Statistica Sinica Preprint No: SS-2019-0190	
Title	Sparseness, consistency and model selection for Markov regime-switching Gaussian autoregressive models
Manuscript ID	SS-2019-0190
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0190
Complete List of Authors	Abbas Khalili and David A. Stephens
Corresponding Author	Abbas Khalili
E-mail	abbas.khalili@mcgill.ca
Notice: Accepted version subject to English editing.	

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

Abstract: We study Markov regime-switching Gaussian autoregressive models which are aimed at capturing temporal heterogeneity exhibited by time series data. In the construction of a Markov regime-switching model, several specifications must be made relating to both the state and observation models; in particular, the complexity of these models must be specified when fitting to a dataset. We propose new regularization methods based on conditional likelihood for simultaneous autoregressive-order and parameter estimation with the number of regimes fixed, and use a regularized Bayesian information criterion for selection of the number of regimes. Unlike the existing information-theoretic approaches, the new methods avoid an exhaustive search of the model space for model selection and thereby are computationally more efficient. We establish large sample properties of the proposed methods for estimation, model selection, and forecasting. We also evaluate finite sample performance of the methods via simulations, and illustrate their applications by analyzing two real datasets.

Key words: Autoregressive models, Markov regime-switching models, Information criteria, Regularization methods, EM algorithm.

1. Introduction

Markov regime-switching models (Hamilton, 1989) are commonly used to incorporate latent structure in time series with the goal of capturing non-stationarity or time-inhomogeneity of real data. There is an extensive literature that discusses the use of these models in econometrics, and many applications relate to representation of economic or business cycles (Hamilton, 2016). Other applications include speech recognition and neurobiology (Krishnamurthy and Yin, 2002).

In a Markov regime-switching model, typically a discrete-state and often first-order Markov 'state' model is used to capture unobserved stochastic variation corresponding to regime changes, and conditional on the latent structure a conventional time series 'observation' model is used to represent the observed data. In practice, complexity of the model – the number of regimes (states) and structure of each regime-specific observation model – must be specified. In this paper, we develop new results based on regularized conditional likelihood that demonstrate that sparse estimation for such two-stage models consistently estimate the parameters of the presumed model under mild conditions. We also establish certain model selection consistency results, including forecasting consistency. Although our technical results apply under general modelling assumptions, our development and exposition focus on Markov regime-switching autoregressive (MSAR) models with Gaussian errors.

A Gaussian MSAR model postulates the existence of a latent process $\{S_t : t =$

$1, 2, \ldots\}$ on a finite set $\{1, \ldots, K\}$ that determines for each time t the Gaussian autoregressive regime that dictates the stochastic behaviour of an observable discrete-time series $\{Y_t : t = 1, 2, \ldots\}$. Specifically, S_t is presumed a first-order Markov chain parameterized through a transition matrix $\mathbb{P}$, and conditional on $S_t = j$, the distribution of Y_t depends on the lagged Y's, say, $Y_{t-1}, \ldots, Y_{t-q_j}$, for some q_j. Such models, in comparison to standard Gaussian autoregressive (AR) processes, are particularly useful when the data exhibit heterogeneity in conditional mean or autocovariance structure.

Maximum likelihood estimation (MLE) is typically used for inference implemented via adaptations of filtering and smoothing using forward-backward algorithms in MSAR models (Frühwirth-Schnatter, 2006; Baum et al., 1970). Krishnamurthy and Rydén (1998), and Douc et al. (2004, 2011) establish consistency and asymptotic normality of MLE when the model complexity – a common AR-order (q) across the regimes and the number of AR-regimes (K) – is fixed. In real applications, however, there may be latent external factors (policy changes, macroeconomic conditions, etc) that dictate that different AR-regimes are in operation, and that these regimes may have different stochastic characteristics as manifested in their mean level, variance or autocovariance. For example, an economy under one regime may be subject to more persistent effects of shock than when under another regime. Hence, our inferential interest centers on the choice of potentially different regime-specific AR-orders $q_1, \ldots, q_K$, the number of AR-regimes K, estimation of AR-coefficients and the transition matrix $\mathbb{P}$, and prediction.

Information criteria such AIC, BIC and their variations (Psaradakis and Spagnolo, 2006) are commonly used for simultaneous selection of the AR-orders and the number of regimes K. Smith et al. (2006) proposed a Markov switching criterion (MSC) as an estimate of a Kullback-Leibler divergence for model selection. However, these methods typically require exhaustive evaluation of 2^{Kq} different models with varying complexity. As illustrated in our simulations, even for moderate values of (q, K), this is computationally rarely feasible.

In addition, such methods can be numerically unstable (Breiman, 1996), and it is difficult to study theoretical properties of the resulting parameter estimators. Regularization techniques such as the LASSO (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) of Fan and Li (2001), and the adaptive LASSO (Zou, 2006) offer a potential solution that we investigate here.

We also study prediction or forecasting, and we demonstrate that consistency in optimal prediction in the sense of mean-squared prediction error can be achieved even when the number of regimes is overestimated. In light of the challenges and limitations of previous approaches, our main contributions are:

1. We develop a new regularized conditional likelihood method which, to the best of our knowledge, is the first work in the field for simultaneous AR-order and parameter estimation in MSAR models, and propose a regularized BIC (RBIC) for choosing the number of regimes K. The advantage of our method compared to the existing methods

is that, given (K, q), it simultaneously estimates AR-orders and parameters without an exhaustive search of 2^{Kq} possible models, and thereby is computationally efficient. This has been also supported by our analysis of average computational time (in seconds) taken by a method to complete per-sample results in simulations, see Section 7.1.

2. We study large sample properties of the methods, and assess their finite sample performance via simulations. Our results show that, under standard regularity conditions, when K is given or consistently estimated, the regularization method is consistent in AR-order and parameter estimation, and achieves consistent prediction of future values of the process. Furthermore, we discuss asymptotic properties of the RBIC in estimating K, and show that the conditional h-step ahead predictive density can be estimated consistently when the number of regimes is estimated by the RBIC.

The rest of the paper is organized as follows. In Section 2, Gaussian MSAR models are introduced. In Section 3, we develop new regularization methods and present their numerical implementation. Section 4 contains prediction in MSAR models. Estimation of the number of AR-regimes is discussed in Section 5. Section 6 contains theoretical study. Our simulation study is given in Section 7. We analyze two real datasets in Section 8. Section 9 contains a summary and discussion. Details of the numerical algorithm, regularity conditions, and the proofs are given in the Supplement.

2. Gaussian MSAR models, and their conditional likelihood

Consider an observable discrete-time series $\{Y_t : t = 1, 2, \ldots\}$ with realized values $\{y_t : t = 1, 2, \ldots\}$, and a latent stochastic process $\{S_t : t = 1, 2, \ldots\}$ taking values in $\{1, \ldots, K\}$ with K being the number of regimes underlying the process. In a MSAR model, the process S_t follows a homogeneous discrete finite-regime (or finite-state) first-order Markov chain with transition matrix $\mathbb{P} = [\alpha_{ij}]$. That is, for each t,

$$\Pr[S_t = j | S_{t-1} = i, S_{t-2} = s_{t-2}, \ldots, S_1 = s_1] = \Pr[S_t = j | S_{t-1} = i] = \alpha_{ij} \ , \ 1 \leq i, j \leq K$$

with initial state distribution $\Pr[S_t = j] = \pi_j \in (0, 1)$, which may, if required, be assumed to be the unique solution of $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbb{P}$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$. Conditional on S_t, the Y_t follows an inhomogeneous Markov process such that for each t conditional distribution of Y_t only depends on the regime indicator $S_t = j$ and lagged Y's, say, $y_{t-1}, \ldots, y_{t-q_j}$, for some q_j, and $j = 1, \ldots, K$. We assume the conditional distribution of $Y_t | (S_t = j, y_{t-1}, \ldots, y_{t-q_j})$ is Gaussian with variance ν_j and mean

$$\mu_{t,j} = \theta_{j0} + \theta_{j1} y_{t-1} + \ldots + \theta_{jq_j} y_{t-q_j} \ ; \ j = 1, \ldots, K. \tag{2.1}$$

For our theoretical study, the Gaussianity assumption can be relaxed and the observation process can be merely assumed to be a linear process driven by a white noise error with appropriate finite moment conditions. It is worth noting that the MSAR models under consideration are rather general. They encompass important special cases including mixture of autoregressive models studied by Wong and Li (2000), and MSAR

models with common AR-orders and AR coefficients across the regimes, i.e. $q_j = q$ and $\theta_{jl} = \theta_l$, for $j = 1, \ldots, K$ and $l = 1, \ldots, q$, discussed in Frühwirth-Schnatter (2006). Stationarity and ergodicity conditions of MSAR models are studied by Yao and Attali (2000) and Francq and Zakoïan (2001). Timmermann (2000) illustrates calculations of variance, higher order moments, and autocovariances of stationary MSAR models.

Let $q^* = \max_{1 \le j \le K} q_j$ denote the maximal AR-order of a stationary MSAR model. Proposition 1 in the Supplement shows that the lag-l population PACF of Y_t is zero for any $l > q^*$, a property shared by a standard AR model of order q^* (Brockwell and Davis, 1991). In practice, the sample PACF of Y_t can be used to provide an estimate of q^* in a MSAR model, but it gives little insight on the regime AR-orders q_j which are also the focus of our inference. We now introduce a conditional likelihood function as the base of our new estimation method described in Section 3.

Conditional likelihood: Let $\{(S_1, Y_1), \ldots, (S_n, Y_n)\} \equiv (S_{1:n}, Y_{1:n})$ be a sample of 'complete' data from a MSAR model. The joint density or complete data likelihood, by the assumptions and for some pre-specified densities g_0 and g_1, can be written as

$$g(s_{1:n}, y_{1:n}) = \Big\{ \Pr[S_1 = s_1] \times \prod_{t=1}^{n-1} \Pr[S_{t+1} = s_{t+1} | s_{1:t}] \Big\} \Big\{ g_0(y_1|s_{1:n}) \prod_{t=2}^{n} g_1(y_t|s_{1:n}, y_{1:(t-1)}) \Big\}$$

$$= \Pr[S_1 = s_1] \times \prod_{t=1}^{n-1} \alpha_{s_t, s_{t+1}} \times \Big\{ g_0(y_1|s_1) \prod_{t=2}^{n} g_1(y_t|s_t, y_{1:(t-1)}) \Big\},$$

where $\alpha_{s_t, s_{t+1}} = \Pr[S_{t+1} = s_{t+1} | s_{1:t}] = \Pr[S_{t+1} = s_{t+1} | S_t = s_t]$, for $1 \le t \le n-1$. The initial probability $\Pr[S_1 = s_1]$ can be incorporated in two ways; it can be either treated

as a separate marginal law that is inferred or conditioned upon during inference; or we may use the stationary distribution $\Pr[S_1 = s_1] = \pi_{s_1}$, $s_1 = 1, \ldots, K$, arising from the Markov chain with transition matrix $\mathbb{P}$ – this renders the probability $\Pr[S_1 = s_1]$ a function of the elements of $\mathbb{P}$. In either case, Ocone and Pardoux (1996), Kleptsyna and Veretennikov (2008), and Douc et al. (2009) demonstrated that, under mild conditions, the influence of the assumptions on $\Pr[S_1 = s_1]$ diminishes at a geometric rate in n.

The incomplete data likelihood $f(y_{1:n})$ is then available by marginalizing $g(s_{1:n}, y_{1:n})$ over the values of $s_{1:n}$. Given a pre-specified value $q \geq q^*$, f may be further factorized as $f_1(y_{1:q}) f_2(y_{q+1:n}|y_{1:q})$. Using a standard conditional approach in time series, we work with f_2 that, by the model assumptions, can be written as

$$f_2\big(y_{q+1:n}\big|y_{1:q}\big) = \sum_{s_1=1}^{K} \cdots \sum_{s_n=1}^{K} f(y_{q+1:n}|y_{1:q}, s_{1:n}) \Pr(s_{1:n}|y_{1:q})$$

$$= \sum_{s_q=1}^{K} \cdots \sum_{s_n=1}^{K} \Big\{ \Pr[S_q = s_q|y_{1:q}] \times \prod_{t=q+1}^{n} \alpha_{s_{t-1},s_t} \Big\} \Big\{ \prod_{t=q+1}^{n} g(y_t|y_{(t-q):(t-1)}, s_t) \Big\} \quad (2.2)$$

with Gaussian density $g(y_t|y_{(t-q):(t-1)}, s_t) = \phi(y_t; \mu_{t,s_t}, \nu_{s_t})$, and $\mu_{t,s_t} = \theta_{s_t,0} + \theta_{s_t,1} y_{t-1} + \ldots + \theta_{s_t,q} y_{t-q}$. Note that in this construction we have used a common AR-order $q(\geq q_j)$ for all the regimes; the regularization method in Section 3 estimates the regime-specific q_j using the data. Treatment of the probability $\Pr[S_q = s_q|y_{1:q}]$ is similar to that of $\Pr[S_1 = s_1]$ as discussed above. To avoid such specification, inspired by Douc et al. (2004), we condition on the state $S_q = s_q$ and work with the conditional density

$$f_3\big(y_{q+1:n}\big|y_{1:q}, s_q, \mathbf{\Phi}_K\big) = \sum_{s_{q+1}=1}^{K} \cdots \sum_{s_n=1}^{K} \Big\{ \prod_{t=q+1}^{n} \alpha_{s_{t-1},s_t} \Big\} \Big\{ \prod_{t=q+1}^{n} \phi(y_t; \mu_{t,s_t}, \nu_{s_t}) \Big\}. \quad (2.3)$$

Finally, the conditional log-likelihood that we use for inference in MSAR models is

$$\ell_n(\mathbf{\Phi}_K|y_{1:q}, s_q) \equiv \ell_n(\mathbf{\Phi}_K; s_q) = \log\{f_3\big(y_{q+1:n}\big|y_{1:q}, s_q, \mathbf{\Phi}_K\big)\}, \tag{2.4}$$

where $\mathbf{\Phi}_K = (\nu_1, \ldots, \nu_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \mathbb{P} = \{\alpha_{ij}\})$, and $\boldsymbol{\theta}_j = (\theta_{j0}, \theta_{j1}, \ldots, \theta_{jq})^\top$.

As discussed in the Introduction, due to the potential regime-specific AR-orders $q_j(\leq q)$, different elements of the vectors $\boldsymbol{\theta}_j$ may be zero, which then results in different sparsity patterns in the $\boldsymbol{\theta}_j$ across AR-regimes. This allows for regime-specific seasonality effects. Alternatively, we may also allow for non-seasonality effects in $\boldsymbol{\theta}_j$ and a decreasing pattern in the $|\theta_{jl}|$ as the lag l increases; see Section 3 for more details.

The marginalization over states s_t in (2.3) is achieved efficiently using standard filtering/prediction recursions utilized in the hidden Markov model literature. Numerical maximization of (2.4) with respect to $\mathbf{\Phi}_K$, and by treating s_t as the missing data, is relatively straightforward via the EM algorithm described in Section 3.

In principle, given (K, q), one could obtain the conditional MLE of $\mathbf{\Phi}_K$ by maximizing $\ell_n(\mathbf{\Phi}_K; s_q)$ in (2.4). However, since all of the estimated AR-coefficients will in general be non-zero, such an approach does not provide a sparse MSAR as postulated. This observation, and the limitations of the existing methods, motivate us to investigate the regularized conditional likelihood methods in later sections.

3. Simultaneous AR-order and parameter estimation

The conditional log-likelihood $\ell_n(\mathbf{\Phi}_K; s_q)$ in (2.4), similar to that of a Gaussian mixture model with unequal component variances ν_j's, diverges to infinity when some ν_j goes to 0. This singularity can be avoided by imposing a positive lower bound on ν_j (Hathaway, 1985) or adding a penalty function to the conditional log-likelihood (Chen et al., 2008). For the implementation convenience, we apply the latter approach and work with

$$\tilde{\ell}_n(\mathbf{\Phi}_K; s_q) = \ell_n(\mathbf{\Phi}_K; s_q) - \sum_{j=1}^{K} p_n(\nu_j), \tag{3.1}$$

where $p_n(\nu_j) \to +\infty$, as $\nu_j \to 0$ or ∞. An example of such penalty is

$$p_n(\nu_j) = \frac{1}{\sqrt{n-q}} \left[\frac{\mathcal{V}_n^2}{\nu_j} + \log\left(\frac{\nu_j}{\mathcal{V}_n^2} \right) \right] \tag{3.2}$$

with $\mathcal{V}_n^2 = (n-q)^{-1} \sum_{t=q+1}^{n} (y_t - \bar{y}_n)^2$ and $\bar{y}_n = (n-q)^{-1} \sum_{t=q+1}^{n} y_t$ as the sample variance and mean of $y_{q+1:n}$. From a Bayesian point of view, (3.2) is a data-dependent Gamma prior on ν_j^{-1} with its mode at $\mathcal{V}_n^{-2}$. With this penalty, we avoid instability of the EM algorithm while obtaining closed-form updates for ν_j's. We refer to (3.1) as the adjusted conditional log-likelihood. We now introduce the new regularization method.

Given (K, q) and any $s_q \in \{1, 2, \ldots, K\}$, we achieve joint AR-order and parameter estimation by maximizing the penalized (adjusted) conditional log-likelihood

$$\mathcal{L}_n(\mathbf{\Phi}_K; s_q, \lambda) = \tilde{\ell}_n(\mathbf{\Phi}_K; s_q) - \mathcal{R}_n(\mathbf{\Phi}_K; \lambda) \tag{3.3}$$

with the penalty (regularization) function

$$\mathcal{R}_n(\boldsymbol{\Phi}_K;\lambda)=\sum_{j=1}^{K}\sum_{l=1}^{q}r_n(\theta_{jl};\lambda). \tag{3.4}$$

Examples of r_n are the LASSO, adaptive LASSO (ADALASSO) and SCAD which are given in Section 1 of the Supplement. Unlike the penalties in information criteria, $r_n(\theta;\lambda)$ is a continuous function of θ and has a spike at $\theta=0$; $\lambda\geq 0$ is a tuning parameter. Given λ, let $\widehat{\boldsymbol{\Phi}}_{n,K,s_q}(\lambda)\equiv\widehat{\boldsymbol{\Phi}}_{n,K,s_q}=\arg\max_{\boldsymbol{\Phi}_K}\{\mathcal{L}_n(\boldsymbol{\Phi}_K;s_q,\lambda)\}$ be the maximum penalized conditional likelihood estimator (MPCLE) of $\boldsymbol{\Phi}_K$. By the properties of r_n and λ (Conditions $\mathbf{C}_1$-$\mathbf{C}_3$ in Section 1 of the Supplement), Theorem 2 shows that irrespective of the initial condition s_q, one can encourage estimates of some θ_{jl} to be zero. Hence, the method performs simultaneous AR-order and parameter estimation without evaluating all candidate MSAR models and thereby is computationally feasible.

In general, the method allows for regime-specific seasonality effects, due to the zero estimates of some θ_{jl}. Using ADALASSO, we also admit no seasonality effects and that the $|\theta_{jl}|$ decline with increasing lag l, as discussed in Section 1 of the Supplement.

Computation: We use a modified EM algorithm for maximization of $\mathcal{L}_n(\boldsymbol{\Phi}_K;s_q,\lambda)$ in (3.3). The core elements of the algorithm are given here; more details including a data-adaptive choice of λ are given in Section 3.2 of the Supplement. In what follows, we fix $s_q\in\{1,\ldots,K\}$, and denote $\boldsymbol{x}_t^\top=(1,y_{t-1},\ldots,y_{t-q})$.

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 $S_t = j$. The complete conditional log-likelihood is

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

where $\mu_{t,j} = \boldsymbol{x}^\top\boldsymbol{\theta}_j$. At $(m+1)$-th iteration, the EM algorithm iterates as follows:

E-step: We compute the conditional expectation of $\ell_n^c(\mathbf{\Phi}_K; s_q)$ with respect to (V_{tij}, U_{tj}), given $(\mathbf{\Phi}_K^{(m)}, s_q, y_{1:n})$. This reduces to the computation of the 'smoothing' probabilities

$$\begin{aligned}
\varpi_{tij}^{(m)} &= E(V_{tij}|y_{1:n}, s_q; \mathbf{\Phi}_K^{(m)}) \equiv \Pr[S_{t-1} = i, S_t = j|y_{1:n}, s_q; \mathbf{\Phi}_K^{(m)}] \quad , \quad 1 \le i, j \le K \\
\omega_{tj}^{(m)} &= E(U_{tj}|y_{1:n}, s_q; \mathbf{\Phi}_K^{(m)}) \equiv \Pr[S_t = j|y_{1:n}, s_q; \mathbf{\Phi}_K^{(m)}]
\end{aligned}$$

for $q+1 \le t \le n$. The probabilities are computed by the forward-backward algorithm of Baum et al. (1970) given in Section 3.1 of the Supplement.

M-step: We maximize a penalization of the conditional expectation of $\ell_n^c(\mathbf{\Phi}_K; s_q)$ computed in **E-step**, with the penalties in (3.2) and (3.4). The maximization with respect to $\boldsymbol{\theta}_j$ is performed using a coordinate descent approach. The parameter estimates are then updated as follows. First, for $1 \le l \le q$ and $1 \le j \le K$, we compute

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

where $\tilde{\mu}_{tj,-l} = \theta_{j0}^{(m)} + \sum_{v=1}^{l-1}\theta_{jv}^{(m+1)} y_{t-v} + \sum_{v>l}^{q}\theta_{jv}^{(m)} y_{t-v}$. We update the θ_{jl} by

$$\theta_{jl}^{(m+1)} = \frac{T(z_{1,jl}; \lambda_{jl})}{z_{2,jl}}, \tag{3.5}$$

where $T(z;\lambda) = \text{sign}(z)(|z| - \lambda)_+$ is the soft-thresholding operator (Donoho and Johnstone, 1994), and λ_{jl} depends on the penalty r_n; for the LASSO, $\lambda_{jl} = \lambda$. The λ_{jl} for the other two penalties are given in Section 3.1 of the Supplement.

The regime-specific intercepts and variances are updated by

$$\theta_{j0}^{(m+1)} = \frac{\sum_{t=q+1}^{n} \omega_{tj}^{(m)}(y_t - \mu_{tj}^{(m+1)})}{\sum_{t=q+1}^{n} \omega_{tj}^{(m)}} \tag{3.6}$$

$$\nu_j^{(m+1)} = \frac{\sum_{t=q+1}^{n} \omega_{tj}^{(m)}(y_t - \boldsymbol{x}_t^\top \boldsymbol{\theta}_j^{(m+1)})^2 + 2\mathcal{V}_n^2/\sqrt{n-q}}{\sum_{t=q+1}^{n} \omega_{tj}^{(m)} + 2/\sqrt{n-q}}, \tag{3.7}$$

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

$$\alpha_{s_q,j}^{(m+1)} = \frac{\sum_{t=q+1}^{n} \varpi_{t,s_q,j}^{(m)}}{\sum_{t=q+1}^{n}\sum_{i=1}^{K} \varpi_{t,s_q,i}^{(m)}} \quad , \quad \alpha_{ij}^{(m+1)} = \frac{\sum_{t=q+2}^{n} \varpi_{tij}^{(m)}}{\sum_{t=q+2}^{n}\sum_{h=1}^{K} \varpi_{tih}^{(m)}} \quad , \quad i \neq s_q,\ 1 \leq i,j \leq K. \tag{3.8}$$

Starting from an initial value $\boldsymbol{\Phi}_K^{(0)}$, the EM algorithm iterates until some convergence criterion is met. We used the stopping rule $\|\boldsymbol{\Phi}_K^{(m+1)} - \boldsymbol{\Phi}_K^{(m)}\| \leq \epsilon$, for a pre-specified small value ε, taken 10^{-5} in our simulations and data analysis. Due to the thresholding structure of the estimates in (3.5), by tuning λ estimates of some θ_{jl} will be exactly zero, which in turn results in simultaneous AR-order and parameter estimation.

4. Prediction

For weakly stationary processes, the conditional expectation of a future observation based on the current data provides an optimal prediction in terms of minimum mean-squared prediction error. In standard AR models, this leads to a straightforward pre-

diction mechanism. In this section, we focus on the predictive density in MSAR models that can also be used to compute the prediction values. Unlike many nonlinear models, the conditional expectation can be easily computed analytically in the MSAR as follows.

Given the observations $y_{1:n}$, we are interested in the joint distribution of the future vector $(Y_{n+1}, \ldots, Y_{n+h}) \equiv Y_{n+1:h}$, or equivalently the h-step ahead predictive density $f_K(y_{n+1:h}|y_{1:n})$. By the model assumptions in Section 2, we have that, for $h = 1, 2$,

$$f_K(y_{n+1}|y_{1:n}) = \sum_{s_{n+1}=1}^{K} \Pr(S_{n+1} = s_{n+1}|y_{1:n})\ \phi(y_{n+1}; \boldsymbol{x}_{n+1}^\top \boldsymbol{\theta}_{s_{n+1}}, \nu_{s_{n+1}}) \tag{4.1}$$

$$f_K(y_{n+1:h}|y_{1:n}) = \sum_{s_{n+1:h}=1}^{K} P(S_{n+1} = s_{n+1}|y_{1:n}) \left[\prod_{j=2}^{h} \alpha_{s_{n+j-1}, s_{n+j}}\right] \left[\prod_{j=1}^{h} \phi(y_{n+j}; \boldsymbol{x}_{n+j}^\top \boldsymbol{\theta}_{s_{n+j}}, \nu_{s_{n+j}})\right], \tag{4.2}$$

where $\boldsymbol{x}_{n+j}^\top = (1, y_{n+j-1}, \ldots, y_{n+j-q})$. The conditional probabilities $P(S_{n+1} = j|y_{1:n}), j = 1, \ldots, K$, are computed recursively using the *prediction* and *filtering* probabilities,

$$\Pr(S_{t+1} = j|y_{1:t}) = \sum_{l=1}^{K} \Pr(S_{t+1} = j|S_t = l, y_{1:t}) P(S_t = l|y_{1:t}) = \sum_{l=1}^{K} \alpha_{lj} \Pr(S_t = l|y_{1:t}),$$

$$\Pr(S_t = l|y_{1:t}) = \frac{f(y_t|y_{1:t-1}, S_t = l) \Pr(S_t = l|y_{1:t-1})}{f_K(y_t|y_{1:t-1})} = \frac{\phi(y_t; \boldsymbol{x}_t^\top \boldsymbol{\theta}_l, \nu_l) \Pr(S_t = l|y_{1:t-1})}{f_K(y_t|y_{1:t-1})}$$

for all $t = n, n-1, \ldots, q+1$. Note that the conditional density $f_K(y_t|y_{1:t-1})$ needed for the filtering probabilities is computed similarly to (4.1). Specifically, for $t = q + 1$,

$$f_K(y_{q+1}|y_{1:q}) = \sum_{l=1}^{K} \Pr(S_{q+1} = l|y_{1:q}) \phi(y_{q+1}; \boldsymbol{x}_{q+1}^\top \boldsymbol{\theta}_l, \nu_l)$$

which requires $\Pr(S_{q+1} = j|y_{1:q}) = \sum_{l=1}^{K} \alpha_{lj} P(S_q = l|y_{1:q})$ and the initial distribution $\{\Pr(S_q = l|y_{1:q}), l = 1, 2, \ldots, K\} \equiv \boldsymbol{\gamma}_q$, to be specified.

Thus, given the data $y_{1:n}$ and upon the specification of $(\boldsymbol{\Phi}_K, \boldsymbol{\gamma}_q)$, the h-step ahead predictive densities of a K-regime MSAR model are available. The effect of the initial distribution $\boldsymbol{\gamma}_q$ on the predictive densities is negligible once n grows (Ocone and Pardoux, 1996; Kleptsyna and Veretennikov, 2008; Douc et al., 2009). For example, one may use a non-informative uniform discrete distribution $\boldsymbol{\gamma}_q = (1/K, \ldots, 1/K)$. The parameter $\boldsymbol{\Phi}_K$ is estimated by its MPCLE $\widehat{\boldsymbol{\Phi}}_{n,s_q,K}$ obtained using the data $y_{1:n}$. We denote the resulting estimated predictive densities (4.1) and (4.2) by $\widehat{f}_K(y_{n+1:h}|y_{1:n})$.

The estimated densities can then be used to compute various quantities such as the conditional expectations for prediction. For example, the optimal one-step prediction value (in the sense of mean-squared prediction error) is given by

$$\widehat{E}^*\{Y_{n+1}|y_{1:n}\} = \sum_{j=1}^{K} \widehat{\Pr}(S_{n+1} = j|y_{1:n})\{\widehat{\theta}_{j0} + \widehat{\theta}_{j1}\ y_n + \ldots + \widehat{\theta}_{jq}\ y_{n+1-q}\}, \tag{4.3}$$

where $(\widehat{\theta}_{j0}, \widehat{\theta}_{jl})$ are the MPCLE, and $E^*\{\cdot\}$ is the expectation under the true model.

5. Choice of the number of AR-regimes, K

The methods in Sections 3 and 4 are used when the number of AR-regimes K is fixed. Typically, K is also required to be chosen using the data. Information criteria such as BIC based on MLE are commonly used for estimating K. We instead propose to use a

regularized BIC (RBIC) based on the MPCLE, which unlike BIC, it does not search the model space for also choosing the AR-orders as this task is performed by the MPCLE.

Consider situations where placing a known upper bound $\mathcal{K}$ on K is feasible. For each $K = 1, \ldots, \mathcal{K}$, we fit a MSAR model with the resulting MPCLE $\widehat{\boldsymbol{\Phi}}_{n,K,s_q}$, for any fixed and arbitrary choice of $s_q \in \{1, \ldots, K\}$. Let $N_K = \sum_{j=1}^{K} \sum_{l=1}^{q} I(\widehat{\theta}_{jl} \neq 0)$ be the total number of non-zero estimated AR-coefficients, and denote

$$\text{RBIC}(\widehat{\boldsymbol{\Phi}}_{n,K,s_q}) = -2\ell_n(\widehat{\boldsymbol{\Phi}}_{n,K,s_q}; s_q) + \log(n-q) \times \{N_K + K(K-1) + 2K\}, \tag{5.1}$$

where $K(K-1) + 2K$ counts the number of parameters $(\nu_j, \theta_{j0}, \alpha_{ij})$, $\ell_n(\cdot; s_q)$ is the conditional log-likelihood in (2.4). The number of AR-regimes is then estimated by

$$\widehat{K}_n = \underset{1 \leq K \leq \mathcal{K}}{\text{argmin}} \ \text{RBIC}(\widehat{\boldsymbol{\Phi}}_{n,K,s_q}). \tag{5.2}$$

We discuss large sample properties of $\widehat{K}_n$ in Section 6. If the penalty in (5.1) is replaced by $2\{N_K + K(K-1) + 2K\}$, we obtain the regularized AIC (RAIC). In our simulations in Section 7.2, we asses finite sample performance of the RAIC, RBIC, and a regularized version of the Markov-switching criterion (MSC) of Smith et al. (2006), which is computed based on the MPCLE and we call it RMSC. It is worth noting that, due to the factor $\log(n-q)$ in (5.1), the penalty in RBIC is more severe than the ones in RAIC and RMSC. Thus, it is expected that in finite sample situations the RBIC may result in models with lower selected orders (underestimation) compared to the models selected by the other two criteria. More discussion is given in Section 7.2 of our simulations.

6. Theoretical study

We first study asymptotic properties of the MPCLE when the true number of AR-regimes K is pre-determined (Theorems 1 and 2). We then study the $\widehat{K}_n$ in (5.2) and discuss the behaviour of the MPCLE when the number of regimes is estimated by $\widehat{K}_n$ (Theorem 3). Regularity conditions $\mathbf{C}_1$-$\mathbf{C}_3$ on the penalty r_n and the tuning parameter λ_n, and the proofs are respectively given in Sections 1 and 2 of the Supplement.

Notation: All the vectors are column vectors and we drop the transpose $^\top$, for convenience. We assume the observed time series is a sample from a MSAR model with K AR-regimes and a d-dimensional true parameter vector $\mathbf{\Phi}^* = (v_1^*, \ldots, v_K^*, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^*, \mathbb{P}^* = \{\alpha_{ij}^*\})$, where $d = K(q+2) + K(K-1)$. The regime-specific AR-coefficient vector is $\boldsymbol{\theta}_j^*$, variance is v_j^* and transition probability is $\alpha_{ij}^* > 0$, $i, j = 1, \ldots, K$. We further assume that $\mathbf{\Phi}^*$ is an interior point of the compact parameter space $\mathbf{\Theta} \subseteq \mathbb{R}^d$. We partition each regime-specific AR-coefficient vector as $\boldsymbol{\theta}_j^* = (\boldsymbol{\theta}_{j1}^*, \boldsymbol{\theta}_{j2}^*)$ so that $\boldsymbol{\theta}_{j1}^*$ and $\boldsymbol{\theta}_{j2}^*$ contain the non-zero and zero AR-coefficients, respectively. We partition the parameter vector $\mathbf{\Phi}^* = (\mathbf{\Phi}_1^*, \mathbf{\Phi}_2^*)$ accordingly so that $\mathbf{\Phi}_2^* = (\boldsymbol{\theta}_{12}^*, \ldots, \boldsymbol{\theta}_{K2}^*) = \mathbf{0}$. The subvector $\mathbf{\Phi}_1^*$ contains all the intercepts θ_{j0}^*, the non-zero θ_{jl}^*, the variances ν_j^*, and the transition probabilities α_{ij}^*. Further, let $\dim(\mathbf{\Phi}_1^*) = d_1 < d$. We partition any candidate parameter as $\mathbf{\Phi} = (\mathbf{\Phi}_1, \mathbf{\Phi}_2)$ following that of $\mathbf{\Phi}^*$. We use $\widehat{\mathbf{\Phi}}_{n,s_q}$ to represent the MPCLE of the vector of parameters of the true MSAR model with K regimes, and for any fixed

$s_q \in \{1, \ldots, K\}$. Let $\mathcal{R}'_n(\cdot;\lambda)$ be the vector of first and $\mathcal{R}''_n(\cdot;\lambda)$ be the matrix of second derivatives of $\mathcal{R}_n(\boldsymbol{\Phi};\lambda)$ with respect to $\boldsymbol{\Phi}$. Also, let $\mathbf{I}_{11}(\boldsymbol{\Phi}_1^*)$ be the Fisher information of the true MSAR model with $\boldsymbol{\Phi}_2^* = \mathbf{0}$. The Euclidean norm is denoted by $\|\cdot\|_2$.

Main results: By conditioning on $y_{1:q}$, the effective sample size is $n-q$. Since $q < \infty$, asymptotically $n \sim n-q$ and thus in what follows we use n instead of $n-q$. Our first result establishes estimation consistency of the MPCLE, irrespective of the choice of s_q.

Theorem 1. *Let $Y_{1:n}$ be a sample from a stationary and ergodic* MSAR *model, and $E|Y_t|^{(4+2\delta)} < \Delta < \infty$, for some $\delta > 0$. Assume λ_n and the penalty r_n satisfy Conditions* $\mathbf{C}_1$-$\mathbf{C}_2$ *in the Supplement. Then, there exists a local maximizer $\widehat{\boldsymbol{\Phi}}_{n,s_q}$ of $\mathcal{L}_n(\boldsymbol{\Phi}; s_q, \lambda_n)$ such that, as $n \to \infty$, $\|\widehat{\boldsymbol{\Phi}}_{n,s_q} - \boldsymbol{\Phi}^*\|_2 = O_p\{n^{-1/2}(1+a_n)\}$, where a_n is given in* $\mathbf{C}_2$.

By Theorem 1, if $a_n = O(1)$, which requires appropriate choices of λ_n and r_n, then $\widehat{\boldsymbol{\Phi}}_{n,s_q}$ is $\sqrt{n}$-consistent. This is the rate for the conditional MLE studied in Douc et al. (2004). For example, to achieve $\sqrt{n}$-consistency of the MPCLE based on the SCAD, it is sufficient that $\lambda_n \to 0$ as $n \to \infty$, since then $a_n = 0$. For the LASSO, $\sqrt{n}$-consistency is achieved if $\lambda_n = O(n^{-1/2})$ (or $o(n^{-1/2})$); and for ADALASSO we need $\sqrt{n}\lambda_n = o(1)$.

In Theorem 2 we show that the $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\Phi}}_{n,s_q}$ has also an oracle property, as defined in Fan and Li (2001). More specifically, consider the partitioning $\widehat{\boldsymbol{\Phi}}_{n,s_q} = (\widehat{\boldsymbol{\Phi}}_{n,s_q,1}, \widehat{\boldsymbol{\Phi}}_{n,s_q,2})$, where $\dim(\widehat{\boldsymbol{\Phi}}_{n,s_q,1}) = \dim(\boldsymbol{\Phi}_1^*) = d_1$ and $\dim(\widehat{\boldsymbol{\Phi}}_{n,s_q,2}) = \dim(\boldsymbol{\Phi}_2^*) = d - d_1$. This partitioning is based on the oracle's perspective.

Theorem 2. *Assume the same conditions of Theorem 1, and that λ_n and the penalty r_n*

in addition satisfy Condition $\mathbf{C}_3$*, and* $a_n = O(1)$*. We have that, for any* $\sqrt{n}$*-consistent estimator* $\widehat{\mathbf{\Phi}}_{n,s_q}$ *of* $\mathbf{\Phi}^*$ *with the above partitioning, as* $n \to \infty$*,*

(i) *Consistency in* AR*-order estimation:* $\Pr(\widehat{\mathbf{\Phi}}_{n,s_q,2} = \mathbf{0}) \longrightarrow 1$.

(ii) *Asymptotic normality:*

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

By Theorems 1 and 2, for the SCAD penalty with $\lambda_n \sim n^{-1/2}\log n$, the MPCLE $\widehat{\mathbf{\Phi}}_{n,s_q}$ is consistent in both parameter and AR-order estimations. With the same choice of λ_n, the MPCLE based on the LASSO is consistent in AR-order estimation but it also introduces bias to the estimators of the true non-zero AR-coefficients, a well-known property of the LASSO in other settings. For ADALASSO, if $\lambda_n \sim n^{-1/2-\psi}$ for a $0 < \psi < \frac{\gamma}{2}$, the resulting MPCLE is consistent in both parameter and AR-order estimations. It is worth noting that, given K and under the conditions of Theorem 1 on Y_t, the standard BIC is consistent in AR-order estimation (Konishi and Kitagawa, 2008). However, compared to the new method, the BIC has higher computational cost of evaluating 2^{Kq} different MSAR models in order to choose a final model.

By consistency of the MPCLE in Theorem 1, from (4.3) we have that, as $n \to \infty$,

$$\widehat{E}^*\{Y_{n+1}|y_{1:n}\} \xrightarrow{p} E^*\{Y_{n+1}|y_{1:n}\}, \tag{6.1}$$

where $E^*\{Y_{n+1}|y_{1:n}\}$ is the optimal one-step prediction. This holds for h-step prediction.

Next, we study properties of the RBIC-based estimator $\widehat{K}_n$ in (5.2), and its effect on

the MPCLE and specifically on the estimated predictive densities $\widehat{f}_\kappa(y_{n+1:h}|y_{1:n})$ when $\kappa = \widehat{K}_n$. We denote $f^*(y_{n+1:h}|y_{1:n})$ as the h-step ahead predictive density based on the true MSAR model with K regimes, and $\mathcal{K}$ is an upper bound for K.

Theorem 3. *Under the conditions of Theorem 2, and in addition, by assuming a compact Euclidean space for parameters $\boldsymbol{\theta}_j$ and ν_j, we have that, as $n \to \infty$,*

(i) $P(\widehat{K}_n \geq K) \to 1$, *where K is the true number of* AR*-regimes.*

(ii) *For any finite $K \leq \kappa \leq \mathcal{K}$, $\widehat{f}_\kappa(y_{n+1:h}|y_{1:n}) \longrightarrow f^*(y_{n+1:h}|y_{1:n})$, almost surely, for all $(y_{1:n+h})$. The result also holds when the number of regimes is estimated by the $\widehat{K}_n$.*

Part (i) indicates that the $\widehat{K}_n$ asymptotically does not underestimate the true number of regimes K. Part (ii) shows that even if the number of regimes is overestimated, we can still obtain consistency of the estimated h-step ahead predictive densities. Hence, for instance, (6.1) still holds. Consistency of the $\widehat{K}_n$ can be established under stronger conditions. For example, for some small constants $\delta > 0$ and $\varepsilon \in (0, 1/2)$, consider the restricted parameter space for the overestimated models with $\kappa > K$ AR-regimes, $\boldsymbol{\Theta}_c = \Big\{\boldsymbol{\Phi} = (v_1, \ldots, v_\kappa, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\kappa, \mathbb{P} = \{\alpha_{ij}\}) : v_j \geq \delta, \alpha_{ij} \in [\varepsilon, 1-\varepsilon]\Big\}$ and the $\boldsymbol{\theta}_j$ belong to a compact Euclidean subspace of $\mathbb{R}^q$. Similar to the results of Keribin (2000) and Lu (2009), the supremum of the log-likelihood ratio statistic of the overestimated models over $\boldsymbol{\Theta}_c$ versus the true model behaves as $O_p(\log\log n)$, as $n \to \infty$. Thus, using the same proof technique of Theorem 3-(i) (Section 2 of the Supplement), the RBIC also prevents overestimation of K, and hence yields consistency of the $\widehat{K}_n$. In our

simulations, in Section 7.2, we find the RBIC performs well in estimating K without any restriction on the parameter space.

7. Simulation study

We study finite sample performance of the proposed methods via simulations. We generated times series data from Gaussian MSAR models with $K = 2, 3$ AR-regimes. For the two-state models, specified parameters are given in Table 1.

$(\alpha_{11}, \alpha_{22})$	$(\nu_1^{1/2}, \nu_2^{1/2})$	$\mu_{t,1}$	$\mu_{t,2}$
$(.80, .70), (.25, .25)$	$(5.0, 3.0)$	$-.60y_{t-1} - .50y_{t-2}$ $.67y_{t-1} - .55y_{t-2}$	$.50y_{t-1} - .70y_{t-2}$ $.45y_{t-1} + .35y_{t-3} - .65y_{t-6}$

Table 1: Simulation parameter settings

For each $(\nu_1^{1/2}, \nu_2^{1/2}, \mu_{t,1}, \mu_{t,2})$, we considered two transition matrices $\mathbb{P}$. This results in four models, **M1-M4**. The fifth model **M5** is a three-state model, and together with its simulation results are given in Section 4 of the Supplement. Our results are based on 300 simulated time series of different sizes n from each model. The computations are done in C++ and on a Mac OS X machine with 2.9 GHz Intel Core i5.

In Section 7.1, given the number of regimes K, we compare the regularization method using the LASSO, ADALASSO, and SCAD with the standard BIC via the measures:

– estimated sensitivity (ES1): the proportion of correctly estimated zero AR-coefficients.

– estimated specificity (ES2): the proportion of correctly estimated non-zero AR-coefficients.

– estimation error: $L_2 = \|\widehat{\psi} - \psi\|_2$ losses of estimates ($\widehat{\psi}$) of parameters (ψ) AR-coefficients, variances, and transition probabilities, separately.

– average computational time (ACT, in seconds) taken to complete per-sample results.

For models **M1-M2** and **M3-M4**, the maximal AR-orders are $q^* = 2, 6$, respectively. To demonstrate the performance of the new method, we set a larger AR-order $q = 10$ in the penalized log-likelihood (3.3) for all models; the parameter λ is chosen using an information criterion given in Section 3.2 of the Supplement. To reduce the computational burden of the BIC for AR-order estimation, we set the smaller common AR-orders $q = 5$ and $q = 6$ respectively for **M1-M2** and **M3-M4**; these orders produce about 961 and 3969 models to be examined by the BIC. We also examine the performance of the new method with the smaller values $q = 5, 6$, and the results are summarized at the end of Section 7.1 below.

In Section 7.2, we evaluate the performance of the RAIC, RBIC and RMSC in estimating the true number of AR-regimes K. We also compare the estimated predictive density $\widehat{f}_K(y_{n+1:h}|y_{1:n})$ when K is correctly specified versus when overestimated.

7.1 Analysis of (ES1, ES2), $(\mathbf{L}_1, \mathbf{L}_2)$, and ACT: K is pre-specified

Table 2 shows the average and standard deviation, over 300 replications, of the ES1 and ES2 values corresponding to models **M1-M4**. Since the results were similar when conditioning on initial state $s_q = 1$ or 2, we report the results for $s_q = 1$.

From Table 2, we see that the average ES1 for the BIC varies between 90.4% to

100%, and for the new method varies between 88% to 100%, across the models **M1-M4**, the sample sizes $n = 150, 250, 500$, and the three penalties. Regarding the average ES2, when $n = 150$, the BIC performs better by correctly identifying the true non-zero θ_{jl}'s about 90% to 100% of the times for different models. These proportions for the LASSO, ADALASSO, and SCAD are respectively about 57% to 100%, 74% to 100%, and 72% to 100%. For $n = 250, 500$, the BIC, ADALASSO, and SCAD perform similarly with the average ES2 of more than 92%, and for the LASSO the average is more than 83%.

We now assess computational efficiency of the methods by comparing the average computational times (ACT, in seconds) reported in Table A1 of the Supplement. The new method based on the LASSO, ADALASSO, and SCAD respectively takes on average .853 to 5.44, .375 to 2.35, and .830 to 3.77 seconds to complete per-sample results, depending on the model and sample size. The BIC takes much longer to complete the same task, as for models **M1-M2** its ACT is 17.4 to 96.6 seconds, and for models **M3-M4** is about 85 to 297 seconds.

Boxplots of the empirical L_2 losses of the parameter estimates based on the BIC, LASSO, ADALASSO, and SCAD, as well as the estimates from the model in which the redundant zero AR-coefficients are removed (the oracle model) are given in Figures A1-A4 of the Supplement. For the smaller sample sizes, the empirical median (and variation) losses of the estimates, particularly those based on LASSO, are higher than those of the estimator under the oracle model. As n increases, performance of all the

estimates improve and are comparable to the oracle estimator.

Similar to the BIC, we also ran the new method with the smaller AR-order $q = 5$ and 6 for models **M1-M2** and **M4-M5**, respectively. The average and standard deviation of the ES1 and ES2 values and boxplots of the empirical L_2 losses are respectively given in Table A8 and Figures A9-A12 of the Supplement. For $n = 150$, the performance of the method (in term of the ES1, ES2, and loss) improves as AR-order upper bound q reduces from 10 to 5 or 6. This is expected as by reducing q the potential number of parameters $K(q+2) + K(K-1)$ to be estimated also decreases. As n increases to $250, 500$, the effect of q is less apparent in each of the models under consideration.

7.2 Estimation of the number of AR-regimes K, and prediction

We first examine the performance of the estimator $\widehat{K}_n$ of K based on the RAIC, RBIC and RMSC described in Section 5. We fit MSAR models with $K = 1, \ldots, 5$, to each simulated sample, and obtain the MPCLE which is then used to compute the RAIC, RBIC and RMSC. We choose $\widehat{K}_n$ as the one that minimizes a criterion. Here we report the results when the MPCLE is obtained using SCAD; the results based on the ADALASSO and LASSO are similar, and are given in Tables A2-A3 of the Supplement.

Table 3 contains the average proportions of times that a number of regimes $K = 1, \ldots, 5$, is selected by a criterion for models **M1-M4**. We can see that, for $n = 150, 250$, RBIC has a higher percentage of underestimation of the true K while RAIC and RMSC tend to overestimate K. As explained at the end of Section 5, this behaviour is expected

since the penalty function in the RBIC in (5.1) is heavier than the ones in the other two criteria, which results in higher percentages of underestimation of the true K by RBIC. As the sample size increases to $n = 500$, the percentages of underestimation of the true K by RBIC tends to zero, supporting the result of Theorem 3-(i). We can see that, when $n = 500$, the RBIC estimates the true K almost 100% in all the four models. For $n = 500$, RMSC estimates the true K about 82% to 92% of the times across the four models, while RAIC estimates K approximately 58% to 81% of the times.

Finally, we examine the finite sample behaviour of the estimated predictive density $\widehat{f}_K(y_{n+1:h}|y_{1:n})$, when K is correctly specified and overestimated. We generated 300 time series of sizes $n + h$ from model **M2** with $K = 2$, where $n = 250, 500, 800, 1000$ and $h = n/10$. For each generated sample, we used the first n observations to fit MSAR models (using the regularization method) with $K = 2, 3, 4, 5$, and the remaining h observations were used to compute $\log[\widehat{f}_K(y_{n+1:h}|y_{1:n})]$. Figure A6 of the Supplement shows boxplots of the log-predictive densities. We see that: 1) overall, the empirical median and interquartile range of the log-predictive density values of the overestimated models ($K \geq 3$) are approximately equal to those of the models with correct $K = K^*$; 2) for the smaller sample size $n = 250$, as expected the variation of the log-predictive density values increases as the number of extra regimes increases; 3) as n increases, the log-predictive density values of the overestimated models ($K \geq 3$) are approximately equal to those for the true model, supporting the result of Theorem 3-(ii).

8. Real data analysis

We illustrate the application of our method via two real data examples. Figures A7 and A8 of the Supplement are used through our analysis below. We use sample PACF to obtain an (approximate) upper bound q, required by our regularization method, for the maximal AR-order $q^* = \max_{1\le j\le K} q_j$. From Figures A7-(b) and A8-(c), $q = 15, 25$, are reasonable choices in Examples 1 and 2, respectively. We report the results based on the SCAD and ADALASSO with the lag-dependent weights $\omega_{jl}(\alpha) = |\tilde{\theta}_{jl}\ \alpha(1-\alpha)^l|^{-1}$, and $\alpha = 0.8$. The fitted models based on the LASSO and ADALASSO with $\omega_{jl} = |\tilde{\theta}_{jl}|^{-1}$ were either similar or outperformed, in terms of log-predictive density values, by the fitted models discussed below and thus not reported. The $\tilde{\theta}_{jl}$ are the (conditional) MLE.

Example 1. Data are the quarterly real gross domestic product (GDP) growth rate, computed as $y_t = 100(\log \text{GDP}_t - \log \text{GDP}_{t-1})$ and adjusted for inflation, of the U.S. over the period of the first quarter of 1947 to the third quarter of 2016 obtained from `https://fred.stlouisfed.org`. Figure A7-(a) is the time series plot of 278 observations. The plot shows that the variation in the series changes over time, which motivated us to consider fitting a MSAR model to y_t. We used 267 observations from the period 1947-2013 for fitting, and 11 observations over the period 2013-2016 for prediction.

We applied the regularization method with $q = 15$ and fitted MSAR with $K = 1, 2, 3, 4$. The RBIC values based on SCAD are: $691.9, \mathbf{658.7}, 690.4, 720.2$, and based on ADALASSO with the lag-dependent weights are: $688.4, \mathbf{665.7}, 693.4, 732.5$. Thus, based

on RBIC we select $\widehat{K} = 2$, and the fitted models are given below; standard errors are in $[\cdot]$. The log-predictive density values computed based on the 11 observations, for the two fitted MSAR models are respectively -7.66 and -7.19.

$(\widehat{\alpha}_{11}, \widehat{\alpha}_{22})$	$(\widehat{\nu}_1^{1/2}, \widehat{\nu}_2^{1/2})$	regime 1: $\widehat{\mu}_{t,1}$	regime 2: $\widehat{\mu}_{t,2}$
SCAD: $(.983, .981)$	$(.471, 1.10)$	$\underset{[.032]}{.521} + \underset{[.022]}{.290} y_{t-2}$	$\underset{[.044]}{.546} + \underset{[.036]}{.365} y_{t-1}$
ADALASSO: $(.985, .981)$	$(.483, 1.12)$	$\underset{[.033]}{.513} + \underset{[.028]}{.133} y_{t-1} + \underset{[.023]}{.158} y_{t-2}$	$\underset{[.045]}{.607} + \underset{[.036]}{.298} y_{t-1}$

Below we provide an analysis of the fitted model based on SCAD; a similar analysis can be performed for the second model. Figure A7-(c) shows classification of y_t's into the two regimes of the model. Most of the observations from around 1950-1984 and 2008-2009 are classified into regime 2, and the remaining observations from around 1984-2007 and 2010-2013 are classified into regime 1. We may interpret the two regimes as follows: regime 1 describes the periods where the growth rate was mostly positive and more stable with a relatively lower variation compared to regime 2 where the growth rate was a combination of mostly large positives and also occasionally large negatives (between 1950-1960, and noticeably around 2008-2009 which was the recent economical crisis) with a much higher variation. From Figure A7-(c), once the economy falls into one of the two regimes it stays in that regime for a long time period which is confirmed by the large diagonal values $(\widehat{\alpha}_{11}, \widehat{\alpha}_{22})$ of the estimated transition probability matrix $\widehat{\mathbb{P}}$.

Example 2. Data are the monthly U.S. unemployment rates (y_t) over the period of

1948 to 2010, obtained from `https://www.bea.org`. The time series plot in Figure A8-(a) shows an increasing-decreasing trend and also high volatility in the series. To remove the trend in the mean, we consider the differences $x_t = y_{t+1} - y_t, t = 1, \ldots, 754$. We used 731 observations from the period 1948-2008 for fitting, and the remaining 24 observations from 2009-2010 were used for computing the log-predictive density.

We use the regularization method with $q = 25$ and fitted MSAR models with $K = 1, 2, 3, 4$, to the data. The RBIC values based on SCAD are: $589.6, \mathbf{565.5}, 609.2, 616.3$. Thus, based on RBIC we select $\widehat{K} = 2$ and the fitted model is

$$\text{regime } 1: \widehat{\mu}_{t,1} = \underset{[.016]}{.053}x_{t-2} + \underset{[.016]}{.094}x_{t-3} - \underset{[.015]}{.082}x_{t-12}\ ,\ \ \widehat{\nu}_1^{1/2} = .136\ ,\ \ \widehat{\alpha}_{11} = .785$$

$$\text{regime } 2: \widehat{\mu}_{t,2} = \underset{[.038]}{.225}x_{t-4} + \underset{[.036]}{.272}x_{t-5} - \underset{[.033]}{.115}x_{t-10} - \underset{[.038]}{.244}x_{t-24}\ ,\ \ \widehat{\nu}_2^{1/2} = .225\ ,\ \ \widehat{\alpha}_{22} = .551.$$

with the log-predictive estimated density value -1.23. The RBIC values based on ADALASSO with the lag-dependent weights are: $645.7, \mathbf{579.6}, 605.7, 618.4$. Thus, we select $\widehat{K} = 2$ and the fitted model is

$$\text{regime } 1: \widehat{\mu}_{t,1} = \underset{[.023]}{-.112}x_{t-1}\ ,\ \ \widehat{\nu}_1^{1/2} = .135\ ,\ \ \widehat{\alpha}_{11} = .975$$

$$\text{regime } 2: \widehat{\mu}_{t,2} = \underset{[.028]}{.129}x_{t-1} + \underset{[.029]}{.109}x_{t-2}\ ,\ \ \widehat{\nu}_2^{1/2} = .238\ ,\ \ \widehat{\alpha}_{11} = .970$$

with the log-predictive density value -3.47. In both models, the estimates of the intercepts θ_{j0}'s were zero. Below, we focus on the SCAD model. Figure A8-(d) shows classification of x_t's into the two regimes of the model. A possible interpretation is that regime 1 captures time periods with relatively low changes in the unemployment rates compared to regime 2 that captures periods with larger jumps or drops in the rates.

9. Summary and discussion

We have developed new regularization methods for AR-order and parameter estimation, as well as selection of the number of AR-regimes in MSAR models. The methods present a substantial computational advantage over AIC, BIC and their variations by avoiding an exhaustive search of the model space, and they also have desirable large sample properties. In addition, we have demonstrated consistency of optimal prediction, in the sense of mean-squared prediction error and predictive density, in cases of correctly and over-specified number of regimes. Simulation results support our theoretical findings.

Our focus has been on the Gaussian case, but similar results hold under milder conditions provided the equivalent moment conditions hold. Extensions to incorporate conditional heteroscedasticity or to general state space models, are avenues of future work. There remain, however, interesting research challenges – for example, under what less restrictive conditions the RBIC provides a consistent estimator of the number of regimes is yet to be further studied.

Acknowledgment. The authors would like to thank the editor, an associate editor, and two anonymous referees for their constructive comments and suggestions that improved the article. The research of the first author was partially supported by a grant (NSERC RGPIN-2015-03805) from the Natural Sciences and Engineering Research Council of Canada.

Table 2: Average (standard deviation), over 300 replications, of estimated Sensitivity (ES1) and Specificity (ES2)[1].

	Model	MSAR Regimes	$n = 150$ ES1	$n = 150$ ES2	$n = 250$ ES1	$n = 250$ ES2	$n = 500$ ES1	$n = 500$ 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	$.933_{(.141)}$	$.565_{(.444)}$	$.965_{(.072)}$	$.858_{(.315)}$	$.988_{(.040)}$	$.995_{(.064)}$
		Reg_2	$.932_{(.119)}$	$.668_{(.355)}$	$.958_{(.091)}$	$.878_{(.270)}$	$.988_{(.036)}$	$.998_{(.029)}$
	M2	Reg_1	$.962_{(.074)}$	$.988_{(.076)}$	$.988_{(.041)}$	$.997_{(.059)}$	$.999_{(.013)}$	$1.00_{(.000)}$
		Reg_2	$.974_{(.067)}$	$.997_{(.058)}$	$.997_{(.018)}$	$.998_{(.029)}$	$1.00_{(.000)}$	$1.00_{(.000)}$
	M3	Reg_1	$.920_{(.114)}$	$.800_{(.377)}$	$.946_{(.077)}$	$.945_{(.223)}$	$.986_{(.042)}$	$1.00_{(.000)}$
		Reg_2	$.880_{(.154)}$	$.779_{(.322)}$	$.936_{(.108)}$	$.916_{(.217)}$	$.989_{(.040)}$	$.999_{(.019)}$
	M4	Reg_1	$.901_{(.147)}$	$.593_{(.464)}$	$.945_{(.091)}$	$.830_{(.363)}$	$.980_{(.050)}$	$.988_{(.104)}$
		Reg_2	$.878_{(.154)}$	$.819_{(.274)}$	$.937_{(.117)}$	$.953_{(.159)}$	$.989_{(.038)}$	$1.00_{(.000)}$
ADALASSO	**M1**	Reg_1	$.930_{(.154)}$	$.738_{(.362)}$	$.973_{(.074)}$	$.928_{(.214)}$	$.997_{(.020)}$	$.997_{(.041)}$
		Reg_2	$.948_{(.130)}$	$.685_{(.322)}$	$.972_{(.069)}$	$.885_{(.230)}$	$.997_{(.019)}$	$.995_{(.050)}$
	M2	Reg_1	$.970_{(.074)}$	$.978_{(.111)}$	$.991_{(.033)}$	$.997_{(.041)}$	$1.00_{(.007)}$	$1.00_{(.000)}$
		Reg_2	$.989_{(.047)}$	$.997_{(.058)}$	$1.00_{(.007)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	$1.00_{(.000)}$
	M3	Reg_1	$.943_{(.110)}$	$.907_{(.261)}$	$.973_{(.064)}$	$.983_{(.122)}$	$.998_{(.014)}$	$1.00_{(.000)}$
		Reg_2	$.914_{(.153)}$	$.794_{(.303)}$	$.962_{(.094)}$	$.938_{(.170)}$	$.997_{(.020)}$	$.999_{(.019)}$
	M4	Reg_1	$.919_{(.133)}$	$.758_{(.387)}$	$.964_{(.081)}$	$.943_{(.213)}$	$.998_{(.018)}$	$.998_{(.029)}$
		Reg_2	$.931_{(.132)}$	$.837_{(.262)}$	$.976_{(.070)}$	$.969_{(.130)}$	$.998_{(.016)}$	$1.00_{(.000)}$
SCAD	**M1**	Reg_1	$.883_{(.207)}$	$.795_{(.315)}$	$.978_{(.085)}$	$.948_{(.187)}$	$.994_{(.032)}$	$.997_{(.041)}$
		Reg_2	$.935_{(.144)}$	$.718_{(.297)}$	$.980_{(.058)}$	$.918_{(.190)}$	$.996_{(.027)}$	$.993_{(.057)}$
	M2	Reg_1	$.974_{(.074)}$	$.976_{(.107)}$	$.996_{(.025)}$	$.995_{(.051)}$	$1.00_{(.007)}$	$1.00_{(.000)}$
		Reg_2	$.981_{(.061)}$	$1.00_{(.000)}$	$.999_{(.010)}$	$1.00_{(.000)}$	$1.00_{(.000)}$	$1.00_{(.000)}$
	M3	Reg_1	$.945_{(.116)}$	$.938_{(.205)}$	$.979_{(.073)}$	$.993_{(.071)}$	$.998_{(.018)}$	$1.00_{(.00)}$
		Reg_2	$.877_{(.196)}$	$.810_{(.274)}$	$.968_{(.105)}$	$.949_{(.143)}$	$.997_{(.020)}$	$1.00_{(.000)}$
	M4	Reg_1	$.921_{(.136)}$	$.798_{(.364)}$	$.977_{(.071)}$	$.967_{(.155)}$	$.996_{(.025)}$	$1.00_{(.000)}$
		Reg_2	$.906_{(.159)}$	$.838_{(.258)}$	$.974_{(.079)}$	$.977_{(.105)}$	$.995_{(.029)}$	$1.00_{(.000)}$

[1] For BIC, $q = 5$ and 6 were used for models **M1-M2** and **M3-M4**, respectively. For the new method based on the three penalties, $q = 10$ was used for all the four models.

Table 3: 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 $K = 2$ are in **bold**.

		$n = 150$			$n = 250$			$n = 500$		
Model	K	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC	RAIC	RBIC	RMSC
M1	1	.022	.561	.068	.000	.144	.004	.000	.004	.004
	2	**.288**	**.432**	**.245**	**.496**	**.848**	**.644**	**.583**	**.996**	**.861**
	3	.194	.007	.094	.216	.008	.072	.166	.000	.045
	4 or 5	.496	.000	.593	.288	.000	.028	.251	.000	.090
M2	1	.000	.014	.000	.000	.000	.000	.000	.000	.000
	2	**.578**	**.972**	**.550**	**.783**	**.993**	**.733**	**.814**	**1.00**	**.823**
	3	.202	.014	.032	.148	.007	.040	.122	.000	.034
	4 or 5	.220	.000	.418	.069	.000	.227	.064	.000	.143
M3	1	.013	.430	.103	.000	.107	.020	.000	.003	.000
	2	**.350**	**.570**	**.283**	**.513**	**.887**	**.663**	**.673**	**.997**	**.860**
	3	.253	.000	.057	.213	.006	.027	.140	.000	.007
	4 or 5	.384	.000	.557	.274	.000	.290	.187	.000	.133
M4	1	.010	.380	.100	.007	.103	.033	.000	.000	.000
	2	**.257**	**.620**	**.237**	**.507**	**.897**	**.650**	**.657**	**1.00**	**.917**
	3	.247	.000	.057	.173	.000	.010	.153	.000	.003
	4 or 5	.486	.000	.606	.313	.000	.307	.190	.000	.080

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

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.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Stat. 24*, 2350–2383.

Brockwell, P. and R. Davis (1991). *Time Series: Theory and Methods* (Second ed.). New York: Springer-Verlag.

Chen, J., X. Tan, and R. Zhang (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica 18*, 443–465.

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, J. Olsson, and R. van Handel (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Stat. 39*, 474–513.

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.

Francq, C. and J. M. Zakoïan (2001). Stationarity of multivariate markov-switching arma models. *Journal of Econometrics 102*, 339–364.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer New York.

Hamilton, J. D. (1989). A new approach to the economic anlaysis of nonstationary time series and the business cycle. *Econometrica 57*, 357–384.

Hamilton, J. D. (2016). Macroeconomic regimes and regime shifts. In J. B. Taylor and H. Uhlig (Eds.), *Handbook of Macroeconomics*, Volume 2, Chapter 3, pp. 163–201. Elsevier. http://www.sciencedirect.com/science/article/pii/S1574004816000057.

Hathaway, R. J. (1985). A constraint formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Stat. 13*, 795–800.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya 62*(Series A), 49–66.

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

Konishi, S. and G. Kitagawa (2008). *Information Criteria and statistical modeling*. Springer.

Krishnamurthy, V. and T. Rydén (1998). Consistent estimation of linear and non-linear autoregressive models with Markov regime. *J. Time Ser. Anal. 19*, 291–307.

Krishnamurthy, V. and G. G. Yin (2002). Recursive algorithms for estimation of hidden markov models and autoregressive models with markov regime. *IEEE Trans. Information Theory 48*, 458–476.

Lu, Z.-H. (2009). Covariate selection in mixture models with the censored response variable. *Comp. Stat. and Data Anal. 53*, 2710–2723.

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.

Psaradakis, Z. and N. Spagnolo (2006). Joint determination of the state dimension and autoregressive order for models with Markov regime switching. *J. Time Ser. Anal. 27*, 753–766.

Smith, A., A. N. Prasad, and C.-L. Tsai (2006). Markov-switching model selection using kullback-leibler divergence. *Journal of Econometrics 134*, 553–577.

Tibshirani, R. (1996). Regression shrinkage and selection via Lasso. *J. Roy. Statist. Soc. B 58*, 267–288.

Timmermann, A. (2000). Moments of markov switching models. *Journal of Econometrics 96*, 75–111.

Wong, C. S. and W. K. Li (2000). On a mixture autoregressive model. *J. Roy. Statist. Soc. B 62*, 95–115.

Yao, J. F. and J. G. Attali (2000). On stability of nonlinear ar processes with markov switching. *Adv. Appl. Prob. 32*, 394–407.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*, 1418–1429.