dim(\boldsymbol{\Phi}) = d, \dim(\boldsymbol{\Phi}_1) = \dim(\boldsymbol{\Phi}_1^*) = d_1 < d$.

By Theorem 1, if $a_n = O(1)$, for any $s_q \in \{1,\ldots,K\}$ there exists a local maximizer $\widehat{\boldsymbol{\Phi}}_{n,s_q}$ of the penalized conditional log-likelihood $\mathcal{L}_n(\boldsymbol{\Phi};s_q,\lambda_n)$ such that $\|\widehat{\boldsymbol{\Phi}}_{n,s_q} - \boldsymbol{\Phi}^*\|_2 = O_p(n^{-1/2})$, as $n \to \infty$.

Lemma 2 *Assume the conditions of Theorem 1 and that λ_n and the penalty r_n in addition satisfy Condition $\mathbf{C}_3$, and $a_n = O(1)$. Then, for any $\mathbf{\Phi}$ such that $\|\mathbf{\Phi} - \mathbf{\Phi}^*\|_2 = O(n^{-1/2})$ and with the above partitioning, and for any $s_q \in \{1, \ldots, K\}$, as $n \to \infty$, with probability tending to one,*

$$\mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q, \lambda_n) - \mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{0}); s_q, \lambda_n) < 0.$$

Proof of Lemma 2. Note that the two partitioning $(\mathbf{\Phi}_1, \mathbf{\Phi}_2)$ and $(\mathbf{\Phi}_1, \mathbf{0})$ of $\mathbf{\Phi}$ share the same regime specific variance values. Thus, $\mathcal{L}_n(\mathbf{\Phi}; s_q, \lambda_n)$ and $\tilde{\ell}_n(\mathbf{\Phi}; s_q)$ have equal-sized adjustment in terms of $p_n(\nu_k)$. Therefore, we have

$$\begin{aligned} \mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q, \lambda_n) - \mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{0}); s_q, \lambda_n) &= \{\ell_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q) - \ell_n((\mathbf{\Phi}_1, \mathbf{0}); s_q)\} - \\ &\quad \{\mathcal{R}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); \lambda_n) - \mathcal{R}_n((\mathbf{\Phi}_1, \mathbf{0}); \lambda_n)\}. \end{aligned} \tag{A.3}$$

We now provide order assessments of the two differences in (A.3).

By a second-order Taylor's expansion,

$$\begin{aligned} \ell_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q) &= \ell_n(\mathbf{\Phi}^*; s_q) + \{\ell_n'(\mathbf{\Phi}^*; s_q)\}^\top(\mathbf{\Phi} - \mathbf{\Phi}^*) + (\mathbf{\Phi} - \mathbf{\Phi}^*)^\top[\ell_n''(\mathbf{\Phi}_n^*; s_q)](\mathbf{\Phi} - \mathbf{\Phi}^*) \\ \ell_n((\mathbf{\Phi}_1, \mathbf{0}); s_q) &= \ell_n((\mathbf{\Phi}_1^*, \mathbf{0}); s_q) + \{\ell_n'((\mathbf{\Phi}_1^*, \mathbf{0}); s_q)\}^\top(\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*) \\ &\quad +(\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*)^\top[\ell_n''((\mathbf{\Phi}_{n1}^*, \mathbf{0}); s_q)](\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*). \end{aligned}$$

where $\mathbf{\Phi}_n^*$ lies between $\mathbf{\Phi}$ and $\mathbf{\Phi}^*$; and $\mathbf{\Phi}_{n1}^*$ lies between $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_1^*$. Furthermore, because $\mathbf{\Phi}_2^* = \mathbf{0}$, we have $\ell_n(\mathbf{\Phi}^*; s_q) = \ell_n((\mathbf{\Phi}_1^*, \mathbf{0}); s_q)$, and also the partitioning

$$\{\ell_n'(\mathbf{\Phi}^*; s_q)\}^\top(\mathbf{\Phi} - \mathbf{\Phi}^*) = \{\partial\ell_n((\mathbf{\Phi}_1^*, \mathbf{0}); s_q)/\partial\mathbf{\Phi}_1\}^\top(\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*) + \{\partial\ell_n((\mathbf{\Phi}_1^*, \mathbf{0}); s_q)/\partial\mathbf{\Phi}_2\}^\top\mathbf{\Phi}_2.$$

Also, since $\mathbf{\Phi}_2^* = \mathbf{0}$, by Lemma 1 we have $\partial\ell_n((\mathbf{\Phi}_1^*, \mathbf{0}); s_q)/\partial\mathbf{\Phi}_2 = O_p(\sqrt{n})$. Thus,

$$\{\ell_n'(\mathbf{\Phi}^*; s_q)\}^\top(\mathbf{\Phi} - \mathbf{\Phi}^*) = \{\partial\ell_n((\mathbf{\Phi}_1^*, \mathbf{0}); s_q)/\partial\mathbf{\Phi}_1\}^\top(\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*) + O_p(\sqrt{n})|\mathbf{\Phi}_2|.$$

Therefore,

$$\begin{aligned} \ell_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q) - \ell_n((\mathbf{\Phi}_1, \mathbf{0}); s_q) &= O_p(\sqrt{n})|\mathbf{\Phi}_2| + n(\mathbf{\Phi} - \mathbf{\Phi}^*)^\top[\ell_n''(\mathbf{\Phi}_n^*; s_q)/n](\mathbf{\Phi} - \mathbf{\Phi}^*) \\ &- n(\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*)^\top[\ell_n''((\mathbf{\Phi}_{n1}^*, \mathbf{0}); s_q)/n](\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*) \end{aligned}$$

By Lemma 1, since $\mathbf{\Phi}_n^* \xrightarrow{p} \mathbf{\Phi}^*$ and $\mathbf{\Phi}_{n1}^* \xrightarrow{p} \mathbf{\Phi}_1^*$, as $n \to \infty$, for all $s_q \in \{1, 2, \ldots, K\}$,

$$n^{-1}\ell_n''(\mathbf{\Phi}_n^*; s_q) \xrightarrow{p} -\mathbf{I}(\mathbf{\Phi}^*) \quad , \quad n^{-1}\ell_n''((\mathbf{\Phi}_{n1}^*, \mathbf{0}); s_q) \xrightarrow{p} -\mathbf{I}_{11}(\mathbf{\Phi}_1^*).$$

Note that $\mathbf{I}_{11}(\mathbf{\Phi}_1^*)$ is a sub-matrix of $\mathbf{I}(\mathbf{\Phi}^*)$. By condition $E|Y_t|^{(4+2\delta)} < \Delta < \infty$, the two matrices in the above two limits are finite. Thus, for large n, we have

$$\begin{aligned}\ell_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q) - \ell_n((\mathbf{\Phi}_1, \mathbf{0}); s_q) &= O_p(\sqrt{n})|\mathbf{\Phi}_2| - n(\mathbf{\Phi} - \mathbf{\Phi}^*)^\top[\mathbf{I}(\mathbf{\Phi}^*)](\mathbf{\Phi} - \mathbf{\Phi}^*)(1 + o_p(1)) \\ &+ n(\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*)^\top[\mathbf{I}_{11}(\mathbf{\Phi}^*)](\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*)(1 + o_p(1)).\end{aligned}$$

By partitioning the matrix $\mathbf{I}(\mathbf{\Phi}^*)$ into four sub-matrices, one of which is $\mathbf{I}_{11}(\mathbf{\Phi}^*)$, we have

$$\begin{aligned}\ell_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q) - \ell_n((\mathbf{\Phi}_1, \mathbf{0}); s_q) &= O_p(\sqrt{n})|\mathbf{\Phi}_2| - n\{2(\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*)^\top[\mathbf{I}_{12}(\mathbf{\Phi}^*)]\mathbf{\Phi}_2 \\ &\quad +\mathbf{\Phi}_2^\top[\mathbf{I}_{22}(\mathbf{\Phi}^*)]\mathbf{\Phi}_2\}(1 + o_p(1))\end{aligned}$$

where $\mathbf{I}_{12}(\mathbf{\Phi}^*)$ and $\mathbf{I}_{22}(\mathbf{\Phi}^*)$ are sub-matrices of $\mathbf{I}(\mathbf{\Phi}^*)$. Since $\|\mathbf{\Phi}_1 - \mathbf{\Phi}_1^*\| = O(n^{-1/2})$ and $\|\mathbf{\Phi}_2\| = O(n^{-1/2})$, thus

$$\ell_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q) - \ell_n((\mathbf{\Phi}_1, \mathbf{0}); s_q) = O_p(\sqrt{n})|\mathbf{\Phi}_2| + n\{O_p(n^{-1/2})|\mathbf{\Phi}_2| + O_p(n^{-1/2})|\mathbf{\Phi}_2|\}(1 + o_p(1))$$

which implies that, for large n,

$$\ell_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q) - \ell_n((\mathbf{\Phi}_1, \mathbf{0}); s_q) = O_p(\sqrt{n})|\mathbf{\Phi}_2| = O_p(\sqrt{n})\sum_{j=1}^{K}\sum_{l \notin \Im_j^*}|\theta_{jl}|.$$

We now study the second difference in (A.3). Since $r_n(0; \lambda_n) = 0$, for any λ_n, we have

$$\mathcal{R}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); \lambda_n) - \mathcal{R}_n((\mathbf{\Phi}_1, \mathbf{0}); \lambda_n) = \sum_{j=1}^{K}\sum_{l \notin \Im_j^*} r_n(\theta_{jl}; \lambda_n).$$

Combining the assessments of the two differences in (A.3), for all $s_q \in \{1, \ldots, K\}$ we arrive at

$$\mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q, \lambda_n) - \mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{0}); s_q, \lambda_n) = \sum_{j=1}^{K}\sum_{l \notin \Im_j^*}\Big\{O_p(\sqrt{n})|\theta_{jl}| - r_n(\theta_{jl}; \lambda_n)\Big\}.$$

Because $\mathbf{\Phi}$ is within $n^{-1/2}$-neighbourhood of $\mathbf{\Phi}^*$, we must have $\theta_{jl} = O(n^{-1/2})$ for all $l \notin \Im_j^*$ (the set of non-zero coefficients). Once θ is in this range, condition $\mathbf{C}_3$ on the penalty r_n is applicable which leads to

$$O_p(\sqrt{n})|\theta_{jl}| - r_n(\theta_{jl}; \lambda_n) < 0$$

for large n. Therefore the conclusion of this lemma: for all s_q, as $n \to \infty$,

$$\mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q, \lambda_n) - \mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{0}); s_q, \lambda_n) < 0$$

with probability tending to one. This completes the proof of Lemma 2. ♠

We now return to the proof of Theorem 2.

Proof of Theorem 2:

Part (i). Let $(\widehat{\mathbf{\Phi}}_{n,s_q,1}, \mathbf{0})$ be the maximizer of the penalized conditional log-likelihood $\mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{0}); s_q, \lambda_n)$, for all $s_q \in \{1, \dots, K\}$. Note that

$$\begin{aligned} \mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q, \lambda_n) - \mathcal{L}_n((\widehat{\mathbf{\Phi}}_{n,s_q,1}, \mathbf{0}); s_q, \lambda_n) &= [\mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q, \lambda_n) - \mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{0}); s_q, \lambda_n)] \\ &- [\mathcal{L}_n((\widehat{\mathbf{\Phi}}_{n,s_q,1}, \mathbf{0}); s_q, \lambda_n) - \mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{0}); s_q, \lambda_n)]. \end{aligned}$$

The second term in the above difference is positive by the definition of $(\widehat{\mathbf{\Phi}}_{n,s_q,1}, \mathbf{0})$. By Lemma 2, with probability tending to one the first term is also negative, whenever $\|\mathbf{\Phi} - \mathbf{\Phi}^*\|_2 = O(n^{-1/2})$. Hence, for any $\mathbf{\Phi}$ such that $\|\mathbf{\Phi} - \mathbf{\Phi}^*\|_2 = O(n^{-1/2})$, we have

$$\mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q, \lambda_n) - \mathcal{L}_n((\widehat{\mathbf{\Phi}}_{n,s_q,1}, \mathbf{0}); s_q, \lambda_n) < 0$$

in probability. This implies that, with probability tending to one as $n \to \infty$, the maximizer of the penalized conditional log-likelihood $\mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{\Phi}_2); s_q, \lambda_n)$ is indeed $(\widehat{\mathbf{\Phi}}_{n,s_q,1}, \mathbf{0})$. On the other hand, by Theorem 1 if $a_n = O(1)$, there exists a root-n consistent estimator $\widehat{\mathbf{\Phi}}_{n,s_q} = (\widehat{\mathbf{\Phi}}_{n,s_q,1}, \widehat{\mathbf{\Phi}}_{n,s_q,2})$ of $\mathbf{\Phi}^*$. This implies **Part (i)** and completes the proof.

Part (ii). By **Part (i)**, the maximizer $\widehat{\mathbf{\Phi}}_{n,s_q} = (\widehat{\mathbf{\Phi}}_{n,s_q,1}, \widehat{\mathbf{\Phi}}_{n,s_q,2})$ of $\mathcal{L}_n(\mathbf{\Phi}; s_q, \lambda_n)$ has its $\widehat{\mathbf{\Phi}}_{n,s_q,2} = 0$, with probability tending to one, as $n \to \infty$. Thus, by considering $\mathcal{L}_n((\mathbf{\Phi}_1, \mathbf{0}); s_q, \lambda_n) \equiv \mathcal{L}_n(\mathbf{\Phi}_1; s_q, \lambda_n)$ as only a function of $\mathbf{\Phi}_1$, Theorem 1 guarantees the existence of a local maximizer $\widehat{\mathbf{\Phi}}_{n,s_q,1}$ of $\mathcal{L}_n(\mathbf{\Phi}_1; s_q, \lambda_n)$ which is a $\sqrt{n}$-consistent estimator of $\mathbf{\Phi}_1^*$. Because $\mathbf{\Phi}_1^*$ is not on the boundary of the parameter space, the local maximizer in its infinitesimal neighborhood must satisfies

$$\frac{\partial \mathcal{L}_n(\mathbf{\Phi}_1; s_q, \lambda_n)}{\partial \mathbf{\Phi}_1}\bigg|_{\mathbf{\Phi}_1 = \widehat{\mathbf{\Phi}}_{n,s_q,1}} = \left\{ \frac{\partial \tilde{\ell}_n(\mathbf{\Phi}_1; s_q)}{\partial \mathbf{\Phi}_1} - \frac{\partial \mathcal{R}_n(\mathbf{\Phi}_1; \lambda_n)}{\partial \mathbf{\Phi}_1} \right\}\bigg|_{\mathbf{\Phi}_1 = \widehat{\mathbf{\Phi}}_{n,s_q,1}} = \mathbf{0}.$$

Using the first order Taylor's expansion, we have

$$\begin{aligned}
\frac{\partial \mathcal{R}_n(\mathbf{\Phi}_1; \lambda_n)}{\partial \mathbf{\Phi}_1}\bigg|_{\mathbf{\Phi}_1 = \widehat{\mathbf{\Phi}}_{n,s_q,1}} &= \mathcal{R}_n'(\mathbf{\Phi}_1^*; \lambda_n) + \left\{ \mathcal{R}_n''(\mathbf{\Phi}_1^*; \lambda_n) + o_p(n) \right\} \left(\widehat{\mathbf{\Phi}}_{n,s_q,1} - \mathbf{\Phi}_1^* \right) \\
\frac{\partial \tilde{\ell}_n(\mathbf{\Phi}_1; s_q)}{\partial \mathbf{\Phi}_1}\bigg|_{\mathbf{\Phi}_1 = \widehat{\mathbf{\Phi}}_{n,s_q,1}} &= \tilde{\ell}_n'(\mathbf{\Phi}_1^*; s_q) + \left\{ \tilde{\ell}_n''(\mathbf{\Phi}_1^*; s_q) + o_p(n) \right\} \left(\widehat{\mathbf{\Phi}}_{n,s_q,1} - \mathbf{\Phi}_1^* \right).
\end{aligned}$$

Thus,

$$\sqrt{n}\left\{ \frac{\tilde{\ell}_n''(\mathbf{\Phi}_1^*; s_q)}{n} - \frac{\mathcal{R}_n''(\mathbf{\Phi}_1^*; \lambda_n)}{n} + o_p(1) \right\} (\widehat{\mathbf{\Phi}}_{n,s_q,1} - \mathbf{\Phi}_1^*) = -\frac{\tilde{\ell}_n'(\mathbf{\Phi}_1^*; s_q)}{\sqrt{n}} + \frac{\mathcal{R}_n'(\mathbf{\Phi}_1^*; \lambda_n)}{\sqrt{n}}.$$

By the definition of $p_n(\nu_k)$, we have that $\tilde{\ell}_n(\mathbf{\Phi}; s_q) - \ell_n(\mathbf{\Phi}; s_q) = O(n^{-1/2})$. Thus, by Lemma 1, as $n \to \infty$,

$$\frac{\tilde{\ell}_n'(\mathbf{\Phi}_1^*; s_q)}{\sqrt{n}} = \frac{\ell_n'(\mathbf{\Phi}_1^*; s_q)}{\sqrt{n}} - o_p(1) \xrightarrow{D} \mathcal{N}(0, \mathbf{I}_{11}(\mathbf{\Phi}_1^*))$$

$$-\frac{\tilde{\ell}_n''(\mathbf{\Phi}_1^*; s_q)}{n} = -\frac{\ell_n''(\mathbf{\Phi}_1^*; s_q)}{n} + o_p(1) = -\frac{\ell_n''(\mathbf{\Phi}_1^*; s_q)}{n} + o_p(1) \xrightarrow{p} \mathbf{I}_{11}(\mathbf{\Phi}_1^*)$$

where $\mathbf{I}_{11}(\mathbf{\Phi}_1^*)$ is the positive definite matrix based on the MSAR model when the zero AR-coefficients $\mathbf{\Phi}_2^* = \mathbf{0}$ are removed from the model. By Slutsky's theorem, as $n \to \infty$,

$$\sqrt{n}\left\{ \left[\mathbf{I}_{11}(\mathbf{\Phi}_1^*) + \frac{\mathcal{R}_n''(\mathbf{\Phi}_1^*; \lambda_n)}{n} \right] (\widehat{\mathbf{\Phi}}_{n,s_q,1} - \mathbf{\Phi}_1^*) + \frac{\mathcal{R}_n'(\mathbf{\Phi}_1^*; \lambda_n)}{n} \right\} \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{I}_{11}(\mathbf{\Phi}_1^*)). \quad \spadesuit$$

Proof of Theorem 3:

Part (i). Let $\widehat{\mathbf{\Phi}}_{n,\kappa,s_q''}$ and $\widehat{\mathbf{\Phi}}_{n,s_q'}^*$ be maximizers of the penalized conditional log-likelihood function $\mathcal{L}_n(\mathbf{\Phi}; s_q, \lambda_n)$ under the MSAR models with κ and K regimes, respectively, for any choices of the initial states $s_q = s_q'' \in \{1, \ldots, \kappa\}$ and $s_q = s_q' \in \{1, \ldots, K\}$. Then,

$$\Pr(\widehat{K}_n < K) \le \sum_{\kappa=1}^{K-1} \Pr\{\text{RBIC}(\widehat{\mathbf{\Phi}}_{n,\kappa,s_q''}) > \text{RBIC}(\widehat{\mathbf{\Phi}}_{n,s_q'}^*)\}.$$

It suffices to show that, for each $\kappa < K$,

$$\text{RBIC}(\widehat{\mathbf{\Phi}}_{n,\kappa,s_q''}) - \text{RBIC}(\widehat{\mathbf{\Phi}}^*_{n,s_q'}) < 0 \tag{A.4}$$

in probability, as $n \to \infty$. We now focus on such a κ.

Let $\mathcal{G}_\kappa$ be the space of the parameter vector $\mathbf{\Phi}_\kappa$ of a MSAR model with exactly κ regimes. For each $\mathbf{\Phi} \in \mathcal{G}_\kappa$, we have $\mathbf{\Phi} \neq \mathbf{\Phi}^*$. By the definition of the conditional likelihood in equation (A.1) of this document, we have that

$$\begin{aligned}\frac{1}{n}\{\ell_n(\mathbf{\Phi}; s_q'') - \ell_n(\mathbf{\Phi}^*; s_q')\} &= \frac{1}{n}\log\left[\frac{f_3(Y_{q+1},\ldots,Y_n|Y_{1:q}, s_q'', \mathbf{\Phi})}{f_3(Y_{q+1},\ldots,y_n|Y_{1:q}, s_q', \mathbf{\Phi}^*)}\right] \\ &= \frac{1}{n}\sum_{t=q+1}^{n}\log\left[\frac{h(Y_t|Y_{1:t-1}, s_q'', \mathbf{\Phi})}{h(Y_t|Y_{1:t-1}, s_q', \mathbf{\Phi}^*)}\right]\end{aligned}$$

where $h(\cdot)$'s are the conditional densities of Y_t given $(Y_{1:t-1}, s_q)$, for $s_q = s_q''$ or $s_q = s_q'$. By Jensen's inequality, the ergodic theorem and the results of Douc et al. (2004) and Xie et al. (2008), for any $\mathbf{\Phi} \in \mathcal{G}_\kappa$, and any choices of s_q'' and s_q' as mentioned above, as $n \to \infty$, we have

$$\frac{1}{n}\{\ell_n(\mathbf{\Phi}; s_q'') - \ell_n(\mathbf{\Phi}^*; s_q')\} \xrightarrow{a.s} E^*\left\{\log\left[\frac{h(Y_t|Y_{-\infty:t-1}, \mathbf{\Phi})}{h(Y_t|Y_{-\infty:t-1}, \mathbf{\Phi}^*)}\right]\right\} < 0,$$

where $E^*\{\cdot\}$ is the expectation under the true distribution. The quantity on the right hand side of the above limit resembles the negative of Kulback-Leibler information, and thus we call it $-KL(\mathbf{\Phi}^*; \mathbf{\Phi})$.

Next, we show that the above inequality is true uniformly on $\mathcal{G}_\kappa$ rather than valid for individual members. This can be done following the same steps of Wald (1949). We outline a few key steps without complete details as follows. For each $\mathbf{\Phi} \in \mathcal{G}_\kappa$ and a positive value $\rho > 0$, define

$$h(y_t|y_{1:t-1}, \mathbf{\Phi}, s_q'', \rho) = \sup\left\{h(y_t|y_{1:t-1}, \tilde{\mathbf{\Phi}}, s_q'') : \|\tilde{\mathbf{\Phi}} - \mathbf{\Phi}\|_2 < \rho\right\}$$

and

$$\ell_n(\mathbf{\Phi}; s_q'', \rho) = \sum_{t=q+1}^{n}\log h(y_t|y_{1:t-1}, \mathbf{\Phi}, s_q'', \rho).$$

When $\rho \downarrow 0$, we have (in obvious notation)

$$KL(\mathbf{\Phi}^*; \mathbf{\Phi}, \rho) = E^* \left\{ \log \left[\frac{h(Y_t | Y_{-\infty:t-1}, \mathbf{\Phi}^*)}{h(Y_t | Y_{-\infty:t-1}, \mathbf{\Phi}, \rho)} \right] \right\} \longrightarrow KL(\mathbf{\Phi}^*; \mathbf{\Phi}) > 0.$$

Thus, for each $\mathbf{\Phi} \in \mathcal{G}_\kappa$, there exists a $\rho > 0$ such that $KL(\mathbf{\Phi}^*; \mathbf{\Phi}, \rho) > 0$. By the ergodic theorem for martingales, the same ρ makes

$$n^{-1}\{\ell_n(\mathbf{\Phi}; s_q'', \rho) - \ell_n(\mathbf{\Phi}^*; s_q')\} < -\epsilon < 0$$

for some $\epsilon > 0$, almost surely, as $n \to \infty$. Due to the assumed compactness of $\mathcal{G}_\kappa$, this result implies

$$\sup\{\ell_n(\mathbf{\Phi}; s_q'') : \mathbf{\Phi} \in \mathcal{G}_\kappa\} - \ell_n(\mathbf{\Phi}^*; s_q') < -n\epsilon$$

for some $\epsilon > 0$, almost surely, for the given $\kappa < K$.

Now we use the above result to prove (A.4). That is, the inequality must be valid after the influences of the penalty and regularization terms are considered. We have

$$\begin{aligned}
\ell_n(\widehat{\mathbf{\Phi}}_{n,\kappa,s_q''}; s_q'') - \ell_n(\widehat{\mathbf{\Phi}}^*_{n,s_q'}; s_q') &= \ell_n(\widehat{\mathbf{\Phi}}_{n,\kappa,s_q''}; s_q'') - \mathcal{L}_n(\widehat{\mathbf{\Phi}}^*_{n,s_q'}; s_q', \lambda_n) - \mathcal{R}_n(\widehat{\mathbf{\Phi}}^*_{n,s_q'}; \lambda_n) - \sum_{j=1}^{K} p_n(\hat{\nu}_j^*) \\
&\leq \{\ell_n(\widehat{\mathbf{\Phi}}_{n,\kappa,s_q''}; s_q'') - \mathcal{L}_n(\mathbf{\Phi}^*; s_q', \lambda_n)\} - \mathcal{R}_n(\widehat{\mathbf{\Phi}}^*_{n,s_q'}; \lambda_n) - \sum_{j=1}^{K} p_n(\hat{\nu}_j^*) \\
&\leq \{\sup \ell_n(\mathbf{\Phi}; s_q'') - \ell_n(\mathbf{\Phi}^*; s_q')\} + \{\mathcal{R}_n(\mathbf{\Phi}^*; \lambda_n) - \mathcal{R}_n(\widehat{\mathbf{\Phi}}^*_{n,s_q'}; \lambda_n)\} \\
&+ \sum_{j=1}^{K} \{p_n(\nu_j^*) - p_n(\hat{\nu}_j^*)\}. \qquad \text{(A.5)}
\end{aligned}$$

For notational simplicity, we have omitted the range of supremum which is $\mathcal{G}_\kappa$. We also used $\hat{\nu}_j^*$ for the corresponding element in $\widehat{\mathbf{\Phi}}^*_{n,s_q'}$.

By Theorem 1, $\widehat{\mathbf{\Phi}}^*_{n,s_q'} \xrightarrow{a.s} \mathbf{\Phi}^*$, as $n \to \infty$. Since $p_n(\nu_k^*) = o(n)$ and $\mathcal{R}_n(\mathbf{\Phi}^*; \lambda_n) = o(n)$, we have

$$\{\mathcal{R}_n(\mathbf{\Phi}^*; \lambda_n) - \mathcal{R}_n(\widehat{\mathbf{\Phi}}^*_{n,s_q'}; \lambda_n)\} + \{\sum_{j=1}^{K} p_n(\nu_j^*) - \sum_{j=1}^{K} p_n(\hat{\nu}_j^*)\} = o(n).$$

Using this assessment, from (A.5), for large n, we find

$$\ell_n(\widehat{\mathbf{\Phi}}_{n,\kappa,s_q''}; s_q'') - \ell_n(\widehat{\mathbf{\Phi}}^*_{n,s_q'}; s_q') \leq -n\epsilon + o(n).$$

Thus, for each $\kappa < K$, and any $s_q'' \in \{1, \ldots, \kappa\}$ and $s_q' \in \{1, \ldots, K\}$, as $n \to \infty$,

$$\begin{aligned}
\text{RBIC}(\widehat{\boldsymbol{\Phi}}_{n,\kappa,s_q''}) - \text{RBIC}(\widehat{\boldsymbol{\Phi}}^*_{n,s_q'}) &= \{\ell_n(\widehat{\boldsymbol{\Phi}}_{n,\kappa,s_q''}; s_q'') - \ell_n(\widehat{\boldsymbol{\Phi}}^*_{n,s_q'}; s_q')\} \\
&- 0.5(\log n)\{\dim(\widehat{\boldsymbol{\Phi}}_{n,\kappa,s_q''}) - \dim(\widehat{\boldsymbol{\Phi}}^*_{n,s_q'})\} < 0
\end{aligned}$$

almost surely because the penalty term is $o(n)$. This completes the proof of this part. ♠

Part (ii). First, we introduce some notation. For any given finite number of regimes κ, where $K \le \kappa \le \mathcal{K}$, consider the one-step ahead predictive density similar to the one given in equation (4.1) of the manuscript but with K replaced by κ. The density can be written as

$$f_\kappa(y_{n+1}|y_{1:n}) = \sum_{j=1}^{\kappa} \Pr(S_{n+1} = j|y_{1:n})\phi(y_{n+1}; \boldsymbol{x}_{n+1}^\top \boldsymbol{\theta}_j, \nu_j) = \int_{\mathbb{R}^{q+1}\times\mathbb{R}^+} \phi(y_{n+1}; \boldsymbol{x}_{n+1}^\top \boldsymbol{\theta}, \nu)\, \mathrm{d}G_{n,\kappa}(\boldsymbol{\theta}, \nu),$$

where

$$G_{n,\kappa}(\boldsymbol{\theta}, \nu) = \sum_{j=1}^{\kappa} \Pr(S_{n+1} = j|y_{1:n})\ \mathrm{I}(\boldsymbol{\theta}_j \le \boldsymbol{\theta}, \nu_j \le \nu) \tag{A.6}$$

and $\mathrm{I}(\cdot)$ is an indicator function. Using this representation, the estimated one-step ahead predictive density, based on a fitted κ-regime MSAR model, is written as

$$\widehat{f}_\kappa(y_{n+1}|y_{1:n}) = \int_{\mathbb{R}^{q+1}\times\mathbb{R}^+} \phi(y_{n+1}; \boldsymbol{x}_{n+1}^\top \boldsymbol{\theta}, \nu)\ \mathrm{d}\widehat{G}_{n,\kappa}(\boldsymbol{\theta}, \nu),$$

where

$$\widehat{G}_{n,\kappa}(\boldsymbol{\theta}, \nu) = \sum_{j=1}^{\kappa} \widehat{\Pr}(S_{n+1} = j|y_{1:n})\ \mathrm{I}(\widehat{\boldsymbol{\theta}}_j \le \boldsymbol{\theta}, \widehat{\nu}_j \le \nu). \tag{A.7}$$

The estimates are based on the MPCLE given in Section 3 of the manuscript. Similarly, the true one-step ahead predictive density has the representation

$$f^*(y_{n+1}|y_{1:n}) = \int_{\mathbb{R}^{q+1}\times\mathbb{R}^+} \phi(y_{n+1}; \boldsymbol{x}_{n+1}^\top \boldsymbol{\theta}, \nu)\ \mathrm{d}G_n^*(\boldsymbol{\theta}, \nu),$$

where

$$G_n^*(\boldsymbol{\theta}, \nu) = \sum_{j=1}^{K} \overset{*}{\Pr}(S_{n+1} = j|y_{1:n})\ \mathrm{I}(\boldsymbol{\theta}_j^* \le \boldsymbol{\theta}, \nu_j^* \le \nu). \tag{A.8}$$

Upon the specification of the (conditional) initial distributions $\{\Pr(S_q = l|y_{1:q}), l = 1, \ldots, \kappa\} \equiv \boldsymbol{\gamma}_q$ and $\{\Pr(S_q = l|y_{1:q}), l = 1, \ldots, K\} \equiv \boldsymbol{\gamma}_q^*$, the conditional probabilities

$\Pr(\cdot|y_{1:n})$ in (A.6), (A.7), and (A.8) are computed using the *prediction* and *filtering* equations given in Section 4 of the manuscript. As discussed in the manuscript, the effect of the initial distribution diminishes at a geometric rate in n, (Ocone and Pardoux, 1996; Kleptsyna and Veretennikov, 2008; Douc et al., 2009).

In the first part of the proof, we show that the distance (defined below) between $\widehat{G}_{n,\kappa}$ and G_n^* converges to zero, almost surely, as $n \to \infty$, for any $K \leq \kappa \leq \mathcal{K}$.

For any distribution of the form (A.6), consider the distance

$$D_n(G_{n,\kappa}, G_n^*) = \int_{y_{1:n+1}} \left\{ \int_{\mathbb{R}^{q+1}\times\mathbb{R}^+} |G_{n,\kappa}(\boldsymbol{\theta},\nu) - G_n^*(\boldsymbol{\theta},\nu)|\, e^{-\{|\boldsymbol{\theta}|_1+|\nu|\}}\, \mathrm{d}\boldsymbol{\theta}\mathrm{d}\nu \right\} f^*(y_{1:n+1})\mathrm{d}y_{1:n+1} \tag{A.9}$$

where $|\cdot|_1$ is the L_1-norm of a vector, and $f^*(y_{1:n+1})$ is the joint density of $Y_{1:n+1}$. For an arbitrary small value $\delta > 0$, consider the family

$$\mathcal{H}_n(\delta) = \{G_{n,\kappa} : D_n(G_{n,\kappa}, G_n^*) > \delta\}$$

of those distributions that are of δ distance from the true distribution G_n^*. Clearly $G_n^* \notin \mathcal{H}_n(\delta)$, and thus we have

$$E^*\left\{ \log\left[\frac{f_\kappa(Y_{n+1}|Y_{1:n})}{f^*(Y_{n+1}|Y_{1:n})}\right]\right\} < 0$$

for any $G_{n,\kappa} \in \mathcal{H}_n(\delta)$, where the expectation is with respect to the true joint density function of $Y_{1:n+1}$. This implies that, for $t = q+1,\ldots,n$,

$$E^*\left\{ \log\left[\frac{f_\kappa(Y_t|Y_{1:t-1})}{f^*(Y_t|Y_{1:t-1})}\right]\right\} < 0$$

where the conditional densities $f_\kappa(Y_t|Y_{1:t-1})$ and $f^*(Y_t|Y_{1:t-1})$ have similar representations as of $f_\kappa(Y_{n+1}|Y_{1:n})$ and $f^*(Y_{n+1}|Y_{1:n})$ described above but with their own distributions $G_{t,\kappa} \in \mathcal{H}_t(\delta)$ and G_t^*, respectively. By the stationarity and ergodicity conditions, we then have that

$$\frac{1}{n-q}\sum_{t=q+1}^{n} \log\left[\frac{f_\kappa(Y_t|Y_{1:t-1})}{f^*(Y_t|Y_{1:t-1})}\right] = \frac{1}{n-q}\log\left[\frac{f_2(Y_{q+1},\ldots,Y_n|Y_{1:q})}{f_2^*(Y_{q+1},\ldots,Y_n|Y_{1:q})}\right] < -\epsilon(\delta) \tag{A.10}$$

for some $\epsilon(\delta) > 0$, and given $\kappa > K$, almost surely, for large n. Here, $f_2(\cdot)$ is the joint conditional distribution given in (2.2) of the manuscript, and $f_2^*(\cdot)$ is the true distribution based on a MSAR with K regimes. These functions may also be re-written as

$$\begin{aligned} f_2(Y_{q+1},\dots,Y_n|Y_{1:q}) &= \int \left\{ \prod_{t=q+1}^{n} \phi(y_t;\boldsymbol{x}_t^\top \boldsymbol{\theta}_t,\nu_t) \right\} \mathrm{d}G_{\text{joint},K}(\boldsymbol{\gamma}) \\ f_2^*(Y_{q+1},\dots,Y_n|Y_{1:q}) &= \int \left\{ \prod_{t=q+1}^{n} \phi(y_t;\boldsymbol{x}_t^\top \boldsymbol{\theta}_t,\nu_t) \right\} \mathrm{d}G^*_{\text{joint}}(\boldsymbol{\gamma}), \end{aligned}$$

where $\boldsymbol{x}_t^\top = (1, y_{t-1},\dots,y_{t-q})$, $\boldsymbol{\gamma} = (\boldsymbol{\theta}_1,\nu_1,\dots,\boldsymbol{\theta}_n,\nu_n)$, and the joint distributions

$$\begin{aligned} G_{\text{joint},\kappa}(\boldsymbol{\gamma}) &= \sum_{s_{q+1}=1}^{\kappa} \dots \sum_{s_n=1}^{\kappa} \Pr(S_{q+1}=s_{q+1}|y_{1:q}) \left\{ \prod_{t=q+2}^{n} \alpha_{s_{t-1},s_t} \prod_{t=q+1}^{n} \mathrm{I}(\boldsymbol{\theta}'_{s_t} \le \boldsymbol{\theta}_t, \nu'_{s_t} \le \nu_t) \right\} \\ G^*_{\text{joint}}(\boldsymbol{\gamma}) &= \sum_{s_{q+1}=1}^{K} \dots \sum_{s_n=1}^{K} \overset{*}{\Pr}(S_{q+1}=s_{q+1}|y_{1:q}) \left\{ \prod_{t=q+2}^{n} \alpha^*_{s_{t-1},s_t} \prod_{t=q+1}^{n} \mathrm{I}(\boldsymbol{\theta}^*_{s_t} \le \boldsymbol{\theta}_t, \nu^*_{s_t} \le \nu_t) \right\}. \end{aligned}$$

Here, the distribution $G_{\text{joint},K}(\boldsymbol{\gamma})$ belongs to the family

$$\mathcal{H}_{n,\text{joint}}(\delta) = \{G_{\text{joint},\kappa} : D_n(G_{\text{joint},\kappa}, G^*_{\text{joint}}) > \delta\},$$

where the distance is defined as in (A.9) with the inner integration with respect to $\boldsymbol{\gamma}$.

On the other hand, since both penalties $p_n(\cdot)$ and $r_n(\cdot;\lambda_n)$ are of order $o(n)$, for large n we can generalize (A.10) to

$$\begin{aligned} \sup_{\mathcal{H}_{n,\text{joint}}(\delta)} \Bigg\{ \log\left[\frac{f_2(Y_{q+1},\dots,Y_n|Y_{1:q})}{f_2^*(Y_{q+1},\dots,Y_n|Y_{1:q})}\right] - \left[\sum_{j=1}^{\kappa} p_n(\nu_j) - \sum_{j=1}^{K} p_n(\nu_j^*)\right] \\ - \Bigg[\mathcal{R}_n(\boldsymbol{\Phi}_\kappa;\lambda_n) - \mathcal{R}_n(\boldsymbol{\Phi}^*;\lambda_n)\Bigg]\Bigg\} < -(n-q)\epsilon(\delta), \end{aligned}$$

where

$$\mathcal{R}_n(\boldsymbol{\Phi}_\kappa;\lambda_n) = \sum_{j=1}^{\kappa}\sum_{l=1}^{q} r_n(\theta_{jl};\lambda_n) \quad , \quad \mathcal{R}_n(\boldsymbol{\Phi}^*;\lambda_n) = \sum_{j=1}^{K}\sum_{l=1}^{q} r_n(\theta^*_{jl};\lambda_n).$$

Hence, the maximum penalized log-likelihood estimator of the joint distribution $G_{\text{joint},\kappa}$ cannot belong to the family $\mathcal{H}_{n,\text{joint}}(\delta)$, almost surely, as $n \to \infty$. Since δ is arbitrarily small, we have $D_n(\widehat{G}_{\text{joint},\kappa}, G^*_{\text{joint},n}) \to 0$, as $n \to \infty$. This implies that $\widehat{G}_{\text{joint},\kappa}$ converges weakly to $G^*_{\text{joint},n}$, as $n \to \infty$. Consequently, the distribution in (A.7) converges to the true distribution G_n^* in (A.8), as $n \to \infty$. Therefore, due to the boundedness of

$\phi(y_{n+1}; \boldsymbol{x}_{n+1}^{\top}\boldsymbol{\theta}, \nu)$ on the assumed compact parameter space, we conclude that, as $n \to \infty$, the one-step ahead predictive density $\widehat{f}_{\kappa}(y_{n+1}|y_{1:n})$ converges to the true one-step ahead predictive density $f^*(y_{n+1}|y_{1:n})$, almost surely for all values of $y_{1:n+1}$. The same proof can be extend to the h-step ahead predictive density.

Finally, we consider the proof when the RBIC-based estimator $\widehat{K}_n$ is used for the number of regimes K. Note that, for any $\delta > 0$,

$$P\Big\{D_n(G_{n,\widehat{K}_n}, G_n^*) > \delta\Big\} = \sum_{\kappa=1}^{\mathcal{K}} P\Big\{D_n(G_{n,\widehat{K}_n}, G_n^*) > \delta, \widehat{K}_n = \kappa\Big\}$$

$$= \sum_{\kappa=1}^{K-1} P\Big\{D_n(G_{n,\kappa}, G_n^*) > \delta, \widehat{K}_n = \kappa\Big\} + \sum_{\kappa=K}^{\mathcal{K}} P\Big\{D_n(G_{n,\kappa}, G_n^*) > \delta, \widehat{K}_n = \kappa\Big\}$$

$$\le \sum_{\kappa=1}^{K-1} P(\widehat{K}_n = \kappa) + \sum_{\kappa=K}^{\mathcal{K}} P\Big\{D_n(G_{n,\kappa}, G_n^*) > \delta\Big\}.$$

By Theorem 3-(i), as $n \to \infty$ the first term on the right hand side goes to zero, and according to the above discussion the second term also goes to zero. The rest of the proof is similar to fixed case K and thus omitted. ♠

3 COMPUTATION VIA EM ALGORITHM

Maximization of the penalized conditional log-likelihood $\mathcal{L}_n(\boldsymbol{\Phi}; s_q, \lambda)$, given in (3.3) of the manuscript, for a MSAR model with K regimes and a maximal AR-order q is an optimization over a space of dimension $K(q + 2) + K(K - 1)$. For example, with $K = 3$ and $q = 10$, the likelihood function involves 42 parameters; this dimension is large, but direct optimization using Nelder-Mead or quasi-Newton methods (via `optim` in `R`) is still possible when a local quadratic approximation (LQA) to the penalty is adopted: following Fan and Li (2001), the LQA

$$r_n(\theta_{jl}; \lambda) \simeq r_n(\theta_{jl}^{(0)}; \lambda) + \frac{r_n'(\theta_{jl}^{(0)}; \lambda)}{2\theta_{jl}^{(0)}}(\theta_{jl}^2 - \theta_{jl}^{2(0)})$$

holds in a neighbourhood of a current value $\theta_{jl}^{(0)}$, and may be used.

In this manuscript we instead use a modified EM algorithm to numerically approximate the maximum point of the function $\mathcal{L}_n(\boldsymbol{\Phi}; s_q, \lambda)$. We apply coordinate descent-based methods for maximization in the **M-step** of the EM algorithm. As in Zou and Li (2008), for folded concave penalties such as SCAD we use the local linear approximation (LLA)

$$r_n(\theta_{jl};\lambda) \simeq r_n(\theta_{jl}^{(0)};\lambda) + r_n'(\theta_{jl}^{(0)};\lambda)(|\theta_{jl}| - |\theta_{jl}^{(0)}|). \tag{A.11}$$

The advantage of the LLA is that when coupled with a coordinate descent method, it leads to thresholding-type updates of the AR-coefficients in the **M-step** of the EM algorithm; more details are given below.

3.1 EM Algorithm

For observation y_t, let V_{tij} equal 1 if $S_{t-1} = i$ and $S_t = j$, and equal 0 otherwise; V_{tij} records the presence of a transition between regime i at time $t-1$ and regime j at time t. Also, let U_{tj} equal 1 if $S_t = j$. Let $\boldsymbol{x}_t^\top = (1, y_{t-1}, \ldots, y_{t-q})$. The complete conditional log-likelihood is

$$\ell_n^c(\mathbf{\Phi}_K; s_q) = \sum_{i=1}^K \sum_{j=1}^K \sum_{t=q+1}^n V_{tij} \log \alpha_{ij} + \sum_{j=1}^K \sum_{t=q+1}^n U_{tj} \left\{ \log \phi(y_t; \mu_{t,j}, \nu_j) \right\},$$

where $\mu_{t,j} = \boldsymbol{x}^\top \boldsymbol{\theta}_j$. Thus, the penalized complete (conditional adjusted) log-likelihood is

$$\mathcal{L}_n^c(\mathbf{\Phi}_K; s_q, \lambda) = \ell_n^c(\mathbf{\Phi}_K; s_q) - \sum_{j=1}^K p_n(\nu_j) - \sum_{j=1}^K \sum_{l=1}^q r_n(\theta_{jl};\lambda).$$

Given the current value of the parameter $\mathbf{\Phi}_K^{(m)}$, at $(m+1)$-th iteration the EM algorithm proceeds as follows.

E-step: In this step, we compute the conditional expectation of the approximated penalized complete conditional log-likelihood with respect to V_{tij} and U_{tj}, given $\mathbf{\Phi}_K^{(m)}$, s_q and the data $y_{1:n}$. The approximation is due to the LLA in (A.11). Thus, at $(m+1)$-th iteration, the EM objective function (up to a constant) is given by

$$\begin{aligned} Q(\mathbf{\Phi}_K; \mathbf{\Phi}_K^{(m)}, s_q) \quad &= \quad \sum_{j=1}^K \sum_{t=q+1}^n \varpi_{t,s_q,j}^{(m)} \log \alpha_{s_q,j} + \sum_{i \neq s_q}^K \sum_{j=1}^K \sum_{t=q+2}^n \varpi_{tij}^{(m)} \log \alpha_{ij} \\ &+ \quad \sum_{j=1}^K \sum_{t=q+1}^n \omega_{tj}^{(m)} \log \phi(y_t; \mu_{tj}, \nu_j) - \sum_{j=1}^K p_n(\nu_j) - \sum_{j=1}^K \sum_{l=1}^q r_n'(\theta_{jl}^{(m)};\lambda)|\theta_{jl}|, \end{aligned}$$

where for each t, j, k,

$$\begin{aligned}
\varpi_{tij}^{(m)} &= E(V_{tij}|y_{1:n}, s_q; \mathbf{\Phi}_K^{(m)}) \equiv P[S_{t-1} = i, S_t = j|y_{1:n}, s_q; \mathbf{\Phi}_K^{(m)}], \qquad \text{(A.12)}\\
\omega_{tj}^{(m)} &= E(U_{tj}|y_{1:n}, s_q; \mathbf{\Phi}_K^{(m)}) \equiv P[S_t = j|y_{1:n}, s_q; \mathbf{\Phi}_K^{(m)}]. \qquad \text{(A.13)}
\end{aligned}$$

These are "smoothing" probabilities that may be computed numerically in a routine fashion using the conventional forward-backward algorithm of Baum et al. (1970) proposed for hidden Markov models. The details are given in the end of this Section.

M-step: Here, by using the penalty $p_n(\nu_k)$ given in equation (3.2) of the manuscript, we maximize the function $Q(\mathbf{\Phi}_K; \mathbf{\Phi}_K^{(m)}, s_q)$ with respect to $\mathbf{\Phi}_K$. The maximization with respect to the AR-coefficients θ_{jl} is performed using a coordinate descent approach. The parameter estimates are then updated as follows. First, we compute the quantities

$$z_{1,jl} = \frac{1}{n-q} \sum_{t=q+1}^{n} \omega_{tj}^{(m)} y_{t-l}(y_t - \tilde{\mu}_{tj,-l}) \text{ and } z_{2,jl} = \frac{1}{n-q} \sum_{t=q+1}^{n} \omega_{tj}^{(m)} y_{t-l}^2,$$

where $\tilde{\mu}_{tj,-l} = \theta_{j0}^{(m)} + \sum_{v=1}^{l-1} \theta_{jv}^{(m+1)} y_{t-v} + \sum_{v>l}^{q} \theta_{jv}^{(m)} y_{t-v}$, $1 \le l \le q$ and $1 \le j \le K$. We then update the AR-coefficients by

$$\theta_{jl}^{(m+1)} = \frac{T(z_{1,jl}; \lambda_{jl})}{z_{2,jl}}, \qquad \text{(A.14)}$$

where $T(z; \lambda) = \text{sign}(z)(|z| - \lambda)_+$ is the soft-thresholding operator (Donoho and Johnstone, 1994), and λ_{jl} varies depending on the penalty $r_n(\theta_{jl}; \lambda)$. For the three penalties considered in this manuscript the values of λ_{jl} are given in the end of this section.

The regime-specific intercepts and variances are updated by

$$\begin{aligned}
\theta_{j0}^{(m+1)} &= \frac{\sum_{t=q+1}^{n} \omega_{tj}^{(m)} (y_t - \mu_{tj}^{(m+1)})}{\sum_{t=q+1}^{n} \omega_{tj}^{(m)}} \qquad \text{(A.15)}\\
\nu_j^{(m+1)} &= \frac{\sum_{t=q+1}^{n} \omega_{tj}^{(m)} (y_t - \boldsymbol{x}_t^\top \boldsymbol{\theta}_j^{(m+1)})^2 + 2\mathcal{V}_n^2/\sqrt{n-q}}{\sum_{t=q+1}^{n} \omega_{tj}^{(m)} + 2/\sqrt{n-q}}, \qquad \text{(A.16)}
\end{aligned}$$

where $\mu_{tj}^{(m+1)} = \sum_{l=1}^{q} \theta_{jl}^{(m+1)} y_{t-l}$. The updated transition probabilities are

$$\alpha_{s_q,j}^{(m+1)} = \frac{\sum_{t=q+1}^{n} \varpi_{t,s_q,j}^{(m)}}{\sum_{t=q+1}^{n} \sum_{i=1}^{K} \varpi_{t,s_q,i}^{(m)}} \quad , \quad \alpha_{ij}^{(m+1)} = \frac{\sum_{t=q+2}^{n} \varpi_{tij}^{(m)}}{\sum_{t=q+2}^{n} \sum_{h=1}^{K} \varpi_{tih}^{(m)}} \qquad \text{(A.17)}$$

for $i \neq s_q$, $1 \leq i, j \leq K$, and any $s_q \in \{1, \ldots, K\}$.

Starting from an initial value $\mathbf{\Phi}_K^{(0)}$, the EM algorithm continues until some convergence criterion is met. We used the stopping rule $\|\mathbf{\Phi}_K^{(m+1)} - \mathbf{\Phi}_K^{(m)}\| \leq \epsilon$, for a pre-specified small value ε, taken 10^{-5} in our simulations and data analysis.

More details:

1. The **soft-thresholding** operator in (A.14):

$$T(z;\lambda) = \text{sign}(z)(|z| - \lambda)_+ = \begin{cases} z - \lambda & ; \quad z > \lambda > 0, \\ 0 & ; \quad |z| \leq \lambda, \\ z + \lambda & ; \quad z < -\lambda, \end{cases}$$

The values of λ_{jl} used in (A.14):

$$\lambda_{jl} = \begin{cases} \lambda & ; \quad \text{LASSO}, \\ \lambda \, \omega_{jl} & ; \quad \text{ADALASSO}, \\ r_n'(\theta_{jl}^{(m)}; \lambda)/(n-q) & ; \quad \text{SCAD}, \end{cases}$$

where ω_{jl} are the weights in the ADALASSO and $r_n'(\cdot;\lambda)$ is the first derivative of the SCAD penalty with respect to $|\theta_{jl}|$.

2. **Forward-backward algorithm for computing** (A.12) **and** (A.13):

Note that all of our computation is done by conditioning on $y_{1:q}$ and s_q. For simplicity in notation, denote the event (or set) $E_q = \{Y_{1:q} = y_{1:q}, S_q = s_q\}$. Thus, the conditional probabilities in (A.12) and (A.13) can be re-written as

$$\varpi_{tij}^{(m)} = P_{E_q}(S_{t-1} = i, S_t = j | y_{q+1:n}; \mathbf{\Phi}_K^{(m)}), \tag{A.18}$$

$$\omega_{tj}^{(m)} = P_{E_q}(S_t = j | y_{q+1:n}; \mathbf{\Phi}_K^{(m)}) \tag{A.19}$$

for $t = q+1, \ldots, n$ and $i, j = 1, 2, \ldots, K$, where $P_{E_q}(\cdot)$ indicates conditioning on $y_{1:q}$ and s_q. Note that $\varpi_{q+1,s_q,j}^{(m)} = \omega_{q+1,j}^{(m)}$, for all $j = 1, 2, \ldots, K$. Numerical computation of these conditional probabilities is done as follows:

Denote the quantities

$$\begin{aligned} a_j(t; E_q) &= P_{E_q}(y_{q+1}, \ldots, y_i, S_i = j; \mathbf{\Phi}_K^{(m)}) \quad , \quad t = q+1, \ldots, n & \text{(A.20)} \\ b_j(t; E_q) &= P_{E_q}(y_{i+1}, \ldots, y_n | S_i = j, \mathbf{\Phi}_K^{(m)}) \quad , \quad t = q+1, \ldots, n-1. & \text{(A.21)} \end{aligned}$$

Then, using the model assumptions we have that

$$\omega_{tj}^{(m)} = \frac{a_j(t; E_q) b_j(t; E_q)}{\sum_{l=1}^{K} a_l(n; E_q)} \tag{A.22}$$

for $t = q+1, \ldots, n$ and $j = 1, 2, \ldots, K$. Similarly,

$$\varpi_{tij}^{(m)} = \frac{\alpha_{ij}^{(m)} \phi(y_t; \mu_{tj}^{(m)}, \nu_j^{(m)}) \times a_i(t-1; E_q) \times b_j(t; E_q)}{\sum_{l=1}^{K} a_l(n; E_q)} \tag{A.23}$$

for $t = q+2, \ldots, n$ and $i, j = 1, 2, \ldots, K$. The quantities in (A.20)-(A.21) are computed recursively using the following forward-backward formulae:

$$\begin{aligned} a_j(t; E_q) &= \sum_{i=1}^{K} a_i(t-1; E_q) \alpha_{ij}^{(m)} \phi(y_t; \mu_{tj}^{(m)}, \nu_j^{(m)}), & \text{(A.24)} \\ b_j(t; E_q) &= \sum_{i=1}^{K} \alpha_{ji}^{(m)} \phi(y_{t+1}; \mu_{t+1,j}^{(m)}, \nu_j^{(m)}) b_i(t+1; E_q), & \text{(A.25)} \end{aligned}$$

where $a_j(q+1; E_q) = \alpha_{s_q,j}^{(m)} \phi(y_{q+1}; \mu_{q+1,j}^{(m)}, \nu_j^{(m)})$ and $b_j(n; E_q) = 1$, for $j = 1, \ldots, K$. Hence, we first compute the sequences in (A.24) and (A.25) which are then used to compute the quantities in (A.22) and (A.23).

3.2 Tuning of λ in $r_n(\theta, \lambda)$

One remaining issue in the implementation of the regularization method is the choice of tuning parameter λ. Given (K, q), we recommend an information criterion together with a grid search scheme as follows.

Consider a pre-chosen grid of λ-values $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ for some M, say, $M = 10$. For each λ_i, we obtain the MPCLE $\widehat{\mathbf{\Phi}}_{n,K,s_q}(\lambda_i)$ using the EM algorithm presented above. We compute the information criterion (IC)

$$\text{IC}(\lambda_i) = \ell_n(\widehat{\mathbf{\Phi}}_{n,K,s_q}(\lambda_i); s_q) - 0.5 \times \text{DF}(\lambda_i) \log(n-q)$$

where $\ell_n(\cdot)$ is the conditional log-likelihood in (A.1), and $\text{DF}(\lambda_i) = \sum_{j=1}^{K} \sum_{l=1}^{q} I(\hat{\theta}_{jl} \neq 0)$ is the total number of estimated non-zero AR-coefficients $\hat{\theta}_{jl}$. This information criterion mimics the one used in generalized linear regression by Zhang et al. (2010). We choose the value of tuning parameter as $\tilde{\lambda} = \text{argmax}_{1 \leq i \leq M}\{\text{IC}(\lambda_i)\}$.

Although theoretical properties of the tuning parameter $\tilde{\lambda}$ chosen by the IC is currently unknown to us, our numerical results suggests that the IC performs reasonably well in selecting appropriate level of the penalty parameter λ.

4 ADDITIONAL SIMULATION RESULTS

This section contains the simulation results for the fifth model, **M5**, which is a three-state ($K = 3$) MSAR with the transition probability matrix and its corresponding stationary distribution specified as

$$\mathbb{P} = \begin{bmatrix} .20 & .40 & .40 \\ .10 & .50 & .40 \\ .70 & .20 & .10 \end{bmatrix} \quad , \quad (\pi_1, \pi_2, \pi_3) = (.316, .376, .308).$$

The state-specific variances and means are, respectively, $(\sigma_1, \sigma_2, \sigma_3) = (2, 2, 4)$, and

$$\mu_{t,1} = .7y_{t-1} - .6y_{t-2} \quad , \quad \mu_{t,2} = -.5y_{t-1} \quad , \quad \mu_{t,3} = 1.5y_{t-1} - .75y_{t-2}.$$

Since the simulation results were similar when conditioning on any initial state $s_q \in \{1, 2, 3\}$, we report the results for $s_q = 1$. For a pre-specified common AR-order $q = 5$, the total number of possible models (2^{Kq}) to be examined by the standard BIC for AR-order estimation is about 29791, which is computationally not feasible for us. Thus, below we only discuss the simulation results for the new method.

Table A4 contains the average (over 300 replications) estimated sensitivity and specificity (ES1, ES2) results using LASSO, ADALASSO and SCAD. We have considered larger sample sizes $n = 250, 500, 800$ for this model since it is more complex compared to the two-state models **M1**–**M4**. We see that the regularization method identifies the true zero AR-coefficients (average estimated sensitivity, ES1) at least 93% to 100% of the times, across different regimes and sample sizes, for the three penalties. Regarding the ES2, for the smaller sample size $n = 250$, the LASSO identifies the true non-zero AR-coefficients approximately 68% to 93% of the times across the three regimes of the model, whereas ADALASSO and SCAD identify the true non-zero coefficients approximately 83% to 95% and 91% to 97% of the times across the three regimes, respectively. Overall, the regularization method using the SCAD and ADALASSO outperform the LASSO for the smaller sample sizes that we have considered. As the sample size increase all the three penalties improve.

Figure A5 shows the boxplots of the empirical L_2 losses of the parameter estimates based on the LASSO, ADALASSO, and SCAD, as well as the estimates based on the oracle model. The results are based on the 300 random samples from model **M5**. For the sample size $n = 250$, the empirical median (and variation) losses of the estimates based on the new method are higher than those of the estimates under the true (oracle) model. This is more evident for the LASSO penalty. As the sample size increases the performance of the new method based on the three penalties improves and it is comparable to the oracle estimator, with SCAD and ADALASSO outperforming the LASSO.

We next examine the performance of the estimator $\widehat{K}_n$ of K obtained by using the three criteria RAIC, RBIC and RMSC described in Section 5 of the manuscript. We fit MSAR models with number of hidden regimes $K = 1, \ldots, 5$, to each simulated sample from model **M5**, and obtain the MPCLE which is then used to compute the criteria RAIC, RBIC and RMSC. We choose $\widehat{K}_n$ as the one that maximizes any of the three criteria. The results corresponding to different sample sizes ($n = 250, 500, 800$) and the regularization penalties are given in Tables A5–A7.

It is seen that, for the smaller sample size, the proportion of underestimation of the correct order $K = 3$ by the RBIC is higher than those by RAIC and RMSC. As the sample size increases this proportion decreases to zero, which is expected by the result of Theorem 3-(i). When $n = 800$, the performance of all the three is very good.

References

Baum, L. E., T. Petrie, G. Soules, and G. Weiss (1970). A maximization technique occuring in the statistical anlaysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41**, 164–171.

Donoho, D. L. and J. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Douc, R., G. Fort, E. Moulines, and P. Priouret (2009). Forgetting the initial distribution for hidden Markov models. *Stoch. Process. Appl* **119**, 1235–1256.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Kleptsyna, M. L. and A. Y. Veretennikov (2008). On discrete time ergodic filters with wrong initial data. *Probab. Theory Relat. Fields* **141**, 411–444.

Ocone, D. and E. Pardoux (1996). Asymptotic stability of the optimal filter with respect to its initial condition. *SIAM J. Control Optim.* **34**, 226–243.

Park, H. and F. Sakaori (2013). Lag weighted lasso for time series model. *Comput. Stat.* **28**, 493–504.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist* **20**, 595–601.

Xie, Y., J. Yu, and B. Ranneby (2008). A general autoregressive model with Markov switching: estimation and consistency. *Mathematical Methods of Statistics* **17**, 228–240.

Zhang, Y., R. Li, and C.-L. Tsai (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105**, 312–323.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509–1566.

Table A1: Average computational time (ACT, in seconds) taken by a method to complete per-sample results discussed in Section 7.1. of the main paper.

	$n = 150$				$n = 250$				$n = 500$			
Model	BIC	LASSO	ADALASSO	SCAD	BIC	LASSO	ADALASSO	SCAD	BIC	LASSO	ADALASSO	SCAD
M1	17.4	1.36	.410	.957	30.4	2.02	.707	1.36	66.2	4.03	1.71	1.98
M2	22	.853	.375	1.05	42.6	1.41	.595	1.71	96.6	4.35	2.35	3.77
M3	85	1.06	.403	.830	111	1.81	.853	1.37	270	2.90	2.17	1.99
M4	90.4	1.59	.830	1.05	144	3.00	1.32	1.78	297	5.44	2.13	2.91

Table A2: Average proportion of times (in 300 replications) that a number of AR-regimes $1 \leq K \leq 5$ is selected by a criterion[1]. Results for the true order $K = 2$ are in **bold**.

		$n = 150$			$n = 250$			$n = 500$		
Models	K	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC
M1	1	.094	.691	.165	.019	.246	.030	.000	.009	.004
	2	**.399**	**.306**	**.385**	**.644**	**.750**	**.731**	**.762**	**.987**	**.942**
	3	.201	.003	.086	.170	.004	.117	.126	.004	.036
	4 or 5	.306	.000	.364	.167	.000	.122	.112	.000	.018
M2	1	.000	.028	.000	.000	.004	.000	.000	.000	.000
	2	**.784**	**.965**	**.794**	**.866**	**.989**	**.870**	**.932**	**1.00**	**.860**
	3	.135	.007	.053	.090	.007	.047	.051	.000	.030
	4 or 5	.081	.000	.153	.044	.000	.083	.017	.000	.110
M3	1	.073	.543	.223	.013	.220	.060	.000	.003	.000
	2	**.507**	**.457**	**.387**	**.703**	**.780**	**.770**	**.810**	**.997**	**.903**
	3	.200	.000	.050	.160	.000	.020	.110	.000	.010
	4 or 5	.220	.000	.340	.124	.000	.150	.080	.000	.087
M4	1	.077	.510	.203	.017	.177	.070	.000	.000	.000
	2	**.390**	**.487**	**.413**	**.703**	**.823**	**.767**	**.790**	**1.00**	**.943**
	3	.230	.003	.043	.143	.000	.010	.100	.000	.010
	4 or 5	.303	.000	.341	.137	.000	.153	.110	.000	.047

[1] Each criterion is computed based on the MPCLE obtained by the ADALASSO penalty with $q = 10$.

Table A3: Average proportion of times (in 300 replications) that a number of AR-regimes $1 \leq K \leq 5$ is selected by a criterion[1]. Results for the true order $K = 2$ are in **bold**.

		$n = 150$			$n = 250$			$n = 500$		
Models	K	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC
M1	1	.201	.752	.205	.057	.371	.072	.004	.031	.004
	2	**.338**	**.248**	**.245**	**.583**	**.621**	**.614**	**.816**	**.969**	**.825**
	3	.169	.000	.076	.155	.008	.068	.085	.000	.013
	4 or 5	.292	.000	.474	.205	.000	.246	.095	.000	.158
M2	1	.000	.064	.011	.000	.004	.000	.000	.000	.000
	2	**.745**	**.929**	**.805**	**.892**	**.989**	**.848**	**.941**	**1.00**	**.823**
	3	.163	.007	.025	.090	.007	.051	.051	.000	.042
	4 or 5	.092	.000	.159	.018	.000	.101	.008	.000	.135
M3	1	.113	.697	.353	.027	.317	.137	.000	.013	.000
	2	**.470**	**.303**	**.210**	**.660**	**.683**	**.610**	**.867**	**.987**	**.800**
	3	.203	.000	.057	.160	.000	.040	.083	.000	.050
	4 or 5	.214	.000	.380	.153	.000	.213	.050	.000	.150
M4	1	.110	.633	.267	053	.330	.147	.000	.013	.003
	2	**.380**	**.367**	**.197**	**.587**	**.667**	**.527**	**.807**	**.987**	**.840**
	3	.233	.000	.017	.157	.003	.030	.117	.000	.023
	4 or 5	.277	.000	.519	.203	.000	.296	.076	.000	.134

[1] Each criterion is computed based on the MPCLE obtained by the LASSO penalty with $q = 10$.

Table A4: Average (standard deviation), over 300 replications, of estimated Sensitivity (ES1) and Specificity (ES2) in model **M5**: $K = 3$ and $q = 10$.

	MSAR	n=250		n=500		n=800	
Penalty	Regimes	ES1	ES2	ES1	ES2	ES1	ES2
LASSO	Reg$_1$	.930$_{(.103)}$	.685$_{(.408)}$	.966$_{(.062)}$	.830$_{(.345)}$	.980$_{(.047)}$	.977$_{(.120)}$
	Reg$_2$	.979$_{(.052)}$	.928$_{(.259)}$	.996$_{(.023)}$	.993$_{(.085)}$	1.00$_{(.000)}$	1.00$_{(.000)}$
	Reg$_3$	.967$_{(.072)}$	.779$_{(.393)}$	.985$_{(.044)}$	.978$_{(.139)}$	.994$_{(.027)}$	.998$_{(.030)}$
ADALASSO	Reg$_1$	.975$_{(.066)}$	.829$_{(.326)}$	.992$_{(.034)}$	.968$_{(.149)}$	1.00$_{(.000)}$	.993$_{(.072)}$
	Reg$_2$	.991$_{(.038)}$	.959$_{(.199)}$	.999$_{(.011)}$	1.00$_{(.000)}$	1.00$_{(.000)}$	1.00$_{(.000)}$
	Reg$_3$	.976$_{(.065)}$	.959$_{(.161)}$	.995$_{(.024)}$	1.00$_{(.000)}$	1.00$_{(.007)}$	1.00$_{(.000)}$
SCAD	Reg$_1$	.977$_{(.077)}$	.909$_{(.241)}$	.997$_{(.020)}$	.987$_{(.078)}$	.999$_{(.017)}$	.998$_{(.030)}$
	Reg$_2$	.994$_{(.029)}$	.976$_{(.153)}$	.998$_{(.013)}$	1.00$_{(.000)}$	1.00$_{(.000)}$	1.00$_{(.000)}$
	Reg$_3$	.989$_{(.051)}$	.979$_{(.108)}$	.998$_{.(015)}$	.996$_{(.060)}$	.997$_{(.023)}$	1.00$_{(.000)}$

Table A5: Average proportion of times (in 300 replications) that a number of AR-regimes $1 \leq K \leq 5$ is selected by a criterion[1]. Results for the true order $K = 3$ are in **bold**.

		$n = 250$			$n = 500$			$n = 800$		
Model	K	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC
M5	1	.000	.007	.000	.000	.000	.000	.000	.000	.000
	2	.084	.650	.303	.003	.182	.020	.000	.003	.000
	3	**.714**	**.343**	**.609**	**.814**	**.814**	**.932**	**.953**	**.997**	**.983**
	4 or 5	.202	.000	.088	.183	.004	.048	.047	.000	.017

[1] Each criterion is computed based on the MPCLE obtained by the ADALASSO penalty with $q = 10$.

Table A6: Average proportion of times (in 300 replications) that a number of AR-regimes $1 \leq K \leq 5$ is selected by a criterion[1]. Results for the true order $K = 3$ are in **bold**.

		$n = 250$			$n = 500$			$n = 800$		
Model	K	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC
M5	1	.000	.007	.003	.000	.000	.000	.000	.000	.000
	2	.215	.842	.684	.057	.574	.233	.003	.127	.013
	3	**.471**	**.151**	**.175**	**.659**	**.426**	**.713**	**.793**	**.873**	**.970**
	4 or 5	.314	.000	.138	.284	.000	.054	.204	.000	.017

[1] Each criterion is computed based on the MPCLE obtained by the LASSO penalty with $q = 10$.

Table A7: Average proportion of times (in 300 replications) that a number of AR-regimes $1 \leq K \leq 5$ is selected by a criterion[1]. Results for the true order $K = 3$ are in **bold**.

		$n = 250$			$n = 500$			$n = 800$		
Model	K	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC
M5	1	.000	.007	.000	.000	.000	.000	.000	.000	.000
	2	.037	.592	.209	.000	.145	.007	.000	.003	.000
	3	**.667**	**.401**	**.667**	**.791**	**.851**	**.916**	**.866**	**.997**	**.980**
	4 or 5	.296	.000	.124	.209	.004	.077	.134	.000	.020

[1] Each criterion is computed based on the MPCLE obtained by the SCAD penalty with $q = 10$.

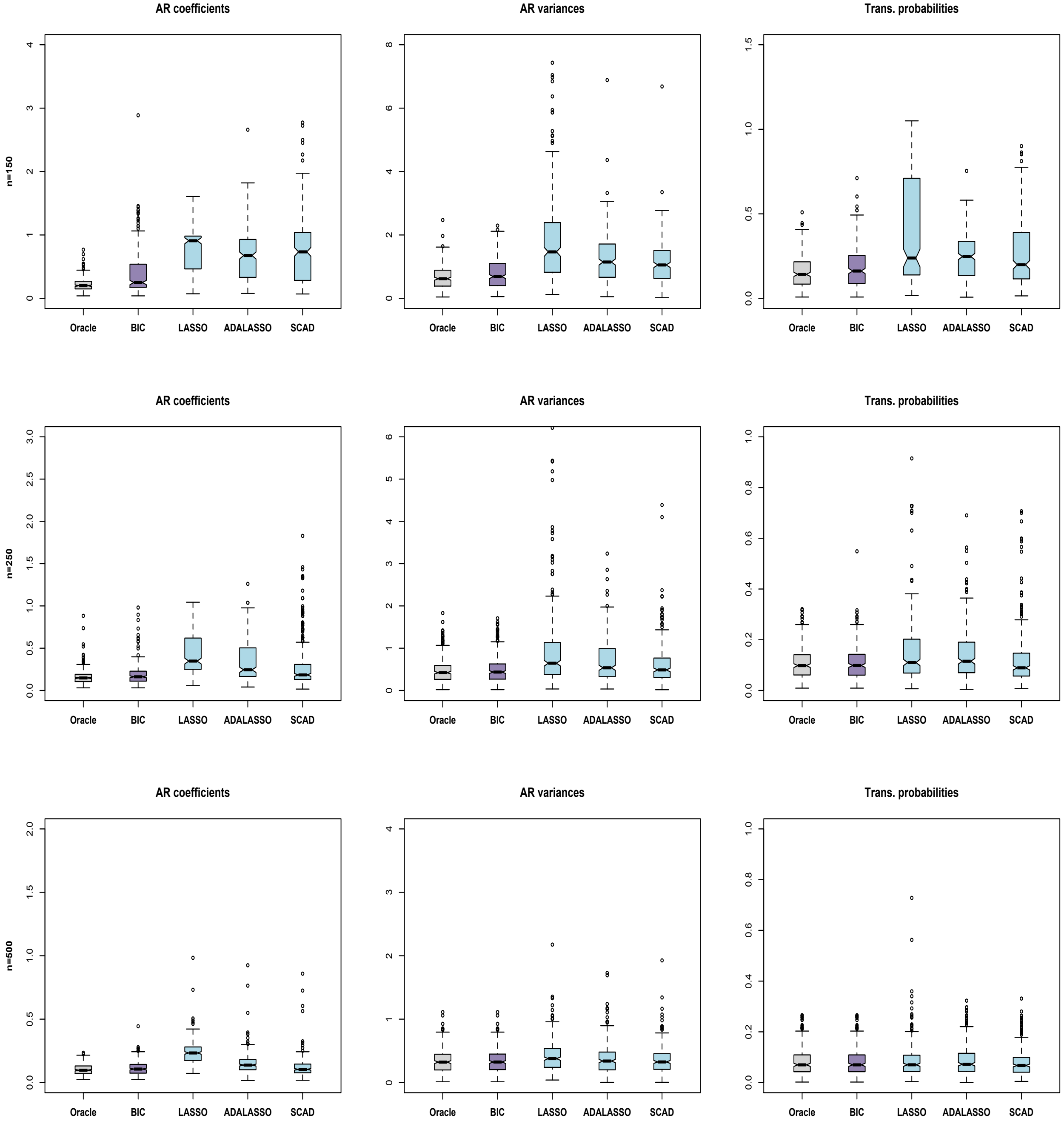

Figure A1: Model **M1**: Empirical L_2-losses of the estimates of AR coefficients, variances, and transition probabilities (represented by the columns), based on the oracle, BIC with $q = 5$, and LASSO, ADALASSO, SCAD with $q = 10$, and sample sizes $n = 150, 250, 500$ (represented by the rows).

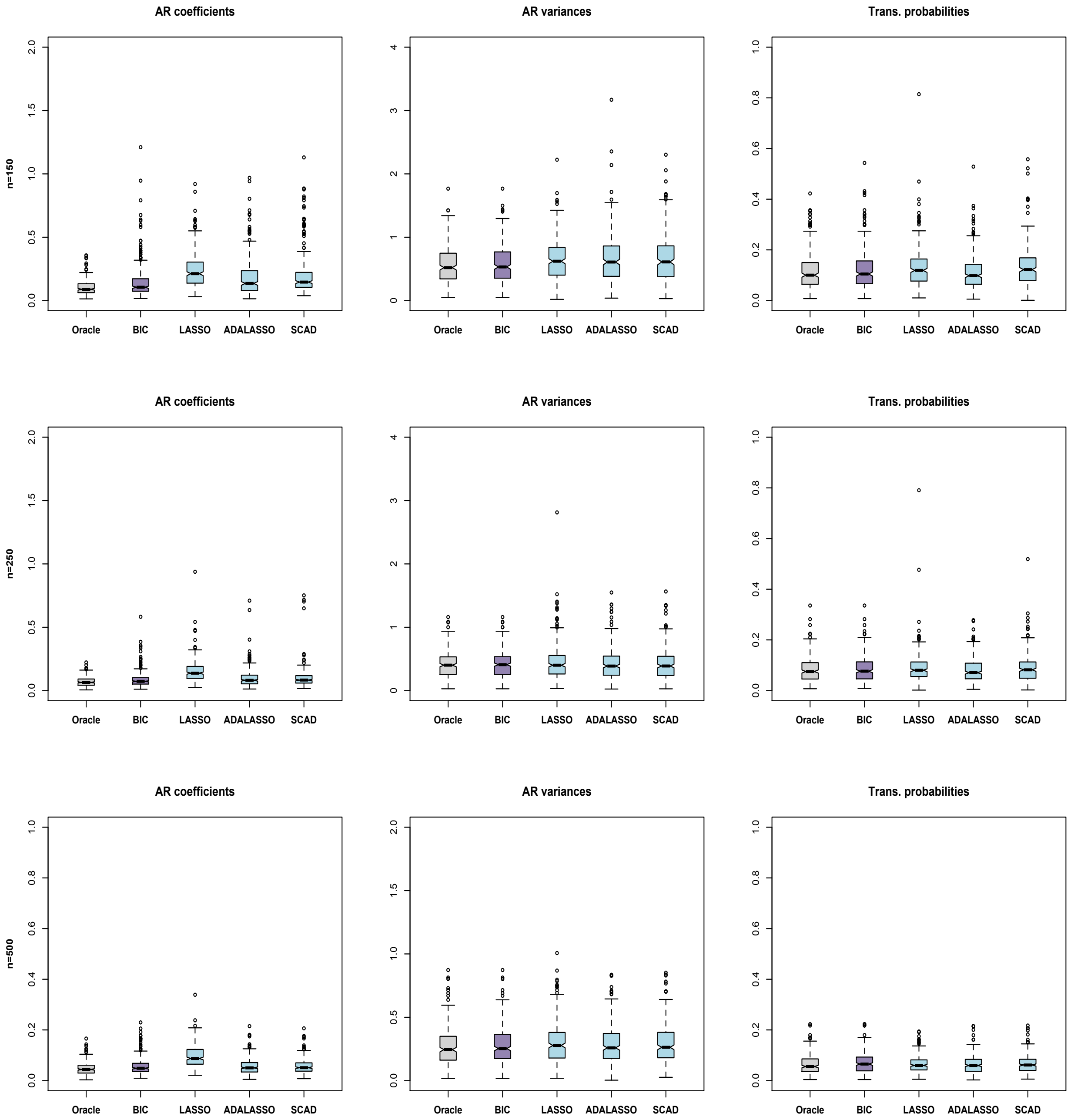

Figure A2: Model **M2**: Empirical L_2-losses of the estimates of AR coefficients, variances, and transition probabilities (represented by the columns), based on the oracle, BIC with $q = 5$, and LASSO, ADALASSO, SCAD with $q = 10$, and sample sizes $n = 150, 250, 500$ (represented by the rows).

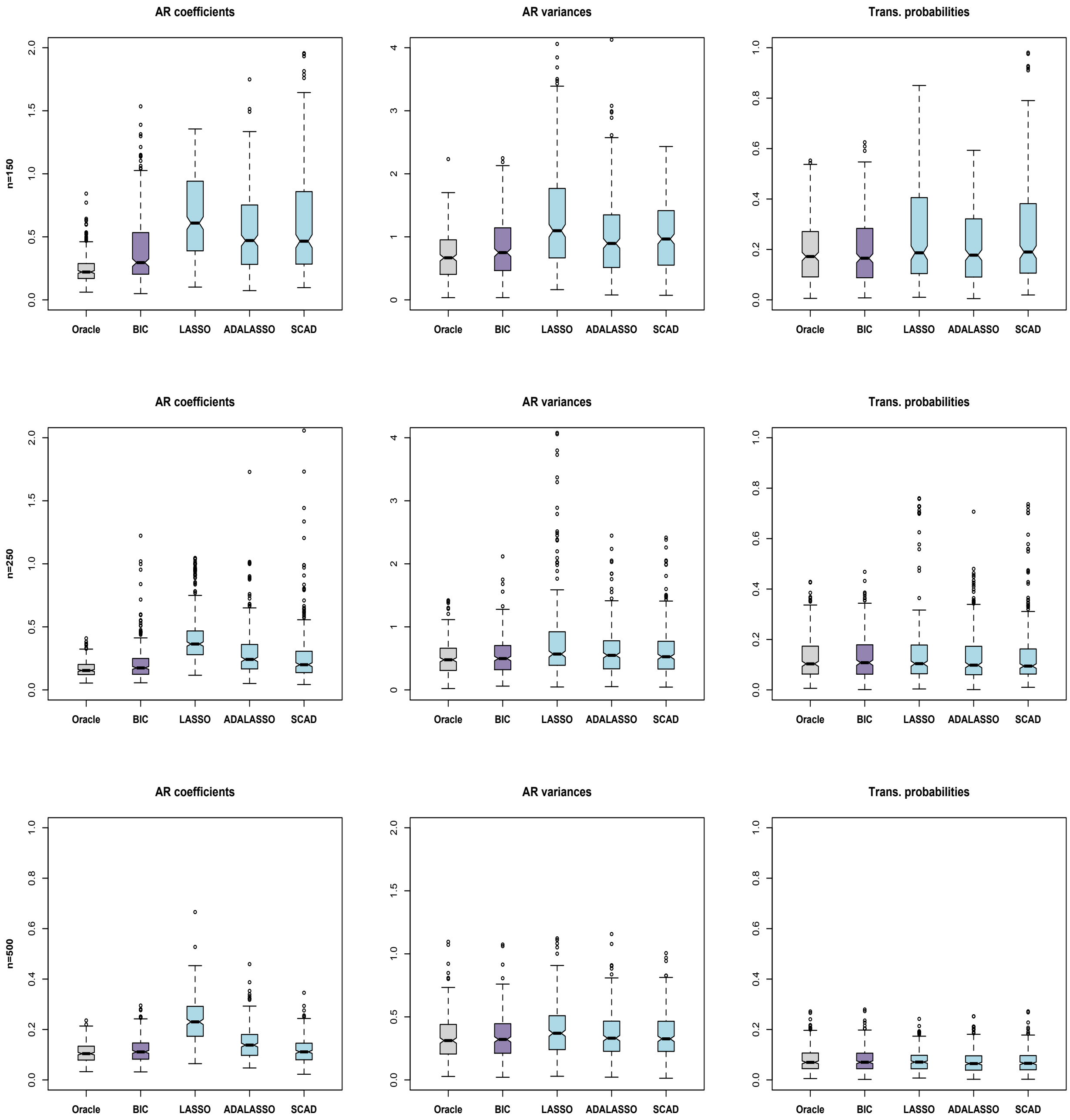

Figure A3: Model **M3**: Empirical L_2-losses of the estimates of AR coefficients, variances, and transition probabilities (represented by the columns), based on the oracle, BIC with $q = 6$, and LASSO, ADALASSO, SCAD with $q = 10$, and sample sizes $n = 150, 250, 500$ (represented by the rows)..

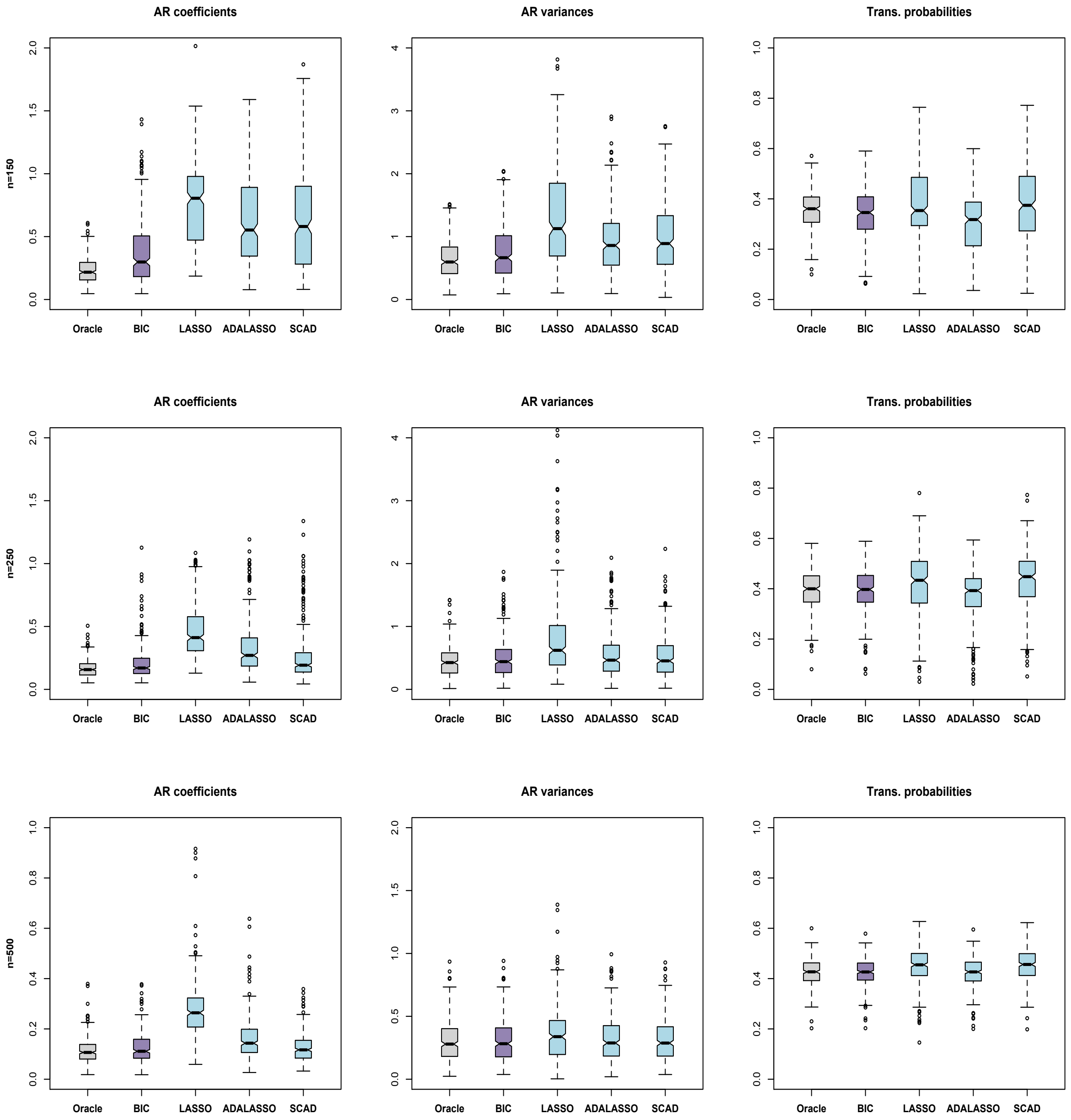

Figure A4: Model **M4**: Empirical L_2-losses of the estimates of AR coefficients, variances, and transition probabilities (represented by the columns), based on the oracle, BIC with $q = 6$, and LASSO, ADALASSO, SCAD with $q = 10$, and sample sizes $n = 150, 250, 500$ (represented by the rows).

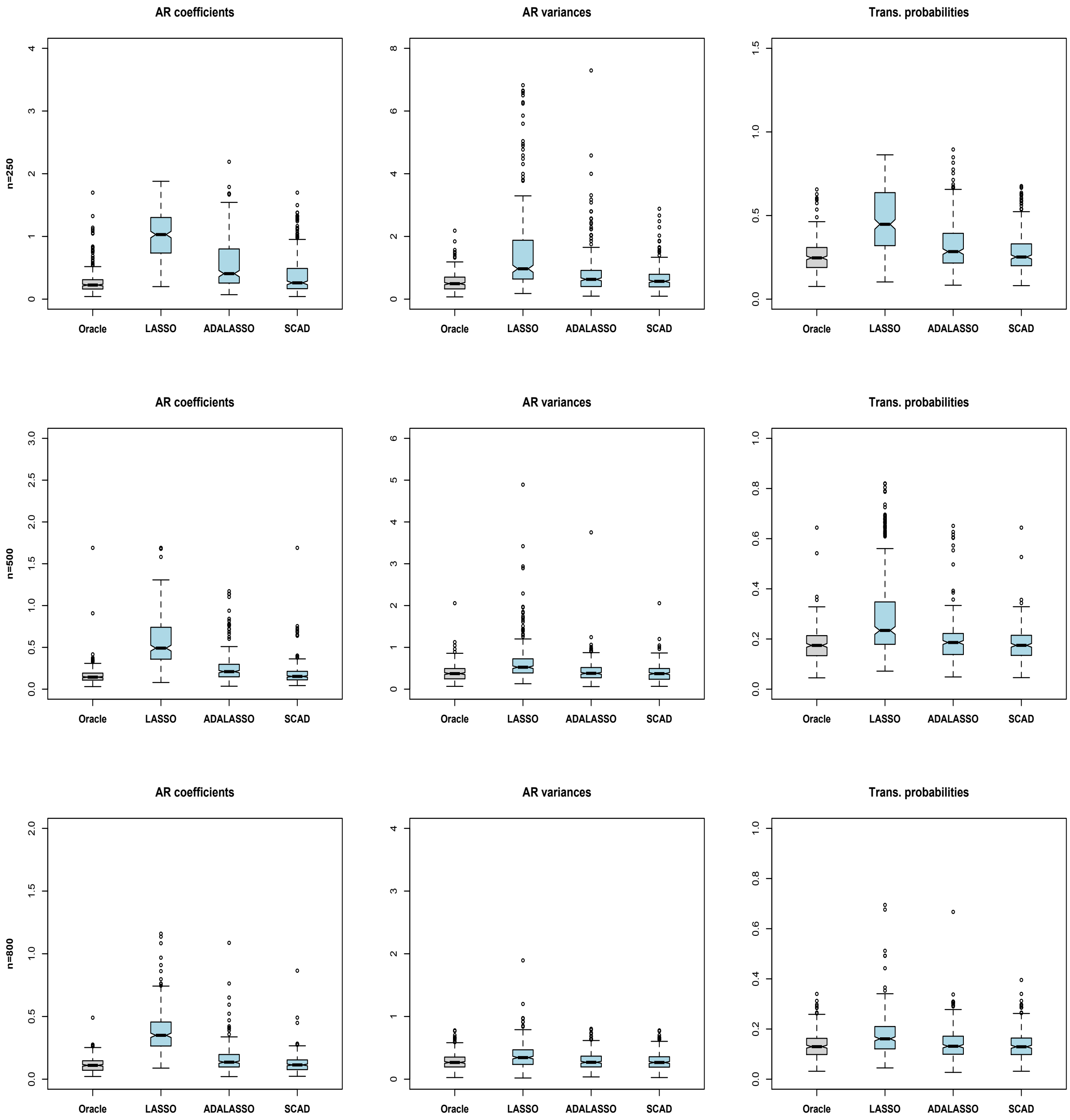

Figure A5: Model **M5**: Empirical L_2-losses of the estimates of AR coefficients, variances, and transition probabilities (represented by the columns), based on the oracle, and LASSO, ADALASSO, SCAD with $q = 10$, and sample sizes $n = 250, 500, 800$ (represented by the rows).

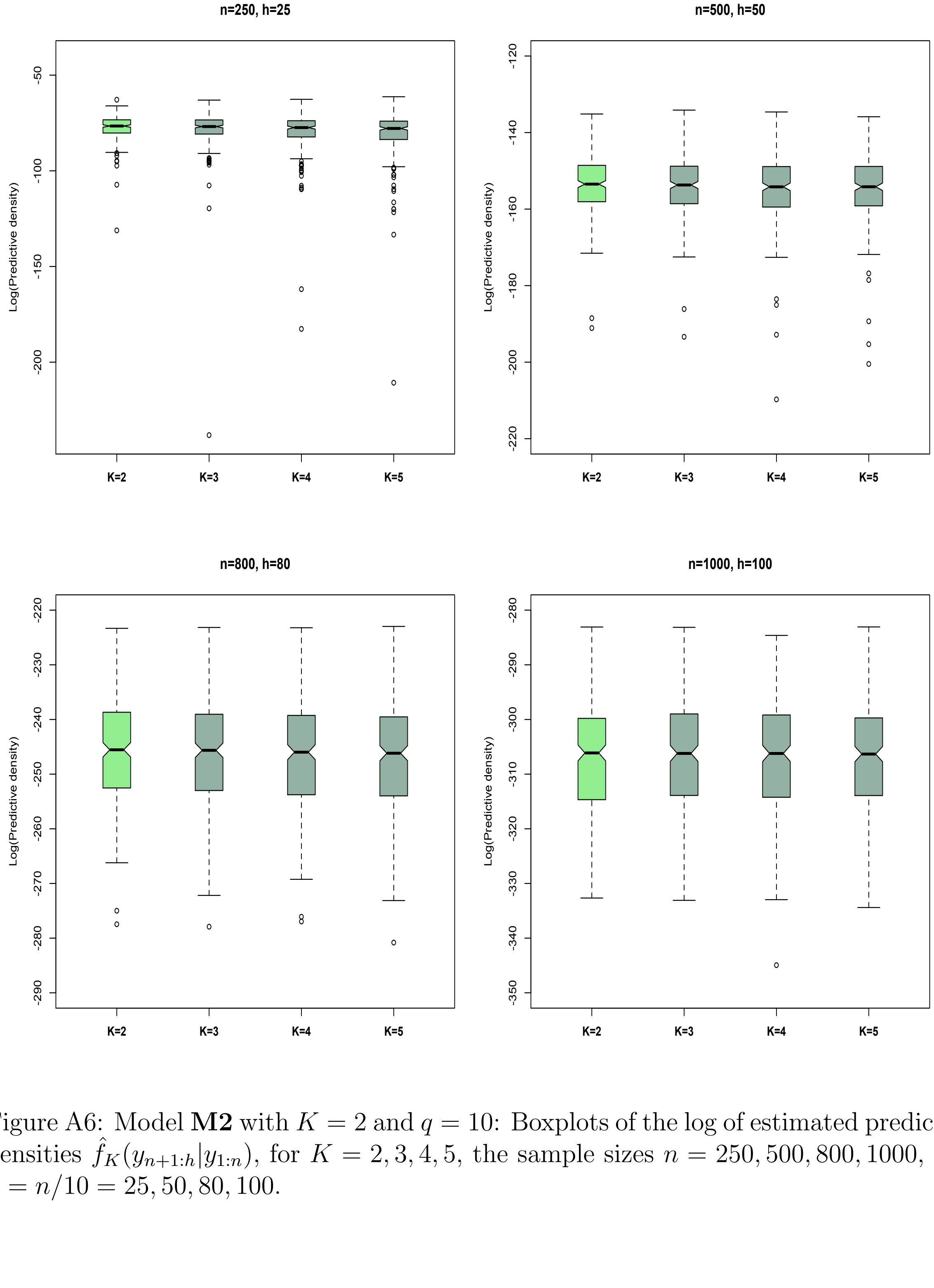

Figure A6: Model **M2** with $K = 2$ and $q = 10$: Boxplots of the log of estimated predictive densities $\hat{f}_K(y_{n+1:h}|y_{1:n})$, for $K = 2, 3, 4, 5$, the sample sizes $n = 250, 500, 800, 1000$, and $h = n/10 = 25, 50, 80, 100$.

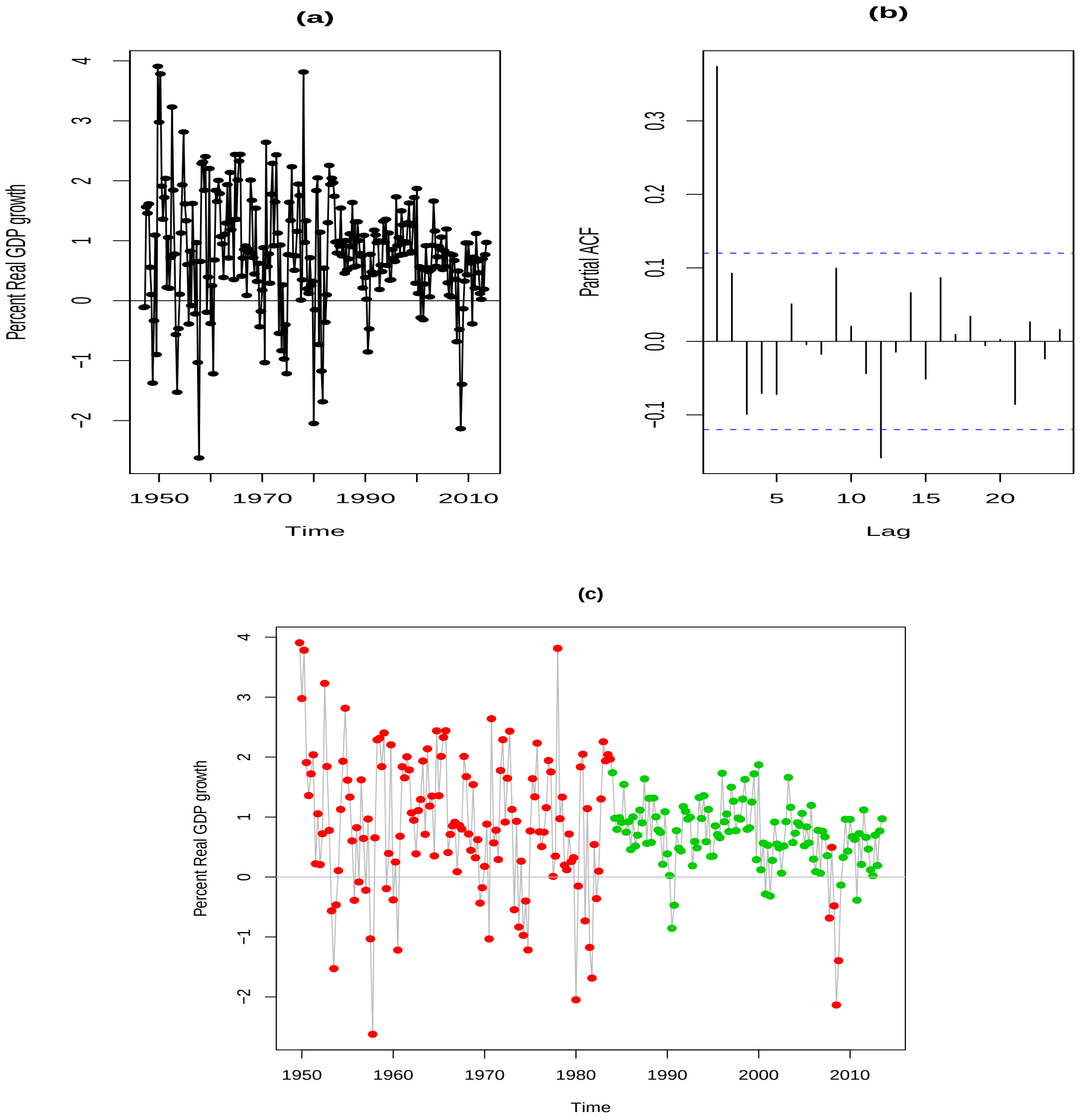

Figure A7: (a) Time series plot of the U.S. quarterly GDP data over the period of 1947-2013. (b) the sample PACF plot of the series. (c) classified $y_t, t = 16, \ldots, 267$, into regime 1 (green) and regime 2 (red) based on the fitted MSAR model using SCAD.

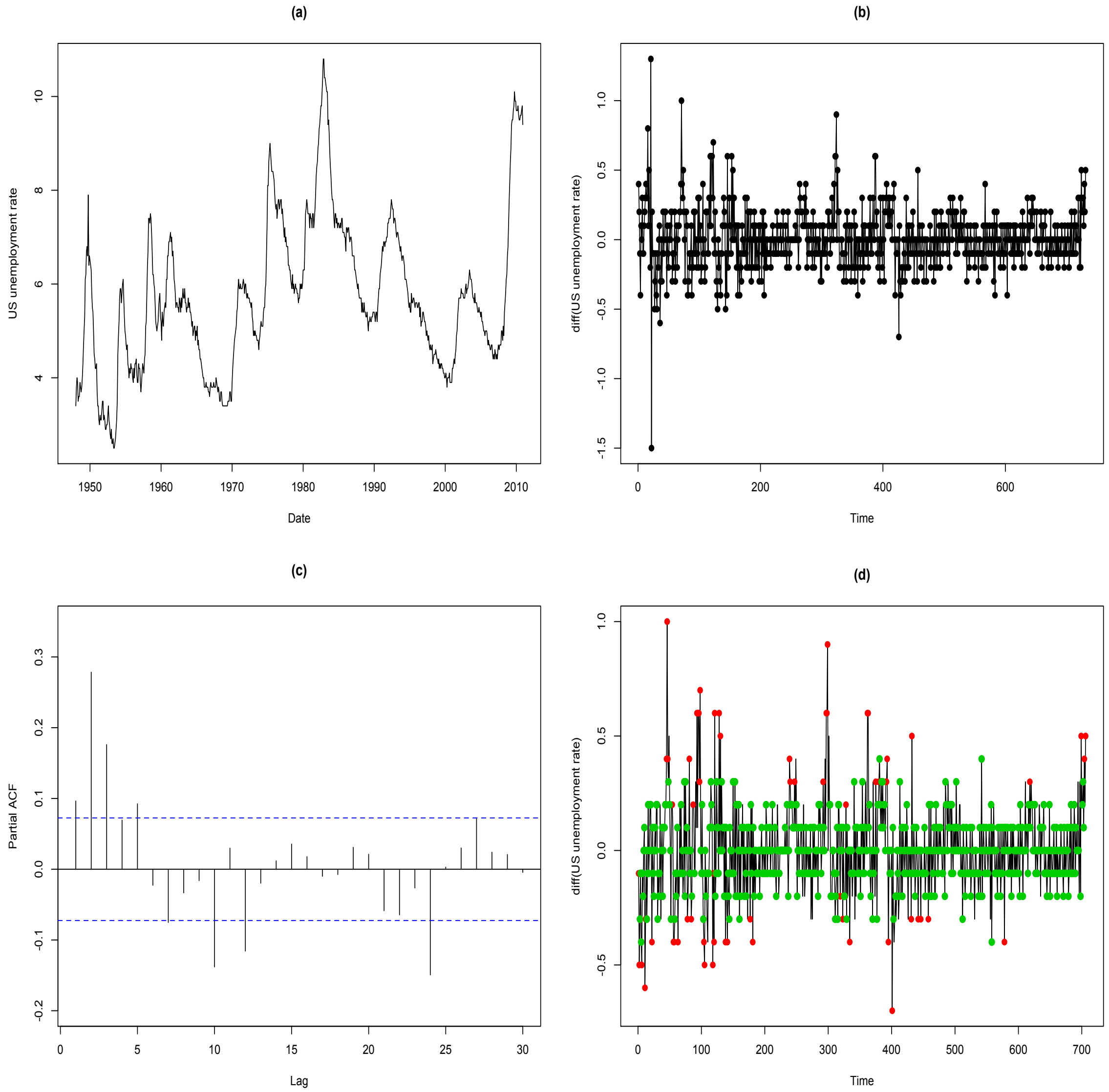

Figure A8: (a) Time series plot of the monthly U.S. unemployment rates over the period of 1948 to 2010. (b) time series plot of the first differences of the series. (c) the sample PACF plot of the first differences. (d) classified $x_t, t = 26, \ldots, 731$, into regime 1 (green) and regime 2 (red) based on the fitted MSAR model using SCAD.

Table A8: Average (standard deviation), over 300 replications, of estimated Sensitivity (ES1) and Specificity (ES2)[1].

		MSAR	$n = 150$		$n = 250$		$n = 500$	
	Model	Regimes	ES1	ES2	ES1	ES2	ES1	ES2
BIC	**M1**	Reg$_1$	$.950_{(.137)}$	$.905_{(.217)}$	$.970_{(.096)}$	$.985_{(.085)}$	$.983_{(.073)}$	$1.00_{(.000)}$
		Reg$_2$	$.929_{(.173)}$	$.908_{(.210)}$	$.972_{(.092)}$	$.992_{(.076)}$	$.980_{(.079)}$	$1.00_{(.000)}$
	M2	Reg$_1$	$.961_{(.114)}$	$.989_{(.072)}$	$.980_{(.085)}$	$1.00_{(.000)}$	$.989_{(.059)}$	$1.00_{(.000)}$
		Reg$_2$	$.977_{(.094)}$	$.998_{(.030)}$	$.987_{(.071)}$	$1.00_{(.000)}$	$.991_{(.055)}$	$1.00_{(.000)}$
	M3	Reg$_1$	$.920_{(.129)}$	$.985_{(.085)}$	$.960_{(.096)}$	$1.00_{(.000)}$	$.982_{(.071)}$	$1.00_{(.000)}$
		Reg$_2$	$.904_{(.167)}$	$.933_{(.142)}$	$.959_{(.110)}$	$.986_{(.068)}$	$.986_{(.068)}$	$1.00_{(.000)}$
	M4	Reg$_1$	$.929_{(.123)}$	$.940_{(.163)}$	$.963_{(.096)}$	$.998_{(.029)}$	$.988_{(.054)}$	$1.00_{(.000)}$
		Reg$_2$	$.940_{(.128)}$	$.962_{(.106)}$	$.970_{(.096)}$	$.994_{(.043)}$	$.979_{(.081)}$	$1.00_{(.000)}$
LASSO	**M1**	Reg$_1$	$.893_{(.179)}$	$.760_{(.374)}$	$.977_{(.081)}$	$.967_{(.149)}$	$.997_{(.029)}$	$1.00_{(.000)}$
		Reg$_2$	$.908_{(.165)}$	$.741_{(.355)}$	$.971_{(.085)}$	$.945_{(.172)}$	$.996_{(.032)}$	$.998_{(.029)}$
	M2	Reg$_1$	$.938_{(.127)}$	$.971_{(.136)}$	$.969_{(.096)}$	$.992_{(.076)}$	$.996_{(.032)}$	$.998_{(.029)}$
		Reg$_2$	$.968_{(.095)}$	$.970_{(.138)}$	$.993_{(.043)}$	$.987_{(.107)}$	$.998_{(.029)}$	$.997_{(.058)}$
	M3	Reg$_1$	$.883_{(.176)}$	$.878_{(.315)}$	$.925_{(.130)}$	$.975_{(.148)}$	$.971_{(.083)}$	$1.00_{(.000)}$
		Reg$_2$	$.851_{(.223)}$	$.880_{(.248)}$	$.926_{(.168)}$	$.968_{(.116)}$	$.979_{(.090)}$	$.999_{(.019)}$
	M4	Reg$_1$	$.879_{(.192)}$	$.710_{(.434)}$	$.923_{(.146)}$	$945_{(.227)}$	$.962_{(.107)}$	$1.00_{(.000)}$
		Reg$_2$	$.859_{(.214)}$	$.900_{(.194)}$	$.909_{(.180)}$	$.988_{(.062)}$	$.979_{(.081)}$	$1.00_{(.000)}$
ADALASSO	**M1**	Reg$_1$	$.923_{(.167)}$	$.843_{(.287)}$	$.977_{(.081)}$	$.967_{(.149)}$	$.997_{(.029)}$	$1.00_{(.000)}$
		Reg$_2$	$.934_{(.156)}$	$.778_{(.327)}$	$.971_{(.085)}$	$.945_{(.172)}$	$.996_{(.032)}$	$.998_{(.029)}$
	M2	Reg$_1$	$.959_{(.107)}$	$.961_{(.151)}$	$.986_{(.065)}$	$.987_{(.090)}$	$1.00_{(.000)}$	$.998_{(.029)}$
		Reg$_2$	$.985_{(.069)}$	$.972_{(.136)}$	$.995_{(.041)}$	$.990_{(.081)}$	$.999_{(.014)}$	$.998_{(.029)}$
	M3	Reg$_1$	$.927_{(.154)}$	$.942_{(.210)}$	$.971_{(.090)}$	$995_{(.064)}$	$.999_{(.014)}$	$1.00_{(.000)}$
		Reg$_2$	$.880_{(.210)}$	$.878_{(.221)}$	$.953_{(.136)}$	$.961_{(.126)}$	$.997_{(.033)}$	$.999_{(.019)}$
	M4	Reg$_1$	$.919_{(.151)}$	$.841_{(330)}$	$958_{(.109)}$	$.970_{(.161)}$	$.998_{(.020)}$	$.997_{(.058)}$
		Reg$_2$	$.920_{(.171)}$	$.914_{(.176)}$	$.962_{(.122)}$	$.988_{(.079)}$	$.997_{(.033)}$	$1.00_{(.000)}$
SCAD	**M1**	Reg$_1$	$.930_{(.165)}$	$.857_{(.276)}$	$.979_{(.078)}$	$.972_{(.136)}$	$.991_{(.047)}$	$1.00_{(.000)}$
		Reg$_2$	$.937_{(.160)}$	$.802_{(.303)}$	$.971_{(.088)}$	$.953_{(.157)}$	$.992_{(.045)}$	$1.00_{(.000)}$
	M2	Reg$_1$	$.958_{(.115)}$	$.965_{(.140)}$	$.988_{(.069)}$	$.985_{(.095)}$	$1.00_{(.000)}$	$.998_{(.029)}$
		Reg$_2$	$.973_{(.101)}$	$.975_{(.124)}$	$.993_{(.052)}$	$.988_{(.104)}$	$.998_{(.029)}$	$.997_{(.058)}$
	M3	Reg$_1$	$.943_{(.127)}$	$.855_{(.308)}$	$.971_{(.092)}$	$.981_{(.131)}$	$.993_{(.045)}$	$1.00_{(.000)}$
		Reg$_2$	$.922_{(.172)}$	$.918_{(.176)}$	$.963_{(.121)}$	$.994_{(.043)}$	$.986_{(.073)}$	$1.00_{(.000)}$
	M4	Reg$_1$	$.943_{(.127)}$	$.855_{(.308)}$	$.971_{(.092)}$	$.982_{(.131)}$	$.993_{(.045)}$	$1.00_{(.000)}$
		Reg$_2$	$.922_{(.172)}$	$.918_{(.176)}$	$.963_{(.121)}$	$.994_{(.043)}$	$.986_{(.073)}$	$1.00_{(.000)}$

[1] For all the methods, we used $q = 5$ and 6 for models **M1**–**M2** and **M3**–**M4**, respectively.

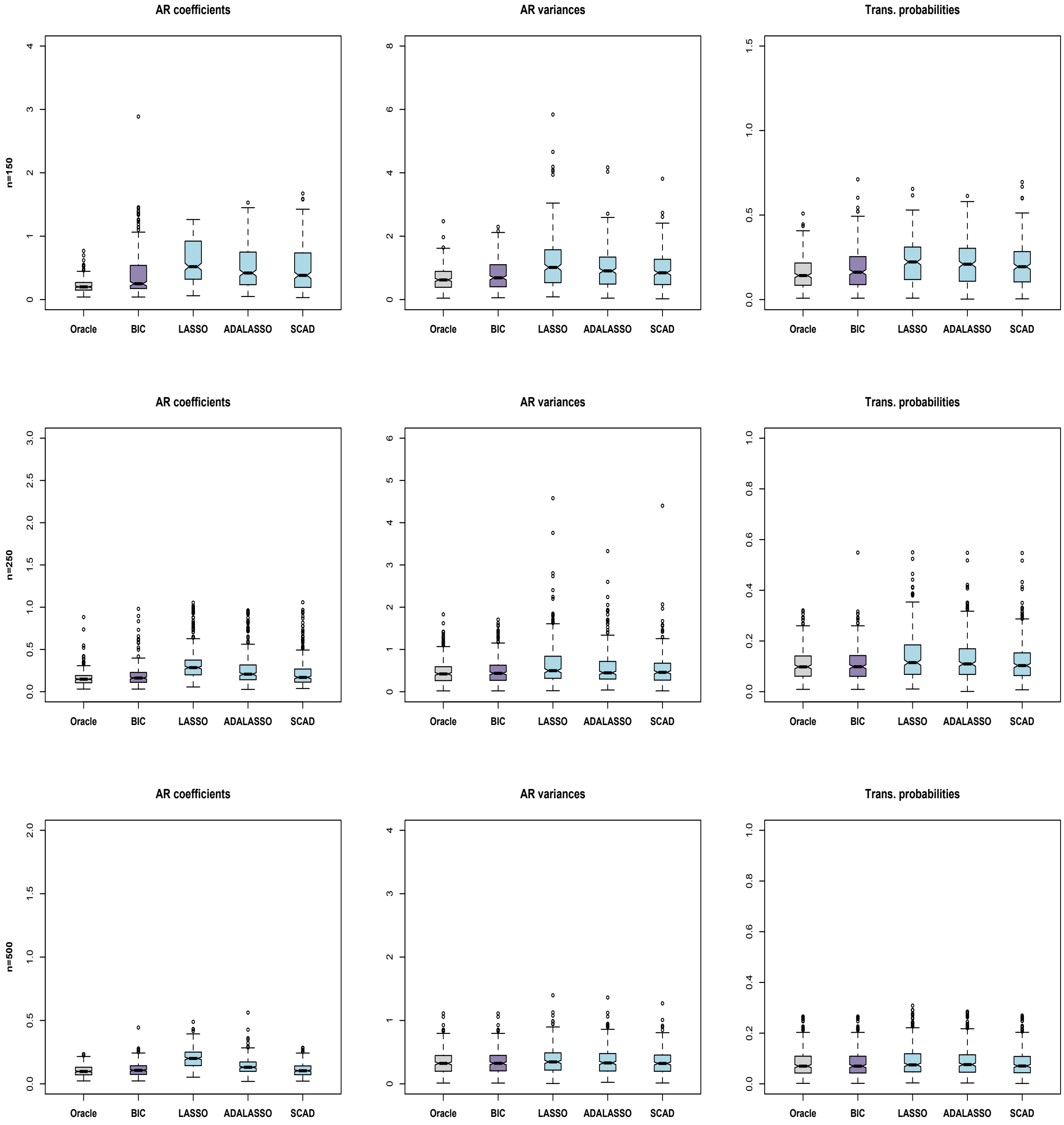

Figure A9: Model **M1**: The empirical L_2-losses for the oracle, BIC, LASSO, ADALASSO, and SCAD estimates of AR coefficients, variances, and transition probabilities (represented by the columns), sample sizes $n = 150, 250, 500$ (represented by the rows), and $q = 5$.

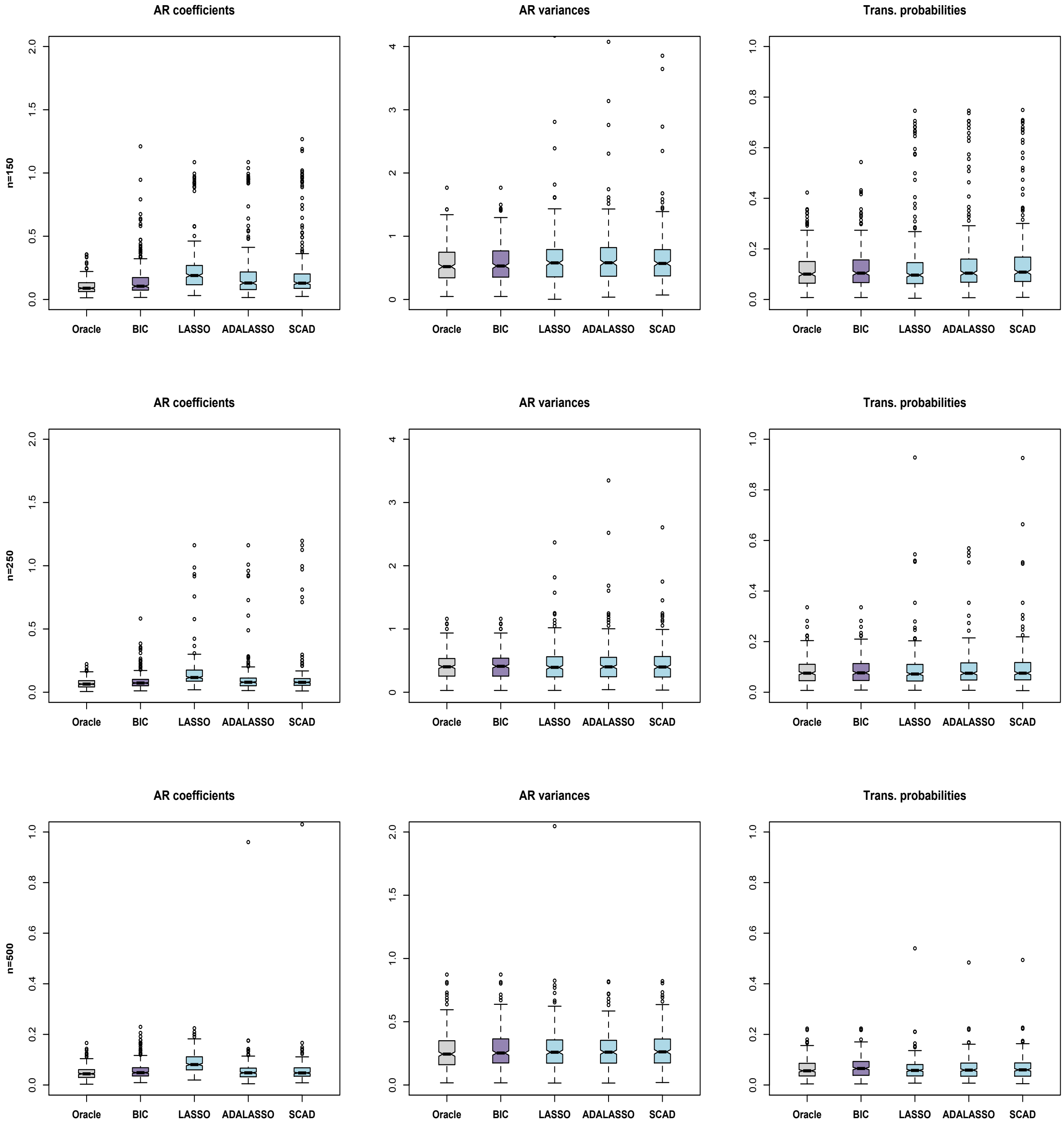

Figure A10: Model **M2**: The empirical L_2-losses for the oracle, BIC, LASSO, ADALASSO, and SCAD estimates of AR coefficients, variances, and transition probabilities (represented by the columns), sample sizes $n = 150, 250, 500$ (represented by the rows), and $q = 5$.

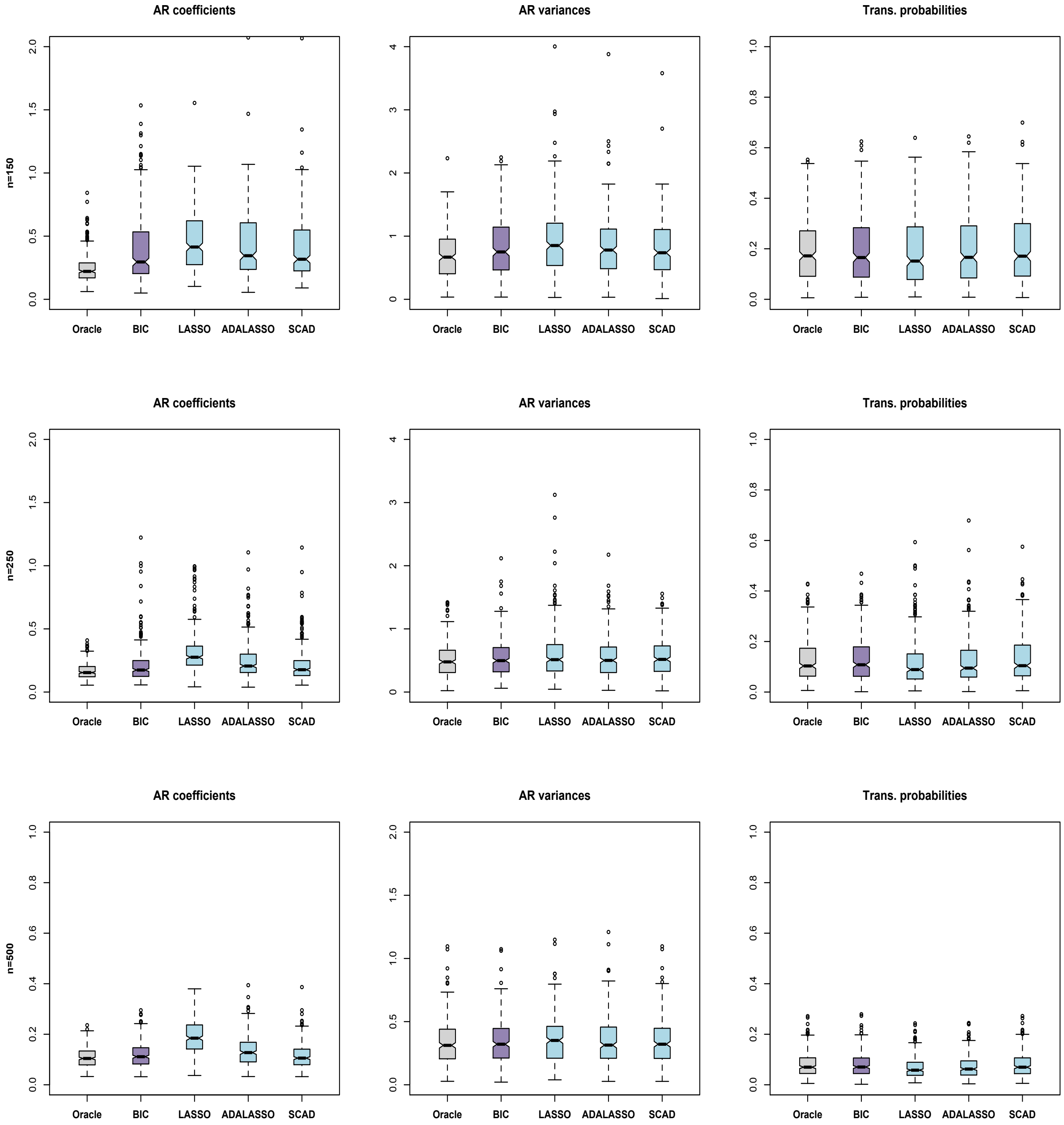

Figure A11: Model **M3**: The empirical L_2-losses for the oracle, BIC, LASSO, ADALASSO, and SCAD estimates of AR coefficients, variances, and transition probabilities (represented by the columns), sample sizes $n = 150, 250, 500$ (represented by the rows), and $q = 6$.

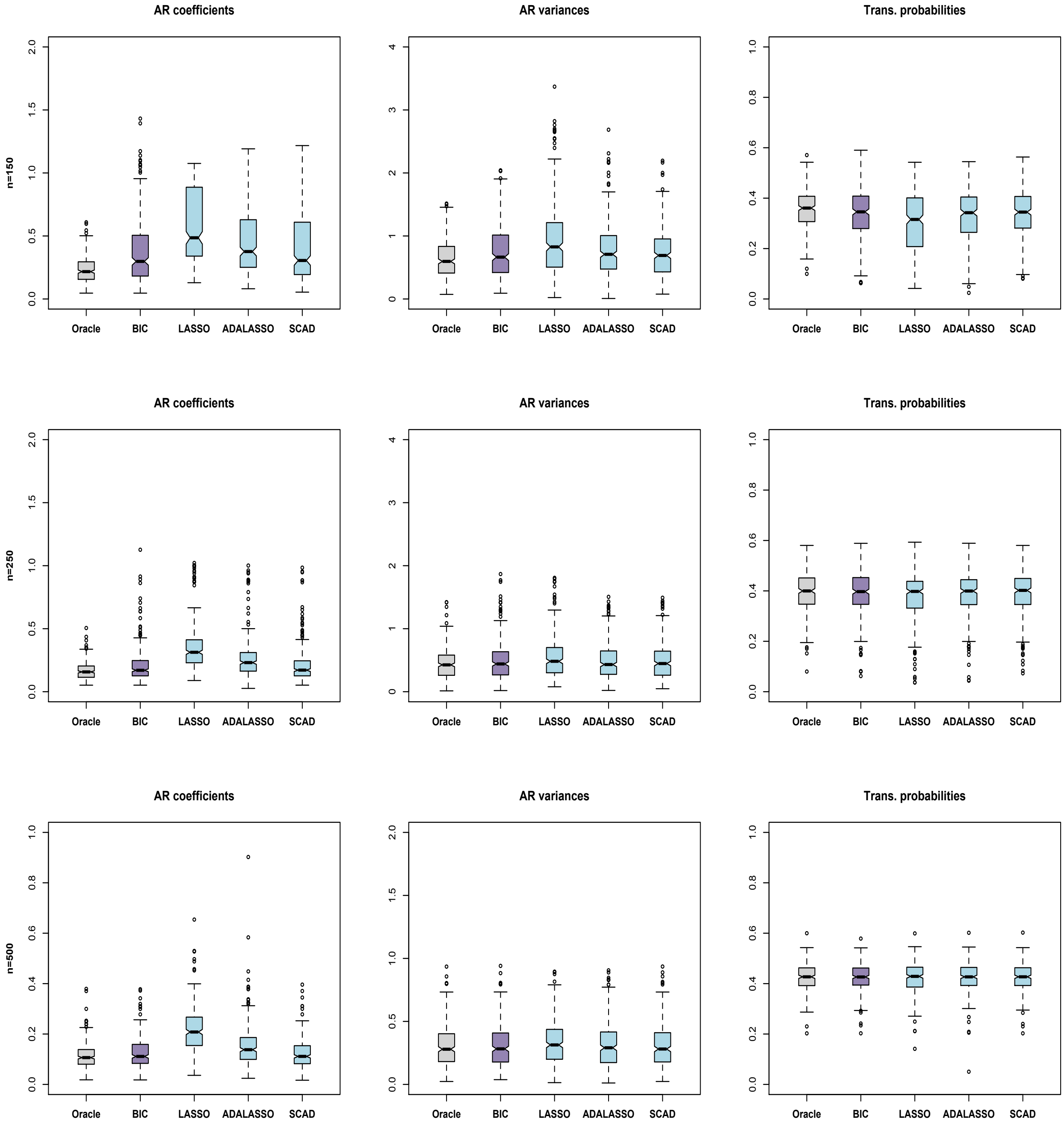

Figure A12: Model **M4**: The empirical L_2-losses for the oracle, BIC, LASSO, ADALASSO, and SCAD estimates of AR coefficients, variances, and transition probabilities (represented by the columns), sample sizes $n = 150, 250, 500$ (represented by the rows), and $q = 6$.