Annals of Operations Research
https://doi.org/10.1007/s10479-025-06585-w

ORIGINAL RESEARCH

On using fuzzy clustering for detecting the number of states in Markov switching models

Edoardo Otranto[1] · **Luca Scaffidi Domianello**[2]

Received: 10 June 2024 / Accepted: 18 March 2025

Abstract
An open problem of Markov switching models is identifying the number of states, generally fixed a priori; it is impossible to apply classical tests due to the issue of the nuisance parameters present only under the alternative hypothesis. In this work, we show, by Monte Carlo simulations, that fuzzy clustering is able to reproduce the parametric state inference derived from the Hamilton filter and that the typical indices used in clustering to determine the number of groups can be used to identify the number of states in this framework. The procedure is very simple to apply, considering that it is performed independently of the data generating process and that the indicators we use are available in most statistical packages. Furthermore, the proposed approach appears to be sufficiently robust to perturbations in the data generating processes. A final application of real data completes the analysis.

Keywords Nuisance parameters · Groups identification · Monte Carlo simulations · Markov chains

1 Introduction

Markov switching (MS) models (Hamilton, 1990) have received increasing attention in time series analysis with several applications in economics (see, for example, Hamilton, 2016), finance (Gallo & Otranto, 2015; Koutmos, 2020), neuroscience (Degras et al., 2022), just to name a few fields. Their main advantage consists in the possibility to consider the existence of several states, interpreted as particular regimes (for example, expansion and contraction in the business cycle, quiet and turmoil periods in the financial markets), which are not observed, but whose dynamics can be represented by an ergodic Markov chain. Thanks to the properties of ergodic Markov chains, it is possible to make inference on the unobserved state, assigning to each state at each time a filtered or smoothed probability obtained from

Edoardo Otranto and Luca Scaffidi Domianello have contributed equally to this work.

✉ Edoardo Otranto
edoardo.otranto@uniroma1.it

Luca Scaffidi Domianello
luca.scaffidi@unict.it

1 Department of Social and Economic Sciences, Sapienza University of Rome, and CRENoS, Piazzale Aldo Moro, 5, 00185 Rome, Italy

2 Department of Economics and Business, University of Catania, Corso Italia, 55, 95122 Catania, Italy

Published online: 02 May 2025

the so-called Hamilton filter (Hamilton, 1990). An open problem is identifying the number of states, which is not feasible with classical tests for the problem of nuisance parameters present only under the alternative hypothesis. Consequently, in practical applications, the number of states is generally set a priori equal to 2 or 3 (rarely 4) without the use of statistical procedures to detect it.

In statistical terms the problem can be summarized as follows: when the model under the null hypothesis is an MS model with k states and the model under the alternative is a model with $k + 1$ states, the transition probabilities referring to the state $k + 1$ are not identified under the null hypothesis (see, for example, Hansen, 1992). As known, in this framework the classical tests, such as Likelihood Ratio (LR), do not follow the standard distributions and the true distribution is unknown.

Specific tests for MS models were developed by Hansen (1992) and Garcia (1998). Both methods are based on the Supremum LR test (Davies, 1977), following the work of Hansen (1996). Both authors attempt to derive the asymptotic distribution of the LR statistic. Hansen (1992) does not derive the exact asymptotic distribution of the statistic, but the upper limit, therefore an approximate distribution; as a consequence, the test proposed is conservative. Furthermore, the distribution is not standard but depends on the model adopted. In particular, the Supremum test requires evaluating different combinations of parameters constrained under the null hypothesis, along with nuisance parameters. Given the substantial computational burden involved, its application is generally limited to MS models whose parameterization is parsimonious. Garcia (1998) extends the Supremum test proposed by Hansen (1996) by excluding the extreme values of the transition probabilities from the range of possible values; in particular, he suggests using the interval [0.15;0.85], excluding the cases of high persistence of a state (probability to stay in the same state greater than 0.9), very frequent in real cases. In practice, given the possibility to derive the (approximated) asymptotic distributions, there are practical drawbacks that have not favoured the diffusion of these approaches. Other approaches follow this line of research, trying to derive the asymptotic distribution of LR tests, always with high computational costs because the results are not standard; however, they are aimed at testing the null hypothesis of linearity against MS with 2 states, not dealing more general cases involving a generic number of states (Cho & White, 2007; Carrasco et al., 2014; Qu & Zhuo, 2021).

Alternative approaches avoid the development of tests, trying to directly identify the number of states by penalized likelihood criteria (Psadarakis & Spagnolo, 2003, 2006) or Kullback–Leibler divergence (Smith et al., 2006). The basic idea is to use, also for MS models, likelihood-based methods, such as AIC or BIC, successfully adopted to detect the number of components of a mixture model (see, e.g., Leroux, 1992). Actually, for the MS case, the extension of this approach is not trivial; the Markov dynamics of the unknown state and the model specification affect the performance of penalized likelihood criteria. For example, Kapetanios (2001) fixes the number of states (in threshold models) and notes that BIC performs better than AIC in detecting the autoregressive order of the Data Generating Process (DGP); Psadarakis and Spagnolo (2003) fix the autoregressive order and try to detect the number k of states, noting a poor performance of BIC, which underestimates k, while AIC performs well if the number of observations T and the difference in switching parameters are large. A similar experiment was performed by Psadarakis and Spagnolo (2006) to detect both the number of states in MS models and the AR order, with even better performance by AIC, but again with high T and large differences in switching parameters. In practice, the performance of AIC and BIC are closely related to the correct specification of the model. Furthermore, consistency of model selection procedures based on penalized likelihood criteria remains an open question (Leroux, 1992; Psadarakis & Spagnolo, 2006; Fuh et al., 2024; Keribin, 2000).

However, the AIC and BIC results obtained with the true model specification can be useful benchmarks to evaluate the performance of alternative models by simulation experiments.

The problem of detecting the number of states has also been widely considered in the machine learning literature (often referring to Hidden Markov Models) and in the Bayesian framework. A recent interesting proposal, based on marginal likelihood, and a review of the literature, both in classical and Bayesian frameworks, are contained in Chen et al. (2024).

A very practical idea is to identify the number of states before the estimation step, emphasizing the fact that in MS models each observation is generated by a mixture density, with the number of components equal to the number of states of the Markov chain. This approach was developed in a nonparametric Bayesian framework by Otranto and Gallo (2002), where the posterior distribution of the number of states is derived using the Gibbs sampler, a well-known Monte Carlo Markov Chain (MCMC) algorithm. This intuition is the basis of our proposal.

First, we note a similarity between the inference on the regime in MS models and the grouping derived from fuzzy clustering (D'Urso, 2015). This last approach can provide a clustering of statistical units in k groups, with a probability of belonging to each of the k clusters. The intuition is that there is an "assonance" between the probability of a certain state at time t, derived from the estimation of MS models, and the membership grade of the fuzzy approach. Both approaches can assign a probability to each observation of the analyzed time series and the inference on the state (i.e. the assignment of an observation to the state) is more effective the greater this probability is in both approaches. The relationship between mixture models and fuzzy clustering has been underlined in several works. For example, Davenport et al. (1988) compare the constrained Maximum Likelihood (ML) estimator of the parameters of a Normal mixture density with another based on fuzzy c-means clustering; an interesting result is that the initialization of the algorithm to maximize the likelihood with fuzzy estimates improves the computation time and the accuracy of the estimates. Hathaway (1986) explores the relationship between some clustering methods and the EM algorithm (also used to derive ML estimates of MS models), by decomposing the likelihood function into a part depending on the hard k-means objective function and a penalty term for soft partitions; this idea was extended to the fuzzy case by Ichihashi et al. (2001). Recently, Serafini et al. (2023) have compared model-based methods, derived from mixtures of Gaussian and t densities, and fuzzy methods for soft clustering.

Our first analysis consists of assessing whether the fuzzy approach provides a grouping similar to the one derived from the inference on regimes of the MS models, using the same assignment criterion (the unit is assigned to the group corresponding to the mode). This exercise is performed using Monte Carlo experiments. After checking the similarity of the results in terms of state inference, we check whether the typical indices used to select the number of clusters are able to identify the true number of states. We use the fuzzy method as a separate step from the estimation procedure because the aim is to identify the number of states before the estimation step; in practice, the fuzzy procedure is part of the model identification procedure, as in Otranto and Gallo (2002), but unlike this, it is very fast and uses tools implemented in the main statistical routines. The practical suggestion is that a nonparametric fuzzy-based procedure could be a useful tool for specifying a nonlinear model, solving an as-yet unsolved problem in a parametric context. This method could be applied to any model that requires fixing the number of states or, more generally, the number of components of a mixture distribution that generates these data. But while in the case of mixtures there are several contributions to determine the number of components,[1] in the MS case this is an open problem.

The structure of the paper is as follows: in the next section the main features of the MS model (Sect. 2.1) and of fuzzy clustering (Sect. 2.2) are briefly recalled; in Sect. 3 we present a large set of Monte Carlo experiments to evaluate the ability of fuzzy clustering to reproduce the inference on the states of MS models (Sect. 3.1) and the performance of several indices in detecting the number of states (Sect. 3.2); the comparison of the proposed method with AIC and BIC, computed under the MS model specification, is explored in the Sect. 3.3, distinguishing *good* cases from *puzzling* cases; the section concludes with simulation experiments related to alternative distributions (Normal, Generalized Error Distribution-GED,[2] presence of asymmetric effects), checking the performance of the proposed fuzzy approach, compared to the classical AIC and BIC, calculated on the marginal distribution without considering the structure of the DGP, similarly to the proposed fuzzy approach (Sect. 3.4). Section 4 describes in detail the proposed procedure, which is applied to a real dataset, referring to the Gross Domestic Product (GDP) series of the U.S. Some final remarks will conclude the paper.

2 MS models and fuzzy clustering

2.1 MS models

Let us consider a time series y_t with $t = 1, \ldots, T$. We say that its DGP follows an MS process with k states if:

$$y_t = f(\boldsymbol{x}; \boldsymbol{\theta}_{s_t}) + \varepsilon_t,$$

where $\boldsymbol{x}$ is a vector of (weakly) exogenous variables, possibly including lagged values of y_t, ε_t $(t = 1, \ldots, T)$ are independent disturbances with zero mean and constant variance. The vector of unknown parameters $\boldsymbol{\theta}_{s_t}$ depends on an unobservable discrete random variable s_t,[3] which can assume values $1, 2, \ldots k$, and whose dynamics is driven by an ergodic Markov chain, with elements of the transition probability matrix $\boldsymbol{P}$:

$$p_{ij} = Pr(s_t = j | s_{t-1} = i), \qquad i, j = 1, \ldots, k,$$

with $\sum_{j=1}^{k} p_{ij} = 1$ for each $i = 1, \ldots, k$. A typical model used in an MS framework is the AutoRegressive (AR) model of order p, called $MS - AR(p)$:

$$y_t = \mu_{s_t} + \sum_{i=1}^{p} \phi_i \left(y_{t-i} - \mu_{s_{t-i}} \right) + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2). \tag{1}$$

Under stationarity constraints, specification (1) involves a time-varying unconditional mean of the process, μ_{s_t}, depending on the state at time t. The transition probabilities p_{ij} are estimated with the other parameters $\boldsymbol{\theta}_{s_t} = (\mu_1, \ldots, \mu_k)'$, $\boldsymbol{\phi} = \left(\phi_1, \ldots, \phi_p\right)'$ and σ^2 by

1 See, for example, Titterington (1997) and the review for the case of the Normal mixture in McLachlan and Rathnayake (2014).

2 GED is a unimodal symmetric distribution belonging to the exponential family. The parameterization of the density function we use was provided by Nelson (1991).

3 The variance could also depend on s_t; we will not consider this case in our experiments, but the generalization is obvious.

Maximum Likelihood Estimator (MLE).[4] Model (1) includes the latent variable s_t and a possible solution to perform the MLE is the use of the EM algorithm. Hamilton (1990) provides such a procedure for MS models and Diebold et al. (1994) detail the steps of the EM procedure for a more general MS model (with time-varying probabilities). However, there is the possibility of deriving the likelihood function using the so-called Hamilton filter (Hamilton, 1990), a sort of Kalman filter with a discrete random variable in the state equation. The same Hamilton filter provides the possibility to specify, at each time t, the conditional probability to fall in a certain state j; given the information set $\mathcal{I}_\tau = (y_\tau, y_{\tau-1}, \dots)$, it is possible to obtain the so-called filtered probabilities $Pr\,(s_t = j|\mathcal{I}_t)$, the predicted probabilities $Pr\,(s_t = j|\mathcal{I}_{t-1})$ and the smoothed probabilities $Pr\,(s_t = j|\mathcal{I}_T)$ (the steps to obtain them are detailed in chapter 22 of Hamilton, 1994). More precisely, the filtered and predictive probabilities are derived from the Hamilton (1990) filter, and the smoothed probabilities from the Kim (1994) smoother. Both are iterative procedures to derive the probability of the unobservable state at any time. The difference is in the information set used to derive them: filtering is a procedure to derive the probability conditional on the information up to the previous (predicted probabilities) and current (filtered probabilities) time, smoothing on the complete information set. In filtering the recursive estimate of the state moves forward through the data, while in smoothing the recursive estimate of the state moves backward through the data.

The predicted probabilities enter the likelihood function; in fact, the conditional density of each y_t is expressed as:

$$f(y_t|\mathcal{I}_{t-1}; \boldsymbol{\theta}_{s_t}, \boldsymbol{\phi}, \sigma^2) = \sum_{j=1}^{k} f(y_t|s_t = j, \mathcal{I}_{t-1}; \boldsymbol{\theta}_{s_t}, \boldsymbol{\phi}, \sigma^2) Pr\,(s_t = j|\mathcal{I}_{t-1})\,, \tag{2}$$

which is a mixture of distributions with weights represented by the predicted probabilities. The typical problem of multiple local maxima in the likelihood function based on mixture distributions can be addressed by adopting several starting values for the maximization algorithm and selecting the final estimates that provide the highest likelihood function; this way of proceeding provides a consistent, asymptotically Normal and efficient estimator (Kiefer, 1978).

Moreover, this procedure provides also an intuitive inference on the regime, assigning to state j the observation at time t with the mode in j, using as a mass distribution the filtered or, more frequently, the smoothed probabilities, which use the full in-sample information.

As said in Sect. 1, identifying the number of states k is not achievable through classical statistical tests for the problem of nuisance parameters present only under the alternative hypothesis. As an example, let us consider the null hypothesis of the linear model (no states or $k = 1$):

$$y_t = \mu + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2),$$

[4] When the hypothesis of normality is not longer valid, for example in the case of financial time series, we can interpret the estimation in Quasi MLE terms, ensuring consistency and asymptotic normality of the estimator (Xie, 2009) Gibbs sampling is another alternative, mainly used in Bayesian frameworks (Albert and Chib, 1993, McCulloch and Tsay, 1994, introduced for MS models by), particularly when the distribution is unknown (Hwu & Kim, 2023). Both MLE and Gibbs sampling methods for MS models are described in Kim and Nelson (1999).

against the alternative of an MS model with $k = 2$ states:

$$y_t = \mu_{s_t} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2), \qquad s_t = 1, 2$$
$$\boldsymbol{P} = \begin{bmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{bmatrix} \tag{3}$$

The linear model can not be obtained from the MS model simply imposing that the switching parameters are equal in the two states ($\mu_1 = \mu_2$) because the transition probabilities p_{11} and p_{22} are not identified under the null hypothesis. Similarly, if the 2-state model (3) is assumed under the null hypothesis, whereas the alternative refers to an MS model with $k = 3$:

$$y_t = \mu_{s_t} + \varepsilon_t, \qquad \varepsilon \sim N(0, \sigma^2), \qquad s_t = 1, 2, 3$$
$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}; \qquad \sum_{j=1}^{3} p_{ij} = 1, \quad i = 1, 2, 3 \tag{4}$$

In this case, 4 probabilities in $\boldsymbol{P}$ are not identified under the null hypothesis.

2.2 Fuzzy clustering

There is some similarity between the MS models and the fuzzy clustering methods. The latter detects the belonging of each statistical unit to a cluster with a certain probability (the *membership grade*), unlike the classical hard clustering, where each unit can belong to exactly one cluster only. Clustering based on the degree of membership in each group can be seen as similar to state inference done using the smoothed probabilities in an MS framework. For example, considering the most popular fuzzy clustering algorithm, the *fuzzy k-means* (Bezdek, 1981), the fuzzy partition in k groups of the observed y_t's ($t = 1, \ldots, T$) is obtained by minimizing:

$$\min_{\boldsymbol{U},\boldsymbol{c}} \sum_{t=1}^{T} \sum_{j=1}^{k} u_{tj}^m d^2\left(y_t, c_j\right), \qquad \sum_{j=1}^{k} u_{tj} = 1; \; u_{tj} \geq 0, \; \forall t, \tag{5}$$

where $\boldsymbol{U} = \left\{u_{t,j}\right\}$ ($t = 1, \ldots, T$; $j = 1, \ldots, k$) is the membership grade matrix, $\boldsymbol{c} = (c_1, \ldots, c_k)'$ is the vector of centroids, $m > 1$ is the fuzziness parameter, which tunes the degree of fuzziness, $d(\cdot, \cdot)$ is a distance measure (we adopt the Euclidean distance). The solution to the above minimization problem can be achieved by iterating the following steps (Bezdek, 1981):

1. Given a threshold $\epsilon > 0$ and fixing $m > 1$, choose an initial membership grade matrix, $\boldsymbol{U}^{(0)}$, that satisfies the constraint illustrated in (5). Let r be the symbol of the iteration ($r = 0, 1, ...,$);
2. update the centroids through the following formula:

$$c_j^{(r+1)} = \frac{\sum_{t=1}^{T} u_{tj}^m y_t}{\sum_{t=1}^{T} u_{tj}^m} \qquad \forall j;$$

3. update the membership grade matrix:

$$u_{tj}^{(r+1)} = \frac{1}{\sum_{i=1}^{k}\left[\frac{d^2\left(y_t, c_j\right)}{d^2\left(y_t, c_i\right)}\right]^{\frac{1}{m-1}}} \qquad \forall t, j;$$

4. Compare $\boldsymbol{U}^{(r)}$ with $\boldsymbol{U}^{(r+1)}$ with a convenient matrix norm: if $\|\boldsymbol{U}^{(r+1)} - \boldsymbol{U}^{(r)}\| \leq \epsilon$ stop, otherwise set $r = r + 1$ and return to step 2.

There is no theoretical criterion for selecting m; it should be chosen in advance. We refer to D'Urso (2015, section 24.2.2), for an exhaustive review of the works dealing with the selection of the fuzziness parameter; we follow his suggestion, also agreed with Bezdek (1981), to set $m = 2$, considering that most authors indicate this value as a parameter that seems to work in several application frameworks.

Each column of $\boldsymbol{U}$ can be interpreted similarly to the smoothed probabilities of an MS model. Based on this insight, we ask whether methods used to detect the number of groups in clustering can be used as a method to identify the number of states in an MS model. The detection of the number of clusters is generally carried out by adopting indicators implemented in the main statistical packages, making this approach very simple and easily usable even by non-experts.

We try to support the intuition about the relationship between fuzzy clustering and the number of states in MS models with two sets of Monte Carlo experiments: first, fixing the *true* k, we compare the smoothed probabilities derived from the estimated MS model with the corresponding grade of membership matrix derived from (5),[5] then we use the main indices to detect the number of clusters checking if they are able to identify the correct number of states of the MS DGP. More specifically, we rely on the following cluster validation indices: Partition Coefficient (PC), Partition Entropy (PE), Modified Partition Coefficient, (MPC), Average Silhouette Width (ASW), Average Silhouette Width Fuzzy (ASWF), and Xie-Beni (XB).

The Partition Coefficient (Bezdek, 1981) is given by:

$$PC = \sum_{t=1}^{T}\sum_{j=1}^{k} u_{tj}^2 / T.$$

It can assume values between $1/k$ and 1, with its maximum value, as a function of k, giving us the optimal number of clusters.

The Partition Entropy (Bezdek, 1981):

$$PE = -\sum_{t=1}^{T}\sum_{j=1}^{k} u_{tj} \ln(u_{tj}) / T,$$

ranges from 0 to $\ln k$ and its lowest value provides the best number of clusters.

The Modified Partition Coefficient (Dave, 1996)

$$MPC = 1 - \frac{k}{k-1}(1 - PC),$$

normalizes the PC index, so that it ranges from 0 to 1, thus eliminating range dependence on k, unlike PC.

[5] Classifying the European Central Bank announcements, Gallo et al. (2021) find very similar results between the classification derived from smoothed probabilities of their MS model and the k-means clustering procedure.

The Average Silhouette Width (Rousseeuw, 1987) is calculated as follows:

$$ASW = (1/T)\sum_{t=1}^{T}\frac{b_t - a_t}{max\{a_t, b_t\}},$$

with a_t the average distance between y_t (belonging to the cluster τ) and the other observations belonging to the same group τ, while b_t is the minimum average distance among y_t and the observations belonging to another cluster $j \neq \tau$. It can assume values between -1 and 1, with the maximum value providing the best number of clusters.

The Average Silhouette Width Fuzzy (Campello & Hruschka, 2006),

$$ASWF = \frac{\sum_{t=1}^{T}(u_{tj_1} - u_{tj_2})^{\lambda} ASW_t}{\sum_{t=1}^{T}(u_{tj_1} - u_{tj_2})^{\lambda}},$$

is a weighted average of the Silhouette, where u_{tj_1} and u_{tj_2} are the elements of the t-th row of $\mathbf{U}$ with first and second largest values, respectively, ASW_t the Silhouette of the $t - th$ observation, and $\lambda \geq 0$. Notice that, as opposed to ASW, it takes into account the membership grade matrix.

Finally, the Xie-Beni index (Xie & Beni, 1991),

$$XB = \frac{\sum_{t=1}^{T}\sum_{j=1}^{k} u_{tj}^2 d^2(y_t, c_j)}{T \min_{i,j} d^2(c_i, c_j)},$$

is minimized to obtain the best partition with k clusters. In this case, the centroids are those detected by minimizing (5).

We compare several partitions with $k = 2, \ldots, 7$, selecting k based on the validation index used.

3 Monte Carlo evidence

The DGPs used for the Monte Carlo experiments are the models in Eq. (3) (call it MS(2)), Eq. (4) (MS(3)) and MS-AR(1) models like (1), with $p = 1$, and with 2 (MS(2)-AR(1)) or 3 (MS(3)-AR(1)) states, adopting the Normal distribution for ε_t. We cover several scenarios, combining a set of parameters that yields the 32 models (8 of each model type) shown in the upper part of Table 1 (labelled to identify them). Moreover, we will also consider four DGPs with $k = 4$ states (lower part of Table 1).

As the distance between the μ_i coefficients increases, the existence of states becomes clearer. This can be better appreciated by looking at Fig. 1, where we show the mixture of Normal distributions obtained using, as mixture weights, the ergodic probability of each state.[6] The parameters of the Normal distributions are given by the unconditional means of y_t, expressed by μ_{s_t} in the MS(2) and MS(3) DGPs, and $(\mu_{s_t} - \phi\mu_{s_{t-1}})/(1-\phi)$ in the case of MS(2)-AR(1) and MS(3)-AR(1).[7] The number of mixtures is indistinguishable when the μ_i

[6] As shown in Eq. (2), each density has different mixture weights, given by the predicted probabilities $Pr(s_t|I_{t-1})$; the vector of ergodic probabilities is the (normalized) eigenvector associated to the unit eigenvalue of $\boldsymbol{P}'$, which can be interpreted as the vector of unconditional probabilities of the state s_t (for details, see, Hamilton, 1994).

[7] This implies that, in the case of an AR(1) model, the number of states can be seen as 2^k (for details, see, Hamilton, 1994).

Table 1 Data Generating Processes (DGPs) used in Monte Carlo experiments

DGP: MS(2) Label	DGP: MS(2)-AR(1) Label	Parameters
MS2-1	MS2AR-1	$\mu_1 = 0; \mu_2 = 1; \sigma = 0.5$
MS2-2	MS2AR-2	$\mu_1 = 0; \mu_2 = 2; \sigma = 0.5$
MS2-3	MS2AR-3	$\mu_1 = 0; \mu_2 = 3; \sigma = 0.5$
MS2-4	MS2AR-4	$\mu_1 = 0; \mu_2 = 4; \sigma = 0.5$
MS2-5	MS2AR-5	$\mu_1 = 0; \mu_2 = 1; \sigma = 0.25$
MS2-6	MS2AR-6	$\mu_1 = 0; \mu_2 = 2; \sigma = 0.25$
MS2-7	MS2AR-7	$\mu_1 = 0; \mu_2 = 3; \sigma = 0.25$
MS2-8	MS2AR-8	$\mu_1 = 0; \mu_2 = 4; \sigma = 0.25$
DGP: MS(3) Label	**DGP: MS(3)-AR(1) Label**	**Parameters**
MS3-1	MS3AR-1	$\mu_1 = 0; \mu_2 = 1; \mu_3 = 2; \sigma = 0.5$
MS3-2	MS3AR-2	$\mu_1 = 0; \mu_2 = 2; \mu_3 = 4; \sigma = 0.5$
MS3-3	MS3AR-3	$\mu_1 = 0; \mu_2 = 3; \mu_3 = 6; \sigma = 0.5$
MS3-4	MS3AR-4	$\mu_1 = 0; \mu_2 = 4; \mu_3 = 8; \sigma = 0.5$
MS3-5	MS3AR-5	$\mu_1 = 0; \mu_2 = 1; \mu_3 = 2; \sigma = 0.25$
MS3-6	MS3AR-6	$\mu_1 = 0; \mu_2 = 2; \mu_3 = 4; \sigma = 0.25$
MS3-7	MS3AR-7	$\mu_1 = 0; \mu_2 = 3; \mu_3 = 6; \sigma = 0.25$
MS3-8	MS3AR-8	$\mu_1 = 0; \mu_2 = 4; \mu_3 = 8; \sigma = 0.25$
DGP: MS(4) Label	**DGP: MS(4)-AR(1) Label**	**Parameters**
MS4-3	MS4AR-3	$\mu_1 = 0; \mu_2 = 3; \mu_3 = 6; \mu_4 = 9; \sigma = 0.5$
MS4-7	MS4AR-7	$\mu_1 = 0; \mu_2 = 3; \mu_3 = 6; \mu_4 = 9; \sigma = 0.25$

The AR(1) coefficient, when present, is equal to 0.7 in all DGPs. The transition probabilities in the MS(2) and MS(2)-AR(1) DGPs are $p_{11} = 0.90$, $p_{12} = 0.10$, $p_{21} = 0.20$, $p_{22} = 0.80$; in the MS(3) and MS(3)-AR(1) DGPs are $p_{11} = 0.90$, $p_{12} = 0.07$, $p_{13} = 0.03$, $p_{21} = 0.15$, $p_{22} = 0.80$, $p_{23} = 0.05$, $p_{31} = 0.10$, $p_{32} = 0.2$, $p_{33} = 0.70$; in the MS(4) and MS(4)-AR(1) DGPs are $p_{11} = 0.85$, $p_{12} = 0.08$, $p_{13} = 0.05$, $p_{14} = 0.02$, $p_{21} = 0.10$, $p_{22} = 0.80$, $p_{23} = 0.06$, $p_{24} = 0.04$, $p_{31} = 0.05$, $p_{32} = 0.15$, $p_{33} = 0.70$, $p_{34} = 0.10$, $p_{41} = 0.03$, $p_{42} = 0.07$, $p_{43} = 0.25$, $p_{44} = 0.65$

coefficients are close and the variance is larger (as in MS2-1, MS2AR-1, MS3-1, MS3AR-1); the presence of states is more evident by increasing the distance between μ_i parameters and decreasing the variance. The presence of the AR parameters increases the number of components of the mixture, but the number of highest peaks is equal to the number of states.

For each DGP we generate 1000 time series of length $T = 100, 500, 1000$. For each series we perform two types of analyses: first, considering the number of states known, we check whether fuzzy clustering provides a similar inference on the states obtained from the estimated MS model (Sect. 3.1); then we check through the validation indices listed at the end of Sect. 2 if the clustering algorithm is able to detect the right number of states (Sect. 3.2). The rest of the section is devoted to the comparison of the fuzzy approach with AIC and BIC, assuming to know the correct family of models (Sect. 3.3) and not knowing it (Sect. 3.4).

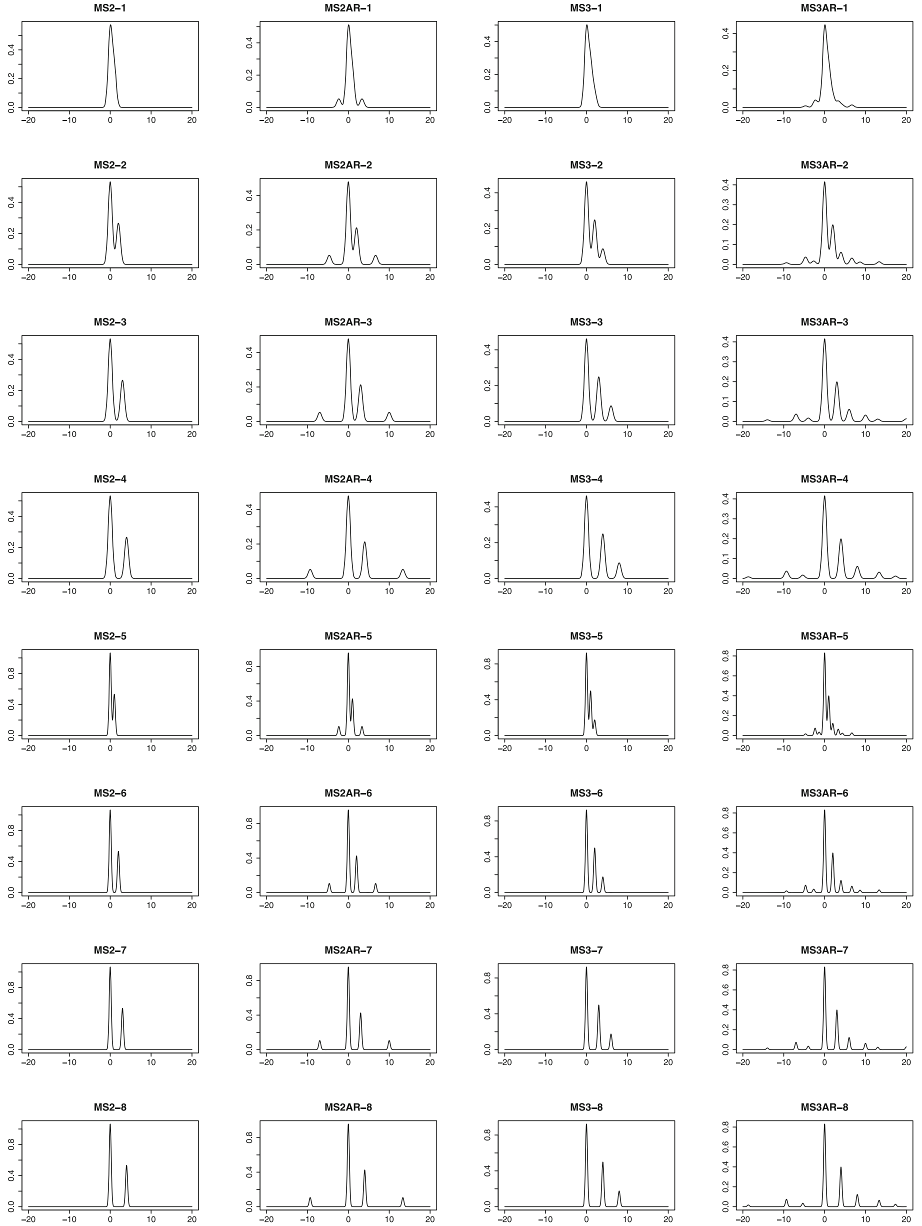

Fig. 1 Normal Mixture density functions relative to 32 DGPs described in Table 1, with weights equal to the corresponding ergodic probabilities

3.1 Inference on the state

As said, the inference on the state of MS models is performed by assigning the observation at time t to the state with the highest smoothed probability. In our first experiment we generate data from the 32 DGPs with 2 and 3 states shown in Table 1, estimate the corresponding MS model, derive the smoothed probabilities, assign each observation to a state, and finally compare the clustering to the true one using the Adjusted Rand index (ARI, Hubert and Arabie, 1985):

$$ARI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{T_{ij}}{2} - \sum_{i=1}^{k}\binom{T_{i\cdot}}{2}\sum_{j=1}^{k}\binom{T_{\cdot j}}{2}/\binom{T}{2}}{1/2\left[\sum_{i=1}^{k}\binom{T_{i\cdot}}{2} + \sum_{j=1}^{k}\binom{T_{\cdot j}}{2}\right] - \sum_{i=1}^{k}\binom{T_{i\cdot}}{2}\sum_{j=1}^{k}\binom{T_{\cdot j}}{2}/\binom{T}{2}}, \tag{6}$$

where T_{ij} is the number of observations belonging to group i in the true clustering and assigned to group j in the MS clustering; $T_{i\cdot}$ and $T_{\cdot j}$ represent the number of observations belonging to group i in the true clustering and to group j in the MS clustering, respectively. The lower bound of ARI is negative; this occurs quite rarely and suggests that the two clusters being compared have less in common than if units were randomly assigned to the clusters, which is the case for an ARI of 0; value 1 corresponds to the case of perfect match between true and MS clustering. We calculated ARI for each simulated time series of each DGP and the results are shown in Table 2 for $T = 100$, in Table 3 for $T = 500$, in Table 4 for $T = 1000$. In each block of columns of these tables, we show the five-number summary (minimum, first quartile, median, third quartile, maximum) of the empirical distribution of ARI for each DGP.

It is interesting to note that, when $T = 100$, even though the estimated model is correctly specified when the means are close to each other-see also Fig. 1-it is not trivial to obtain the correct classification of the observations with the parametric approach (see first block of columns in Table 2, models MS2-1, MS3-1), and the problem is more evident when there is an autoregressive term (MS2AR-1, MS2AR-2, MS3AR-1, MS3AR-2, MS3-AR5). However, as expected in a parametric approach, the correct identification of the number of states increases as the length of the T series increases (first block of columns in Tables 3 and 4). The ARI increases when the variance of each state is smaller (models labelled with numbers from 4 to 8). However, excluding the MS2-1, MS2AR-1, MS3-1 and the MS3AR-1 cases, in each DGP more than 50% of the simulations provide an ARI index greater than 0.9. The 100% correct classifications are achieved when the distance among the means of each state is larger and the variance is smaller.

Considering the fuzzy classification, obtained using the squared Euclidean distance in (5), and comparing this classification with the true one (second block of columns in Table 2), we note a certain difference concerning the previous comparison only in the minimum value of the ARI, due to the failure of a small number of cases. However, it increases as the length of the T series increases, with less variability in the distribution of the ARI; this is not a trivial result for the fuzzy approach, which ignores the estimation step. It is also able to achieve the 100% correct classification, already for $T = 100$, for models MS2-7, MS2-8, MS2AR-7, MS2AR-8, and for other 12 DGPs from the second quartile. Cases MS2-4, MS2-6, MS3-4, MS3-6, MS3-7, MS3AR-7, MS3AR-8 reach a minimum ARI of 0.99 when $T = 500$ and $T = 1000$.

In practice, the classification obtained in a nonparametric way and not considering the model that generates the data shows very similar results to the case using the right model (not known in practical cases), which requires the estimation step. This is confirmed by observing the third block of columns of Tables 2, 3 and 4, where the MS and the fuzzy approach are compared (in Eq. (6) the index i now refers to the MS classification and j to the fuzzy classification); classification differences are reduced when the distance between means is larger and the variability is smaller and when T is greater.

As said in Sect. 1, in general, the number of states k considered in MS models is 2 or 3, especially due to the increasing estimation problems when k is large (flat zones in the log-likelihood function, transition probabilities near to the boundary, singular Hessian). As a consequence, a parametric approach could be affected by these drawbacks, while, in principle, the fuzzy approach works by being preliminary to the estimation step. To compare the performance of the two approaches in this case, we repeated the simulation experiments for a limited number of DGPs, again with $T = 100, 500, 1000$. Table 5 summarizes the results. Similar comments concerning the $k = 2$ and $k = 3$ cases apply: the performance of the parametric and non-parametric approaches is similar after the first quartile; the fuzzy procedure shows greater variability of the results for $T = 100$, which are reduced for T equal to 500 and 1000.

3.2 Identification of the number of states

Having empirical evidence of consistent classifications between the typical parametric MS procedure and the fuzzy nonparametric clustering, the next step is to check if it is possible to use the usual criteria for detecting the number of clusters to identify the number of states in MS models. We use the same series generated for the experiment illustrated in Sect. 3.1 to apply the criteria for detecting the number of clusters based on the indices PC, PE, MPC, ASW, ASWF, and XB. In Tables 6, 7, 8 we show the percentage of correct identification of the number of states for each index for series of length T equal to 100, 500, 1000 respectively.

Considering DGPs with 2 states and $T = 100$ (Table 6), in general, all indices are able to detect the correct number of states with a high percentage of success, excluding the cases MS2-1 and MS2AR-1, where the difference between the two means is smaller and the variance is larger. PE and PC show a good performance also in these last two cases, with a success rate close to 100%. The other indices are often able to detect the correct number of states, even with a percentage equal to or very close to 100% when the DGP does not contain AR dynamics.

Cases with 3 states present greater uncertainty in detecting the correct number of groups, especially for small differences in the means (MS3-1, MS3-2, MS3AR-1, MS3AR-2), but in this case, the lower variance seems to promote better performance. The presence of the autoregressive term in DGP seems to worsen the success rate.

In general, the PE index shows the best performance for the 2-state case, presenting the highest percentage of correct detection in the 16 corresponding DGPs (always equal or close to 100%). In the 3-state case there is no clear preferred index: PC and PE seem less performing than the other four indices. By summing the success rate in each column of Table 6, the ASW index shows the highest score with ASWF being second best.

The results are sensitive to the length of the series: the cases of $T = 500$ (Table 7) and $T = 1000$ (Table 8) show a clear improvement in the rate of success; in particular ASW and ASWF show excellent performance in almost all cases.

The experiment has been replicated for the 4 DGPs with 4 states and the results are shown in Table 9. Once again we notice a very good performance of the indices, increasing with T; the only exception is for PC and PE in the MS4AR-3 case.

Table 2 Summary Statistics of the Adjusted Rand Index for MS Inference on the state and Fuzzy k-means clustering in 1000 Monte Carlo experiments and in relation to 16 Data Generating Processes (DGP) with 2 states and 16 DGPs with 3 states. $T = 100$

DGP	Adjusted Rand Index														
	MS-True State					Fuzzy-True State					Fuzzy-MS				
	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.
MS2-1	0.00	0.54	0.65	0.75	1.00	0.00	0.29	0.38	0.48	0.74	0.00	0.31	0.46	0.60	1.00
MS2-2	0.77	0.92	0.96	1.00	1.00	0.02	0.84	0.91	0.95	1.00	0.01	0.86	0.92	0.96	1.00
MS2-3	0.92	1.00	1.00	1.00	1.00	0.04	1.00	1.00	1.00	1.00	0.04	1.00	1.00	1.00	1.00
MS2-4	0.96	1.00	1.00	1.00	1.00	0.06	1.00	1.00	1.00	1.00	0.06	1.00	1.00	1.00	1.00
MS2-5	0.77	0.92	0.96	1.00	1.00	0.02	0.84	0.91	0.95	1.00	0.01	0.86	0.92	0.96	1.00
MS2-6	0.96	1.00	1.00	1.00	1.00	0.06	1.00	1.00	1.00	1.00	0.06	1.00	1.00	1.00	1.00
MS2-7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2AR-1	0.00	0.06	0.19	0.35	0.86	0.00	0.12	0.20	0.33	0.77	0.00	0.09	0.33	0.54	1.00
MS2AR-2	0.01	0.84	0.96	1.00	1.00	0.01	0.57	0.70	0.80	1.00	0.01	0.57	0.72	0.81	1.00
MS2AR-3	0.84	1.00	1.00	1.00	1.00	0.01	0.91	0.96	1.00	1.00	0.01	0.91	0.96	1.00	1.00
MS2AR-4	1.00	1.00	1.00	1.00	1.00	0.03	1.00	1.00	1.00	1.00	0.03	1.00	1.00	1.00	1.00
MS2AR-5	0.01	0.84	0.96	1.00	1.00	0.01	0.57	0.70	0.80	1.00	0.01	0.57	0.72	0.81	1.00
MS2AR-6	1.00	1.00	1.00	1.00	1.00	0.03	1.00	1.00	1.00	1.00	0.03	1.00	1.00	1.00	1.00
MS2AR-7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2AR-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS3-1	0.03	0.57	0.67	0.78	1.00	0.02	0.26	0.31	0.39	1.00	0.02	0.28	0.35	0.46	0.91
MS3-2	0.75	0.94	0.96	1.00	1.00	0.05	0.75	0.87	0.92	1.00	0.07	0.76	0.88	0.93	1.00
MS3-3	0.90	1.00	1.00	1.00	1.00	0.09	0.98	1.00	1.00	1.00	0.09	1.00	1.00	1.00	1.00
MS3-4	1.00	1.00	1.00	1.00	1.00	0.10	1.00	1.00	1.00	1.00	0.10	1.00	1.00	1.00	1.00
MS3-5	0.56	0.94	0.96	1.00	1.00	0.05	0.75	0.87	0.92	1.00	0.05	0.76	0.87	0.93	1.00
MS3-6	0.96	1.00	1.00	1.00	1.00	0.10	1.00	1.00	1.00	1.00	0.10	1.00	1.00	1.00	1.00
MS3-7	1.00	1.00	1.00	1.00	1.00	0.10	1.00	1.00	1.00	1.00	0.10	1.00	1.00	1.00	1.00

Table 2 continued

DGP	Adjusted Rand Index														
	MS-True State					Fuzzy-True State					Fuzzy-MS				
	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.
MS3-8	1.00	1.00	1.00	1.00	1.00	0.10	1.00	1.00	1.00	1.00	0.10	1.00	1.00	1.00	1.00
MS3AR-1	0.00	0.18	0.30	0.45	0.85	0.00	0.16	0.22	0.29	0.96	0.00	0.21	0.30	0.40	1.00
MS3AR-2	0.13	0.81	0.93	1.00	1.00	0.09	0.39	0.55	0.72	1.00	0.02	0.39	0.55	0.74	1.00
MS3AR-3	0.78	1.00	1.00	1.00	1.00	0.12	0.84	0.93	0.98	1.00	0.12	0.84	0.93	0.98	1.00
MS3AR-4	1.00	1.00	1.00	1.00	1.00	0.13	1.00	1.00	1.00	1.00	0.13	1.00	1.00	1.00	1.00
MS3AR-5	0.13	0.81	0.93	1.00	1.00	0.09	0.39	0.56	0.72	1.00	0.02	0.39	0.55	0.73	1.00
MS3AR-6	1.00	1.00	1.00	1.00	1.00	0.12	0.98	1.00	1.00	1.00	0.12	0.98	1.00	1.00	1.00
MS3AR-7	1.00	1.00	1.00	1.00	1.00	0.13	1.00	1.00	1.00	1.00	0.13	1.00	1.00	1.00	1.00
MS3AR-8	1.00	1.00	1.00	1.00	1.00	0.13	1.00	1.00	1.00	1.00	0.13	1.00	1.00	1.00	1.00

The summary statistics are the minimum (Min), the first quartile (Q1), the median (Q2), the third quartile (Q3), the maximum (Max) of the ARI for each set of Monte Carlo experiments. MS refers to the classification obtained by the MS inference on the state, Fuzzy to the classification obtained by the Fuzzy k-means procedure of clustering, True State is the correct classification of data

Table 3 Summary Statistics of the Adjusted Rand Index for MS Inference on the state and Fuzzy k-means clustering in 1000 Monte Carlo experiments and in relation to 16 Data Generating Processes (DGP) with 2 states and 16 DGPs with 3 states. $T = 500$

DGP	Adjusted Rand Index														
	MS-True State					Fuzzy-True State					Fuzzy-MS				
	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.
MS2-1	0.45	0.65	0.69	0.73	0.84	0.18	0.36	0.41	0.45	0.58	0.16	0.40	0.47	0.53	0.76
MS2-2	0.88	0.95	0.96	0.98	1.00	0.78	0.88	0.90	0.92	0.98	0.79	0.90	0.92	0.94	0.98
MS2-3	0.98	0.99	1.00	1.00	1.00	0.97	0.99	1.00	1.00	1.00	0.97	0.99	1.00	1.00	1.00
MS2-4	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
MS2-5	0.88	0.95	0.96	0.98	1.00	0.78	0.88	0.90	0.92	0.98	0.79	0.90	0.92	0.94	0.98
MS2-6	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
MS2-7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2AR-1	0.00	0.20	0.29	0.36	0.62	0.06	0.18	0.22	0.27	0.41	0.00	0.18	0.36	0.51	0.75
MS2AR-2	0.69	0.88	0.93	0.96	1.00	0.46	0.63	0.68	0.73	0.88	0.42	0.65	0.70	0.75	0.94
MS2AR-3	0.96	1.00	1.00	1.00	1.00	0.74	0.91	0.94	0.95	1.00	0.74	0.91	0.94	0.95	1.00
MS2AR-4	0.99	1.00	1.00	1.00	1.00	0.95	0.98	0.99	1.00	1.00	0.95	0.98	0.99	1.00	1.00
MS2AR-5	0.69	0.88	0.93	0.96	1.00	0.46	0.63	0.68	0.73	0.88	0.42	0.65	0.70	0.75	0.94
MS2AR-6	0.99	1.00	1.00	1.00	1.00	0.95	0.98	0.99	1.00	1.00	0.95	0.98	0.99	1.00	1.00
MS2AR-7	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
MS2AR-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS3-1	0.51	0.70	0.73	0.77	0.89	0.22	0.29	0.32	0.36	0.50	0.20	0.30	0.34	0.39	0.67
MS3-2	0.91	0.96	0.97	0.98	1.00	0.33	0.88	0.90	0.91	0.97	0.34	0.88	0.91	0.92	0.99
MS3-3	0.98	1.00	1.00	1.00	1.00	0.96	0.99	1.00	1.00	1.00	0.97	0.99	1.00	1.00	1.00
MS3-4	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00

Table 3 continued

DGP	Adjusted Rand Index														
	MS-True State					Fuzzy-True State					Fuzzy-MS				
	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.
MS3-5	0.91	0.96	0.97	0.98	1.00	0.33	0.88	0.90	0.91	0.97	0.34	0.88	0.91	0.92	0.99
MS3-6	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
MS3-7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS3-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS3AR-1	0.02	0.27	0.35	0.42	0.69	0.09	0.19	0.21	0.24	0.35	0.04	0.24	0.29	0.34	0.72
MS3AR-2	0.70	0.89	0.92	0.95	1.00	0.29	0.50	0.60	0.67	0.87	0.29	0.51	0.61	0.69	0.87
MS3AR-3	0.95	1.00	1.00	1.00	1.00	0.36	0.91	0.93	0.95	1.00	0.36	0.91	0.93	0.95	1.00
MS3AR-4	0.99	1.00	1.00	1.00	1.00	0.57	0.99	0.99	1.00	1.00	0.57	0.99	0.99	1.00	1.00
MS3AR-5	0.70	0.89	0.92	0.95	1.00	0.29	0.50	0.60	0.67	0.87	0.29	0.51	0.61	0.69	0.87
MS3AR-6	0.99	1.00	1.00	1.00	1.00	0.57	0.99	0.99	1.00	1.00	0.57	0.99	0.99	1.00	1.00
MS3AR-7	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
MS3AR-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The summary statistics are the minimum (Min), the first quartile (Q1), the median (Q2), the third quartile (Q3), the maximum (Max) of the ARI for each set of Monte Carlo experiments. MS refers to the classification obtained by the MS inference on the state, Fuzzy to the classification obtained by the Fuzzy k-means procedure of clustering, True State is the correct classification of data

Table 4 Summary Statistics of the Adjusted Rand Index for MS Inference on the state and Fuzzy k-means clustering in 1000 Monte Carlo experiments and in relation to 16 Data Generating Processes (DGP) with 2 states and 16 DGPs with 3 states. $T = 1000$

DGP	Adjusted Rand Index														
	MS-True State					Fuzzy-True State					Fuzzy-MS				
	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.
MS2-1	0.59	0.67	0.70	0.73	0.81	0.25	0.38	0.41	0.44	0.54	0.27	0.42	0.46	0.51	0.67
MS2-2	0.90	0.95	0.96	0.97	0.99	0.85	0.89	0.91	0.92	0.96	0.84	0.90	0.92	0.93	0.97
MS2-3	0.98	1.00	1.00	1.00	1.00	0.98	0.99	1.00	1.00	1.00	0.98	0.99	1.00	1.00	1.00
MS2-4	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2-5	0.90	0.95	0.96	0.97	0.99	0.85	0.89	0.91	0.92	0.96	0.84	0.90	0.92	0.93	0.97
MS2-6	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2-7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2AR-1	0.00	0.26	0.32	0.37	0.56	0.11	0.20	0.23	0.26	0.41	0.00	0.23	0.38	0.49	1.00
MS2AR-2	0.75	0.90	0.92	0.95	1.00	0.51	0.66	0.69	0.73	0.83	0.52	0.67	0.71	0.74	0.85
MS2AR-3	0.97	1.00	1.00	1.00	1.00	0.81	0.92	0.94	0.95	0.99	0.81	0.92	0.94	0.95	0.99
MS2AR-4	1.00	1.00	1.00	1.00	1.00	0.95	0.99	0.99	1.00	1.00	0.95	0.99	0.99	1.00	1.00
MS2AR-5	0.75	0.90	0.92	0.95	1.00	0.51	0.66	0.69	0.73	0.83	0.52	0.67	0.71	0.74	0.85
MS2AR-6	1.00	1.00	1.00	1.00	1.00	0.95	0.99	0.99	1.00	1.00	0.95	0.99	0.99	1.00	1.00
MS2AR-7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS2AR-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS3-1	0.60	0.72	0.74	0.77	0.85	0.24	0.30	0.32	0.35	0.46	0.25	0.31	0.34	0.37	0.54
MS3-2	0.93	0.96	0.97	0.98	0.99	0.82	0.89	0.90	0.91	0.95	0.81	0.89	0.91	0.92	0.96
MS3-3	0.99	1.00	1.00	1.00	1.00	0.98	0.99	0.99	1.00	1.00	0.97	0.99	1.00	1.00	1.00

Table 4 continued

DGP	Adjusted Rand Index														
	MS-True State					Fuzzy-True State					Fuzzy-MS				
	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.
MS3-4	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS3-5	0.93	0.96	0.97	0.98	0.99	0.82	0.89	0.90	0.91	0.95	0.81	0.89	0.91	0.92	0.96
MS3-6	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS3-7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MS3-8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The summary statistics are the minimum (Min), the first quartile (Q1), the median (Q2), the third quartile (Q3), the maximum (Max) of the ARI for each set of Monte Carlo experiments. MS refers to the classification obtained by the MS inference on the state, Fuzzy to the classification obtained by the Fuzzy k-means procedure of clustering, True State is the correct classification of data

Table 5 Summary Statistics of the Adjusted Rand Index for MS Inference on the state and Fuzzy k-means clustering in 1000 Monte Carlo experiments and in relation to 4 Data Generating Processes (DGP) with 4 states. $T = 100; 500; 1000$

DGP	MS-True State					Fuzzy-True State					Fuzzy-MS				
	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.	Min.	Q1	Q2	Q3	Max.
$T = 100$															
MS4-3	0.93	1.00	1.00	1.00	1.00	0.45	0.98	1.00	1.00	1.00	0.45	0.99	1.00	1.00	1.00
MS4-7	1.00	1.00	1.00	1.00	1.00	0.55	1.00	1.00	1.00	1.00	0.55	1.00	1.00	1.00	1.00
MS4AR-3	0.75	1.00	1.00	1.00	1.00	0.42	0.87	0.94	0.97	1.00	0.42	0.87	0.94	0.97	1.00
MS4AR-7	1.00	1.00	1.00	1.00	1.00	0.52	1.00	1.00	1.00	1.00	0.52	1.00	1.00	1.00	1.00
$T = 500$															
MS3-4	0.98	0.99	1.00	1.00	1.00	0.97	0.99	0.99	1.00	1.00	0.97	0.99	1.00	1.00	1.00
MS3-7	1.00	1.00	1.00	1.00	1.00	0.77	1.00	1.00	1.00	1.00	0.77	1.00	1.00	1.00	1.00
MS3AR-4	0.95	1.00	1.00	1.00	1.00	0.68	0.92	0.93	0.95	0.99	0.68	0.92	0.93	0.95	0.99
MS3AR-7	1.00	1.00	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00
$T = 1000$															
MS3-4	0.99	1.00	1.00	1.00	1.00	0.98	0.99	0.99	1.00	1.00	0.98	0.99	1.00	1.00	1.00
MS3-7	1.00	1.00	1.00	1.00	1.00	0.73	1.00	1.00	1.00	1.00	0.73	1.00	1.00	1.00	1.00
MS3AR-4	0.97	1.00	1.00	1.00	1.00	0.86	0.92	0.93	0.95	0.97	0.86	0.92	0.93	0.95	0.97
MS3AR-7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

The summary statistics are the minimum (Min), the first quartile (Q1), the median (Q2), the third quartile (Q3), the maximum (Max) of the ARI for each set of Monte Carlo experiments. MS refers to the classification obtained by the MS inference on the state, Fuzzy to the classification obtained by the Fuzzy k-means procedure of clustering, True State is the correct classification of data

Table 6 Percentage of correct identification, with different validation indices, of the number of states out of 1000 Monte Carlo experiments and in relation to 16 Data Generating Processes (DGP) with 2 states and 16 DGPs with 3 states. $T = 100$

DGP	Indices					
	PC	PE	MPC	ASW	ASWF	XB
MS2-1	97.6	99.9	6.4	51.6	34.7	10.9
MS2-2	99.8	100	96	99.2	99.3	98.4
MS2-3	99.9	99.9	99.8	99.8	99.8	99.8
MS2-4	100.0	100.0	99.9	99.9	99.9	99.9
MS2-5	99.8	100.0	96.0	99.2	99.3	98.4
MS2-6	100.0	100.0	99.9	99.9	99.9	99.9
MS2-7	100.0	100.0	100.0	100.0	100.0	100.0
MS2-8	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-1	96.1	99.5	4.5	38.9	25.4	6.9
MS2AR-2	98.3	100.0	45.7	81.7	76.4	59.6
MS2AR-3	99.8	99.9	96.3	99.1	99.4	98.4
MS2AR-4	99.9	99.9	99.9	99.9	99.9	99.9
MS2AR-5	98.3	100.0	45.7	81.7	76.4	59.6
MS2AR-6	99.9	99.9	99.9	99.9	99.9	99.9
MS2AR-7	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-8	100.0	100.0	100.0	100.0	100.0	100.0
MS3-1	2.6	0.0	7.4	14.1	11.4	13.3
MS3-2	18.7	6.7	57.9	53.6	55.3	59.7
MS3-3	63.4	42.7	87.3	88.4	86.8	85.1
MS3-4	95.6	92.7	97.2	96.7	96.3	97.2
MS3-5	18.7	6.7	57.9	53.6	55.3	59.7
MS3-6	83.1	70.9	92.5	92.8	92.2	92.1
MS3-7	93.3	86.5	95.4	95	94.9	95.3
MS3-8	95.6	92.7	97.2	96.7	96.3	97.2
MS3AR-1	5.9	0.7	8.0	15.7	14.1	18.2
MS3AR-2	5.9	1.4	21.8	22.6	22.8	25.7
MS3AR-3	30.9	15.6	66.8	65.1	64.6	62.6
MS3AR-4	93.3	87.9	95.6	95.2	94.8	95.7
MS3AR-5	5.9	1.4	21.8	22.6	22.8	25.7
MS3AR-6	64.2	45.4	87.3	88.4	87.3	83.3
MS3AR-7	88.2	78.6	93.5	93.9	93.7	93.2
MS3AR-8	93.3	87.9	95.6	95.2	94.8	95.7

3.3 AIC/BIC (based on DGP structure) versus fuzzy

A natural way to evaluate the performance of the fuzzy procedure, in terms of correctly detecting the number of states, is to compare its results with those obtained by applying the AIC and BIC criteria, assuming that the correct DGP is known (we will relax this assumption in Sect. 3.4). Unfortunately, using these approaches involves computational effort since

Table 7 Percentage of correct identification, with different validation indices, of the number of states out of 1000 Monte Carlo experiments and in relation to 16 Data Generating Processes (DGP) with 2 states and 16 DGPs with 3 states. $T = 500$

DGP	Indices					
	PC	PE	MPC	ASW	ASWF	XB
MS2-1	100.0	100.0	13.9	95.4	78.8	4.5
MS2-2	100.0	100.0	100.0	100.0	100.0	100.0
MS2-3	100.0	100.0	100.0	100.0	100.0	100.0
MS2-4	100.0	100.0	100.0	100.0	100.0	100.0
MS2-5	100.0	100.0	100.0	100.0	100.0	100.0
MS2-6	100.0	100.0	100.0	100.0	100.0	100.0
MS2-7	100.0	100.0	100.0	100.0	100.0	100.0
MS2-8	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-1	100.0	100.0	2.5	80.0	42.7	0.6
MS2AR-2	100.0	100.0	88.7	99.9	99.7	85.7
MS2AR-3	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-4	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-5	100.0	100.0	88.7	99.9	99.7	85.7
MS2AR-6	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-7	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-8	100.0	100.0	100.0	100.0	100.0	100.0
MS3-1	0.0	0.0	8.8	2.8	3.5	3.4
MS3-2	6.9	0.4	89.7	83.9	89.6	95.4
MS3-3	91.4	62.4	100.0	100.0	100.0	100.0
MS3-4	99.2	94.8	100.0	100.0	100.0	100.0
MS3-5	6.9	0.4	89.7	83.9	89.6	95.4
MS3-6	99.2	94.8	100.0	100.0	100.0	100.0
MS3-7	100.0	99.7	100.0	100.0	100.0	100.0
MS3-8	100.0	100.0	100.0	100.0	100.0	100.0
MS3AR-1	0.0	0.0	6.1	9.7	8.4	5.3
MS3AR-2	0.1	0.0	25.9	9.4	14.4	35.2
MS3AR-3	22.8	2.6	95.9	94.1	95.9	97.3
MS3AR-4	85.5	53.3	99.7	99.8	99.9	99.7
MS3AR-5	0.1	0.0	25.9	9.4	14.4	35.2
MS3AR-6	85.5	53.3	99.7	99.8	99.9	99.7
MS3AR-7	99.6	97.2	100.0	100.0	100.0	100.0
MS3AR-8	99.9	99.8	100	100	100	100

they require estimating several models with different numbers of states to compare.[8] For this reason, we limit this analysis only to 16 DGPs: 8 DGPs are related to *good* cases where both the MS and fuzzy approaches seem to perform well (MS2-2, MS2-5, MS2-6, MS2AR-2, MS2AR-5, MS2AR-6, MS3-6, MS3-AR6); 8 DGPs are related to the most *puzzling* cases, which deserve a more in-depth analysis (MS2-1, MS2AR-1, MS3-1, MS3-2, MS3-5, MS3AR-1, MS3AR-2, MS3AR-5).

[8] Just to have an idea, the computational time to apply the loss function based approach to 1000 simulated series is about 190 min for the processes with the AR term and about 84 min when the AR term is not present, while for the fuzzy approach, it takes about 1 min and 30 s.

Table 8 Percentage of correct identification, with different validation indices, of the number of states out of 1000 Monte Carlo experiments and in relation to 16 Data Generating Processes (DGP) with 2 states and 16 DGPs with 3 states. $T = 1000$

DGP	Indices					
	PC	PE	MPC	ASW	ASWF	XB
MS2-1	100.0	100.0	18.7	99.2	95.0	2.9
MS2-2	100.0	100.0	100.0	100.0	100.0	100.0
MS2-3	100.0	100.0	100.0	100.0	100.0	100.0
MS2-4	100.0	100.0	100.0	100.0	100.0	100.0
MS2-5	100.0	100.0	100.0	100.0	100.0	100.0
MS2-6	100.0	100.0	100.0	100.0	100.0	100.0
MS2-7	100.0	100.0	100.0	100.0	100.0	100.0
MS2-8	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-1	100.0	100.0	1.8	96.8	59.3	0.1
MS2AR-2	100.0	100.0	98.7	100.0	100.0	96.2
MS2AR-3	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-4	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-5	100.0	100.0	98.7	100.0	100.0	96.2
MS2AR-6	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-7	100.0	100.0	100.0	100.0	100.0	100.0
MS2AR-8	100.0	100.0	100.0	100.0	100.0	100.0
MS3-1	0.0	0.0	6.8	0.5	0.9	1.1
MS3-2	3.2	0.1	98.1	95.3	98.6	99.9
MS3-3	97.7	69.5	100.0	100.0	100.0	100.0
MS3-4	100.0	99.0	100.0	100.0	100.0	100.0
MS3-5	3.2	0.1	98.1	95.3	98.6	99.9
MS3-6	100.0	99.0	100.0	100.0	100.0	100.0
MS3-7	100.0	100.0	100.0	100.0	100.0	100.0
MS3-8	100.0	100.0	100.0	100.0	100.0	100.0
MS3AR-1	0.0	0.0	5.0	3.5	5.5	3.7
MS3AR-2	0.0	0.0	23.7	4.2	10.0	40.7
MS3AR-3	16.9	0.5	99.5	99.6	99.6	99.7
MS3AR-4	94.0	57.5	100.0	100.0	100.0	100.0
MS3AR-5	0.0	0.0	23.7	4.2	10.0	40.7
MS3AR-6	94.0	57.5	100.0	100.0	100.0	100.0
MS3AR-7	100.0	99.4	100.0	100.0	100.0	100.0
MS3AR-8	100.0	100.0	100.0	100.0	100.0	100.0

In these new simulation experiments, to detect the number of states with AIC and BIC, we estimate MS models with a number of states from 1 to 4.

In the left panel of Table 10 we show the percentage of correct identifications with AIC and BIC for the 8 *good* DGPs; they are very high, especially with BIC reaching 100% successes in most of the 2-state cases. However, comparing them with the results in Table 6, the fuzzy approach also performs similarly, in some cases better than AIC. Again, larger differences are found in the 3-state cases, but, considering that the fuzzy approach is non-parametric and does not use MS model estimation, the differences are not so dramatic. In summary, for these cases, the simple fuzzy procedure seems to have an overall good performance.

Table 9 Percentage of correct identification, with different validation indices, of the number of states out of 1000 Monte Carlo experiments and in relation to 4 Data Generating Processes (DGP) with 4 states. $T = 100; 500; 1000$

DGP	PC	PE	MPC	ASW	ASWF	XB
$T = 100$						
MS4-3	82.7	64.5	89.4	90.2	89.7	89.6
MS4-7	95.4	94.0	95.8	95.5	95.1	95.8
MS4AR-3	39.0	13.3	74.3	69.0	72.8	71.6
MS4AR-7	94.1	91.3	94.7	94.5	94.4	94.6
$T = 500$						
MS4-3	100.0	99.3	100.0	100.0	100.0	100.0
MS4-7	99.9	99.9	99.9	99.9	99.9	99.9
MS4AR-3	60.6	5.5	99.7	98.3	99.6	99.8
MS4AR-7	99.9	99.8	99.9	99.9	99.9	99.9
$T = 1000$						
MS4-3	100.0	99.9	100.0	100.0	100.0	100.0
MS4-7	99.9	99.9	99.9	99.9	99.9	99.9
MS4AR-3	69.7	1.5	100.0	99.9	99.9	100.0
MS4AR-7	100.0	100.0	100.0	100.0	100.0	100.0

Table 10 Percentage of correct identification, with AIC and BIC of the number of states out of 1000 Monte Carlo experiments and in relation to 16 DGPs. $T = 100$

DGP	AIC	BIC	DGP	AIC	BIC
MS2-2	95.4	100	MS2-1	89.9	72.6
MS2-5	95.9	100	MS2AR-1	21.1	1.9
MS2-6	99.8	100	MS3-1	47.8	8.8
MS2AR-2	94.6	97.2	MS3-2	94.3	88.2
MS2AR-5	94.0	97.0	MS3-5	94.5	88.2
MS2AR-6	100	100	MS3AR-1	9.0	0.0
MS3-6	97.5	97.4	MS3AR-2	86.7	38.1
MS3AR-6	97.4	97.4	MS3AR-5	86.4	38.2

The left panel refers to 8 *good* DGPs where both MS and fuzzy approaches provide good results; right panel refers to more *puzzling* cases

To investigate the reasons why the fuzzy approach seems to fail in some cases, we consider the empirical distribution of the number of states identified with each of the six indices for the 8 *puzzling* DGPs mentioned above with $T = 100$ (Fig. 2).[9] In most cases the procedure is in doubt between 2 and 3 states; in particular PC and PE clearly favour 2 states, while the other indices present two modes corresponding to 2 and 3 states, excluding the cases of MS2-1, MS2AR-1, MS3-1 and MS3AR-1, which are the most confounding also in terms of marginal distributions (see Fig. 1). Looking at Fig. 1, it is clear that these two cases have densities with only one mode, so a clustering algorithm fails to detect groups.

Given that the problem is caused by particular types of DGP and, as a consequence, by the shape of the corresponding mixture density, it could be interesting to check if AIC and BIC, under the assumption of MS model with k states, might be more suitable. The right panel of Table 10 shows the percentage success rate of these parametric methods in the 8 puzzling cases. These criteria also have difficulty identifying the correct number of states in

9 Graphs for all other DGPs are available upon request.

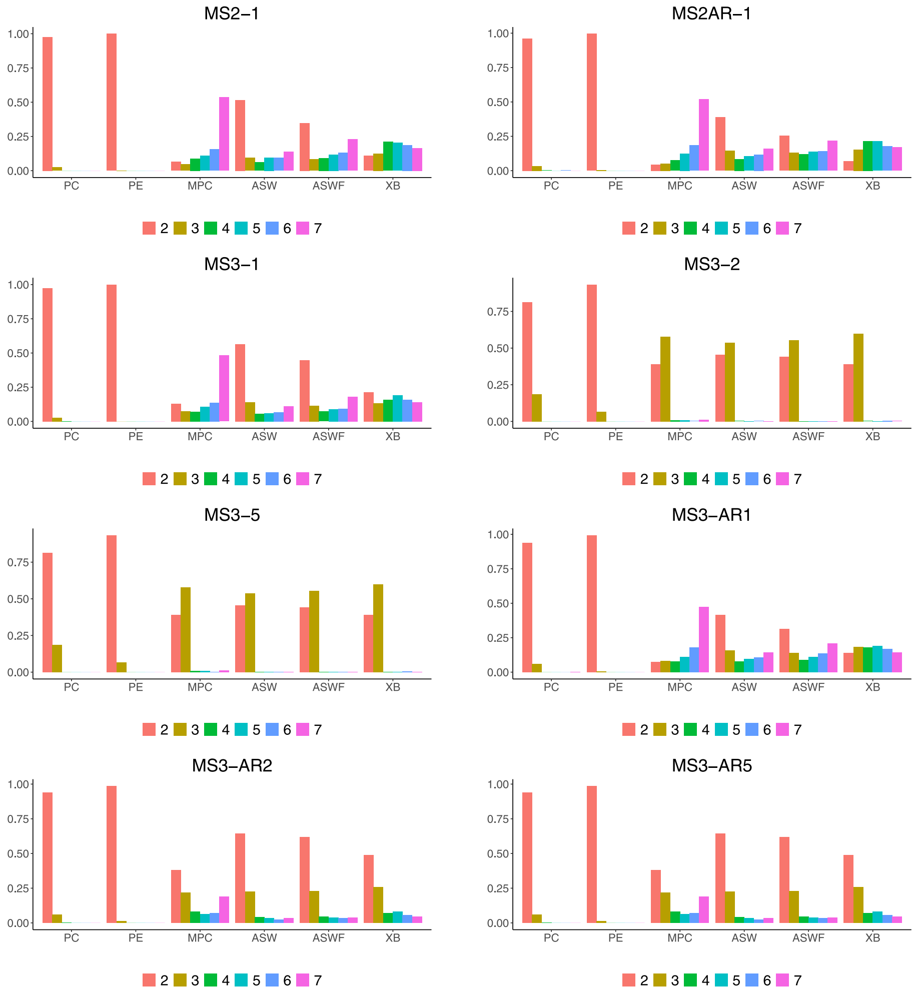

Fig. 2 Empirical distributions of the identified number of partitions with six indices in 1000 Monte Carlo experiments and 8 DGPs. $T = 100$

the MS2AR-1, MS3-1 and MS3AR-1 cases, where the components of the mixture overlap. In other cases, when the difference among the three components of the mixtures is sharper, AIC works much better than BIC, consistently with the results of Psadarakis and Spagnolo (2003). Comparing these results with those of the fuzzy approach, shown in Table 6, we notice very peculiar behaviors. In the MS2-1 and MS2AR-1 cases, PC and PE outperform AIC and BIC; the other four validation indices are more similar to AIC for the MS2AR-1 case. All approaches show problems in identifying the correct number of states for the MS3-1 and MS3AR-1 cases, while AIC clearly performs better in the MS3-2, MS3-5, MS3AR-2 and MS3-AR5 DGPs.

After these results, the suspicion is that these puzzling cases are due to the overlap of the mixture components, suggesting the identification of an MS model with an incorrect number of states; but the question is whether fitting the wrong model has similar accuracy

Table 11 Rejection rate of the Diebold–Mariano test under the null hypothesis of in-sample equal predictive ability of the MS model with number of states identified by the fuzzy procedure compared to the MS model with correct number of states in 1000 Monte Carlo experiments; Loss: MSE; $H_0 : MSE_{fuzzy} = MSE_{true}$; $H_a : MSE_{fuzzy} > MSE_{true}$; significance levels: 1%, 5%

DGP	Significance level	
	1%	5%
MS2-1	0.0	0.0
MS2AR-1	0.0	0.0
MS3-1	1.9	16.0
MS3-2	1.7	23.1
MS3-5	0.4	19.9
MS3AR-1	0.0	6.3
MS3AR-2	0.0	3.2
MS3AR-5	0.0	3.3

compared to the right model. For this purpose, for the 1000 series simulated with $T = 100$, we estimate the model with the real number of states and the one with the number of states derived from the fuzzy approach (according to the PC index, which presents an apparently very bad performance for the six 3-states DGPs in Table 6). We then compare the MSE of the two models in terms of Diebold and Mariano (1995) tests to check if they are not significantly different. Table 11 shows the percentage of cases in which the null hypothesis of equal MSE is rejected in favour of a lower MSE of the model with the true number of states. At a significance level of 1% the null hypothesis is almost always not rejected (always in the case of DGPs with an AR component); at a significance level of 5% the percentage increases significantly for non-autoregressive cases, but with at least 23% rejections. For the DGPs with 2 states, the null hypothesis is never rejected. These results suggest that the fuzzy approach can provide a suitable identification of MS model states in terms of MSE.

3.4 AIC/BIC (based on marginal distribution) versus fuzzy

In most of the above cases, AIC and BIC seem to outperform the proposed fuzzy approach in detecting the number of states; in those experiments, AIC and BIC were calculated assuming that the model is known (MS specification and presence/non-presence of the AR component), with only the number of states unknown. The exercise is useful because, in this way, AIC and BIC are benchmarks for evaluating the proposed non-parametric approach. In this subsection, we compare the fuzzy approach with AIC and BIC computed on the marginal distribution (Normal mixture), without using the DGP structure, with information similar to the fuzzy case.

We perform another set of simulation experiments, where we generate data from alternative distributions. The data are generated using the MS2-6, MS2AR-6, MS3-6 and MS3AR-6 DGP parameters (Table 1), where we have sufficient distance between the coefficients μ_i. The alternative distributions are[10]

[10] To simulate the data from GED, we use the R library *fGarch* (Wurtz et al., 2006); in practice, we use the standardized GED proposed by Nelson (1991) to generate data, which are multiplied by the value of σ and then the μ_i shown in Table 1 are added, obtaining mixture data components with the same mean and variance as the Normal cases generated for the simulations illustrated in the previous subsections. The shape parameter ν controls the thickness of the tails.

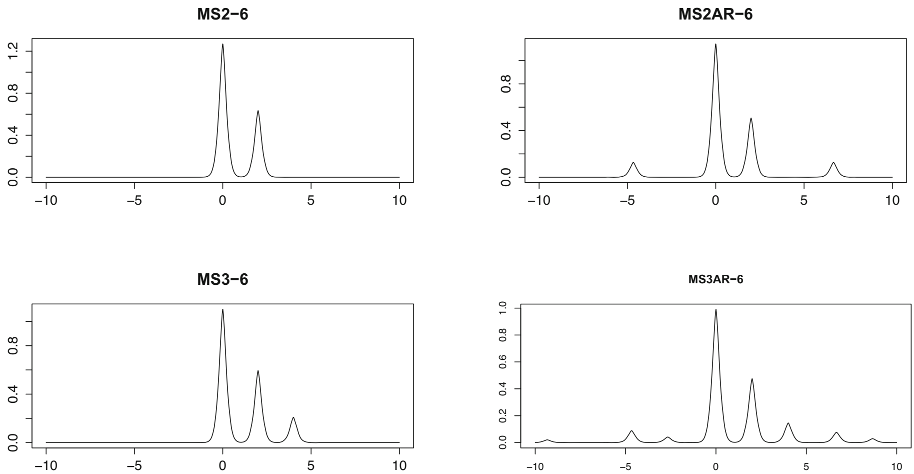

Fig. 3 GED ($\nu = 1.5$) Mixture density functions relative to MS2-6, MS2AR-6, MS3-6 and MS3AR-6 DGPs described in Table 1, with weights equal to the corresponding ergodic probabilities

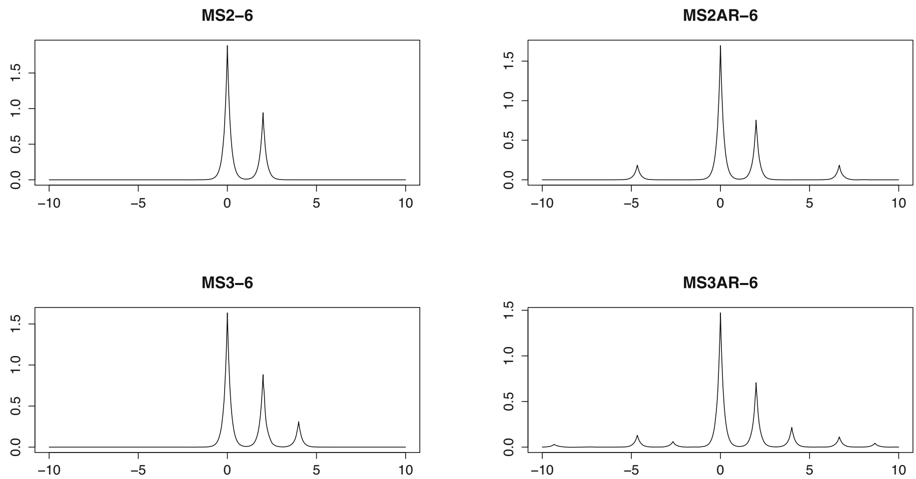

Fig. 4 GED ($\nu = 1$) Mixture density functions relative to MS2-6, MS2AR-6, MS3-6 and MS3AR-6 DGPs described in Table 1, with weights equal to the corresponding ergodic probabilities

- Normal distribution;
- GED with shape parameter $\nu = 1.5$; it is a symmetric leptokurtic distribution. The corresponding mixture densities, with weights equal to the ergodic probabilities of the Markov chain, are shown in Fig. 3;
- GED with $\nu = 1$, corresponding to the Laplace distribution. The corresponding mixture densities, with weights equal to the ergodic probabilities of the Markov chain, are shown in Fig. 4;
- the three previous distributions but inserting asymmetric effects in the model.

The asymmetry is obtained by adding a new parameter (γ) in the MS2AR-6 and MS3AR-6 AR models, which considers a different effect between negative and positive news (as usually happens in financial time series); in practice it is a sort of GJR[11] parameterization (Glosten et al., 1993) extended to the AR(1) case. More in detail:

$$y_t = \mu_{s_t} + (\phi + \gamma D_{t-1})\left(y_{t-1} - \mu_{s_{t-1}}\right) + \varepsilon_t,$$
$$D_t = \begin{cases} 1 \text{ if } \left(y_t - \mu_{s_t}\right) < 0 \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

In the simulations, the parameter γ is set equal to 0.3.

Table 12 shows the results of these new experiments. Starting from the Normal case, AIC and BIC work quite well, with the highest performances in the MS2-6 case and a clear worsening considering the AR cases; this is easily explained because, as shown in Fig. 1, the presence of the AR dynamics increases the number of modes in the marginal distribution. Similar comments apply to the fuzzy approach, which performs better than AIC and BIC in the MS2AR-6 case; similarly to previous experiments, PC and PE show a worsening in the 3-state cases.

If the distribution that generates the data is not the Normal one (the two symmetric GED cases), the number of correct identifications for AIC and BIC decreases (consistently with the results of, Psadarakis and Spagnolo 2004, 2009); on the other hand the fuzzy approach seems robust, in the sense that the performances are very similar to the previous Normal case, with a better behavior than the parametric methods (100% successes) in the 2-state cases.

Finally, the analysis of the asymmetric cases confirms the difficulties of AIC and BIC in detecting the correct number of states when the model specification is wrong; in particular, BIC shows a strong suffering in the 3-state cases. The fuzzy approach maintains the robustness of the results in the 2-state cases, showing a clear failure in the 3-state cases (but with a better performance than BIC).

In summary, in a framework where the correct DGP is unknown, the fuzzy approach appears to be more robust than the AIC and BIC methods.

4 The fuzzy procedure to identify the number of states in MS models

Given the results obtained by the Monte Carlo experiments and the above similarities between the membership grade matrix and the smoothed probabilities derived from the estimation of the MS model, we propose the following procedure for identifying the number of states in MS models:

1. Check for states in the time series y_t. This can be done using the homogeneity test procedure proposed by Hennig and Lin (2015) (more details in the next subsection);
2. Group the observed y_t using the *fuzzy k-means* procedure proposed by Bezdek (1981) for different numbers of clusters, (we propose $k = 2, ..., 7$);
3. Choose the number of clusters, k, that optimizes a given cluster validation index;
4. Estimate the MS model with a number of states corresponding to the value of k identified in step 3.;
5. assign each observation to the group that maximizes the degree of membership, i.e. the maximum value of each row $\boldsymbol{U}$;
6. Compare with the ARI the groups obtained in step 4. through smoothed inference and those obtained in step 5. by clustering.

[11] GJR comes from the first letter of the names of the three authors who proposed the corresponding model, namely Glosten, Jagannathan and Runkle.

Table 12 Percentage of correct identification, with AIC, BIC and fuzzy validation indices, of the number of states out of 1000 Monte Carlo experiments and in relation to MS2-6, MS2AR-6, MS3-6 and MS3AR-6 DGPs, with different distributions. AIC and BIC are detected by estimating mixture models wit 1,2,3,4 components. $T = 100$

		AIC	BIC	PC	PE	MPC	ASW	ASWF	XB
Normal	MS2-6	99.1	100	100	100	99.9	99.9	99.9	99.9
	MS2AR-6	90.8	97.3	99.9	99.9	99.9	99.9	99.9	99.9
	MS3-6	97.1	97.5	83.1	70.9	92.5	92.8	92.2	92.1
	MS3AR-6	91.0	94.4	64.2	45.4	87.3	88.4	87.3	83.3
GED $\nu = 1.5$	MS2-6	85.8	93.0	100	100	99.9	99.9	99.9	99.9
	MS2AR-6	89.4	97.4	100	100	100	100	100	100
	MS3-6	95.0	97.3	82.3	71.3	92.7	93.5	92.9	93.0
	MS3AR-6	88.2	93.6	63.0	45.1	85.4	86.4	85.6	81.6
GED $\nu = 1$	MS2-6	88.9	94.7	100	100	100	100	100	100
	MS2AR-6	86.5	95.1	100	100	100	100	100	100
	MS3-6	93.3	95.5	84.0	74.8	93.3	93.4	93.1	93.7
	MS3AR-6	85.8	93.0	66.6	49.6	85.9	87.0	85.3	83.6
Normal+Asym	MS2AR-6	48.7	58.7	93.3	96.6	69.6	84.4	81.4	79.2
	MS3AR-6	53.7	39.2	25.2	13.9	46.5	48.6	46.9	46.0
GED ν=1.5+Asym.	MS2AR-6	46.9	56.7	93.3	96.7	70.5	82.9	80.3	79.3
	MS3AR-6	53.4	38.5	24.7	12.7	46.9	47.2	48.1	45.1
GED ν=1+Asym.	MS2AR-6	45.0	56.4	92.7	97.3	67.3	80.7	77.6	75.2
	MS3AR-6	55.4	40.4	26.3	14.1	45.6	47.3	46.5	48.3

The last step of the procedure is to check the consistency of the results of the fuzzy procedure and the classical parametric inference on the state of the MS models.

Now we illustrate the procedure with a macroeconomic example.

4.1 An empirical application: identification of the number of regimes in U.S. business cycle

The seminal papers that sanctioned the success of MS models concern the analysis of business cycle (Hamilton, 1989, 1990). The baseline models adopted may vary: Hamilton (1989) estimates an AR(4) model of quarterly GDP; Kim (1994) uses the state-space representation of the Lam (1990) model, which extends the previous Hamilton (1989) model with a general autoregressive component; Chauvet and Hamilton (2006) adopt a simple model as the one shown in Eq. (3). The common element is the number of states, set a priori equal to 2 because the aim is to interpret them as the two phases of the business cycle (expansion and contraction respectively).

We consider the quarterly annualized U.S. GDP growth rate[12] (in percentage terms) from 1947:II to 2020:I, resulting in 292 observations.[13] The reference model is that of Chauvet

[12] Data are taken from the Federal Reserve Economic Data database.

[13] We consider 2020:I as the last observation to avoid the outlier in correspondence with the COVID pandemic, which affects dramatically the stability of the estimates (as underlined by Hamilton in his Econbrowser blog: *https://econbrowser.com/recession-index*).

Table 13 Best values of six clustering validation indices and p-values values under the null hypothesis of homogeneity

	PC	PE	MPC	ASW	ASWF	XB
Best value	0.924	0.136	0.886	0.748	0.796	0.103
p value	0.002	0.001	0.002	0.002	0.002	0.078

and Hamilton (2006), which corresponds to the model in Eq. (3), an MS model without autoregressive terms. Indeed, in the case of states and regime changes, very often the data show spurious autocorrelation and even spurious unit root (see the seminal work of, Perron, 1989). There are several examples where estimating an MS model without autoregressive structures, the residual series is white noise.[14] Other examples of applications of the MS model without AR terms are illustrated in Engel and Hamilton (1990) and Cheung and Erlandsson (2005), who apply this model to exchange rate series.

The presence of states (groups) in GDP growth rates is well established in the literature, where, as mentioned, 2-state models are generally adopted.[15]

In our opinion, and always following the parallelism with the clustering approach, it is a good practice to check for clusters in the dataset (step 1. of the proposed procedure). For this purpose we adopt the homogeneity test proposed by Hennig and Lin (2015) to check if data are grouped or not. The procedure consists of simulating the data from a homogeneous (clusterless) DGP with parameters derived from the observed data and calculating the cluster validation index on each of them, obtaining the empirical distribution of the index under the null hypothesis of homogeneity. The comparison of the validation index calculated on real data with the critical value derived from the simulated distribution will provide evidence of homogeneity or the presence of clusters.

We simulate 2000 series from a Normal distribution with parameters equal to the sample mean and variance of the GDP series, obtaining the *p*-values of the best indices (which identify 3 states) shown in Table 13. The rejection of the null hypothesis of homogeneity, based on all indices except XB (which shows a *p* value between 0.05 and 0.10), is strongly significant. This result supports the idea of the presence of states in the GDP growth rate series.

The identification of the number of states with the fuzzy approach and the AIC/BIC using the DGP structure and the marginal distribution is shown in Table 14. Using fuzzy clustering, all six validation indices identify the number of states $k = 3$. AIC and BIC (calculated again by comparing the models with 1, 2, 3, and 4 states) show somewhat contrasting results: if we consider the DGP structure, AIC identifies 4 states and BIC 3, while if we base their calculation on the marginal distribution, AIC identifies 3 states and BIC the linear model (1 state).

To analyze the fit of the alternative models, we compare the linear model (let us call it MS(1)) and the MS models with 2, 3 and 4 states in terms of two loss functions: Mean Squared Error (MSE) and Mean Absolute Error (MAE). The results are shown in the left part of Table 15. The numerical values of the two loss functions favor the models with more states, but this was expected: as is known, overparameterized models fit better than parsimonious

[14] This also happens for the series studied in this paper, adopting the selected MS(3) model of equation (4). Results are available on request.

[15] The study of GDP dynamics is the typical application of MS models in economic literature. It is emblematic that in the Hamilton (1994) time series book, where ch.22 is devoted to MS models, the example used to illustrate the theoretical framework is related to the quarterly GDP of the United States. The states are easily interpreted as growth and recession and MS models fit very well (the smoothed probabilities are close to 0 or 1).

Table 14 Number of states identified using the fuzzy approach and the AIC and BIC criteria for the quarterly time series of US GDP

	Fuzzy approach					
	PC	PE	MPC	ASW	ASWF	XB
Number of states	3	3	3	3	3	3
	DGP structure			Marginal distribution		
	AIC	BIC		AIC	BIC	
Number of states	4	3		3	1	

Table 15 Loss functions and MCS procedure for 4 MS models estimated on the quarterly time series of US GDP

Model	MSE	MAE	MCS-SE		MCS-AE	
			Model	p value	Model	p value
MS(1)	13.91	2.69	MS(2)	0.0003	MS(2)	0.0013
MS(2)	13.25	2.65	MS(1)	0.0088	MS(1)	0.0344
MS(3)	12.17	2.52	MS(3)	0.503	MS(4)	0.944
MS(4)	12.02	2.52	MS(4)	1	MS(3)	1

MS(1) represents the model without states. MCS-SE refers to the MCS procedure applied to the squared errors; MCS-AE refers to the MCS procedure applied to the absolute errors. In the MCS columns the first row represents the first model removed, down to the best performing model in the last row. The statistic used for the MCS procedure is the semi-quadratic statistic; the number of bootstrap replications to obtain the variance of the errors is 10,000 (for details of the MCS procedure, see Hansen et al., 2011)

models (see e.g. Hansen and Timmerman, 2015). The real question is whether the models with more states fit the data significantly better than other models with a smaller number of states. For this purpose, we use the Model Confidence Set (MCS) approach (Hansen et al., 2011), which selects, with an iterative test-based procedure, models with the same fitting performance. From the right side of Table 15 we note that MS(3) and MS(4) do not have a significantly different fit, both in terms of squared errors and absolute errors at all typical significance levels, while they significantly outperform linear and 2-state models (for absolute errors at 5% significance level). Since no significant differences in fit between MS(3) and MS(4) are found, based on the parsimony criterion, we identify the 3-state model as the preferred one. This result was suggested by all the validation indices in the fuzzy approach; the identification of the linear model by the BIC based on the marginal distribution seems consistent with the results of Psadarakis and Spagnolo (2003), who underline the tendency of the BIC to underestimate the identification of the number of states in the MS models.

Finally, we compare the empirical results of the MS(3) model we identified - equation (4) - with those derived from the model proposed by Chauvet and Hamilton (2006), corresponding to (3). Estimation results, using the Quasi MLE, are shown in Table 16.

We note that, in the case of 2 states, the expansion periods show an average growth rate of 3.85% and -1.67% in the contraction periods. Persistence in the expansion periods is longer than in contraction periods, as shown by the probability of staying in the same state (p_{11} and p_{22}); in this case the average duration of state i is given by $1/(1-p_{ii})$ (see, Hamilton, 1994), so the average duration of the expansion period is 20 quarters and 3.1 quarters for contraction periods. The first state of the MS(3) case shows a very large mean (8.03), while the other 2 states are more in line with the two states of the MS(2) model. Their interpretation is immediate: state 2 is the expansion phase, state 3 the contraction phase and state 1 captures

Table 16 Parameter estimates with robust standard errors in parentheses for growth rates series of U.S. quarterly GDP (1947:II-2000:I)

μ_1	μ_2	μ_3	p_{11}	p_{12}	p_{21}	p_{22}	p_{32}	p_{33}	σ
MS(2)									
3.85	−1.67		0.95			0.68			3.19
(0.34)	(1.91)		(0.03)			(0.13)			(0.19)
MS(3)									
8.03	2.91	−2.40	0.69	0.31	0.03	0.91	0.11	0.66	2.50
(0.73)	(0.44)	(4.00)	(0.10)	(0.10)	(0.08)	(0.08)	(0.56)	(0.42)	(0.47)

The transition probabilities p_{12} and p_{21} in the 2-state cases are obtained as $p_{ij} = 1 - p_{ii}$ $(i, j = 1, 2;$ $i \neq j)$. The transition probabilities p_{13}, p_{23} and p_{31} are obtained as $1 - \sum_{j \neq i} p_{ij}$. Robust standard errors are computed by using the so-called 'sandwich matrix' (White, 1982), which is the asymptotic variance-covariance matrix of the Quasi MLE when the latter is asymptotically Gaussian. It is given by a quadratic combination of the information matrices derived from the Hessian of the log-likelihood and the outer product of the scores

the highest peaks of the series so that it can be interpreted as a boom period. The average duration of the boom is 3.2 quarters, 11.1 quarters for the expansion phase and 2.9 quarters for the contraction phase. The decrease of the variance parameter in MS(3) compared to MS(2) is clearly due to the presence of more homogeneous growth rates within each state (in practice the largest peaks are not included in the expansion state as in the MS(2) model).

Previous comments are supported by Fig. 5, where we show the inference on the state derived from the smoothed probabilities, both for MS(2) (upper panel) and MS(3) (lower panels). It is clear that state 2 in MS(2) is equivalent to state 3 in MS(3) and corresponds to the contraction periods, while state 1 in MS(2)[16] is decomposed in state 1 (corresponding to the highest peaks) and state 2 (moderate positive values) in MS(3). It is also interesting to note that GDP shows a different behavior in the second half of the time series, with only a couple of troughs.[17] However, the MS model is able to detect this different behavior compared to the first half of the series, where changes are more frequent; the smoothed probabilities follow the profile of the series, assigning most of the observations to state 1 (in the MS(2) model) and state 2 (in the MS(3) model). This result is achieved thanks to the filtering and smoothing procedures, which consider both the Markov transition probabilities and the densities of the studied variable (see Hamilton, 1994).

Finally, we calculate the ARI to compare the inference on the state derived from the MS(3) model and fuzzy clustering. The value is 0.48, which is consistent with the MS3-1 case illustrated in Table 2. Of course, the great variability of the data affects this result, in particular for the presence of abrupt changes in the dynamics of the time series. Smoothing the original series with moving averages, the ARI grows rapidly: a 3-term moving average implies an ARI of 0.63, while a 5-term moving average shows an ARI equal to 0.82, confirming the evidence for the presence of 3 states.

[16] For the 2-state model, the probabilities of state 1 can be obtained as 1 minus the probabilities of state 2.

[17] This behavior might suggest that a promising alternative could be to adopt a time-varying transition probability (TVTP) approach (see, e.g., Diebold et al., 1994) In this case, we need to identify some exogenous variables that drive the time-varying dynamics of $p_{ij,t}$. Identifying exogenous variables might be a critical issue; it can be bypassed by replacing them with a latent variable, as in Otranto (2008).

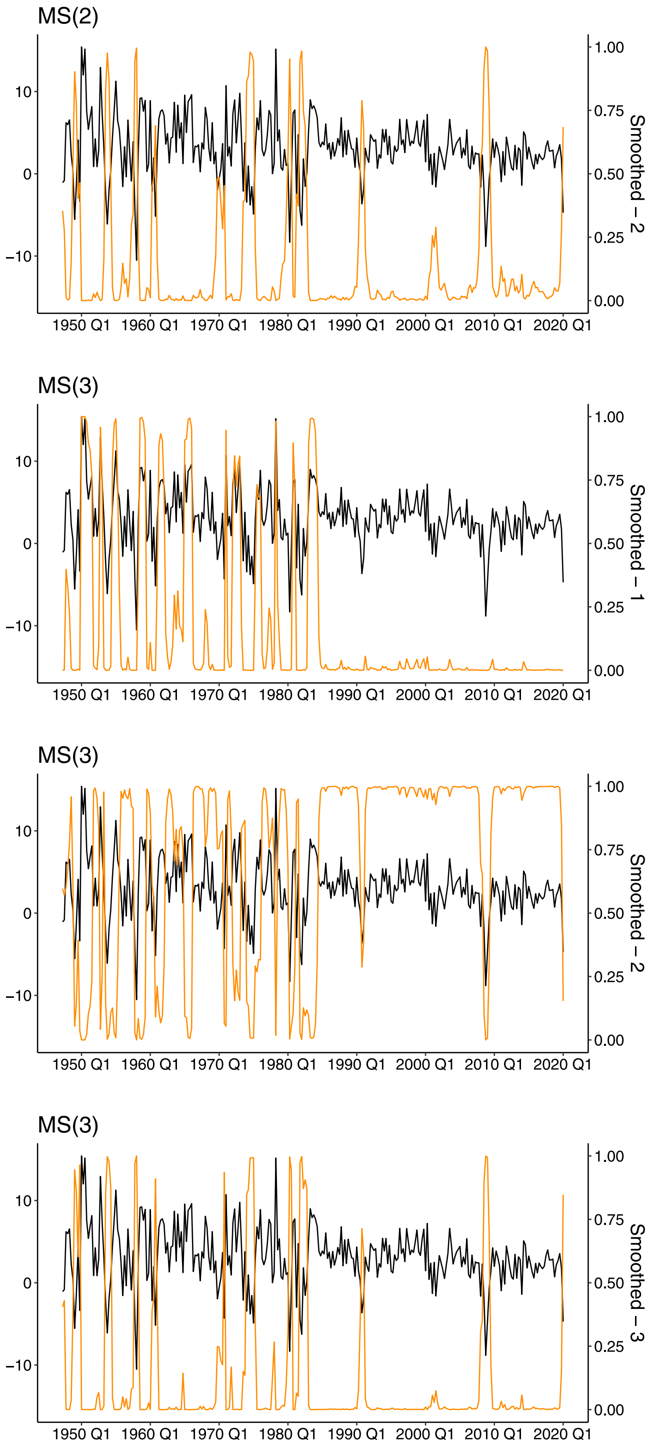

Fig. 5 US GDP growth rate (black line) and smoothed probabilities (orange line) derived from the MS(2) model (first panel from top; smoothed probabilities refer to state 2) and the MS(3) model (second panel, state 1; third panel, state 2; fourth panel, state 3). Sample: 1947:II 2020:I

5 Final remarks

Detecting the number of states k is a crucial task in estimating MS models. This problem is often circumvented, by fixing k a priori, due to the impossibility of applying the classical statistical tests for the problem of nuisance parameters present only under the alternative hypothesis. In this work, we underline the similarity between the idea of inference on the state, obtained from MS models, and fuzzy clustering. After checking the similarity of the grouping with the two procedures through Monte Carlo experiments, on the same simulated series, we investigated the ability of the main clustering validation indices to detect the correct number of states. Our experiments show a good performance, in particular of two indices (ASW and ASWF), also when the dynamics of the time series is affected by autocorrelation. The performance of the indices increases when the differences between the means within states are greater and the variability is lower and when the length of time series increases. Some puzzling cases were analyzed in more depth, showing that when the components of the mixture density generating the data overlap, it is easy to detect an incorrect number of states, but the corresponding model has a similar fitting, in terms of MSE, compared to the model with the correct number of states. Our simulation experiments show that the usual AIC and BIC criteria are good methods for detecting the number of states of MS models when they are derived from the correct model specification, but the fuzzy approach seems more robust when the DGP is unknown.

The results seem to support the possibility of adopting clustering validation indices as a tool to identify, before the estimation step, the number of states of MS models in a non-parametric way. The advantage of this procedure is even more evident if we consider that the calculation of these indices is implemented in the main statistical packages.[18]

We think that when different validation indices identify the same number of states, there is strong support for this result. In the last section, we detail this procedure applying it to the US GDP growth rate series, identifying 3 states with all six indices, against the usual practice of setting (a priori) $k = 2$ for this variable. Comparing the linear and MS models with 2, 3, and 4 states in terms of loss functions and MCS procedure, the 3-state model is preferred.

As future research, it might be interesting to consider other distance measures, such as that of Mahalanobis, or medoid-based approaches as opposed to the centroid-based ones proposed here. It would thus be possible to compare the different fuzzy clustering algorithms to evaluate which one works better.

Furthermore, it could be interesting to investigate whether the *fuzzy c regression model*, adopted by Hathaway and Bezdek (1993) to estimate switching regression models, can improve the proposed fuzzy approach; in this proposal, the authors use as centroids the fitted data implicit in the estimated parameters obtained by minimizing a suitable objective function.

From another perspective, it might be interesting to compare fuzzy clustering with the regime inference obtained from the MS model with fuzzy regimes, recently proposed by Gallo and Otranto (2018), where the separation between regimes is not clear-cut, but with the possibility of overlapping states. Furthermore, the procedure proposed above can be extended to a Time Contrastive Learning framework (Hyvärinen et al., 2023) to identify the number of latent states.

[18] We used the R library *fclust* (Ferraro et al., 2019) to perform the fuzzy clustering and to calculate the validation indices. All codes to estimate the MS models and to implement the fuzzy procedure are available upon request. The algorithm used to iteratively maximize the log-likelihood is the BFGS, implemented within the R function `optim`.

Finally, a further challenge could be to evaluate the performance of the proposed approach with different and more complex MS specifications, such as, for example, the MS Zero-Drift GARCH model proposed by Shi (2023), or with non-Markovian regime switching models, such as the multivariate GARCH model subject to multiple regimes used by Cerqueti et al. (2024).

Acknowledgements Edoardo Otranto acknowledges financial support from the Italian PRIN 2022 grant "Methodological and computational issues in large-scale time series models for economics and finance" (20223725WE), funded by the European Union - Next Generation EU, Mission 4 Component 1 CUP 2J53D23003960006. Luca Scaffidi Domianello acknowledges financial support from PNRR MUR project PE0000013-FAIR. We thank an Associate Editor and two anonymous reviewers for their valuable comments and suggestions which improved the contribution of the first version of the paper. Thanks are due also to the participants at SIS 2024 in Bari (June 17–20, 2024) for their valuable comments.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors report there are no Conflict of interest to declare.

References

Albert, J. H., & Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business & Economic Statistics, 11*(1), 1–15.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press.

Campello, R. J., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems, 157*(21), 2858–2875.

Carrasco, M., Hu, L., & Ploberger, W. (2014). Optimal test for Markov switching parameters. *Econometrica, 82*(2), 765–784.

Cerqueti, R., Gatfaoui, H., & Rotundo, G. (2024). Resilience for financial networks under a multivariate Garch model of stock index returns with multiple regimes. *Annals of Operations Research*, 1–27 .

Chauvet, M., & Hamilton, J. D. (2006). Dating business cycle turning points. In C. Milas, P. Rothman, & D. Van Dijk (Eds.), *Nonlinear time series analysis of business cycles, volume 128 of advances in computers* (pp. 1–54). Emerald Group Publishing Limited.

Chen, Y., Fuh, C. D., & Kao, C. L. M. (2024). Determine the number of states in hidden Markov models via marginal likelihood. *Journal of Machine Learning Research, 24*, 1–58.

Cheung, Y. W., & Erlandsson, U. (2005). Exchange rates and Markov switching dynamics. *Journal of Business & Economic Statistics, 23*, 314–320.

Cho, J. S., & White, H. L. (2007). Testing for regime switching. *Econometrica, 75*, 1671–1720.

Dave, R. N. (1996). Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letters, 17*(6), 613–623.

Davenport, J., Bezdek, J., & Hathaway, R. (1988). Parameter estimation for finite mixture distributions. *Computers & Mathematics with Applications, 15*(10), 819–828.

Davies, R. (1977). Hypotheses testing when a nuisance parameter is present only under the alternative. *Biometrika, 174*, 247–254.

Degras, D., Ting, C. M., & Ombao, H. (2022). Markov-Switching state-space models with applications to neuroimaging. *Computational Statistics and Data Analysis, 174*, 107525.
Diebold, F., Lee, J. H., & Weinbach, G. (1994). Regime switching with time-varying transition probabilities. In C. P. Hargreaves (Ed.), *Nonstationary time series analysis and cointegration* (pp. 283–302). Oxford University Press.
Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 134–144.
D'Urso, P. (2015). Fuzzy clustering. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 545–566). Chapman and Hall.
Engel, C., & Hamilton, J. (1990). Long swings in the dollar: Are they in the data and do markets know it? *The American Economic Review, 80*, 689–713.
Ferraro, M. B., Giordani, P., Serafini, A., et al. (2019). fclust: an R package for fuzzy clustering. *The R Journal* 1–18.
Fuh, C. D., Kao, C. L. M., & Pang, T. (2024). Kullback–Leibler divergence and Akaike information criterion in general hidden Markov models. *IEEE Transactions on Information Theory, 70*(8), 5888–5909.
Gallo, G., & Otranto, E. (2015). Forecasting realized volatility with changing average volatility levels. *International Journal of Forecasting, 31*, 620–634.
Gallo, G., & Otranto, E. (2018). Combining sharp and smooth transitions in volatility dynamics: A fuzzy regime approach. *Journal of the Royal Statistical Society, Series C: Applied Statistics, 67*, 549–573.
Gallo, G. M., Lacava, D., & Otranto, E. (2021). On classifying the effects of policy announcements on volatility. *International Journal of Approximate Reasoning, 134*, 23–33.
Garcia, R. (1998). Asymptotic null distribution of the likelihood ratio test in Markow-switching models. *International Economic Review, 39*, 763–788.
Glosten, L., Jagannathan, R., & Runkle, D. (1993). On the relation between expected value and the volatility of the nominal excess return on stocks. *Journal of Finance, 48*, 1779–1801.
Haas, M. (2009). Value-at-risk via mixture distributions reconsidered. *Applied Mathematics and Computation, 6*(215), 2103–2119.
Haas, M., Mittnik, S., & Paolella, M. (2004). A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics, 2*(4), 493–530.
Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica, 57*, 357–384.
Hamilton, J. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics, 45*, 39–70.
Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.
Hamilton, J. D. (2016). Chapter 3—Macroeconomic regimes and regime shifts. In J. B. Taylor & H. Uhlig (Eds.), *Handbook of macroeconomics* (Vol. 2, pp. 163–201). Elsevier.
Hansen, B. (1992). The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. *Journal of Applied Econometrics, 7*, S61–S82.
Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica, 64*, 413–430.
Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica, 79*(2), 453–497.
Hansen, P. R., & Timmerman, A. (2015). Realized GARCH models: Simpler is better. *Journal of Business and Economic Statistics, 33*, 17–21.
Hathaway, R. J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters, 4*(2), 53–56.
Hathaway, R. J., & Bezdek, J. C. (1993). Switching regression models and fuzzy clustering. *IEEE Transactions on Fuzzy Systems, 1*(3), 195–204.
Hennig, C., & Lin, C. J. (2015). Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. *Statistics and Computing, 25*, 821–833.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.
Hwu, S. T., & Kim, C. J. (2023). Markov-switching models with unknown error distributions: Identification and inference within the Bayesian framework. *Studies in Nonlinear Dynamics & Econometrics* .
Hyvärinen, A., Khemakhem, I., & Morioka, H. (2023). Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns 4*(10) .
Ichihashi, H., Miyagishi, K., & Honda, K. (2001). Fuzzy c-means clustering with regularization by K-L information. In *10th IEEE international conference on fuzzy systems* (Cat. No. 01CH37297) (Vol. 2, pp. 924–927). IEEE.
Kapetanios, G. (2001). Model selection in threshold models. *Journal of Time Series Analysis, 22*(6), 631–756.
Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 49–66 .

Kiefer, N. M. (1978). Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrica, 46*, 427–434.
Kim, C. (1994). Dynamic linear models with Markov switching. *Journal of Econometrics, 60*, 1–22.
Kim, C. J., & Nelson, C. (1999). *State-space models with regime switching: Classical and Gibbs-sampling approaches with applications*. The MIT Press.
Koutmos, D. (2020). Market risk and bitcoin returns. *Annals of Operations Research, 294*(1), 453–477.
Lam, P. S. (1990). The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series. *Journal of Monetary Economics, 26*, 409–432.
Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics, 20*(3), 1350–1360.
McCulloch, R. E., & Tsay, R. S. (1994). Statistical analysis of economic time series via Markov switching models. *Journal of Time Series Analysis, 15*, 523–539.
McLachlan, G. J., & Rathnayake, S. I. (2014). On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4 .
Nelson, D. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica, 59*, 349–370.
Otranto, E. (2008). A time varying hidden Markov model with latent information. *Statistical Modelling, 8*(4), 347–366.
Otranto, E., & Gallo, G. M. (2002). A nonparametric Bayesian approach to detect the number of regimes in Markov switching models. *Econometric Reviews, 21*, 477–496.
Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. *Econometrica, 57*(6), 1361–1401.
Psadarakis, Z., & Spagnolo, N. (2003). On the determination of the number of regimes in Markov-switching autoregressive models. *Journal of Time Series Analysis, 24*, 237–252.
Psadarakis, Z., & Spagnolo, N. (2006). Joint determination of the state dimension and autoregressive order for models with Markov regime switching. *Journal of Time Series Analysis, 27*, 753–766.
Qu, Z., & Zhuo, F. (2021). Likelihood ratio-based tests for Markov regime switching. *The Review of Economic Studies, 88*(2), 937–968.
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.
Serafini, A., Scrucca, L., Alfò, M., Giordani, P., & Ferraro, M. B. (2023). *Fuzzy and model based clustering methods: Can we fruitfully compare them?, Models for Data Analysis: SIS 2018, Palermo, Italy, June 20–22, 283–304*. Springer.
Shi, Y. (2023). A simulation study on the Markov regime-switching zero-drift GARCH model. *Annals of Operations Research, 330*(1), 1–20.
Smith, A., Naikb, P. A., & Tsai, C. L. (2006). Markov-Switching model selection using Kullback–Leibler divergence. *Journal of Econometrics, 134*, 553–577.
Titterington, D. M. (1997). *Mixture distributions (update)* (pp. 399–407). Wiley.
White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–25.
Wurtz, D., Chalabi, Y., & Luksan, L. (2006). Parameter estimation of ARMA models with GARCH/APARCH errors an R and SPlus software implementation. *Journal of Statistical Software, 55*(2), 28–33.
Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 13*(08), 841–847.
Xie, Y. (2009). Consistency of maximum likelihood estimators for the regime-switching GARCH model. *Statistics, 43*(2), 153–165.