Computational Statistics (2022) 37:1399–1433
https://doi.org/10.1007/s00180-021-01162-8

ORIGINAL PAPER

Bayesian analysis of mixture autoregressive models covering the complete parameter space

Davide Ravagli[1] · **Georgi N. Boshnakov**[1]

Received: 30 July 2020 / Accepted: 20 September 2021 / Published online: 1 November 2021

Abstract
Mixture autoregressive (MAR) models provide a flexible way to model time series with predictive distributions which depend on the recent history of the process and are able to accommodate asymmetry and multimodality. Bayesian inference for such models offers the additional advantage of incorporating the uncertainty in the estimated models into the predictions. We introduce a new way of sampling from the posterior distribution of the parameters of MAR models which allows for covering the complete parameter space of the models, unlike previous approaches. We also propose a relabelling algorithm to deal a posteriori with label switching. We apply our new method to simulated and real datasets, discuss the accuracy and performance of our new method, as well as its advantages over previous studies. The idea of density forecasting using MCMC output is also introduced.

Keyword Mixture autoregressive model · Stationarity · MCMC methods · Model selection · Forecasting

1 Introduction

Mixture autoregressive (MAR) models (Wong and Li 2000) provide a flexible way to model time series with predictive distributions which depend on the recent history of the process. Not only do the predictive distributions change over time, but they are also different for different horizons for predictions made at a fixed time point. As a consequence, they inherently accommodate asymmetry, multimodality and heteroskedasticity. For this reason, mixture autoregressive models have been considered

✉ Davide Ravagli
davide.ravagli@manchester.ac.uk

Georgi N. Boshnakov
georgi.boshnakov@manchester.ac.uk

[1] Department of Mathematics, The University of Manchester, Room 1.122, Alan Turing Building, Manchester M13 9PY, UK

a valuable alternative to other models for time series, such as the SETAR model (Tong 1990), the Gaussian transition mixture distribution model (Le et al. 1996), or the widely used class of GARCH models (Nelson 1991). Another useful feature of MAR models is that they model jointly the conditional mean and autocovariance. Moreover, the autocovariances are zero on a subspace of the parameters. So, if an uncorrelated (weak white noise) model is required, as is often the case for financial time series, the parameters can be restricted to that subspace.

MAR models can be thought of as random coefficient autoregressive models (Boshnakov 2011). Similarly to the usual autoregressions, there is a stationarity region for the parameters, outside which the MAR models are explosive and thus not generally useful.

Wong and Li (2000) considered estimation of MAR models based on the EM algorithm (Dempster et al. 1977). That method is particularly well suited for mixture-type models and works well. On the other hand, a Bayesian approach can offer the advantage of incorporating the uncertainty in the estimated models into the predictions.

Sampietro (2006) presented the first Bayesian analysis of MAR models. In his work, reversible jump MCMC (Green 1995) is used to select the autoregressive orders of the components in the mixture, and models with different number of components are compared using methods by Chib (1995) and Chib and Jeliazkov (2001), which exploit the marginal likelihood identity. In addition, he derives analytically posterior distributions for all parameters in the selected model.

The Bayesian updates of the autoregressive parameters are problematic, because the parameters need to be kept in the stationarity region, which is very complex, and so cannot really be updated independently of each other. In the case of autoregressive (AR) models, it is routine to use parametrisation in terms of partial autocorrelations (Jones 1987), which are subject only to the restriction to be in the interval $(-1, 1)$. Sampietro (2006) adapted this neatly to MAR models by parameterising the autoregressive parameters of each component of the MAR model with the partial autocorrelations of an AR model with those parameters.

A major drawback of Sampietro's sampling algorithm for the autoregressive parameters, is that it restricts the parameters of each component to be in the stationarity region of an autoregressive model. While this guarantees that the MAR model is stationary, it excludes from consideration considerable part of the stationarity region of the MAR model (Wong and Li 2000, p. 98; Boshnakov 2011). Depending on the mixture probabilities, the excluded part can be substantial. For example, most examples in Wong and Li (2000, p. 98) cannot be handled by Sampietro's approach, see also the examples in Sect. 4.

Lau and So (2008) proposed an infinite mixture of autoregressive models and used a semi-parametric approach based on a Dirichlet process (Ferguson et al. 1973) and the so called Gibbs version of the weighted Chinese restaurant process (Lo 2005) to select the optimal number of mixture components and assign observations to those. However, they do not assess conditions for second order stationarity of the model.

Wood et al. (2011) used data segmentation for estimation of a variant of the MAR models—they divide the data into segments and assign each segment to one mixture component. Their approach is aimed at time series which are piecewise autoregressions

(for example as a result of structural changes), has a different field of applications, and is not directly comparable to the MAR model considered here.

Hossain (2012) developed a full analysis (model selection and sampling), which reduced the constraints of Sampietro's analysis. Using Metropolis–Hastings algorithm and a truncated Gaussian proposal distribution for the moves, he directly simulated the autoregressive parameters from their posterior distribution. This method still imposes a constraint on the autoregressive parameters through the choice of boundaries for the truncated Gaussian proposal. While the truncation is used to keep the parameters in the stationarity region, the choice of boundaries is arbitrary and can leave out a substantial part of the stationarity region of the model. In addition, his reversible jump move for the autoregressive order seems conservative, as it uses functions which always prefer jumps towards low autoregressive orders (this will be seen in Sect. 3.5).

A common problem associated with mixtures is label switching (see for instance Celeux 2000), which derives from symmetry in the likelihood function. If no prior information is available to distinguish components in the mixture, then the posterior distribution will also be symmetric. It is essential that label switching is detected and handled properly in order to obtain meaningful results. A common way to deal with this, also used by Sampietro (2006) and Hossain (2012), is to impose identifiability constraints. However, it is well known that such constraints may lead to bias and other problems. In the case of MAR models, Hossain (2012) showed that these constraints may affect convergence to the posterior distribution.

We develop a new procedure which resolves the above problems. We propose an alternative Metropolis–Hastings move to sample directly from the posterior distribution of the autoregressive components. Our method covers the complete parameter space. We also propose a way of selecting optimal autoregressive orders using reversible jump MCMC for choosing the autoregressive order of each component in the mixture, which is less conservative than that of Hossain. We propose the use of a relabelling algorithm to deal a posteriori with label switching.

We apply the new methodology to both simulated and real datasets, and discuss the accuracy and performance of our algorithm, as well as its advantages over previous studies. Real data examples include two comparisons with previous literature (the IBM common stock closing prices, and the Canadian Lynx data, thoroughly analysed in Wong and Li 2000), and a previously unexplored dataset, which allows to introduce and discuss further practical aspects of parameter estimation and prediction with MAR models.

Finally, we briefly introduce the idea of density forecasting using MCMC output.

The structure of the paper is as follows. In Sect. 2 we introduce the mixture autoregressive model and the notation we need. In Sect. 3 we give detailed description of our method for Bayesian analysis of MAR models, including model selection, full description of the sampling algorithm, and the relabelling algorithm to deal with label switching. Section 4 shows results from application of our method to simulated and real dataset. Section 5 introduces the idea of density forecast using MCMC output.

2 The mixture autoregressive model

A process $\{y_t\}$ is said to follow a Mixture autoregressive (MAR) process if its distribution function, conditional on past information and parameter vector $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\sigma}, \boldsymbol{\phi})$, can be written as

$$F(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k F_k \left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k} \right), \tag{1}$$

where

- $\mathcal{F}_{t-1}$ is the sigma field generated by the process up to (and including) $t-1$. Informally, $\mathcal{F}_{t-1}$ denotes all the available information at time $t-1$, the most immediate past.
- g is the total number of autoregressive components.
- $\pi_k > 0, k = 1, \ldots, g$, are the mixing weights or proportions, specifying a discrete probability distribution. So, $\sum_{k=1}^{g} \pi_k = 1$ and $\pi_g = 1 - \sum_{k=1}^{g-1} \pi_k$. We will denote the vector of mixing weights by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$.
- F_k is the distribution function (CDF) of a standardised distribution with location parameter zero and scale parameter one.
  The corresponding density function will be denoted by f_k.
- $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kp_k})$ is the vector of autoregressive parameters for the kth component, with ϕ_{k0} being the shift. Here, p_k is the autoregressive order of component k and we define $p = \max(p_k)$ to be the largest order among the components. A useful convention is to set $\phi_{kj} = 0$, for $p_k + 1 \leq j \leq p$.
- $\sigma_k > 0$ is the scale parameter for the kth component. We denote by $\boldsymbol{\sigma} = (\sigma_1, \ldots \sigma_g)$ the vector of scale parameters. Furthermore, we define the precision, τ_k, of the kth component by
  $\tau_k = 1/\sigma_k^2$.
- If the process starts at $t = 1$, then Eq. (1) holds for $t > p$.

We will refer to the model defined by Eq. (1) as MAR$(g; p_1, \ldots, p_g)$ model. The following notation will also be needed. Let

$$\mu_{tk} = \phi_{k0} + \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}.$$

The error term associated with the kth component at time t is defined by

$$e_{tk} = y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i} = y_t - \mu_{tk}. \tag{2}$$

A useful alternative expression for ν_{tk} is the following mean corrected form:

$$\mu_{tk} = \mu_k + \sum_{i=1}^{p_k} \phi_{ki} (y_{t-i} - \mu_k).$$

Comparing the two representations we get

$$\phi_{k0} = \mu_k \left(1 - \sum_{i=1}^{p_k} \phi_{ki}\right).$$

If $\sum_{i=1}^{p_k} \phi_{ki} \neq 0$, we also have

$$\mu_k = \frac{\phi_{k0}}{1 - \sum_{i=1}^{p_k} \phi_{ki}}. \tag{3}$$

A nice feature of this model is that the one-step predictive distributions are given directly by the specification of the model with Eq. (1). The h-steps ahead predictive distributions of y_{t+h} at time t can be obtained by simulation (Wong and Li 2000) or, in the case of Gaussian and α-stable components, analytically (Boshnakov 2009).

We focus here on mixtures of Gaussian components. In this case, using the standard notations $\boldsymbol{\Phi}$ and $\boldsymbol{\phi}$ for the CDF and PDF of the standard Normal distribution, we have $F_k \equiv \boldsymbol{\Phi}$ and $f_k \equiv \boldsymbol{\phi}$, for $k = 1, \ldots, g$. The model in Eq. (1) can hence be written as

$$F(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \boldsymbol{\Phi}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k}\right) \tag{4}$$

or, alternatively, in terms of the conditional pdf

$$f(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \frac{\pi_k}{\sigma_k} \boldsymbol{\phi}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}}{\sigma_k}\right) \tag{5}$$

Conditional mean and variance of y_t are

$$\begin{aligned} \mathrm{E}[y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}] &= \sum_{k=1}^{g} \pi_k \left(\phi_{k0} + \sum_{i=1}^{p} \phi_{ki} y_{t-i}\right) = \sum_{k=1}^{g} \pi_k \mu_{tk} \\ \mathrm{Var}(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) &= \sum_{k=1}^{g} \pi_k \sigma_k^2 + \sum_{k=1}^{g} \pi_k \mu_{tk}^2 - \sum_{k=1}^{g} (\pi_k \mu_{tk})^2 \end{aligned} \tag{6}$$

The correlation structure of a stable MAR process with maximum order p is similar to that of an $AR(p)$ process. At lag h we have:

$$\begin{aligned} \rho_h &= \sum_{k=1}^{g} \pi_k \sum_{i=1}^{p} \phi_{ki} \rho_{|h-i|}, \qquad h \geq 1 \\ &= \sum_{i=1}^{p} \left(\sum_{k=1}^{g} \pi_k \phi_{ki}\right) \rho_{|h-i|} \end{aligned}.$$

Setting $a_i = \left(\sum_{k=1}^{g} \pi_k \phi_{ki}\right)$ for $i = 1, \ldots, p$, we see that these are analogous to the Yule-Walker equations for an $AR(p)$ model.

2.1 Stability of the MAR model

Stationarity conditions for MAR time series have some similarity to those for autoregressions with some notable differences. Below we give the results we need, see Boshnakov (2011) and the references therein for further details.

A matrix is stable if and only if all of its eigenvalues have moduli smaller than one (equivalently, lie inside the unit circle). Consider the companion matrices

$$A_k = \begin{bmatrix} \phi_{k1} & \phi_{k2} & \dots & \phi_{k(p-1)} & \phi_{kp} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}, \quad k = 1, \dots, g.$$

We say that the MAR model is stable if and only if the matrix

$$A = \sum_{k=1}^{g} \pi_k A_k \otimes A_k$$

is stable ($\otimes$ is the Kronecker product). If a MAR model is stable, then it can be used as a model for stationary time series. The stability condition is sometimes called stationarity condition.

If $g = 1$, the MAR model reduces to an AR model and the above condition states that the model is stable if and only if $A_1 \otimes A_1$ is stable, which is equivalent to the same requirement for A_1. For $g > 1$, it is still true that if all matrices $A_k, \dots, A_k$, $k = 1, \dots, g$, are stable, then A is also stable. However. the inverse is no longer true, i.e. A may be stable even if one or more of the matrices A_k are not stable.

What the above means is that the parameters of some of the components of a MAR model may not correspond to stationary AR models. It is convenient to refer to such components as "non-stationary".

Partial autocorrelations are often used as parameters of autoregressive models because they transform the stationarity region of the autoregressive parameters to a hyper-cube with sides $(-1, 1)$. The above discussion shows that the partial autocorrelations corresponding to the components of a MAR model cannot be used as parameters if coverage of the entire stationary region of the MAR model is desired.

3 Bayesian analysis of mixture autoregressive models

3.1 Likelihood function and missing data formulation

Given data $y_1, \dots, y_n$, the likelihood function for the MAR model in the case of Gaussian mixture components takes the form of (5)

$$L(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\pi}|\boldsymbol{y}) = \prod_{t=p+1}^{n} \sum_{k=1}^{g} \frac{\pi_k}{\sigma_k} \boldsymbol{\phi}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki}\, y_{t-i}}{\sigma_k}\right).$$

The likelihood function is not very tractable and a standard approach is to resort to the missing data formulation (Dempster et al. 1977).

Let $\boldsymbol{Z}_t = \left(Z_{t1}, \ldots, Z_{tg}\right)$ be a latent allocation random variable, where $\boldsymbol{z}_t$ is a g-dimensional vector with entry k equal to 1 if y_t comes from the kth component of the mixture, and 0 otherwise. We assume that the $\boldsymbol{z}_t$s are realisations of discrete random variables, independently drawn from the discrete distribution:

$$P(z_{tk} = 1|g, \boldsymbol{\pi}) = \pi_k, \qquad k = 1, \ldots, g. \tag{7}$$

and such that exactly one entry is 1, while the remaining entries are 0. This setup, widely exploited in the literature (see, for instance Dempster et al. 1977; Diebolt and Robert 1994) allows to rewrite the likelihood function in a much more tractable way as follows:

$$L(\boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{z}) = \prod_{t=p+1}^{n} \prod_{k=1}^{g} \left(\frac{\pi_k}{\sigma_k} \boldsymbol{\phi}\left(\frac{y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki}\, y_{t-i}}{\sigma_k}\right)\right)^{z_{tk}} \tag{8}$$

where $\boldsymbol{z}$ is a $(n-p) \times g$ matrix which rows are the vectors $\boldsymbol{z}_{p+1}, \ldots, \boldsymbol{z}_n$. In practice, the $\boldsymbol{z}_t$s are not available. We adopt a Bayesian approach to deal with this. We set suitable prior distributions on the latent variables and the parameters of the model and develop a methodology for obtaining posterior distributions of the parameters and dealing with other issues arising in the model building process.

3.2 Priors setup and choice of hyperparameters

The setup of prior distributions is based on Sampietro (2006) and Hossain (2012). In the absence of any relevant prior information it is natural to assume a priori that each data point is equally likely to be generated from any component, i.e. $\pi_1 = \cdots = \pi_g = 1/g$. This is a discrete uniform distribution, which is a particular case of the multinomial distribution. The conjugate prior of the latter is the Dirichlet distribution. We therefore set the prior for the mixing weigths vector, $\boldsymbol{\pi}$, to

$$\boldsymbol{\pi} \sim D\left(w_1, \ldots, w_g\right), \qquad w_1 = \cdots = w_g = 1. \tag{9}$$

The prior distribution on the component means is a normal distribution with common fixed hyperparameters ζ for the mean and κ for the precision, i.e.

$$\mu_k \sim N(\zeta, \kappa^{-1}), \qquad k = 1, \ldots, g. \tag{10}$$

For the component precisions, τ_k, a hierarchical approach is adopted, as suggested in Richardson and Green (1997). Here, for a generic kth component the prior is a

Gamma distribution with hyperparameters c (fixed) and λ, which itself follows a gamma distribution with fixed hyperparameters a and b. We have therefore

$$\begin{aligned} c &- \text{fixed} \\ \lambda &\sim Ga(a, b) \\ \tau_k \mid \lambda &\sim Ga(c, \lambda), \qquad k = 1, \ldots, g. \end{aligned} \tag{11}$$

The main difference between our approach and that of Sampietro (2006) and Hossain (2012) is in the treatment of the autoregressive parameters.

Sampietro (2006) exploits the one-to-one relationship between partial autocorrelations and autoregressive parameters for autoregressive models described in Jones (1987). Namely, he parameterises each MAR component with partial autocorrelations, draws samples from the posterior distribution of the partial autocorrelations via Gibbs-type moves and converts them to autoregressive parameters using the functional relationship between partial autocorrelations and autoregressive parameters. Of course, the term "partial autocorrelations" doe not refer to the actual partial autocorrellations of the MAR process, they are simply transformed parameters. The advantage of this procedure is that the stability region for the partial autocorrelation parameters is just a hyper-cube with marginals in the interval $(-1, 1)$, while for the AR parameters it is a body whose boundary involves non-linear relationships between the parameters.

A drawback of the partial autocorrelations approach in the MAR case is that it covers only a subset of the stability region of the model. Depending on the other parameters, the loss may be substantial.

Hossain (2012) overcomes the above drawbacks by simulating the AR parameters directly. He uses Random Walk Metropolis, while applying a constraint to the proposal distribution (a truncated Normal). The truncation is chosen as a compromise that ensures that most of the stability region is covered, while keeping a reasonable acceptance rate. Although effective with "well behaved" data, there are scenarios, especially concerning financial examples, in which the loss of information due to a pre-set truncation becomes significant, as will be shown later on. In this paper, we choose Random Walk Metropolis for simulation from the posterior distribution of autoregressive parameters, while exploiting the stability condition to avoid restraining the parameter space a priori.

With the above considerations, for the autoregressive parameters we choose a multivariate uniform distribution with range in the stability region of the model, and independence between parameters is assumed. Hence, for a generic $\boldsymbol{\phi}_k$ prior distribution is such that:

$$p(\boldsymbol{\phi} \mid \boldsymbol{\pi}) \propto \mathcal{I}\{Stable\}$$

where $\mathcal{I}$ denotes the indicator function assuming value 1 if the condition is satisfied and 0 otherwise. In other words, what we propose is a flat (uniform) prior over the stability region of the model. This uniform prior allows for better exploration of the parameter space than a Normal prior and does not mask multimodality.

Choice of hyperparameters Here we discuss the settings for the hyperparameters ζ, κ, a, b, and c. We have already discussed that the hyperparameters for the Dirichlet prior distribution on the mixing weights (all equal to 1). Also, λ is a hyperparameter but it is a random variable with distribution which will be fully specified once a and b are.

Following Richardson and Green (1997), let $\mathcal{R}_y = \max(y) - \min(y)$ be the length of the interval variation of the dataset. Also fix the two hyperparameters $a = 0.2$ and $c = 2$. The remaining hyperparameters are set as follows:

$$\zeta = \min(y) + \frac{\mathcal{R}_y}{2} \qquad \kappa = \frac{1}{\mathcal{R}_y} \qquad b = \frac{100a}{c\mathcal{R}_y^2} = \frac{10}{\mathcal{R}_y^2}$$

3.3 Posterior distributions and acceptance probability for RWM

Following Sampietro (2006) and Hossain (2012), posterior distributions for all but the autoregressive parameters are as follows:

$$\begin{aligned}
P(z_{tk} = 1 \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\tau}, \lambda, \boldsymbol{y}) &= \frac{\pi_k \boldsymbol{\phi}\left(\frac{e_{tk}}{\sigma_k}\right)}{\sum_{l=1}^{g} \pi_l \boldsymbol{\phi}\left(\frac{e_{tl}}{\sigma_l}\right)} \\
\boldsymbol{\pi} \mid \boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{y}, \boldsymbol{z} &\sim D\left(1 + n_1, \ldots, 1 + n_g\right) \\
\mu_k \mid \boldsymbol{\mu}_{-\mu_k}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{y}, \boldsymbol{z} &\sim N\left(\frac{\tau_k n_k \bar{e}_k b_k + \kappa\zeta}{\tau_k n_k b_k^2 + \kappa}, \frac{1}{\tau_k n_k b_k^2 + \kappa}\right) \\
\lambda \mid \boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{y}, \boldsymbol{z} &\sim Ga\left(a + gc,\ b + \sum_{k=1}^{g} \tau_k\right) \\
\tau_k \mid \boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\tau}_{-\tau_k}, \lambda, \boldsymbol{\pi}, \boldsymbol{y}, \boldsymbol{z} &\sim Ga\left(c + \frac{n_k}{2},\ \lambda + \frac{1}{2}\sum_{t=p+1}^{n} e_{tk}^2 z_{tk}\right)
\end{aligned} \tag{12}$$

where for $k = 1, \ldots, g$,

$$e_{tk} = y_t - \nu_{tk}, \quad n_k = \sum_{t=p+1}^{n} z_{tk}, \quad b_k = 1 - \sum_{i=1}^{p_k} \phi_{ki}, \quad \bar{e}_k = \frac{1}{n_k}\sum_{t=p+1}^{n} e_{tk} z_{tk}.$$

and $\boldsymbol{z}$ is the matrix of allocation random variables as defined in Sect. 3.1.

All these parameters are updated via a Gibbs-type move. Similarly, $\boldsymbol{z}_t$s are simulated from a multinomial distribution with associated posterior probabilities.

To update autoregressive parameters, let $\boldsymbol{\phi}_k, k = 1, \ldots, g$, be the set of current states of the autoregressive parameters, i.e. a set of draws from the posterior distribution of $\boldsymbol{\phi}_k$. We can simulate $\boldsymbol{\phi}_k^*$ from a proposal $MVN(\boldsymbol{\phi}_k, \Gamma_k^{-1})$ distribution, denoted by $q(\boldsymbol{\phi}_k^*, \boldsymbol{\phi}_k)$, with $\Gamma_k = \gamma_k I_{p_k}$, where I_{p_k} is the identity matrix of size p_k.

Here $\gamma_k, k = 1, \ldots, g$ is a tuning parameter, chosen in such way that the acceptance rate of RWM is optimal (20–25%) for component k. We allow γ_k to change between components, but to be constant within the same component. Notice the difference between our proposal and the two-step approach by Sampietro (2006), or the truncated Normal proposal chosen by Hossain (2012). The probability of accepting a move to the proposed $\boldsymbol{\phi}_k^*$ is

$$\alpha\left(\boldsymbol{\phi}_k, \boldsymbol{\phi}_k^*\right) = \min\left\{1, \frac{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^*\right) p(\boldsymbol{\phi}_k^*) q\left(\boldsymbol{\phi}_k, \boldsymbol{\phi}_k^*\right)}{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k\right) p(\boldsymbol{\phi}_k) q\left(\boldsymbol{\phi}_k^*, \boldsymbol{\phi}_k\right)}\right\}, \tag{13}$$

where $q\left(\boldsymbol{\phi}_k, \boldsymbol{\phi}_k^*\right) = q\left(\boldsymbol{\phi}_k^*, \boldsymbol{\phi}_k\right)$, due to the symmetry in the Normal proposal. Therefore, the acceptance probability will only depend on the likelihood ratio of the new set of parameters over the current set of parameters, i.e.

$$\alpha\left(\boldsymbol{\phi}_k, \boldsymbol{\phi}_k^*\right) = \min\left\{1, \frac{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^*\right)}{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k\right)}\right\} \tag{14}$$

where

$$\frac{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^*\right)}{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k\right)} = \frac{\prod_{\substack{t=p+1 \\ z_{tk}=1}}^{n} \exp\left\{-\frac{1}{2\sigma_k^2}\left(y_t - \phi_{k0}^* - \sum_{i=1}^{p_k} \phi_{ki}^* y_{t-i}\right)^2\right\}}{\prod_{\substack{t=p+1 \\ z_{tk}=1}}^{n} \exp\left\{-\frac{1}{2\sigma_k^2}\left(y_t - \phi_{k0} - \sum_{i=1}^{p_k} \phi_{ki} y_{t-i}\right)^2\right\}}$$

The priors are absent from the above formula, since their ratio is 1, due to the flat priors on the autoregressive parameters.

This means that the likelihood ratio for the kth component is independent of current values of parameters for the remaining components. This enables to calculate likelihood ratios separately for each component.

The procedure described builds a candidate model with updated mixing weights, shift, scale and autoregressive parameters. However, because stability of such model does not only depend on the autoregressive parameters, we must ensure that the stability condition of Sect. 2.1 is satisfied. If this is not the case, the candidate model and all its parameters are rejected, and the current state of the chain is set to be the same as at the previous iteration.

3.4 Dealing with label switching

Once the samples have been drawn, label switching is dealt with using a k-means clustering algorithm proposed by Celeux (2000). Identifiability constraints such as $\pi_1 > \pi_2 > \cdots > \pi_g$ are commonly used to make mixtures identifiable, but it is well known that this choice may be problematic. Examples are given in the discussion

to the paper by Richardson and Green (1997). In addition, Hossain (2012) showed that applying an identifiability constraint such as $\pi_1 > \pi_2 > \cdots > \pi_g$ may in some cases affect convergence of the chain, and they are not recommended particularly when there is evidence that two or more of the mixing weight may be equal. With our approach instead, we do not interfere with the chain during the simulation, and hence convergence is not affected.

Our algorithm works by first choosing the first m simulated values of the output after convergence. The value m shall be chosen small enough for label switching to not have occurred yet, and large enough to be able to calculate reliable initial values of cluster centres and their respective variances.

Let $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_g\right)$ be a subset of model parameters of size g, and N the size of the converged sample. The requirement on subsetting is that corresponding paramters of the different mixture components must be chosen, for instance $\boldsymbol{\theta} \equiv \left(\pi_1, \ldots, \pi_g\right)$ or $\boldsymbol{\theta} \equiv \left(\mu_1, \ldots, \mu_g\right)$ among other choices. For any centre coordinate $\theta_i, i = 1, \ldots, q$ we calculate the mean and variance, based on the first m simulated values, respectively as:

$$\bar{\theta}_i = \frac{1}{m}\sum_{j=1}^{m}\theta_i^{(j)} \qquad \bar{s}_i^2 = \frac{1}{m}\sum_{j=1}^{m}\left(\theta_i^{(j)} - \bar{\theta}_i\right)^2$$

We set this to be the "true" permutation of the components, i.e. we now have an initial center $\bar{\boldsymbol{\theta}}^{(0)}$ with variances $\bar{s}_i^{(0)^2}, i = 1, \ldots, q$. The remaining $g!-1$ permutations can be obtained by simply permuting these centres.

From these initial estimates, the rth iteration ($r = 1, \ldots, N - m$) of the procedure consists of two steps:

- the parameter vector $\boldsymbol{\theta}^{(m+r)}$ is assigned to the cluster such that the normalised squared distance

$$\sum_{i=1}^{g}\frac{\left(\theta_i^{(m+r)} - \bar{\theta}_i^{(m+r-1)}\right)^2}{\left(s_i^{(m+r-1)}\right)^2} \tag{15}$$

  is minimised, where $\bar{\theta}_i^{(m+r-1)}$ is the ith centre coordinate and $s_i^{(m+r-1)}$ its standard deviation, at the latest update $m + r - 1$.
- Centre coordinates and their variances are respectively updated as follows:

$$\bar{\theta}_i^{(m+r)} = \frac{m+r-1}{m+r}\bar{\theta}_i^{(m+r-1)} + \frac{1}{m+r}\theta_i^{(m+r)} \tag{16}$$

and

$$\begin{aligned}(s_i^{(m+r)})^2 = {} & \frac{m+r-1}{m+r}(s_i^{(m+r-1)})^2 + \frac{m+r-1}{m+r}\left(\bar{\theta}_i^{(m+r-1)} - \bar{\theta}_i^{(m+r)}\right)^2 \\ & + \frac{1}{m+r}\left(\theta_i^{(m+r)} - \bar{\theta}_i^{(m+r)}\right)^2\end{aligned} \tag{17}$$

for $i = 1, \ldots, g$.

For the mixture autoregressive case, it is not always clear which subset of the parameters should be used. In fact, group separation might seem clearer in the mixing weights at times, as well as in the scale or shift parameters. Therefore this method requires graphical assistance, i.e. checking the raw output looking for clear group separation. However, it is advisable not to use the autoregressive parameters, especially when the orders are different.

Once the selected subset has been relabelled, labels for the remaining parameters can be switched accordingly.

3.5 Reversible Jump MCMC for choosing autoregressive orders

For this step, we use Reversible Jump MCMC (Green 1995). At each iteration, one component k is randomly chosen from the model. Let p_k be the current autoregressive order of this component, and set p_{max} to be the largest possible value p_k may assume. For the selected component, we propose to increase or decrease its autoregressive order by 1 with probabilities

$$p_k^* = \begin{cases} p_k - 1 & \text{with probability } d(p_k) \\ p_k + 1 & \text{with probability } b(p_k) \end{cases}$$

where $b(p_k) = 1 - d(p_k)$, and such that $d(1) = 0$ and $b(p_{max}) = 0$. Notice that $d(p_k)$ (or equivalently $b(p_k)$) may be any function defined in the interval $[0, 1]$ satisfying such condition. For instance, Hossain (2012) introduced two parametric functions for this step. However, in absence of relevant prior information, we choose $b(p_k) = d(p_k) = 0.5$ in our analysis, while presenting the method in the general case.

Finally, it is necessary to point out that in both scenarios we have a 1-1 mapping between current and proposed model, so that the resulting Jacobian is always equal to 1.

Given a proposed move, we proceed as follows:

- If the proposal is to move from p_k to $p_k^* = p_k - 1$, we simply drop ϕ_{kp_k}, and calculate the acceptance probability by multiplying the likelihood ratio and the proposal ratio, i.e.

$$\alpha\left(\mathcal{M}_{p_k}, \mathcal{M}_{p_k^*}\right) = \min\left\{1, \frac{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k^*}\right) p(\boldsymbol{\phi}_k^{p_k^*})}{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k}\right) p(\boldsymbol{\phi}_k^{p_k})} \times \left[\frac{b\left(p_k^*\right)}{d\left(p_k\right)} \times \boldsymbol{\phi}\left(\frac{\phi_{kp_k} - \phi_{kp_k}}{1/\sqrt{\gamma_k}}\right)\right]\right\} \quad (18)$$

where $\boldsymbol{\phi}\left(\frac{\phi_{kp_k} - \phi_{kp_k}}{1/\sqrt{\gamma_k}}\right)$ is the density of the parameter dropped out of the model, according to its proposal distribution.

If the candidate model is not stable, then it is automatically rejected, i.e. $\alpha\left(\mathcal{M}_{p_k}, \mathcal{M}_{p_k^*}\right) = 0$.

- If the proposed move is from p_k to $p_k^* = p_k + 1$, we proceed by simulating the additional parameter from a suitable distribution. In absence of relevant prior information, the choice is to simulate a value from a uniform distribution centred in 0 and with appropriate range, so that values both close and far apart from 0, both positive and negative, are taken into consideration.
  These considerations lead to draw $\phi_{kp_k^*} \sim \mathcal{U}(-1.5, 1.5)$
  The acceptance probability in this case is

$$\alpha\left(\mathcal{M}_{p_k}, \mathcal{M}_{p_k^*}\right) = \min\left\{1, \frac{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k^*}\right) p(\boldsymbol{\phi}_k^{p_k^*})}{f\left(\boldsymbol{y} \mid \boldsymbol{\phi}_k^{p_k}\right) p(\boldsymbol{\phi}_k^{p_k^*})} \times \left[\frac{d\left(p_k\right)}{b\left(p_k^*\right)} \times 3\right]\right\} \tag{19}$$

  where 3 is the inverse of the $\mathcal{U}(-1.5, 1.5)$ density.
  Once again, if the candidate model is not stable, $\alpha\left(\mathcal{M}_{p_k}, \mathcal{M}_{p_k^*}\right) = 0$ and the current model is retained.

Notice that, similarly to the sampler for autoregressive parameters, the prior ratio in both cases is equal to 1 and therefore omitted.

3.6 Choosing the number of components

To select the appropriate number of autoregressive components in the mixture, we apply the methods proposed by Chib (1995) and Chib and Jeliazkov (2001), respectively, for use of output from Gibbs and Metropolis–Hastings sampling. Both make use of the marginal likelihood identity.

From Bayes' theorem, we know that

$$p(g|\boldsymbol{y}) \propto f(\boldsymbol{y} \mid g) p(g), \tag{20}$$

where $p(g)$ is the prior distribution on g, and $f(y \mid g)$ is the marginal likelihood function, defined as

$$f(\boldsymbol{y} \mid g) = \sum_p \int f(\boldsymbol{y} \mid \boldsymbol{\theta}, p, g) p(\boldsymbol{\theta}, p \mid g) d\boldsymbol{\theta} \tag{21}$$

with $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau})$ being the parameter vector of the model.

For any values $\boldsymbol{\theta}^*$, p^*, number of components g and observed data $\boldsymbol{y}$, we can use the marginal likelihood identity to decompose the marginal likelihood into parts that are know or can be estimated

$$\begin{aligned} f(\boldsymbol{y}|g) &= \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}^*, p^*, g) p\left(\boldsymbol{\theta}^*, p^* \mid g\right)}{p\left(\boldsymbol{\theta}^*, p^* \mid \boldsymbol{y}, g\right)} \\ &= \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}^*, p^*, g) p\left(\boldsymbol{\theta}^* \mid p^*, g\right) p(p^* \mid g)}{p\left(\boldsymbol{\theta}^* \mid p^*, \boldsymbol{y}, g\right) p(p^* \mid \boldsymbol{y}, g)} \end{aligned} \tag{22}$$

Notice that the only quantity not readily available in the above equation is $p\left(\boldsymbol{\theta}^* \mid p^*, \boldsymbol{y}, g\right)$. However, this can be estimated by running reduced MCMC simulations for fixed p^* (which can be obtained by the RJMCMC method described in Section 5.1), as follows:

$$\begin{aligned}\hat{p}\left(\boldsymbol{\theta}^* \mid p^*, \boldsymbol{y}, g\right) = &\hat{p}\left(\boldsymbol{\phi}^* \mid \boldsymbol{y}, p^*, g\right) \\ &\hat{p}\left(\boldsymbol{\mu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right) \\ &\hat{p}\left(\boldsymbol{\tau}^* \mid \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right) \\ &\hat{p}\left(\boldsymbol{\pi}^* \mid \boldsymbol{\tau}^*, \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right)\end{aligned} \tag{23}$$

Once these quantities are estimated (see 25, 26, 27, 28), plug them in Eq. (22), together with the other known quantities, to obtain the marginal likelihood for the model with fixed number of components g.

For higher accuracy of results, it is suggested to compare marginal likelihood with different g at points of high density in the posterior distribution of $\boldsymbol{\theta}^*$. We will use the estimated highest posterior density values.

Estimation of $\hat{p}(\boldsymbol{\phi}^* \mid \boldsymbol{y}, \boldsymbol{p}^*, \boldsymbol{g})$

Suppose we want to estimate $\hat{p}\left(\boldsymbol{\phi}_{\boldsymbol{k}}^* \mid p^*, \boldsymbol{y}, g\right)$, for $k = 1, \ldots, g$. We partition the parameter space into two subsets, namely $\Psi_{k-1} = \left(p, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{k-1}, g\right)$ and $\Psi_{k+1} = \left(\boldsymbol{\phi}_{k+1}, \ldots, \boldsymbol{\phi}_g, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}\right)$, where parameters belonging to Ψ_{k-1} are fixed (known or already selected high density values).

First, produce a reduced chain of length N_j to obtain $\boldsymbol{\phi}_{\boldsymbol{k}}^*$, the highest density value for $\boldsymbol{\phi}_{\boldsymbol{k}}$, using the sampling algorithm in Section 4.3, applied to the non-fixed set of parameters only. Define Ψ_{k^*}, the set of known (fixed) parameters with the addition of $\boldsymbol{\phi}_{\boldsymbol{k}}^*$. From a second reduced chain of length N_i, simulate $\{\tilde{\Psi}_{k+1}^{(i)}, \tilde{z}^{(i)} \mid \Psi_{k^*}, \boldsymbol{y}\}$, as well as new draws $\tilde{\boldsymbol{\phi}}_k^{(i)}$ from the proposal density in Equation 10, centred in $\boldsymbol{\phi}_{\boldsymbol{k}}^*$.

Now, let $\alpha(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_{\boldsymbol{k}}^*)$ and $\alpha(\boldsymbol{\phi}_{\boldsymbol{k}}^*, \tilde{\boldsymbol{\phi}}_k^{(i)})$ denote acceptance probabilities respectively of the first and second chain. We can finally estimate the value of the posterior density at $\boldsymbol{\phi}_k^*$ as

$$\hat{p}\left(\boldsymbol{\phi}_k^* \mid p^*, \boldsymbol{\phi}_1^*, \ldots, \boldsymbol{\phi}_{k-1}^*, g\right) = \frac{\frac{1}{N_j}\sum_{j=1}^{N_j} \alpha(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_{\boldsymbol{k}}^*) q\left(\boldsymbol{\phi}_k^{(j)}, \boldsymbol{\phi}_k^*\right)}{\frac{1}{N_i}\sum_{i=1}^{N_i} \alpha(\boldsymbol{\phi}_{\boldsymbol{k}}^*, \tilde{\boldsymbol{\phi}}_k^{(i)})} \tag{24}$$

Repeat this procedure for all $k = 1, \ldots, g$ and multiply the single densities to obtain

$$\hat{p}\left(\boldsymbol{\phi}^* \mid \boldsymbol{y}, p^*, g\right) = \prod_{k=1}^{g} \hat{p}\left(\boldsymbol{\phi}_k^* \mid p^*, \boldsymbol{\phi}_1^*, \ldots, \boldsymbol{\phi}_{k-1}^*, g\right). \tag{25}$$

Note that there are no requirements on what N_i and N_j should be, granted the first chain is long enough to have reached the stationary distribution.

Estimation of $\hat{p}\left(\boldsymbol{\mu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right)$

Run a reduced chain of length N. At each iteration i, generate draws $\boldsymbol{z}^{(i)}$, $\boldsymbol{\pi}^{(i)}$, $\boldsymbol{\tau}^{(i)}$, $\boldsymbol{\mu}^{(i)}$. Set $\boldsymbol{\mu}^* = \left(\mu_1, \ldots, \mu_g\right)$, the parameter vector of highest posterior density. The posterior density at $\boldsymbol{\mu}^*$ can be estimated as

$$\hat{p}\left(\boldsymbol{\mu}^* \mid \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{g} p\left(\mu_k^* \mid \boldsymbol{\phi}^*, \boldsymbol{\tau}^{(i)}, \boldsymbol{\pi}^{(i)}, \boldsymbol{y}, \boldsymbol{z}^{(i)}, p^*, g\right). \quad (26)$$

Estimation of $\hat{p}\left(\boldsymbol{\tau}^* \mid \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right)$

Run a reduced chain of length N_i. At each iteration i, generate draws $\boldsymbol{z}^{(i)}$, $\boldsymbol{\pi}^{(i)}$, $\boldsymbol{\tau}^{(i)}$. Set $\boldsymbol{\tau}^* = \left(\tau_1, \ldots, \tau_g\right)$, the parameter vector of highest posterior density. Posterior density at $\boldsymbol{\tau}^*$ can be estimated as

$$\hat{p}\left(\boldsymbol{\tau}^* \mid \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{g} p\left(\tau_k^* \mid \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{\pi}^{(i)}, \boldsymbol{y}, \boldsymbol{z}^{(i)}, p^*, g\right). \quad (27)$$

Estimation of $\hat{p}\left(\boldsymbol{\pi}^* \mid \boldsymbol{\tau}^*, \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right)$

Run a reduced chain of length N. At each iteration i, generate draws $\boldsymbol{z}^{(i)}$, $\boldsymbol{\pi}^{(i)}$. Set $\boldsymbol{\pi}^* = \left(\pi_1, \ldots, \pi_g\right)$, the parameter vector of highest posterior density. Posterior density at $\boldsymbol{\pi}^*$ can be estimated as

$$\hat{p}\left(\boldsymbol{\pi}^* \mid \boldsymbol{\tau}^*, \boldsymbol{\mu}^*, \boldsymbol{\phi}^*, \boldsymbol{y}, p^*, g\right) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{g} p\left(\pi_k^* \mid \boldsymbol{y}, \boldsymbol{z}^{(i)}, p^*, g\right). \quad (28)$$

3.7 Label switching and marginal likelihood

We discuss here the possible effect of incorrect label switching on the methodology in Sect. 3.6 for calculation of the marginal likelihood of the data. Recall the formula:

$$p\left(\boldsymbol{y} \mid x\right) = \frac{L\left(\boldsymbol{\theta}^* \mid x\right) p\left(\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}^* \mid y, x\right)}$$

where $\boldsymbol{\theta}^*$ is a point of high density (ideally of highest density) according to its posterior distribution.

For mixture models, we have that the likelihood function $L\left(\boldsymbol{\theta}^* \mid x\right)$ is a product of sums. For simplicity, suppose the model is a mixture of two components, and

$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. It follows that the conditional likelihood is

$$L\left(\boldsymbol{\theta}^* \mid x\right) = \prod_{i=1}^{n} \pi_1 f\left(\boldsymbol{y} \mid \theta_1\right) + \pi_2 f\left(\boldsymbol{y} \mid \theta_2\right) = \prod_{i=1}^{n} \pi_2 f\left(\boldsymbol{y} \mid \theta_2\right) + \pi_2 f\left(\boldsymbol{y} \mid \theta_1\right)$$

which means that the likelihood will be the same, regardless of the permutation of $\boldsymbol{\theta}$. Clearly, this holds for any number of components, g.

Under the same example, prior and posterior distributions for $\boldsymbol{\theta}$ may somewhat be affected by label switching. For prior distributions, this will happen when the practician sets up the experiment with informative priors, as this would bring the risk of evaluating a parameter under the wrong prior distribution. However, informative priors have the purpose of creating enough separation so that label switching does not in fact occur, as they incorporate prior belief on the distribution of the parameters (see Celeux 2000). In the examples presented here, prior distributions are the same across all components for corresponing parameters (for instance, all precisions follow a priori the same Gamma distribution), and therefore label switching will not affect the result.

Posterior distributions are most affected by label switching. However, we point few remarks in favor of the effectiveness of Chib (1995) and Chib and Jeliazkov (2001), even in the case of undetected label switching:

- The authors reassure that the methodology works effectively with a range of high density values under their respective posterior distributions. Returning to the two-component mixture example, suppose that there is undetected switching. The corresponding parameters in the two components, for example π_1 and pi_2, will show two modes. These modes will however correspond to the two highest density values, respectively, of π_1 and π_2. Therfore, it makes sense to believe that, ultimately, the choice of π_1^* and π_2^* will not change significantly, and high density values will be selected regardless.
- From the equations in Sect. 3.6, it is clear that undetected label switching could cause issues in evaluation of the posterior density of $\boldsymbol{\theta}^*$. This brings forward two considerations: first of all, label switching may occur due to little separation between the groups, meaning the two posterior distributions shall not be too dissimilar and a wrong labelling of a few iterations may not affect significantly the evaluation. Secondly, even when incorrect labelling does have an effect, each iteration is dampened by a $1/N$ factor since we take an average over the entire sample.
- It is important to recall that the algorithm sequentially fixes a set of parameters to their highest density values. This implies that, after very few parameters are fixed, label switching will definitely not occur for the remaining parameters. Going back to the two-component example, it is obvious that once we fix $\boldsymbol{\theta}_1^*$, there can no longer be label switching, since now we only draw a sample from $\boldsymbol{\theta}_2$.
- Finally, we must take into account that the contribution of the posterior distribution towards $p\left(y \mid x\right)$ will in general be rather small compared to that of $L\left(\boldsymbol{\theta} \mid \boldsymbol{y}, x\right)$, which is "immune" to label switching.

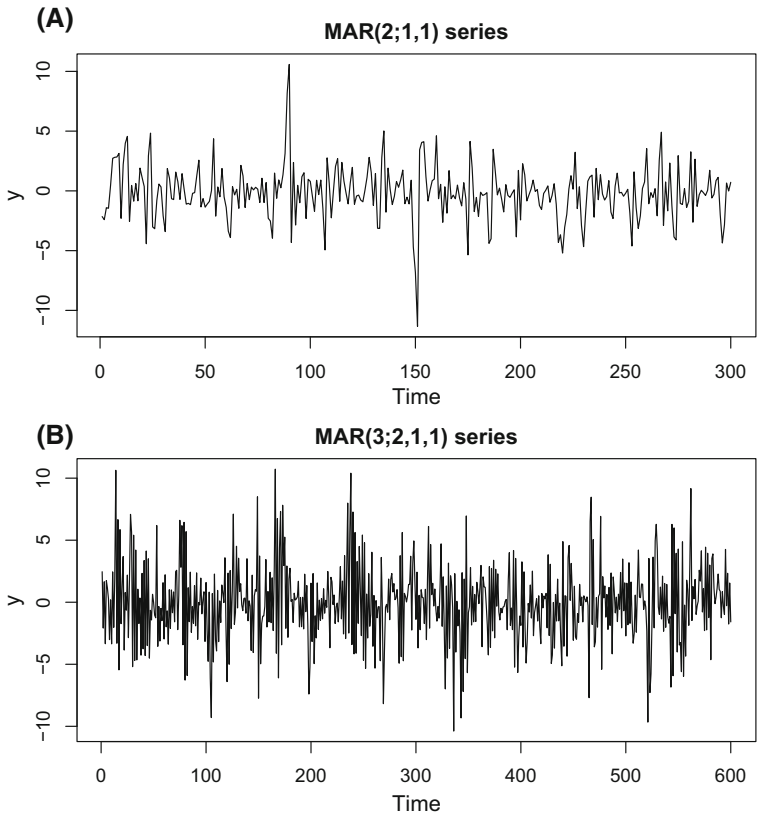

Fig. 1 Simulated series from Model **A** (top) and **B** (bottom)

While handled correctly throughout every example presented, we suggest that the effect of label switching could in general be neglectible when it comes to model selection with marginal likelihood (Figs. 1, 2).

4 Application

4.1 Simulation example

For comparative and demonstrative purposes, we show applications of our method using two simulated datasets from **(A)**

$$F(y_t|\mathcal{F}_{t-1}) = 0.5\mathbf{\Phi}\left(\frac{y_t + 0.5y_{t-1}}{1}\right) + 0.5\mathbf{\Phi}\left(\frac{y_t - y_{t-1}}{2}\right)$$

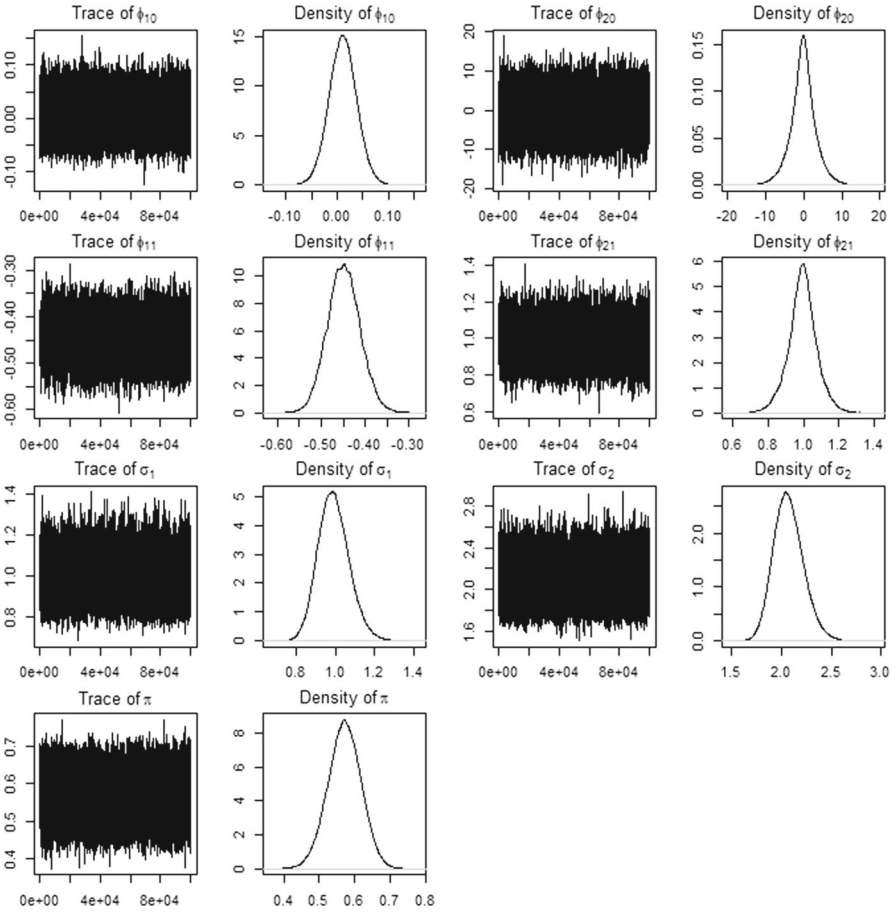

Fig. 2 Trace and density plots of parameters from (**A**). Sample size is 100,000, after discarding 50,000 draws as burn-in period

and **(B)**

$$F(y_t|\mathcal{F}_{t-1}) = 0.5\mathbf{\Phi}\left(\frac{y_t + 0.5y_{t-1} - 0.5y_{t-2}}{1}\right) + 0.3\mathbf{\Phi}\left(\frac{y_t + 0.4y_{t-1}}{2}\right) + 0.2\mathbf{\Phi}\left(\frac{y_t - y_{t-1}}{4}\right),$$

respectively with 300 and 600 observations. Process **(A)** is similar to the one considered by Hossain (2012) and Wong and Li (2000), while **(B)** was chosen to illustrate in practice how label switching is dealt with. The issue of label switching for **(B)** can be seen in Fig. 3, where we show the raw MCMC output with signs of label switch between components 2 and 3 (green and red lines), and the relabelled output after applying the algorithm.

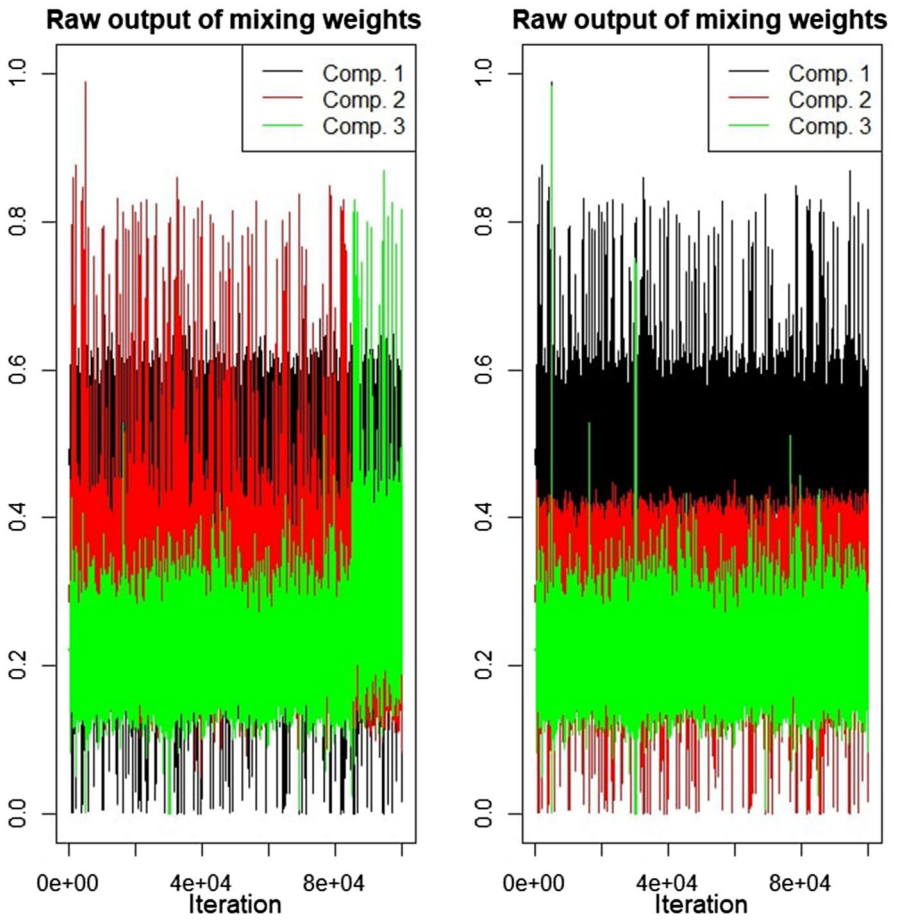

Fig. 3 Comparison of raw output (left) and output adjusted for label switching of mixing weights from (**B**). We notice the effectiveness of the relabelling algorithm applied to our MCMC

The algorithm then proceeds as described in Algorithm 1 below.

Algorithm 1

1: **for** $g \leftarrow 2, \ldots, g_{max}$ **do**
2: *RJMCMC and determine* $p_1^*, \ldots, p_k^*$
3: *Calculate* $f(\boldsymbol{y} \mid g)$
4: *Select* $g^* = \max f(\boldsymbol{y} \mid g)$, $g = 2, \ldots, g_{max}$
5: *Simulate* $f\left(\boldsymbol{\theta} \mid y, g^*, \boldsymbol{p}^*\right)$

As we can see from Tables 1, 2 and 3, and Figs. 2 and 4, the "true" model is chosen in both cases, as it has the largest marginal log-likelihood. In addition, true values of the parameters are found in high density regions of their respective posterior distributions (Tables 4, 5).

Table 1 Results from simulation studies

Model (A)	Preference	Marg. log-lik
MAR(2; 1, 1)	0.7399	−611.8113
MAR(3; 1, 1, 1)	0.1819	−613.0888
MAR(4; 1, 1, 1, 4)	0.0382	−923.1585
Model (B)	**Preference**	**Marg. log-lik**
MAR(2; 2, 1)	0.6258	−1468.628
MAR(3; 2, 1, 1)	0.2937	−1383.061
MAR(4; 2, 1, 2, 1)	0.0491	−1470.543

"Preference" is the proportion of times the model was retained against all models with same number of components

Table 2 Results of simulation from posterior distribution of the parameters under model (A)

Model A	True value	Posterior mean	Standard error	90% HPDR
ϕ_{10}	0	0.011	0.0268	(−0.032, 0.055)
ϕ_{20}	0	−0.183	3.273	(−5.672, 5.206)
ϕ_{11}	−0.5	−0.449	0.037	(−0.511, −0.389)
ϕ_{21}	1	0.994	0.079	(0.869, 1.136)
σ_1	1	0.992	0.079	(0.862, 1.119)
σ_2	2	2.069	0.149	(1.825, 2.311)
π	0.5	0.571	0.046	(0.494, 0.647)

Table 3 Results of simulation from posterior distribution of the parameters under model (B)

Model B	True value	Posterior mean	Standard error	90% HPDR
ϕ_{10}	0	0.001	0.018	(−0.009, 0.007)
ϕ_{20}	0	0.005	0.253	(−0.078, 0.091)
ϕ_{30}	0	0.102	2.133	(−3.145, 3.405)
ϕ_{11}	−0.5	−0.483	0.038	(−0.536, −0.427)
ϕ_{12}	0.5	0.498	0.034	(0.450, 0.547)
ϕ_{21}	−0.4	−0.461	0.105	(−0.596, −0.327)
ϕ_{31}	1	0.731	0.264	(0.432, 1.058)
σ_1	1	1.035	0.246	(0.804, 1.156)
σ_2	2	2.035	0.439	(1.625, 2.522)
σ_3	4	4.074	0.341	(3.559, 4.573)
π_1	0.5	0.495	0.056	(0.411, 0.568)
π_2	0.3	0.293	0.064	(0.207, 0.395)
π_3	0.2	0.212	0.041	(0.148, 0.275)

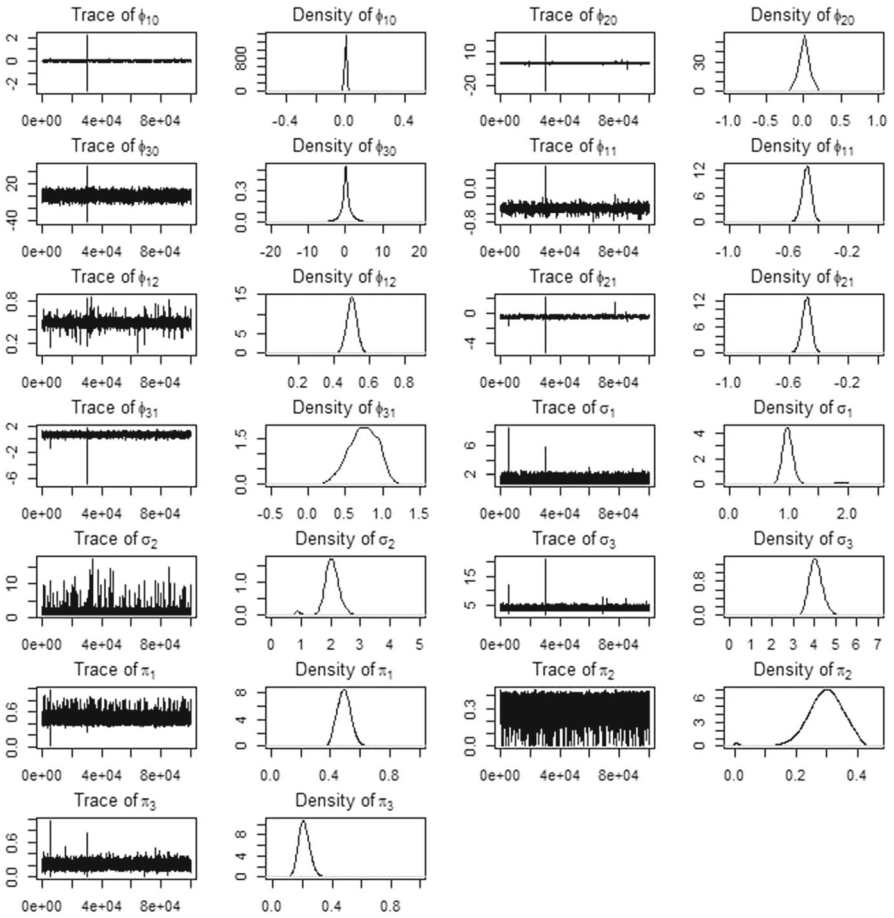

Fig. 4 Trace and density plots of parameters from (**B**). Sample size is 100,000, after discarding 50,000 draws as burn-in period

To show consistency of the method, the experiment on model **(A)** was replicated several times. Details on that are available in the Appendix.

4.2 The IBM common stock closing prices

The IBM common stock closing prices (Box and Jenkins 1976) is a financial time series widely explored several times in the literature (see, for instance Wong and Li 2000). It contains 369 observations from May 17th 1961 to November 2nd 1962. Original and difference series can be seen in Fig. 5.

Following previous studies, we consider the series of first order differences. To allow direct comparison with Wong and Li (2000) and Hossain (2012), we set $\phi_{k0} = 0,\ k = 1, \ldots, g$.

Table 4 Summary statistics for sample of size 100,000 from parameter posterior distributions of the selected model for the log-lynx data

Parameter	MLE	HDvalue	Standard error	90% HPDR
ϕ_{10}	0.4957	0.4962	1.6897	(−1.2599, 3.4341)
ϕ_{20}	2.5728	1.6945	1.2663	(−0.0138, 3.8897)
ϕ_{11}	0.9901	1.0779	0.0667	(0.9893 1.1320)
ϕ_{21}	1.5042	1.7205	0.1594	(1.4717, 1.9866)
ϕ_{22}	−0.8984	−0.7966	0.1528	(−1.0578, −0.5604)
σ_1	0.2313	0.3553	0.1846	(0.2162, 0.6451)
σ_2	0.4828	0.6010	0.1006	(0.4933, 0.7478)
π	0.2358	0.3280	0.1247	(0.1536, 0.5555)

Table 5 Summary statistics for sample of size 100,000 from parameter posterior distributions of the selected model for the daily temperature range data

Parameter	MLE	HD value	Standard Error	90% HPDR
ϕ_{10}	2.0554	1.9856	0.1125	(1.4010, 2.7046)
ϕ_{20}	2.5631	2.5978	0.1084	(2.0934, 3.1083)
ϕ_{11}	0.4967	0.4956	0.0655	(0.3866, 0.6034)
ϕ_{12}	0.2784	0.2951	0.0692	(0.1625, 0.3901)
ϕ_{21}	0.1989	0.2013	0.0544	(0.1161, 0.2939)
σ_1	1.8699	1.8772	0.1125	(1.7104, 2.080)
σ_2	1.1497	1.1710	0.1084	(1.0129, 1.3678)
π	0.5585	0.5602	0.0698	(0.4359, 0.6656)

With the procedure outlined in Algorithm 1 our method chooses a MAR(3; 4, 1, 1) to best fit the data, amongst all 2, 3, and 4 component models of maximum order $p_k = 5,\ k = 1, \ldots, g$. The RJMCMC algorithm selects this model roughly 25% of the time, ahead of MAR(3; 3, 1, 1) with 13%. The marginal log-likelihood for this model is −1245.51, which is larger than that of the best 2 and 4 component models, a MAR(2; 1, 1) and a MAR(4; 1, 1, 1, 1), which respectively have a value of marginal log-likelihood equal to −1248.921 and −1252.381. We immediately notice that this is different from the selected model in Wong and Li (2000). Such difference may occur as the frequentist approach fails to capture the multimodality in the distribution of certain parameters, which we can clearly see from Fig. 6. In fact, by attempting to fit a MAR(3; 4, 1, 1) model by EM-Algorithm from several different starting points, we concluded that this would actually provide a better fit than the MAR(3; 1, 1, 1) chosen by Wong and Li.

With one of the mixture components having a larger autoregressive order, label switching could only arise between the two components with autoregressive order 1. However, no signs of label switching were detected, and therefore no relabelling was required.

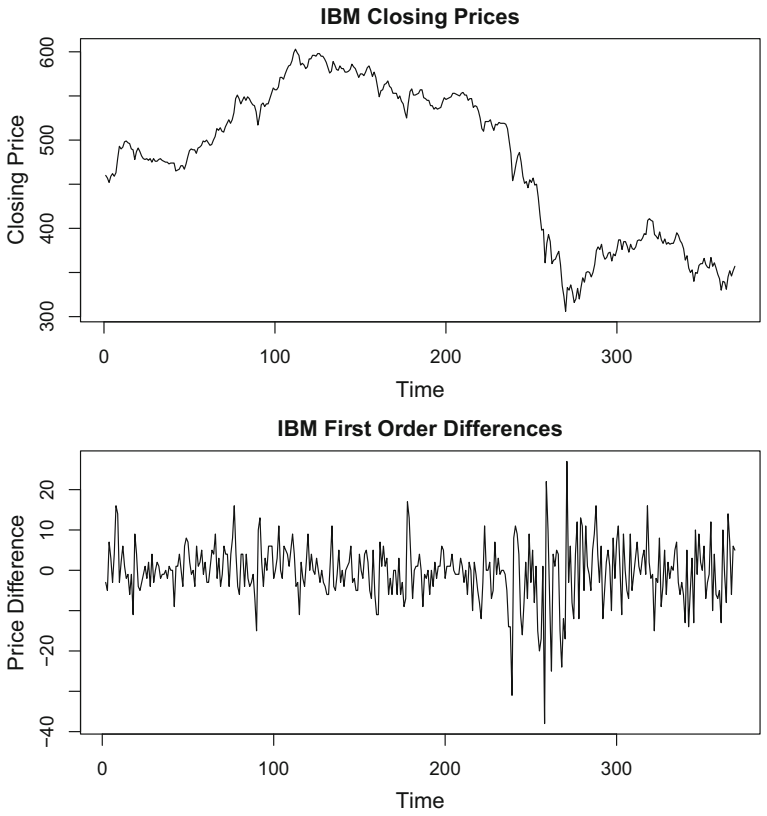

Fig. 5 Times series of IBM closing prices (top) and series of the first order differences (bottom)

Figure 7 shows once again the time series of first order differences of IBM closing prices, with the addition of two lines representing prediction intervals. Specifically, the red lines delimit the 95% highest density region of the average one step prediction densities, calculated using the sample from the parameter posterior distributions (see Sect. 5) for each y_t for $t > 4$. The blue lines denote instead the 95% prediction interval, calculated as the average one step point predictor $\pm$ twice the average conditional standard error recorded for the predictor. Point predictions and corresponding standard error are defined in (6). It appears from the picture that there is indeed an advantage in using prediction density over point prediction. While there is not a substantial difference between the two predictors in periods of relatively low volatility, as the very start of the series shows, the interval calculated using density prediction seem to provide more certainty in periods of higher volatility. This can be seen around observations 250–280, a period of high volatility for the series, where we can see several spikes, and therefore a large prediction interval, for the blue lines, while density prediction seems to accommodate well the sudden jumps in the series. Overall, it appears that, using the highest density region of density forecasts, a MAR model is able to account for the time-dependent volatility and its persistence in the IBM

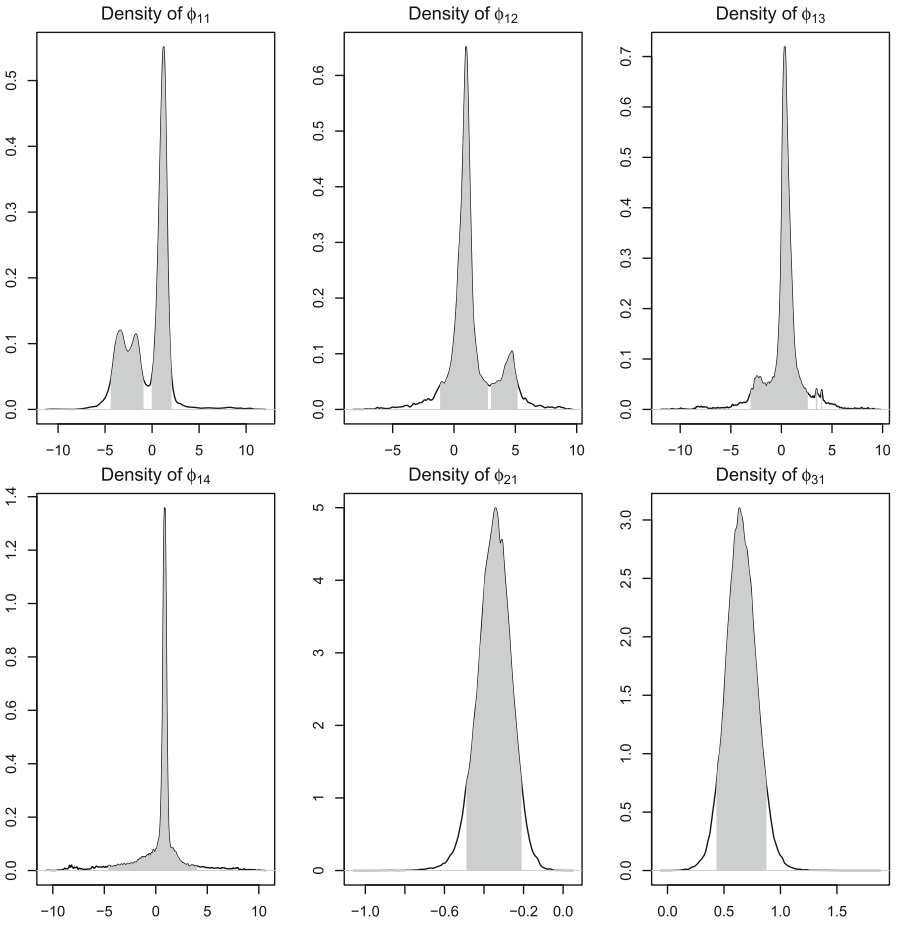

Fig. 6 Posterior distributions of autoregressive parameters from selected model MAR(3; 4, 1, 1), with 90% HPDR highlighted. We can clearly see multimodality occurring for certain parameters. Sample of 300,000 simulated values post burn-in

difference series. Furthermore, if we decided for a narrower prediction interval, the density forecast method would allow us to detect presence of multiple modes, so that the highest density region may no longer be continuous. This feature will be seen in Sect. 5.

4.3 The Canadian lynx data

Another dataset widely explored in time series literature, and particularly by Wong and Li (2000), is the annual record of Canadian lynx trapped in the Mackenzie River district in Canada between 1821 and 1934. This dataset, listed by Elton and Nicholson (1942), includes 111 observations.

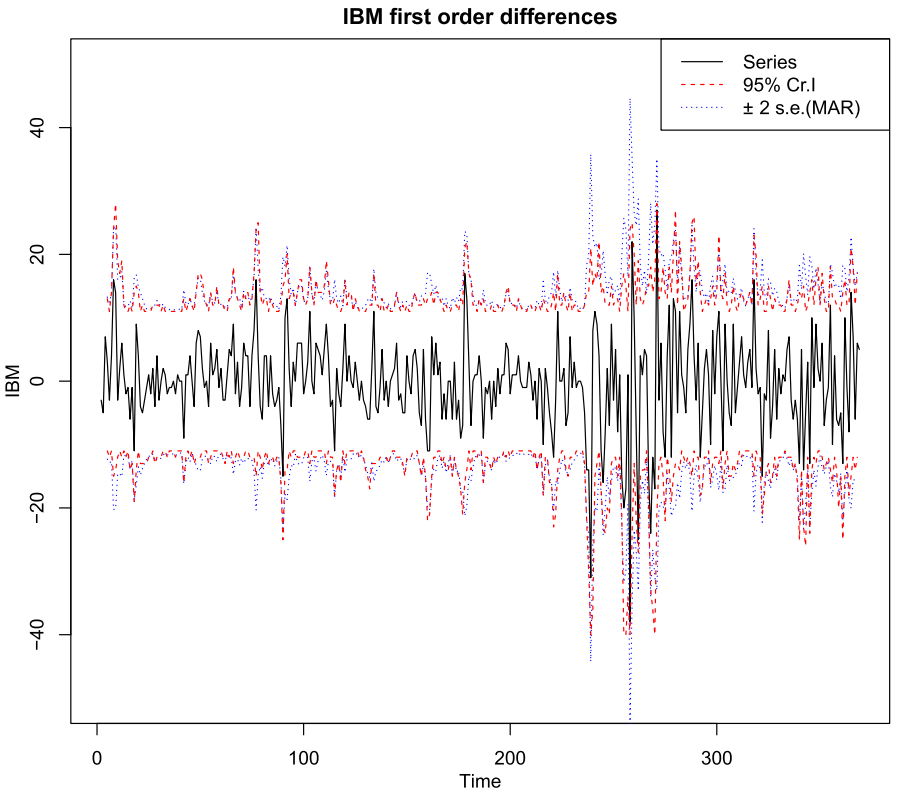

Fig. 7 IBM first order differences with 95% prediction interval from (mean) density forecast (red) and point prediction ± twice the (mean) standard error with fitted MAR(3; 4, 1, 1) model (blue)

Following previous studies, we consider the natural logarithm of the data, which presents a typical autoregressive correlation structure with 10 years cycles. We notice the presence of multimodality in the log-data, with two local maxima (see Fig. 8). This suggest that the series may be in fact generated by a mixture of two components.

In their analysis, Wong and Li (2000) choose a MAR(2; 2, 2) as best model to fit the data. However, their choice was based on the minimum BIC criterion, which the authors themselves acknowledge as not always reliable for MAR models, particularly with small datasets.

Aiming to have a better insight about the data, we apply our Bayesian method. The selected model is in this case a MAR(2; 1, 2), preferred over a MAR(2; 2, 2) by the algorithm, and to all 2, 3 and 4 component models with autoregressive order $p = 1, 2, 3, 4$. In particular, RJMCMC selects MAR(2; 1, 2) about 38% of the time, against 20% for $MAR(2; 2, 2)$. The marginal log-likelihood for this model is -131.0381, which is larger than that of other candidate models MAR(3; 1, 2, 2) with -176.4684 and MAR(4; 1, 2, 2, 1) with -154.9989.

We generated a sample of size 100,000 from the posterior distribution of the parameters of the selected MAR(2; 1, 2) model. It is noticed that, for most parameters, the

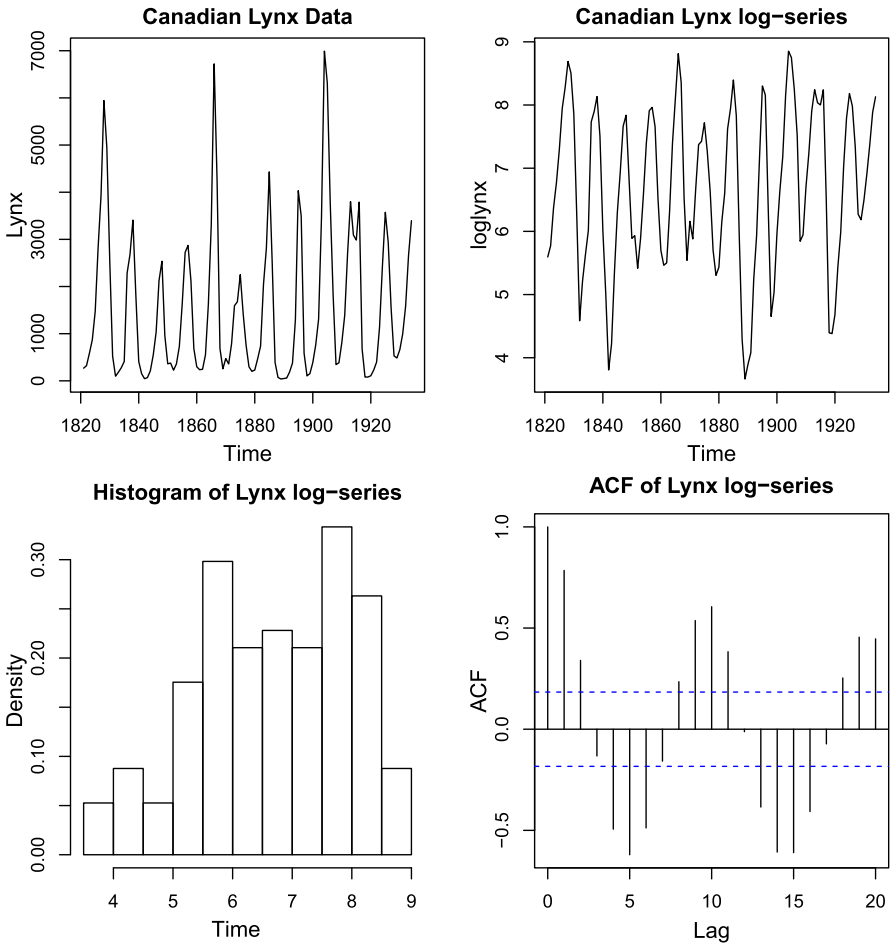

Fig. 8 Original time series of Canadian lynx (top left), series of natural logarithms (top right), histogram of log-data (bottom left) and autocorrelation plot of log-data (bottom right). The data presents a typical autoregressive correlation structure, as well as multimodality

90% credibility region includes the MLEs obtained by Wong and Li (2000). The only exception stands for the scale parameters, which seem to be slightly larger than such MLEs. However, this may be due to our model containing one fewer AR parameter. On the other hand, these results are in line with the estimates obtained by fitting a MAR(2; 1, 2) using the EM algorithm, since all estimates are well within the corresponding 90% highest posterior density region.

Figure 9 displays the raw output of the sample from the posterior distributions of the parameters obtained via MCMC simulation. Due to the two mixture components having different autoregressive order, and the aid of the trace and density plots, we conclude that label switching has not occured, so that relabelling is not required.

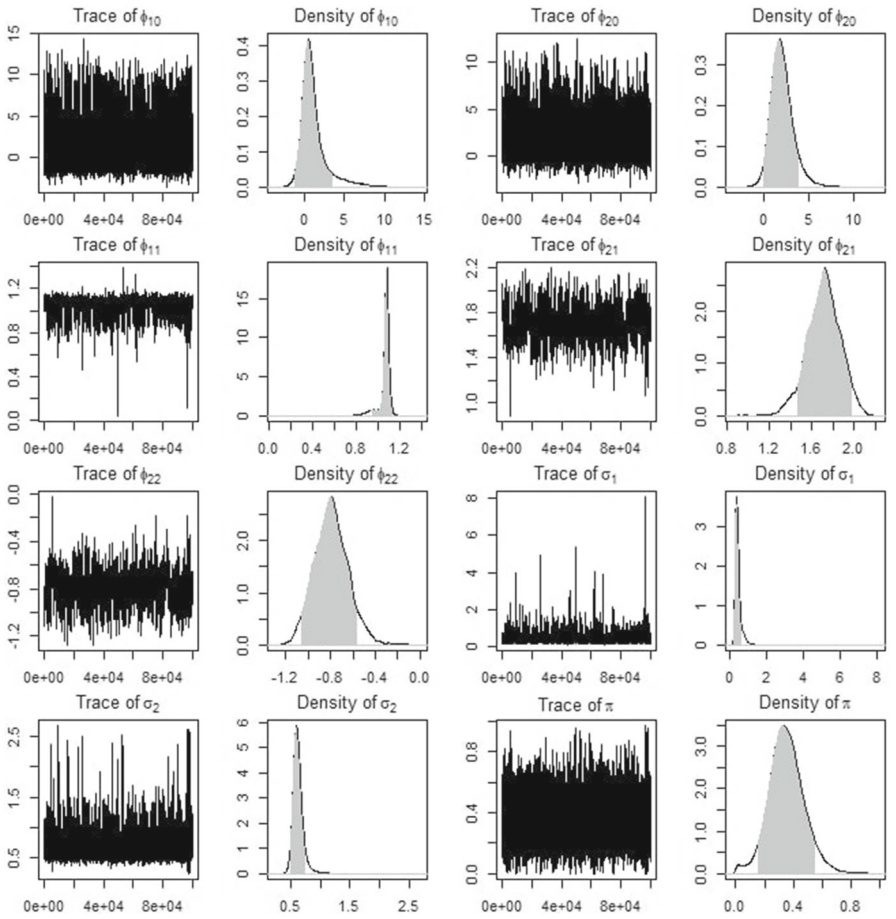

Fig. 9 Posterior trace plots and density of selected MAR(2; 1, 2) model for the natural logarithm of Canadian lynx data. For all parameters, the credibility region contains the estimated values from Wong and Li (2000). Sample size is 100,000, after 50,000 burn-in iterations

4.4 Daily temperature in Manchester city centre

This last example is a dataset of recorded air temperature in Manchester city centre between January 1st 1985 and April 1st 1986. During each day, temperature was recorded between 06:00h and 21:00h, and up to a maximum of 4 times between 22:00h and 05:00h of the following day. The data is available on the CEDA Archive (Met Office 2019).

Here we consider the time series of daily air temperature range, calculated as the difference between the maximum and minimum recorded temperature within a day. The result is a series of 456 observations, which by construction contains only positive values (or equal to 0 as a limit case). The series can be seen in Fig. 10.

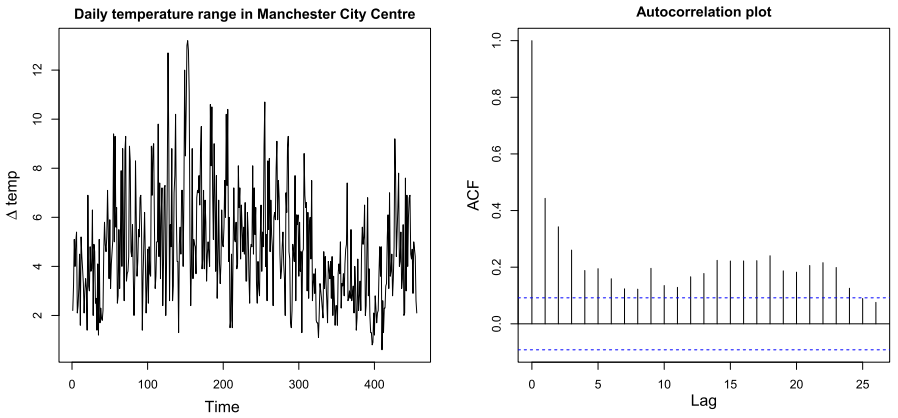

Fig. 10 Time series of daily temperature range in Manchester City Centre

Some interesting dynamics occurred while analysing this dataset, which are worth pointing out. When attempting to simulate parameters from models with $g > 2$ mixture components, the mixing weights of all but two of these components eventually converged towards 0. This suggests that the correct number of mixture components is $g = 2$, as virtually no observation is allocated to the remaining $g - 2$ components. This is in line with the theoretical properties discussed by Rousseau and Mengersen (2011) about asymptotic behavior of the posterior distribution of the mixing weights in a mixture of distributions. The authors derive analytically a results which states that, under certain choices of hyperparameters on the Dirichlet prior for the mixing weights, any redundant mixture components will see their corresponding weights converge to 0, as a sign that the component should not be included in the mixture.

Following the above considerations, we present here analysis for $g = 2$. RJMCMC selcted MAR(2; 2, 1), selected around 55% of the time amongst all 2 component models with maximum autoregressive order $p = 5$. Full conditional posterior distributions of the parameters can be seen in Fig. 11. This is the raw output from the MCMC sampling scheme, which does not show any signs of label switching.

In addition to parameter distributions, Fig. 12 also shows the original series together with three different prediction intervals. The red line is the 95% credibility interval, which is essentially the region of the highest posterior density region of the conditional predictive distribution of y_t under the assumption of MAR model. The blue line is a prediction interval for the predictor of y_t, calculated as $\hat{y}_{t|t-1} \pm 2\sqrt{\text{Var}\,(y_t \mid \mathcal{F}_{t-1})}$. Finally, the green line is a prediction interval for the predictor of y_t by fitting an AR(3) model. For the first, one predictive distribution is calculated for each sample from the posterior distribution of the parameters and for each time t; in this way we obtain a sample of the predictive density. We then calculate the "average" density as the mean of this sample, and finally we extract the highest posterior density region of this. For the remaining two, one prediction is calculated for each sample from the posterior distribution of the parameters and for all t, as well as the corresponding conditional variance. Once again, we ultimately calculate the mean of all predictors and of all conditional variances at each time t.

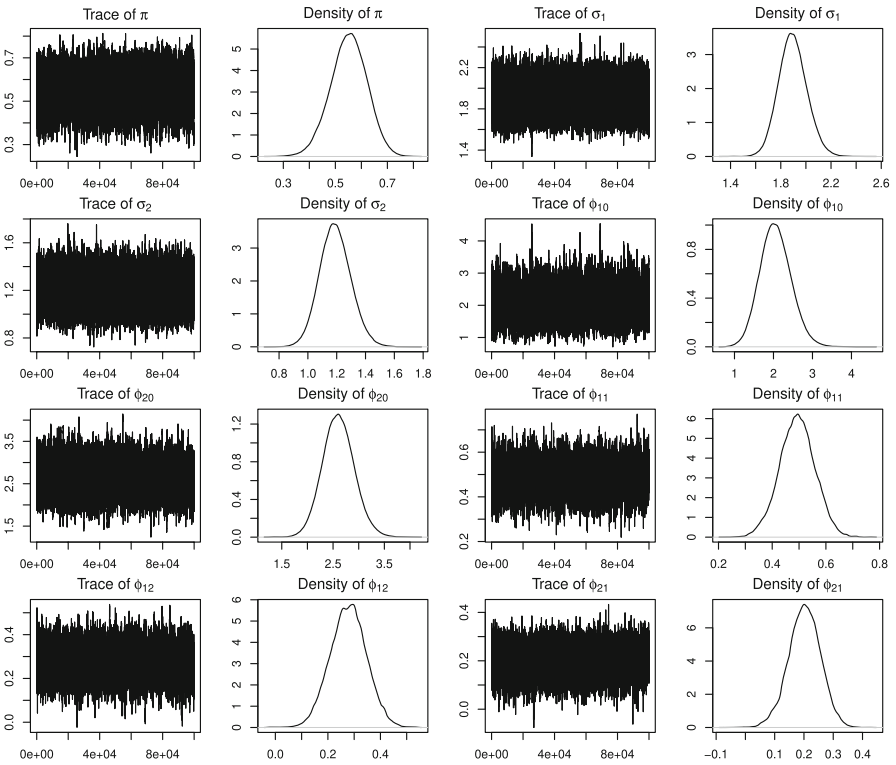

Fig. 11 Trace and density plots of parameter posterior distributions of MAR(2; 2, 1) model for daily temperature range in Manchester City Centre

Figure 12 shows one of the advantages of using density forecasts to obtain a prediction interval. A density forecast, in fact, automatically rules out values of the predictor that are not in the domain of the random variable. In this case, we stated at the beginning that, due to its definition, the temperature range is necessarily larger ≥ 0, and the credibility interval (red dashed line) indeed satisfies this condition. Furthermore, full conditional predictive distributions are available for each data point, which could provide additional information on the forecast where necessary. On the contrary, both the other two intervals considered (blue dashed line for MAR prediction interval and green dashed line for AR prediction interval) contain values that are smaller than 0, which of course violates the assumpion.

5 Bayesian density forecasts with mixture autoregressive models

Once a sample from the posterior is obtained, it is useful to use it to make predictions on future (or off-set) observations.

In the context of mixture models, density forecasts are often more attractive than point predictors and prediction intervals. This is because the qualitative features of a

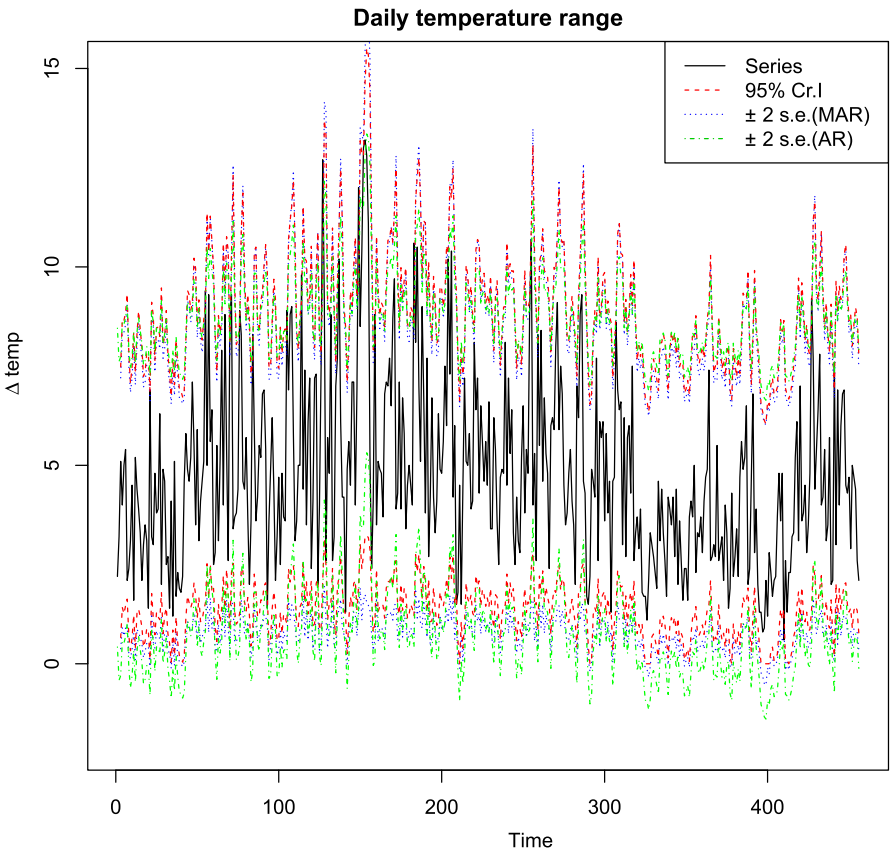

Fig. 12 Time series of daily temperature range in Manchester City Centre with 95% prediction interval from (mean) density forecast (red), point prediction ± twice the (mean) standard error with fitted MAR(2; 2, 1) model (blue) and point prediction ± twice standard error under $AR(2)$ model

predictive distribution, such as multiple modes or skewness, are more intuitive and useful than just a forecast and its associated prediction interval. Think for example of the point prediction for a symmetric bimodal density: a point prediction would fall exactly between the two modes, in a point of lower density, and would therefore be misleading. In addition, when the predictive distribution is available, prediction intervals can easily be obtained by extracting the quantiles of interest (Boshnakov 2009; Lawless and Fredette 2005).

Wong and Li (2000) and Boshnakov (2009) respectively introduced a simulation based and an analytical method for for density forecasts assuming a MAR model. The first method relies on Monte Carlo simulations, while the second derives exact h-step ahead predictive distributions of a given observation.

On one hand, we could estimate density forecasts using the highest posterior density values (i.e. the peak of the posterior distribution). However, it is better in this case to exploit the entire simulated sample as follows:

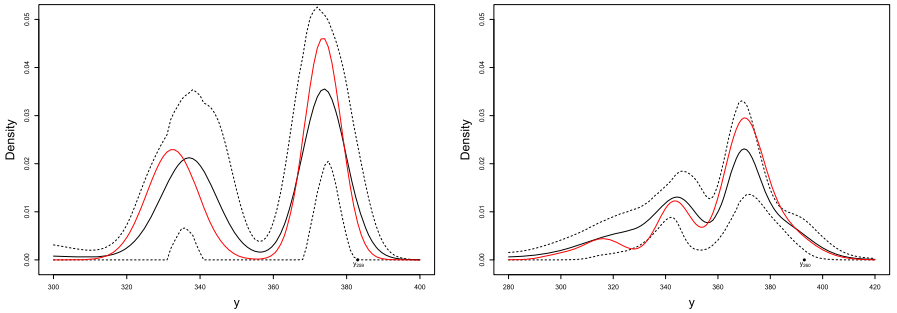

Fig. 13 Density of 1 and 2 step ahead predictor at $t = 258$ for the IBM data. The solid black line represents our Bayesian methodology, with the 90% credible interval identified by the dashed lines. The solid red line represents the predicted density using parameter values from EM estimation by Wong and Li

1. Label each simulation from 1 to N, e.g. $\boldsymbol{\theta}^{(i)}$, $i = 1, \ldots, N$.
2. Calculate density forecast $f^{(i)}\left(y_{t+h} \mid \mathcal{F}_t, \boldsymbol{\theta}^{(i)}\right)$.
3. Estimate the density forecast

$$\hat{f}\left(y_{t+h} \mid \mathcal{F}_t\right) = \frac{1}{N}\sum_{i=1}^{N} f^{(i)}\left(y_{t+h} \mid \mathcal{F}_t, \boldsymbol{\theta}^{(i)}\right)$$

In this way, we obtain a sample from the h-step ahead density forecast of an observation of interest. We then average the density at each point over its sample size, to obtain a "mean" density forecast. Furthermore, the fact that we are using the entire MCMC sample makes this method "immune" to bias due to label switching. The predictive density is a sum, and therefore commutative, making the order label permutation irrelevant towards prediction. Thus, detection of label switching has the sole purpose of providing identifiability and interpretation of the model.

We estimate the 1-step and 2-step predictive distributions of the IBM data at $t = 258$ using the analytical method by Boshnakov (2009), and compare them to the ones obtained by EM algorithm (see Fig. 13). The solid red lines represent the density obtained by Boshnakov (2009) using EM estimates and the exact method. Results of our method are represented by the solid black lines, with the dashed lines as 90% credibility region. The figure also shows how quickly the uncertainty on the predictions grows as we move further in the future, with the 2-step predictive density looking much flatter.

We can see that there are no substantial differences in the shape of these predictive distributions. However, we notice that, particularly for the 2-step predictor, averaging seems to "stabilise" the density line.

We notice from the plots that, clearly for the 1-step predictor and slighlty for the 2-step predictor, the density obtained by MCMC attaches higher density to the observations of interest y_{259} and y_{260}.

6 Conclusion

We presented an innovative fully Bayesian analysis of mixture autoregressive models with Gaussian components, in particular a new methodology for simulation from the posterior distribution of the autoregressive parameters, which covers the whole stationarity region, compared to previous approaches that constrained it in one way or another. Our approach allowed us to better capture presence of multimodality in the posterior distribution of model parameters. We also introduced a way of dealing with label switching that does not interfere with convergence to the posterior distribution of the model parameters. This consisted in using a relabelling algorithm a posteriori.

Simulations indicate that the methodology works well. We presented results for two simulated data sets. In both cases the "true" model was selected, and posterior distributions showed high densities regions around the "true" values of the parameters.

The ability of our methodology to explore the complete stationarity region of the autoregressive parameters allows it to capture better multimodality of distributions. This was illustrated with the IBM and the Canadian Lynx datasets. In the former (Fig. 6) we saw how multimodality in the posterior distribution of autoregressive parameters was captured, aspects which were missed in the analyses of Hossain (2012), see for example Figures 3.10 and 3.11. For this example, it was also noticed that modes of posterior distributions of the autoregressive parameters roughly correspond to point estimates obtained by EM estimation. In the latter (Fig. 9), we found the mode of ϕ_{21} to be quite distant from 0, with values close to 2 lying in the credibility interval. In this case, the risk with Hossain's methodology would be to truncate the Normal proposal at points such that a significant part of the stationarity region of the model is not covered. Sampietro's methodology would have failed to detect such a mode, since it is outside the interval $[-1, 1]$.

Furthermore, we analysed a dataset of daily temperature range in Manchester (UK) city centre. This example gave us further insights on an alternative way of finding the best model for the data under particular circumstances. in addition, it allowed us to show the advantages of using conditional predictive densities to extrapolate information, such as credibility intervals, about the predictor.

In conclusion, we may say that our algorithm provides accurate and informative estimation, and therefore may result in more accurate predictions.

Further work could be done to improve the efficiency of our methodogy. Possible improvements include a different algorithm for sampling of autoregressive parameters.

In particular, acceptance rates for the Random Walk Metropolis moves used for sampling the autoregressive parameters can be rather low for mixtures of large number of components or for components with large autoregressive orders, making the algorithm slow at times, with the added risk of it not being able to explore the complete parameter space efficiently. A different procedure, such as the Metropolis Adjusted Langevin Algorithm (MALA), may be considered to improve the efficiency. This would also help reducing the autocorrelation in the MCMC sample, which was found to be quite large and persistent in some cases. Notice however that all the examples displayed run long enough chains to account for this.

Gaussian mixtures are very flexible but alternatives are worth considering. In particular, components with standardised t-distribution could allow modelling heavier tails with small number of components.

Funding Not applicable.

Availability of data and material The IBM and Canadian lynx data are openly available in R, respectively in package **fma** (Hyndman 2017) and in the base package.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability The analysis uses functions from the R package **mixAR** (Boshnakov and Ravagli 2020), available on CRAN.

Appendix

We explain here how consistency of the method was assessed, with application to data generated from Model **(A)** in Sect. 4.

For this experiment, we simulate 400 different datasets of length $n = 300$ from the underlying MAR process in Model **(A)**, and proceeded as follows:

1. For each dataset, we simulate a sample of size 100,000 from the posterior distribution of the parameters, after allowing 10,000 iterations as burn-in period.
2. For each parameter, we find the overall minimum and maximum over the 400 samples, say l and u. From here, we identify a grid of 512 equally spaced values in the range $[l, u]$, and evaluate the density of such points under each posterior.
3. Finally, we average for each of the points to obtain a unique average density.

Figure 14 summarises results of applying this procedure. As we can see, the densities are well in line with the true values of the parameters.

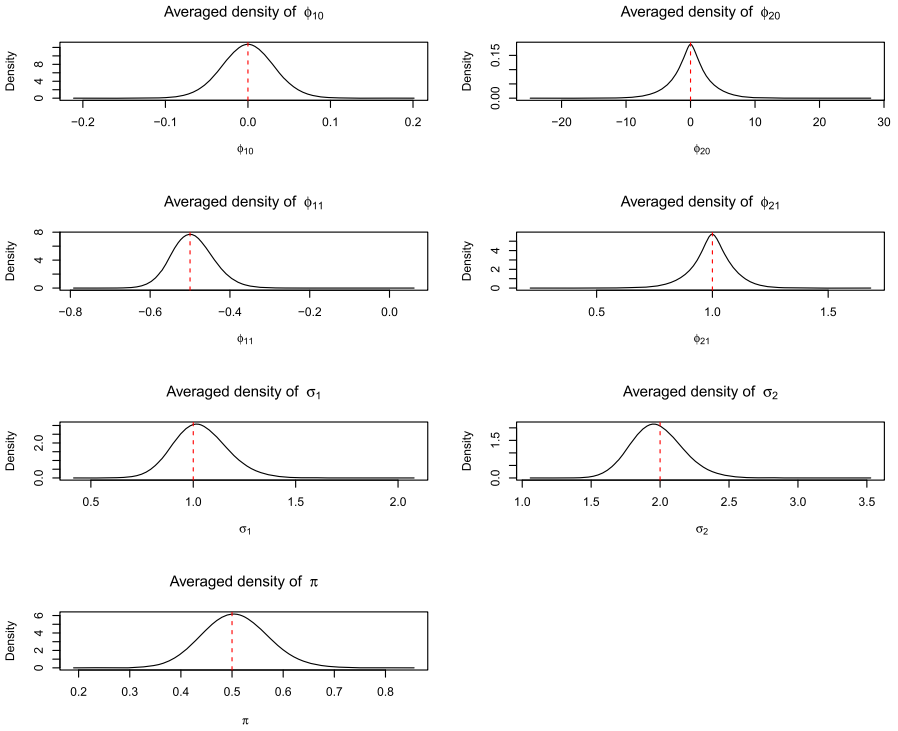

Fig. 14 Average densities of the parameters over 400 simulated datasets of length $n = 300$. Each simulation is a sample of size 100,000 from the posterior distribution of the parameters

References

Boshnakov GN (2009) Analytic expressions for predictive distributions in mixture autoregressive models. Stat Probab Lett 79(15):1704–1709

Boshnakov GN (2011) On first and second order stationarity of random coefficient models. Linear Algebra Appl 434(2):415–423

Boshnakov GN, Ravagli D (2020) mixAR: mixture autoregressive models. R package version 0.22.4. https://CRAN.R-project.org/package=mixAR

Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control/George E.P. Box and Gwilym M. Jenkins, rev. ed. edn, Holden-Day San Francisco

Celeux G (2000) Bayesian inference of mixture: the label switching problem. In: Payne R, Green P (eds) COMPSTAT. Physica, Heidelberg

Chib S (1995) Marginal likelihood from the Gibbs output. J Am Stat Assoc 90(432):1313–1321

Chib S, Jeliazkov I (2001) Marginal likelihood from the Metropolis–Hastings output. J Am Stat Assoc 96(453):270–281

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc Ser B (Methodol) 39:1–38

Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. J R Stat Soc Ser B (Methodol) 56:363–375

Elton C, Nicholson M (1942) The ten-year cycle in numbers of the lynx in Canada. J Anim Ecol 11(2):215–244

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1(2):209–230

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732

Hossain AS (2012) Complete Bayesian analysis of some mixture time series models. PhD thesis, Probability and Statistics Group, School of Mathematics, University of Manchester

Hyndman RJ (2017) fma: data sets from "Forecasting: Methods and Applications" by Makridakis, Wheelwright & Hyndman (1998). R package version 2.3. https://CRAN.R-project.org/package=fma

Jones MC (1987) Randomly choosing parameters from the stationarity and invertibility region of autoregressive-moving average models. J R Stat Soc Ser C (Appl Stat) 36(2):134–138

Lau JW, So MK (2008) Bayesian mixture of autoregressive models. Comput Stat Data Anal 53(1):38–60

Lawless JF, Fredette M (2005) Frequentist prediction intervals and predictive distributions. Biometrika 92(3):529–542

Le ND, Martin R, Raftery AE (1996) Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. J Am Stat Assoc 91(436):1504–1515

Lo AY (2005) Weighted Chinese restaurant processes. COSMOS 01(01):107–111. https://doi.org/10.1142/S0219607705000073

Met Office Centre for Environmental Data Analysis: 2019, Met office midas open: UK land surface stations data (1853-current). data retrieved from the CEDA Archive, http://catalogue.ceda.ac.uk/uuid/dbd451271eb04662beade68da43546e1

Nelson DB (1991) Conditional heteroskedasticity in asset returns: a new approach. Econom J Econom Soc 59:347–370

Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components. J R Stat Soc Ser B Stat Methodol 59(4):731–792

Rousseau J, Mengersen K (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. J R Stat Soc Ser B (Stat Methodol) 73(5):689–710. https://doi.org/10.1111/j.1467-9868.2011.00781.x

Sampietro S (2006) Bayesian analysis of mixture of autoregressive components with an application to financial market volatility. Appl Stoch Models Bus Ind 22(3):242

Tong H (1990) Non-linear time series: a dynamical system approach. Oxford University Press, Oxford

Wong CS, Li WK (2000) On a mixture autoregressive model. J R Stat Soc Ser B Stat Methodol 62(1):95–115

Wood S, Rosen O, Kohn R (2011) Bayesian mixtures of autoregressive models. J Comput Graph Stat 20(1):174–195. https://doi.org/10.1198/jcgs.2010.09174