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

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

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 that capture temporal heterogeneity exhibited by time series data. Constructing a Markov regime-switching model requires making several specifications related to the state and observation models. In particular, the complexity of these models must be specified when fitting to a data set. We propose new regularization methods based on a conditional likelihood for simultaneous autoregressive-order and parameter estimation, with the number of regimes fixed. We use a regularized Bayesian information criterion to select the number of regimes. Unlike existing information-theoretic approaches, the proposed methods avoid an exhaustive search of the model space for model selection, and thus are computationally more efficient. We establish the large-sample properties of the proposed methods for estimation, model selection, and forecasting. We also evaluate the finite-sample performance of the methods using simulations, and apply them to analyze two real data sets.

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

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

1. Introduction

Markov regime-switching models (Hamilton, 1989) are commonly used to incorporate the latent structure of a time series in order to capture the nonstationarity or time-inhomogeneity in real data. An extensive body of literature discusses the use of these models in econometrics, with many applications related to representations of economic or business cycles (Hamilton, 2016). Other applications include those related to speech recognition and neurobiology (Krishnamurthy and Yin, 2002).

In a Markov regime-switching model, we typically use a discrete-state and often first-order Markov “state” model to capture unobserved stochastic variation corresponding to regime changes. In addition, conditional on the latent structure, we use a conventional time series “observation” model to represent the observed data. In practice, we must specify the complexity of the model, that is, the number of regimes (states) and the structure of each regime-specific observation model. In this study, we develop new results based on a regularized conditional likelihood that demonstrate that sparse estimations 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 modeling 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 $\{S_t : t =$

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

$1, 2, \ldots\}$ on a finite set $\{1, \ldots, K\}$ that determines, for each time t, the Gaussian autoregressive regime that dictates the stochastic behavior of an observable discrete-time series $\{Y_t : t = 1, 2, \ldots\}$. Specifically, S_t is presumed to be a first-order Markov chain, parameterized through a transition matrix $\mathbb{P}$. Conditional on $S_t = j$, the distribution of Y_t depends on the lagged Y, 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 the conditional mean or autocovariance structure.

A maximum likelihood estimation (MLE) is typically used for inferences in MSAR models, implemented using adaptations of filtering-smoothing and forward-backward algorithms (Frühwirth-Schnatter, 2006; Baum et al., 1970). Krishnamurthy and Rydén (1998) and Douc et al. (2004, 2011) establish the consistency and asymptotic normality of the 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 which 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 a 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, estimations of the AR-coefficients and the transition matrix $\mathbb{P}$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

prediction.

Information criteria such the AIC (Akaike, 1973), BIC (Schwarz, 1978), and their variations (Psaradakis and Spagnolo, 2006) are commonly used to simultaneously select 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 rarely computationally feasible.

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

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

1. We develop a new regularized conditional likelihood method that, to the best of our knowledge, is the first in the field for simultaneous AR-order and parameter estimation in MSAR models, and propose a regularized BIC (RBIC) for choosing the

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

number of regimes K. Compared with existing methods, given (K, q), the proposed method simultaneously estimates the AR-orders and parameters without an exhaustive search of 2^{Kq} possible models, and is thus computationally efficient. This is supported by our analysis of the average computational time (in seconds) taken by each method to complete the per-sample results in our simulations; see Section 7.1.

2. We study the large-sample properties of the methods, and assess their finite-sample performance using simulations. Our results show that, under standard regularity conditions, when K is given or consistently estimated, the regularization method is consistent in terms of the AR-order and parameter estimation, and achieves consistent predictions of future values of the process. Furthermore, we discuss the 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 using the RBIC.

The rest of the paper is organized as follows. In Section 2, we introduce Gaussian MSAR models. In Section 3, we develop new regularization methods and present their numerical implementation. Section 4 discusses predictions using MSAR models. In Section 5, we estimate the number of AR-regimes. Section 6 contains a theoretical discussion. Our simulations are presented in Section 7. We analyze two real data sets in Section 8. Section 9 concludes the paper.

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

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 $\{1, \ldots, K\}$, where K is 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 an 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, Y_t follows an inhomogeneous Markov process, such that for each t, the conditional distribution of Y_t depends only on the regime indicator $S_t = j$ and the lagged Y, 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 assumed to be a linear process driven by a white-noise error with appropriate finite-moment conditions. Note that the MSAR models under consideration are rather general. They encompass important special cases, including the mixture of autoregressive models studied by Wong and Li (2000), and the MSAR models with

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

common AR-orders and AR coefficients across the regimes; that is, $q_j = q$ and $\theta_{jl} = \theta_l$, for $j = 1, \ldots, K$ and $l = 1, \ldots, q$, as discussed in Frühwirth-Schnatter (2006). The stationarity and ergodicity conditions of MSAR models are studied by Yao and Attali (2000) and Francq and Zakoïan (2001). Timmermann (2000) provides calculations for the 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 Section 1 of the Supplementary Material (henceforth referred to as the Supplement) shows that the lag-l population PACF of Y_t is zero for any $l > q^*$, a property shared by the standard AR model of order q^* (Brockwell and Davis, 1991). In practice, the sample PACF of Y_t can be used to estimate q^* in an MSAR model, but it gives little insight into the regime AR-orders q_j, which are also the focus of our inferences. We now introduce a conditional likelihood function as the basis for 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 an MSAR model. The joint density or complete data likelihood, by the assumptions and for some prespecified 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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

initial probability $\Pr[S_1 = s_1]$ can be incorporated in two ways: it can be treated as a separate marginal law that is inferred or conditioned upon during an inference, or we may use the stationary distribution $\Pr[S_1 = s_1] = \pi_{s_1}$, for $s_1 = 1, \ldots, K$, arising from the Markov chain with the 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) show 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 prespecified 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, which by the model assumptions, can be written as

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

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

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 of the regimes; the regularization method in Section 3 estimates the regime-specific q_j using the data. The treatment of the probability $\Pr[S_q = s_q|y_{1:q}]$ is similar to that of $\Pr[S_1 = s_1]$, discussed above. To avoid this specification, inspired by Douc

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

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 inferences in MSAR models is

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

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

The marginalization over the states s_t in (2.3) is achieved efficiently using the standard filtering/prediction recursions employed in the hidden Markov model literature. A numerical maximization of (2.4) with respect to $\mathbf{\Phi}_K$, treating s_t as the missing data, is relatively straightforward using the expectation-maximization (EM) algorithm (Dempster et al., 1977) 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, in general, all of the estimated AR-coefficients are nonzero. Thus, such an approach does not provide a sparse MSAR, as postulated. This observation, and the limitations of the existing methods, motivate us to investigate

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

regularized conditional likelihood methods.

3. Simultaneous AR-order and parameter estimation

The conditional log-likelihood $\ell_n(\mathbf{\Phi}_K; s_q)$ in (2.4), similarly to that of a Gaussian mixture model with unequal component variances ν_j, diverges to infinity when some ν_j goes to zero. This singularity can be avoided by imposing a positive lower bound on ν_j (Hathaway, 1985) or by adding a penalty function to the conditional log-likelihood (Chen et al., 2008). For convenience in the implementation, 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 a 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 the mean of $y_{q+1:n}$, respectively. 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. Refer to (3.1) as the adjusted conditional log-likelihood. We now introduce the new regularization method.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

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

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

using the penalty (regularization) function

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

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 $\widehat{\mathbf{\Phi}}_{n,K,s_q}(\lambda) \equiv \widehat{\mathbf{\Phi}}_{n,K,s_q} = \arg\max_{\mathbf{\Phi}_K} \{\mathcal{L}_n(\mathbf{\Phi}_K; s_q, \lambda)\}$ be the maximum penalized conditional likelihood estimator (MPCLE) of $\mathbf{\Phi}_K$. 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 a simultaneous AR-order and parameter estimation without evaluating all candidate MSAR models, and thus is computationally feasible.

In general, the method allows for regime-specific seasonality effects, owing to the zero estimates of some θ_{jl}. Using ADALASSO, we admit no seasonality effects. Furthermore, $|\theta_{jl}|$ deceases with an increase in the lag l, as discussed in Section 1 of the Supplement.

Computation: We use a modified EM algorithm for the maximization of $\mathcal{L}_n(\mathbf{\Phi}_K; s_q, \lambda)$ in (3.3). The core elements of the algorithm are given here; more details, including a

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

data-adaptive choice of λ, are given in Section 3.2 of the Supplement. In what follows, we fix $s_q \in \{1, \dots, K\}$, and denote $\boldsymbol{x}_t^\top = (1, y_{t-1}, \dots, y_{t-q})$.

For observation y_t, let V_{tij} equal one if $S_{t-1} = i$ and $S_t = j$, and zero otherwise; V_{tij} records the presence of a transition between regime i at time $t-1$ and regime j at time t. In addition, let U_{tj} equal one if $S_t = j$. The complete conditional log-likelihood is

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

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

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

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

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

M-step: We maximize the penalization of the conditional expectation of $\ell_n^c(\boldsymbol{\Phi}_K; s_q)$ computed in **E-step** using 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 $1 \le j \le K$, compute

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

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

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

where $T(z; \lambda) = \mathrm{sign}(z)(|z| - \lambda)_+$ 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 using

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

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

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

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

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

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

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 the minimum mean squared prediction error. In standard AR models, this leads to a straightforward prediction mechanism. In this section, we focus on the predictive density in MSAR models, which can also be used to compute the prediction values. Unlike many nonlinear models, the conditional expectation can easily be 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 $P(S_{n+1} = j|y_{1:n})$, for $j = 1, \ldots, K$, are computed recursively using the *prediction* and *filtering* probabili-

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

ties,

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

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

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

Thus, given the data $y_{1:n}$ and 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 noninformative uniform discrete distribution $\boldsymbol{\gamma}_q=(1/K,\ldots,1/K)$. The parameter $\boldsymbol{\Phi}_K$ is estimated using its MPCLE $\widehat{\boldsymbol{\Phi}}_{n,s_q,K}$, obtained from 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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

value (in the sense of the 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 $(\widehat{\theta}_{j0}, \widehat{\theta}_{jl})$ denotes 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 chosen using the data. Information criteria, such as the BIC based on the MLE, are commonly used to estimate K. We instead propose using a regularized BIC (RBIC) based on the MPCLE. Unlike the BIC, this does not search the model space when choosing the AR-orders, because 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 an MSAR model with the resulting MPCLE $\widehat{\mathbf{\Phi}}_{n,K,s_q}$, 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 nonzero estimated AR-coefficients, and denote

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

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

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

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 the finite-sample performance of the RAIC, RBIC, and a regularized version of the Markov-switching criterion (MSC) of Smith et al. (2006). The latter is computed based on the MPCLE, and we call it the RMSC. Note that, owing to the factor $\log(n-q)$ in (5.1), the penalty in the RBIC is more severe than those in the RAIC and RMSC. Thus, it is expected that in finite-sample situations, the RBIC may result in models with lower selected orders (underestimation) than those of selected by the other two criteria; see Section 7.2.

6. Theoretical study

We first study the asymptotic properties of the MPCLE when the true number of AR-regimes K is predetermined (Theorems 1 and 2). We then study $\widehat{K}_n$ in (5.2), and discuss the behavior of the MPCLE when this is used to estimate the number of regimes (Theorem 3). The regularity conditions $\mathbf{C}_1$–$\mathbf{C}_3$ on the penalty r_n and the tuning parameter λ_n, as well as the proofs of Theorems 1–3, are given in Sections 1 and 2, respectively, of the Supplement.

Notation: All vectors are column vectors, and we drop the transpose $^\top$, for convenience. We assume the observed time series is a sample from an MSAR model with K AR-regimes and a d-dimensional true parameter vector $\mathbf{\Phi}^* = (v_1^*, \dots, v_K^*, \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*, \mathbb{P}^* =$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

$\{\alpha^*_{ij}\}$), where $d = K(q+2) + K(K-1)$. The regime-specific AR-coefficient vector is $\boldsymbol{\theta}^*_j$, the variance is v^*_j, and the transition probability is $\alpha^*_{ij} > 0$, for $i, j = 1, \ldots, K$. We further assume that $\boldsymbol{\Phi}^*$ is an interior point of the compact parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$. We partition each regime-specific AR-coefficient vector as $\boldsymbol{\theta}^*_j = (\boldsymbol{\theta}^*_{j1}, \boldsymbol{\theta}^*_{j2})$, such that $\boldsymbol{\theta}^*_{j1}$ and $\boldsymbol{\theta}^*_{j2}$ contain the nonzero and zero AR-coefficients, respectively. We partition the parameter vector $\boldsymbol{\Phi}^* = (\boldsymbol{\Phi}^*_1, \boldsymbol{\Phi}^*_2)$ such that $\boldsymbol{\Phi}^*_2 = (\boldsymbol{\theta}^*_{12}, \ldots, \boldsymbol{\theta}^*_{K2}) = \mathbf{0}$. The subvector $\boldsymbol{\Phi}^*_1$ contains all intercepts θ^*_{j0}, the nonzero θ^*_{jl}, the variances ν^*_j, and the transition probabilities α^*_{ij}. Furthermore, let $\dim(\boldsymbol{\Phi}^*_1) = d_1 < d$. We partition any candidate parameter as $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)$, following $\boldsymbol{\Phi}^*$. We use $\widehat{\boldsymbol{\Phi}}_{n,s_q}$ 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 derivatives, and $\mathcal{R}''_n(\cdot; \lambda)$ be the matrix of the second derivatives of $\mathcal{R}_n(\boldsymbol{\Phi}; \lambda)$ with respect to $\boldsymbol{\Phi}$. In addition, 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 $n - q$. Because $q < \infty$, $n \sim n - q$ asymptotically. Thus, in what follows, we use n instead of $n - q$. Our first result establishes the 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*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

$\mathbf{C}_1$–$\mathbf{C}_2$ *in the Supplement. Then, there exists a local maximizer* $\widehat{\mathbf{\Phi}}_{n,s_q}$ *of* $\mathcal{L}_n(\mathbf{\Phi}; s_q, \lambda_n)$ *such that, as* $n \to \infty$, $\|\widehat{\mathbf{\Phi}}_{n,s_q} - \mathbf{\Phi}^*\|_2 = O_p\{n^{-1/2}(1 + a_n)\}$, *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 $\widehat{\mathbf{\Phi}}_{n,s_q}$ 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 for the MPCLE based on the SCAD, it is sufficient that $\lambda_n \to 0$ as $n \to \infty$, because 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 the ADALASSO, we need $\sqrt{n}\lambda_n = o(1)$.

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

Theorem 2. *Assume the same conditions of Theorem 1,* (λ_n, r_n) *satisfy Condition* $\mathbf{C}_3$, *and* $a_n = O(1)$. *We have, 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 the* 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 the parameter and the AR-order estimations. With the same

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

choice of λ_n, the MPCLE based on the LASSO is consistent in the AR-order estimation, but introduces bias to the estimators of the true nonzero AR-coefficients, a well-known property of the LASSO in other settings. For the ADALASSO, if $\lambda_n \sim n^{-1/2-\psi}$, for $0 < \psi < \frac{\gamma}{2}$, the resulting MPCLE is consistent in both the parameter and the AR-order estimations. Note that, given K and the conditions of Theorem 1 on Y_t, the standard BIC is consistent in the AR-order estimation (Konishi and Kitagawa, 2008). However, compared with the new method, the BIC has a higher computational cost of evaluating 2^{Kq} different MSAR models in order to choose a final model.

By the 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 the h-step-ahead prediction.

Next, we study the properties of the RBIC-based estimator $\widehat{K}_n$ in (5.2), and its effect on the MPCLE and, specifically, 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 by assuming a compact Euclidean space for the 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;*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

(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* $(y_{1:n+h})$. *This result also holds when the number of regimes is estimated using* $\widehat{K}_n$.

Part (i) indicates that $\widehat{K}_n$ does not asymptotically underestimate the true number of regimes K. Part (ii) shows that even if the number of regimes is overestimated, we can still obtain consistency in the estimated h-step-ahead predictive densities. Hence, for instance, (6.1) still holds. The consistency of $\widehat{K}_n$ 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, $\mathbf{\Theta}_c = \Big\{\mathbf{\Phi} = (v_1, \ldots, v_\kappa, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\kappa, \mathbb{P} = \{\alpha_{ij}\}) : v_j \geq \delta, \alpha_{ij} \in [\varepsilon, 1-\varepsilon]\Big\}$, where $\boldsymbol{\theta}_j$ belongs to a compact Euclidean subspace of $\mathbb{R}^q$. Similarly to the results of Keribin (2000) and Lu (2009), the supremum of the log-likelihood ratio statistic of the overestimated models over $\mathbf{\Theta}_c$ versus the true model behaves as $O_p(\log\log n)$, as $n \to \infty$. Thus, using the same proof technique as that of Theorem 3-(i) (Section 2 of the Supplement), the RBIC prevents an overestimation of K and, hence, yields the consistency of $\widehat{K}_n$. In our simulations in Section 7.2, we find that the RBIC performs well in estimating K without any restrictions on the parameter space.

7. Simulation study

We evaluate the finite-sample performance of the proposed methods using simulations. We generated times series data from Gaussian MSAR models with $K = 2, 3$ AR-regimes.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

For the two-state models, the specified parameters are given in Table 1.

Table 1: Simulation parameter settings

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

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; this model and 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 performed in C++ 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 using the following measures:

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

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

– estimation error: $L_2 = \|\widehat{\psi} - \psi\|_2$ losses of the estimates $(\widehat{\psi})$ of the 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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

$q = 10$ in the penalized log-likelihood (3.3) for all models; the parameter λ is chosen using the information criterion given in Section 3.2 of the Supplement. To reduce the computational burden of the BIC for the AR-order estimation, we set the smaller common AR-orders $q = 5$ and $q = 6$ for **M1**–**M2** and **M3**–**M4**, respectively; these orders produce about 961 and 3969 models, respectively, to be examined by the BIC. We also examine the performance of the new method with the smaller values $q = 5, 6$; 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 it is overestimated.

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

Table 2 shows the average and standard deviation, over 300 replications, of the ES1 and ES2 values corresponding to models **M1**–**M4**. Because the results are similar when conditioning on an 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% and 100%, and for the new method varies between 88% and 100%, across the models **M1**–**M4**, sample sizes $n = 150, 250, 500$, and three penalties. For the average ES2, when $n = 150$, the BIC performs better by correctly identifying the true nonzero θ_{jl} between 90% and 100% of the time for different models. These proportions for the LASSO, ADALASSO, and SCAD are about 57% to 100%, 74% to 100%, and 72% to 100%, respectively. For

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

$n = 250, 500$, the BIC, ADALASSO, and SCAD perform similarly, with an average ES2 of more than 92%; for the LASSO, the average is more than 83%.

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

Box plots of the empirical L_2 losses of the parameter estimates based on the BIC, LASSO, ADALASSO, and SCAD, as well as 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 the LASSO, are higher than those of the estimator under the oracle model. As n increases, the performance of all of the estimates improves, and is comparable to that of the oracle estimator.

Similarly to the BIC, we 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 box plots of the empirical L_2 losses are given in Table A8 and Figures A9–A12, respectively, of the Supplement. For $n = 150$, the performance of

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

the method (in term of the ES1, ES2, and loss) improves as the AR-order upper bound q reduces from 10 to 5 or 6. This is expected, because reducing q decreases the potential number of parameters $K(q+2)+K(K-1)$ to be estimated. As n increases to $250, 500$, the effect of q becomes 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 $\widehat{K}_n$ to minimizes the criterion. Here, we report the results when the MPCLE is obtained using the 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 $n = 150, 250$, the RBIC has a higher percentage of underestimation of the true K, while the RAIC and RMSC tend to overestimate K. As explained at the end of Section 5, this behavior is expected because the penalty function in the RBIC in (5.1) is heavier than those in the other two criteria, which results in higher percentages of underestimation of the true K by the RBIC. As the sample size increases to $n = 500$, the percentages of underestimation of the true K by the RBIC tend to zero, supporting the result of Theorem 3-(i). We can see that, when $n = 500$, the RBIC estimates the true K almost

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

100% of the time in all four models. For $n = 500$, the RMSC estimates the true K about 82% to 92% of the time across the four models, while the RAIC estimates K approximately 58% to 81% of the time.

Finally, we examine the finite-sample behavior 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 to compute $\log[\widehat{f}_K(y_{n+1:h}|y_{1:n})]$. Figure A6 of the Supplement shows box plots of the log-predictive densities. We find the following: 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; and 3) as n increases, the log-predictive density values of the overestimated models ($K \geq 3$) are approximately equal to those of the true model, supporting the result of Theorem 3-(ii).

8. Real-data analysis

We illustrate the application of our method using two real-data examples. Figures A7 and A8 of the Supplement are used through our analysis below. We use the sample PACF

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

to obtain an (approximate) upper bound q, required by our regularization method, for the maximal AR-order $q^* = \max_{1\leq j\leq K} q_j$. 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 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 $\omega_{jl} = |\tilde{\theta}_{jl}|^{-1}$ performed similarly or worse, in terms of log-predictive density values, than those discussed below, and thus are not reported here. The $\tilde{\theta}_{jl}$ is the (conditional) MLE.

Example 1. The 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 United States for the period from the first quarter of 1947 to the third quarter of 2016, obtained from `https://fred.stlouisfed.org`. Figure A7-(a) shows a time series plot of the 278 observations. The plot shows that the variation in the series changes over time, which motivated us to consider fitting an 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 the MSAR with $K = 1, 2, 3, 4$. The RBIC values based on the SCAD are 691.9, $\mathbf{658.7}$, 690.4, and 720.2, respectively. Those based on the ADALASSO with the lag-dependent weights are 688.4, $\mathbf{665.7}$, 693.4, and 732.5, respectively. Thus, based on the RBIC, we select $\widehat{K} = 2$. The fitted models are given below; the standard errors are shown in brackets. The log-predictive density

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

values computed based on 11 observations for the two fitted MSAR models are −7.66 and −7.19, respectively.

	$(\widehat{\alpha}_{11}, \widehat{\alpha}_{22})$	$(\widehat{\nu}_1^{1/2}, \widehat{\nu}_2^{1/2})$	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 analyze the fitted model based on the SCAD; a similar analysis can be performed for the second model. Figure A7-(c) shows the classification of y_t into the two regimes of the model. Most of the observations from around 1950–1984 and 2008–2009 are classified into regime 2, with the remaining observations from around 1984–2007 and 2010–2013 classified into regime 1. We interpret the two regimes as follows: regime 1 describes periods when the growth rate was mostly positive and more stable, with a relatively lower variation than that of regime 2, where the growth rate is a combination of mostly large positives and, occasionally, large negatives (between 1950–1960 and, noticeably, around 2008–2009, the recent economic crisis), with a much higher variation. Figure A7-(c) shows that once the economy falls into one of the two regimes, it stays in that regime for a long period. This is confirmed by the large diagonal values $(\widehat{\alpha}_{11}, \widehat{\alpha}_{22})$ of the estimated transition probability matrix $\widehat{\mathbb{P}}$.

Example 2. The data are monthly U.S. unemployment rates (y_t) for the period 1948 to 2010, obtained from `https://www.bea.org`. The time series plot in Figure A8-(a)

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

shows an increasing–decreasing trend and 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 to compute the log-predictive density.

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

$$\begin{aligned}
\text{regime 1} : \widehat{\mu}_{t,1} &= \underset{[.016]}{.053}x_{t-2} + \underset{[.016]}{.094}x_{t-3} - \underset{[.015]}{.082}x_{t-12} \ , \ \ \widehat{\nu}_1^{1/2} = .136 \ , \ \ \widehat{\alpha}_{11} = .785 \\
\text{regime 2} : \widehat{\mu}_{t,2} &= \underset{[.038]}{.225}x_{t-4} + \underset{[.036]}{.272}x_{t-5} - \underset{[.033]}{.115}x_{t-10} - \underset{[.038]}{.244}x_{t-24} \ , \ \widehat{\nu}_2^{1/2} = .225 \ , \ \widehat{\alpha}_{22} = .551.
\end{aligned}$$

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

$$\begin{aligned}
\text{regime 1} : \widehat{\mu}_{t,1} &= -\underset{[.023]}{.112}x_{t-1} \ , \ \ \widehat{\nu}_1^{1/2} = .135 \ , \ \ \widehat{\alpha}_{11} = .975 \\
\text{regime 2} : \widehat{\mu}_{t,2} &= \underset{[.028]}{.129}x_{t-1} + \underset{[.029]}{.109}x_{t-2} \ , \ \ \widehat{\nu}_2^{1/2} = .238 \ , \ \ \widehat{\alpha}_{11} = .970,
\end{aligned}$$

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

9. Conclusion

We have developed new regularization methods for AR-order and parameter estimation, as well as for the selection of the number of AR-regimes in MSAR models. The methods present a substantial computational advantage over the AIC, BIC, and their variations by avoiding an exhaustive search of the model space, as well as having desirable large-sample properties. In addition, we have demonstrated the consistency of the optimal prediction, in terms of the mean squared prediction error and predictive density, when the number of regimes is either specified correctly or overspecified. 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 left to future work. There remain, however, interesting research challenges, such as determining the less restrictive conditions under which the RBIC provides a consistent estimator of the number of regimes.

Supplementary Material.

The online Supplementary Material contains four sections. Some preliminary results and regularity conditions are given in Section 1. Proofs of Theorems 1–3 are given in Section 2. Details of the numerical algorithm, including a data-adaptive tuning parameter selection method, are given in Section 3. Additional simulation results are

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

given in Section 4. All tables and figures related to the simulations, and the figures for the real-data analysis can be found at the end of the Supplement Material.

Acknowledgments

The authors would like to thank the editor, associate editor, and two anonymous referees for their constructive comments and suggestions. 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.

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

Table 2: 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$_1$	.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 four models.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

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 $q = 10$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrox and F. Caski. (Eds.), *Second International Symposium on Information Theory*, Budapest:, pp. 267. Akademiai Kiado.

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.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B* **39**, 1–38.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0190

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.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

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.

Statistica Sinica