Hidden Markov Models with state-dependent mixtures: Minimal representation, model testing and applications to clustering

Hajo Holzmann and Florian Schwaiger

Fakultät für Mathematik und Informatik, Philipps-Universität Marburg, Germany

Finite-state hidden Markov models (HMMs), also called Markov-dependent finite mixtures, form a popular, frequently used model class for serially dependent observations with unobserved heterogeneity. We consider HMMs in which the state-dependent distributions are themselves finite mixtures. In such models, the parametrization is not unique, since components from the state-dependent mixtures may also be represented as states in the underlying Markov chain. We determine a unique (up to label switching) representation for the HMM in which the Markov chain has a minimal number of states. Further, we propose a likelihood-ratio test for the hypothesis that the number of states in the Markov chain can be reduced without changing the distribution of the time-series model.

Our method has important applications in cluster analysis and model selection. After highlighting the relevance of serial dependence for clustering, we propose two-step clustering algorithms. Starting with a BIC choice for a standard HMM (with simple state-dependent distributions), in the first step we determine the minimal representation of the HMM by testing, and in the second step we merge components in the resulting state-dependent finite mixtures by using either a local entropy or a modality-based criterion. The states in the resulting Markov chain, potentially split according to the remaining state-dependent components, are then interpreted as clusters. For model selection, we illustrate our method on a series of logarithmic returns of gold prices using normal HMMs. The AIC choice is a six-state HMM, while the BIC choice has four states. When starting with the AIC choice, successive testing results in a four-state Markov chain, with two state-dependent distributions consisting of two-component normal mixtures.

Keywords: hidden Markov models, merging states, clustering, likelihood-ratio test, financial log-returns

1 Introduction

A finite state hidden Markov model (HMM) is a bivariate process $(X_t, S_t)_{t\in\mathbb{N}}$, where $(S_t)_t$ is a unobservable finite state Markov chain with $k \in \mathbb{N}$ states, the observable process $(X_t)_t$ is independent given the Markov chain $(S_t)_t$ and the conditional distribution of each X_t depends on S_t only. Finite-state HMMs, also called Markov-dependent finite mixtures, form a popular, frequently used model class for serially dependent observations with unobserved heterogeneity, with areas of application such as speech recognition, modeling of financial time series or biological sequence analysis. For a comprehensive treatment of theoretical properties of HMMs see Cappé et al. (2005), Zucchini and MacDonald (2009) is a more basic introduction with applications and further references.

Typically, the state-dependent distributions of an HMM, that is, the conditional distributions of the X_t given the S_t, are assumed to belong to a standard parametric family such as the Poisson or the (multivariate) normal distribution. If these are not flexible enough, finite mixtures as state-dependent distributions may provide a more appropriate choice. Ajmera and Wooters (2003) used HMMs with normal mixtures as state-dependent distributions for speaker segmentation in the context of speech recognition. Geweke and Amisano (2011) analyzed such models in a Bayesian framework and gave an application to modeling S&P 500 log returns. Chiu et. al (2011) formulate the EM algorithm for HMMs with state-dependent finite normal mixtures, and use these to analyze epileptic seizure dynamics. Volant et. al (2013) propose a criterion for selecting the number of states in the Markov chain together with the number of components in each mixture, in particular for the purpose of cluster analysis, and also formulate the EM algorithm.

In this paper, we analyze the structure of HMMs with state-dependent finite mixtures in detail and give applications to clustering and model selection. On the methodological side, we show that the parametrization is not unique, since components from the state-dependent mixtures may also be represented as states in the underlying Markov chain. However, we show that there is a unique (up to label switching) representation for the HMM in which the underlying Markov chain has a minimal number of states. Moreover, we propose a likelihood-ratio test for the hypothesis that the number of states in the Markov chain can be reduced without changing the distribution of the HMM.

Our methodology has important applications in cluster analysis and model selection. After highlighting the relevance of serial dependence for clustering, we propose a two-step clustering algorithm. Starting with a BIC choice for a standard HMM (with simple state-dependent distributions), in the first step we determine the minimal representation of the HMM by a backward selection based on testing. Given the minimal representation, we can make certain that no states in the Markov chain are merged for which relevant dependence information is lost. Thus, in the second step we restrict ourselves to merging components in the resulting state-dependent finite mixtures, using either a local entropy or a modality-based criterion. Finally, the states in the resulting Markov chain, potentially split according to the remaining state-dependent components, are interpreted as clusters. For model selection, we illustrate our method on a series of logarithmic returns of gold prices using normal HMMs. The AIC choice is a six-state HMM, while the BIC choice has four states. When starting with the AIC choice, successive testing results in a four-state Markov chain, with two state-dependent distributions consisting of two-component normal mixtures.

The outline of the paper is as follows. Section 2 contains the theoretical methodology. Section 3 presents our clustering algorithm and compares it to that of Volant et al. (2013). Section 4 has some additional simulations on the levels of our proposed test, as well as on the performance of the backward selection. This is investigated both in a correctly specified setting, as well as in a misspecified setting where data are generated from a two-state skew-normal HMM, but ordinary normal HMMs are used in the analysis. Section 5 finally gives an application of the proposed methodology in the context of model selection to a series of logarithmic returns of daily gold prices. All proofs are deferred to the appendix. The supplementary material Holzmann and Schwaiger (2014) contains some further numerical results. An implementation of the proposed algorithm is provided in the R package `mergeHMM`, which can be downloaded from the homepage of the university of Marburg[1].

2 Methodology for HMMs with state-dependent mixtures

In this section we present our methodology. Section 2 analyzes Markov chains under restrictions on the dependence structure. This is used in Section 2.2 to determine the distinct representations of an HMM with state-dependent finite mixtures, and in particular to determine its unique (up to label switching) representation with minimal number of states. Finally, Section 2.3 develops a likelihood-ratio test for the hypothesis that states in the Markov-chain may be represented as mixture components.

2.1 Markov chains under dependence structure restrictions

We start by analyzing Markov chains under restrictions on the dependence structure. Let $(S_t)_t$ be a stationary k-state Markov chain with ergodic transition probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{i,j})_{i,j=1,\ldots,k}$ having the stationary distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$. In the following we always assume $\pi_j > 0$ for $j = 1, \ldots, k$.
For a (disjoint) partition $\mathcal{G} = \{G_1, \ldots, G_r\}$ of the state space (into non-empty sets), we let $G(j)$ be the function which maps a state $j \in \{1, \ldots, k\}$ onto its group, i.e. for $j \in G_l$ we have $G(j) = G_l$. If $P\big(S_{t-1} = i,\ S_t \in G(j)\big) > 0)$ we have the general formula

$$\gamma_{i,j} = P\big(S_t \in G(j) \big| S_{t-1} = i\big) \cdot P\big(S_t = j \big| S_{t-1} = i,\ S_t \in G(j)\big). \tag{1}$$

Define the reduced t.p.m. $\lambda_{\mathcal{G}}(\Gamma)$ by

$$\big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{i,j} = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big), \qquad i, j = 1, \ldots, k. \tag{2}$$

Proposition 1. *The matrix $\lambda_{\mathcal{G}}(\mathbf{\Gamma})$ is a t.p.m., and the following statements are equivalent.*

1. *We have*
$$\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}. \tag{3}$$

2. *For $i, j = 1, \ldots, k$ it holds*
$$P\big(S_t \in G(j) \big| S_{t-1} = i\big) = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \tag{4}$$

[1] http://www.unimarburg.de/fb12/stoch/research/rpackage

and if $P\big(S_t \in G(j), S_{t-1} = i\big) > 0$ *also*

$$P\big(S_t = j \big| S_t \in G(j), S_{t-1} = i\big) = P\big(S_t = j \big| S_t \in G(j)\big).$$

3. *There exists a t.p.m.* $(\nu_{l,m})_{l,m} \in \mathbb{R}^{r\times r}$ *and* $(p_1, \ldots, p_k) \in \mathbb{R}^k$*, with* $p_j \geq 0$*,* $\sum_{g\in G_l} p_g = 1$*,* $l = 1, \ldots, r$*, such that*

$$\gamma_{i,j} = \nu_{a(i),a(j)} \cdot p_j, \qquad i, j = 1, \ldots, k$$

where $a : \{1, \ldots, k\} \to \{1, \ldots, r\}$ *and* $a(g) = l :\Leftrightarrow g \in G_l$*.*

Note that condition 3. in particular implies that the rows with indices in the same element G_l of the partition are all equal.

In the Markov-chain literature, the notion of lumpabilty (cf. Kemeny and Snell 1960) refers to the possibility of merging states of a Markov chain w.r.t. a partition while retaining a Markovian dependence structure. Indeed, lumpability is equivalent to the condition (4), however, (3) is a stronger requirement as the additional conditions in the lemma show. For example, consider the state space $\{1, 2, 3\}$ with partition $\mathcal{G} = \big\{\{1, 2\}, \{3\}\big\}$, and the t.p.m.

$$\mathbf{\Gamma} = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/2 & 1/2 & 0 \end{pmatrix}, \qquad \text{then} \qquad \lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \begin{pmatrix} 3/8 & 3/8 & 1/4 \\ 3/8 & 3/8 & 1/4 \\ 1/2 & 1/2 & 0 \end{pmatrix},$$

so that (3) does not hold, however, $\mathbf{\Gamma}$ is lumpable w.r.t. $\mathcal{G}$.

Next, we show that there is a unique partition $\mathcal{G}^*_{\mathbf{\Gamma}}$ fulfilling (3) and having a minimal number of sets. Note that when $\mathcal{G}$ is a partition with r sets, a k-state Markov chain satisfying $\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ can be parametrized by $r^2 - 2 \cdot r + k$ parameters, thus, a partition with a minimal number of sets provides a parametrization of the t.p.m. with a minimal number of parameters.

Theorem 2. *There exists a unique partition* $\mathcal{G}^*_{\mathbf{\Gamma}}$ *of the state space, which has a minimal number of sets and fulfills* $\lambda_{\mathcal{G}^*_{\mathbf{\Gamma}}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$*.*

We call the partition $\mathcal{G}^*_{\mathbf{\Gamma}}$ the *independence partition* of the Markov chain $(S_t)_t$ or of the transition probability matrix $\mathbf{\Gamma}$.

When $\mathcal{G} = \{G_1, \ldots, G_r\}$ and $\mathcal{H} = \{H_1, \ldots H_q\}$ are two partitions of the state space $\{1, \ldots, k\}$ such that $r > q$ and each set $G_l \in \mathcal{G}$ is a subset of a certain set in $\mathcal{H}$, we call $\mathcal{G}$ a *refinement* of $\mathcal{H}$ or $\mathcal{H}$ a *coarsening* of $\mathcal{G}$. We remark that for any refinement $\mathcal{G}$ of the independence partition $\mathcal{G}^*_{\mathbf{\Gamma}}$, the restriction $\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ also holds. To show this one first uses the property that each set of $\mathcal{G}$ is a subset of one set in $\mathcal{G}^*_{\mathbf{\Gamma}}$, yielding equal rows in $\mathbf{\Gamma}$ for indices in the same set of $\mathcal{G}$. Then the statement follows with the same arguments as used at the end of the proof of Lemma 6 (in the appendix).

2.2 Representations of HMMs with state-dependent mixtures

Let $(X_t, S_t)_t$ be a stationary k-state HMM with state space $\{1, \ldots, k\}$, state dependent densities $f_j(x) = f_{X_t|S_t=j}(x)$, $j = 1, \ldots, k$, t.p.m. $\mathbf{\Gamma}$ and stationary distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$.

Definition 1 (Reducing states to mixture components in an HMM)**.** Let $\mathcal{G} = \{G_1, \ldots, G_r\}$ be a partition of $\{1, \ldots, k\}$. Call *reducing states to mixture components with respect to* $\mathcal{G}$ the mapping of the HMM $(X_t, S_t)_t$ onto the new HMM $(X_t^{(\mathcal{G})}, S_t^{(\mathcal{G})})_t$, the distribution of which is determined by the t.p.m. $\mathbf{\Gamma}^{(\mathcal{G})}$,

$$\big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l,m} := P\big(S_t \in G_m \big| S_{t-1} \in G_l\big), \qquad l, m = 1, \ldots, r$$

of the Markov chain $(S_t^{(\mathcal{G})})_t$ (on the state space $\{1, \ldots, r\}$), and the state-dependent densities

$$f_l^{(\mathcal{G})}(x) := f_{X_t^{(\mathcal{G})}|S_t^{(\mathcal{G})}=l}(x) := f_{X_t|S_t \in G_l}(x), \qquad x \in \mathbb{R}^d,\ l = 1, \ldots, r.$$

of the observable process $(X_t^{(\mathcal{G})})_t$. $\diamond$

The parameters of the reduced HMM are easily determined as follows. For $l, m = 1, \ldots, r$ we have that

$$\begin{aligned}
\big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l,m} &= P\big(S_t \in G_m \big| S_{t-1} \in G_l\big) = \sum_{g \in G_m} P\big(S_t = g \big| S_{t-1} \in G_l\big) \\
&= \sum_{g \in G_m} \sum_{h \in G_l} \left(\frac{P\big(S_t = h\big)}{P\big(S_t \in G_l\big)} \cdot P\big(S_t = g \big| S_{t-1} = h\big) \right) \\
&= \sum_{g \in G_m} \sum_{h \in G_l} \left(\frac{\pi_h}{\sum_{a \in G_l} \pi_a} \cdot \gamma_{h,g} \right)
\end{aligned}$$

and for $x \in \mathbb{R}^d$ that

$$f_l^{(\mathcal{G})}(x) = \sum_{g \in G_l} P(S_t = g | S_t \in G_l) \cdot f_g(x) = \sum_{g \in G_l} \frac{\pi_g}{\sum_{a \in G_l} \pi_a} f_g(x).$$

Thus, the state dependent distributions are indeed given by mixtures of the original state dependent distributions. We say that states in each element of the partition $\mathcal{G}$ are *reduced to mixture components*.

Theorem 3. *The distribution of the observable process* $\big(X_t^{(\mathcal{G})}\big)_t$ *after reducing states to mixture components w.r.t. the partition* $\mathcal{G}$ *is the same as that of an HMM with t.p.m.* $\lambda_{\mathcal{G}}(\mathbf{\Gamma})$ *(on the original state space* $\{1, \ldots, k\}$*) and state-dependent densities* $f_j(x)$*,* $j = 1, \ldots, k$*. In particular, if* $\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ *we have that* $(X_t)_t \overset{(d)}{=} \big(X_t^{(\mathcal{G})}\big)_t$*.*

Corollary 4. *Let* $(X_t, S_t)_t$ *be a stationary* k*-state HMM with t.p.m.* $\mathbf{\Gamma}$ *and state-dependent densities* f_j *belonging to a parametric family, i.e.* $f_j(x) = f(x; \theta(j))$*,* $\theta(j) \in M \subset \mathbb{R}^p$*. If* k *component-mixtures in this parametric family are identifiable, then the independence partition* $\mathcal{G}_{\mathbf{\Gamma}}^*$ *of* $\mathbf{\Gamma}$ *of the set* $\{1, \ldots, k\}$ *is the unique partition with minimal number of states for which we may reduce states within each member of the partition to mixture components, i.e. for which* $(X_t)_t \overset{(d)}{=} \big(X_t^{(\mathcal{G})}\big)_t$*.*

We call $\mathcal{G}_{\mathbf{\Gamma}}^*$ the independence partition of the HMM and the elements of the independence partition $\mathcal{G}_{\mathbf{\Gamma}}^* = \{G_1, \ldots, G_r\}$ of the HMM its *independence clusters*. The corollary follows from Theorems 2 and 3, since identifiability of k-component mixtures guarantees identifiability of the parameters of the HMM.

In the literature, a notion of lumpability w.r.t. a partition of the state space, similar to that of Markov chains has also been developed for HMMs, see White et al. (2000). As for Markov chains, the process with reduced state-space retains a possibly distinct HMM structure, which is in contrast to our notion.

2.3 Testing the validity of reducing states to mixture components

Suppose that the state-dependent densities $f_1(\cdot), \ldots, f_k(\cdot)$ belong to a parametric family, i.e. $f_j(x) = f(x; \theta(j))$, $\theta(j) \in M \subset \mathbb{R}^p$, $j = 1, \ldots, k$. We denote the complete parameter vector by $\boldsymbol{\eta} = \big((\gamma_{i,j})_{i,j=1,\ldots,k}, \theta(1), \ldots, \theta(k)\big) \in \Theta \subset \mathbb{R}^d$. Given a parameter vector $\boldsymbol{\eta}$ we denote its t.p.m. by $\mathbf{\Gamma}_{\boldsymbol{\eta}}$ and the associated stationary distribution by $\boldsymbol{\pi}_{\boldsymbol{\eta}}$, the state dependent parameters by $\theta_{\boldsymbol{\eta}}(j)$, and the log-likelihood function of the observable part by

$$L_T(\boldsymbol{\eta}) = \log\big(p_{\boldsymbol{\eta}}(X_1, \ldots, X_T)\big),$$

where $p_{\boldsymbol{\eta}}$ denotes the density function of $(X_1, \ldots, X_T)$ given parameter $\boldsymbol{\eta}$.

In the following we denote the true, unknown parameter by $\boldsymbol{\eta}_0$ and always assume $\boldsymbol{\pi}_{\boldsymbol{\eta}_0} > 0$. Giudici, Rydén and Vandekerkhove (2000) extend the asymptotic Chi-square distribution of the likelihood-ratio for i.i.d. models to hidden Markov models. We are interested in testing hypotheses on the dependence structure, i.e. whether the hidden Markov chain fulfills the restriction introduced in Section 2.1. Specifically, for a given partition $\mathcal{G} = \{G_1, \ldots, G_r\}$ of the state space consider

$$H_0 : \lambda_{\mathcal{G}}(\mathbf{\Gamma}_{\boldsymbol{\eta}}) = \mathbf{\Gamma}_{\boldsymbol{\eta}} \quad \text{versus} \quad H_1 : \lambda_{\mathcal{G}}(\mathbf{\Gamma}_{\boldsymbol{\eta}}) \neq \mathbf{\Gamma}_{\boldsymbol{\eta}},$$

or equivalently $H_0 : \boldsymbol{\eta} \in \Theta_{0,\mathcal{G}}$ versus $H_1 : \boldsymbol{\eta} \in \Theta \setminus \Theta_{0,\mathcal{G}}$ with

$$\Theta_{0,\mathcal{G}} = \{\boldsymbol{\eta} \in \Theta : \lambda_{\mathcal{G}}(\mathbf{\Gamma}_{\boldsymbol{\eta}}) = \mathbf{\Gamma}_{\boldsymbol{\eta}}\}.$$

An essential condition for the asymptotic chi-square distribution of the LRT is for the null parameter to be an interior point of the parameter space. In our context, we require $P(S_t \in G_l | S_{t-1} \in G_m) > 0$ for $1 \leq l, m \leq r$.

Theorem 5. *Assume the Markov chain $(S_t)_t$ to be ergodic, the MLE $\hat{\boldsymbol{\eta}}_T$ to be strongly consistent, Assumptions 1-3 in the Appendix to hold and the Fisher information $\mathcal{J}(\boldsymbol{\eta}_0)$ of the HMM to be nonsingular. If $\boldsymbol{\eta}_0 \in \Theta_{0,\mathcal{G}}$, $P_{\boldsymbol{\eta}_0}(S_t \in G_l | S_{t-1} \in G_m) > 0$ for $l, m = 1, \ldots, r$, and $\theta_{\eta_0}(i)$ lies in the interior of M, $i = 1, \ldots, k$, then*

$$2 \cdot \Big(\sup_{\boldsymbol{\eta} \in \Theta} L_T(\boldsymbol{\eta}) - \sup_{\boldsymbol{\eta} \in \Theta_{0,\mathcal{G}}} L_T(\boldsymbol{\eta}) \Big) \xrightarrow{d} \chi^2_{h(k,r)}, \ \text{as } T \to \infty,$$

with $h(k,r) = k^2 - 2k - r^2 + 2r$ and $\mathcal{G} = \{G_1, \ldots, G_r\}$.

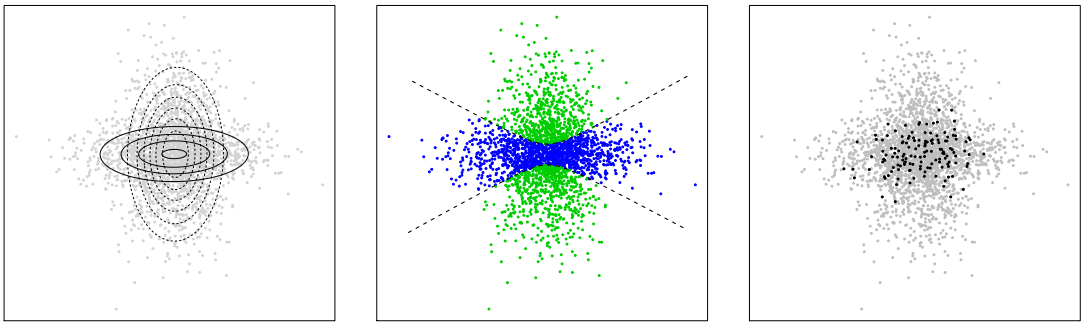
3 Clustering serially-dependent observations

3.1 Importance of dependence for state decoding

Let us start with a simulated example which shows the importance of serial dependence for clustering and state-estimation. We simulate a sequence of length $T = 2.500$ from a two-state HMM with state dependent bivariate normal distributions $f_{X_t|S_t=j}(x) = \varphi(x; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2$, where the parameters are chosen as

$$\boldsymbol{\mu}_1^{(1)} = \boldsymbol{\mu}_2^{(1)} = \begin{pmatrix} 0 & 0 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1^{(1)} = \begin{pmatrix} 10 & 0 \\ 0 & 1.5 \end{pmatrix}, \; \boldsymbol{\Sigma}_2^{(1)} = \begin{pmatrix} 3 & 0 \\ 0 & 11 \end{pmatrix}, \quad \boldsymbol{\Gamma} = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix}.$$

The stationary distribution of the observable part $(X_t)_t$ is the two-component mixture of normals with the above parameters and weight vector $\boldsymbol{\pi} = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$. Figure 1a illustrates contour lines of the state dependent densities and gives a scatter plot of the data.



(a) state dependent densities (b) i.i.d. clustering (c) misclassified HMM

Figure 1: Clustering a sample of a two-state hidden Markov model of bivariate normals.

First we fit an independent two-component normal mixture by ML and determine states by maximum-a-posteriori. Figure 1b visualizes the result, where data assigned to the first (resp. second) component are colored blue (resp. green). There is a hard border between the two clusters which is depicted by the dashed line, and while the majority of the points are correctly classified (647 of 2.500 observations are wrongly classified, which corresponds to 25.88% of the data), in the overlap between the two components no appropriate discrimination is possible.

In contrast, when using a serially-dependent HMM we can separate the two groups very well. When first estimating the parameters by ML and then performing global decoding using the Viterbi-algorithm, i.e. finding the sequence of states $s_1, \ldots, s_T$ which maximizes $P(S_1 = s_1, \ldots, S_T = s_T | X_1 = x_1, \ldots, X_T = x_T)$ only 140 observations (5.6%) are wrongly classified. Thus, even observations in the heavily overlapping area (around $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$) can be well separated. The intuitive reason is that each observation also helps to classify additional observations which are close in time, by the serial dependence.

3.2 Merging states in HMMs

For clustering independent data, there is a substantial literature on the appropriate choice of cluster shapes, see e.g. Hennig (2010) for a recent discussion. When using (independent) finite mixtures and a maximum-a-posteriori analysis directly, the component distributions need to correspond well to what cluster shapes are supposed to look like.

As an illustration, consider the example in Section 3.1. When fitting independent normal mixtures and choosing the number of components by the BIC, the result is the two-component fit as illustrated in Figure 1. However, without serial dependence, there are good reasons to assign all data to a single cluster, one reason being that the two-component normal mixture is actually a unimodal density.

As a solution, one can merge the components of an independent mixture, which correspond to a single cluster, into a single component. Baudry et al. (2010) propose an entropy-based criterion for selecting the candidate components for merging in independent finite mixtures. Hennig (2010) proposes further, in particular density-based (more specifically: modality-based) methods, and compares the distinct methods in a simulation study.

For dependent data modeled by HMMs, the situation is somewhat more involved. Scatter plots as in Figure 1 only refer to the marginal distribution, and miss the additional information provided by serial dependence. However, as Figure 1c illustrates, taking advantage of serial dependence, one can very well separate components which marginally form a (typically unimodal) scale mixture. An important class of data examples are time series of financial log-returns, where means are always close to zero, but where different scales correspond to different volatility phases of the market and must be discriminated. See Section 5 for an application.

Therefore, we propose merging and clustering algorithms for HMMs which focus on retaining the full dependence information. Only states within the same elements of the independence partition of the HMM are allowed to be merged.

First, we formulate a corresponding algorithm based on the local-decoding entropy of the HMM, similar in spirit to the proposal by Baudry et al. (2010) for independent mixtures.

Input: The observed series $x_1, \ldots, x_T$ and the parametric family of the state dependent densities $f(\cdot, \theta)$.

Step 1 Select and fit an appropriate finite-state HMM with state dependent densities from $f(\cdot, \theta)$, e.g. by using the BIC (or possibly the AIC). Denote the number of states of the selected HMM by k.

Step 2 Determine the independence partition $\mathcal{G}^* = \{G_1, \ldots, G_r\}$ of the selected HMM with k states using a backward selection based on the p-values of the test in Theorem 5, according to a certain significance level (say 0.05 or 0.1). Details for the backward selection algorithm are given in Section 4.2.

We let $\hat{\Gamma}$ and $\hat{\theta}_1, \ldots, \hat{\theta}_k$ denote the parameters of the ML-fit under the independence restrictions given by $\mathcal{G}^*$, so that $\hat{\Gamma}$ is a $k \times k$-t.p.m. for which $\lambda_{\mathcal{G}^*}(\hat{\Gamma}) = \hat{\Gamma}$, and we let $(X_t, S_t)_t$ denote a k-state HMM with these parameters.

Step 3 Initialize $\mathcal{H}_0 = \{\{1\}, \ldots, \{k\}\}$, $i = 0$. Compute the local decoding entropy $LDE(0)$ of the HMM $(X_t^{(\mathcal{H}_0)}, S_t^{(\mathcal{H}_0)})_t = (X_t, S_t)_t$ via

$$LDE(0) := -\sum_{t=1}^{T}\sum_{j=1}^{k} \phi_{t,j}(\mathcal{H}_0) \cdot \log\big(\phi_{t,j}(\mathcal{H}_0)\big),$$
$$\phi_{t,j}(\mathcal{H}_0) := P\big(S_t^{(\mathcal{H}_0)} = j \big| X_1^{(\mathcal{H}_0)} = x_1, \ldots, X_T^{(\mathcal{H}_0)} = x_T\big),$$

for $j = 1, \ldots, k,\ t = 1, \ldots, T$.

Iteration If $i + 1 > k - r$, stop, otherwise

For each partition $\mathcal{H}$ which is a coarsening of $\mathcal{H}_i$ with one element less than $\mathcal{H}_i$, but a refinement of the independence partition $\mathcal{G}^*$, compute the local decoding entropy of the HMM $(X_t^{(\mathcal{H})}, S_t^{(\mathcal{H})})_t$

$$LDE(\mathcal{H}) := -\sum_{t=1}^{n}\sum_{j=1}^{k-(i+1)} \phi_{t,j}(\mathcal{H}) \cdot \log\big(\phi_{t,j}(\mathcal{H})\big),$$
$$\phi_{t,j}(\mathcal{H}) := P\big(S_t^{(\mathcal{H})} = j \big| X_1^{(\mathcal{H})} = x_1, \ldots, X_T^{(\mathcal{H})} = x_T\big),$$

for $j = 1, \ldots, k - (i + 1),\ t = 1, \ldots, T$. Choose $\mathcal{H}_{i+1} = \mathcal{H}$ for which $LDE(\mathcal{H}) =: LDE(i + 1)$ is minimal, and continue iteration with $i + 1$.

Choosing the clusters We obtain a nested sequence of partitions

$$\{\{1\}, \ldots, \{k\}\} = \mathcal{H}_0, \mathcal{H}_1, \ldots, \mathcal{H}_{k-r} = \mathcal{G}^*,$$

together with the local decoding entropies

$$LDE(0) \geq LDE(1) \geq \ldots \geq LDE(k - r),$$

and choose $0 \leq i^* \leq k - r$ appropriately, e.g. as an elbow in the entropy plot or such that the relative reduction in the entropy from step i^* to $i^* + 1$ falls below a certain threshold.

The elements in $\mathcal{H}_{i^*}$ correspond to clusters, while the states within each element of $\mathcal{H}_{i^*}$ are merged.

Next, we propose to replace Step 3 by using a modality-based method. Specifically, we use the unimodal ridgeline or the ridgeline ratio method in Hennig (2010). Underlying this method is the fact that all local extrema of Gaussian mixtures occur along the so-called ridgeline, see Ray and Lindsay (2005) or Hennig (2010). The ridgeline ratio for a two-component normal mixture is then defined as 1 if the mixture is unimodal, and as the ratio between the minimum of the density along the ridgeline between the exterior modes and the value of the density at the minimal mode in case of more than one mode.

Step 3 For the states in each element of the independence partition $\mathcal{G}^*$, we apply separately one of the ridgeline merging algorithms of Henning (2010), either the ridgeline unimodal method or the ridgeline ratio method. Thus, only merges within each element of the independence partition are allowed.

Volant et al. (2013) propose related hierarchical merging and clustering algorithms, based on three different model-selection based criteria (BIC and two versions of the ICL by Biernacki et al. 2000). Our method focuses more strongly on the dependence structure of the underlying Markov chain: Only states within the same element of the independence partition are allowed to be merged, as motivated in Section 3.1. In the next section we shall compare the numerical performance of the two methods. For future work, a version of the ICL could also be built into our algorithm.

3.3 Numerical Illustrations

We present two numerical illustrations of the above algorithms, which we implemented in the R package `mergeHMM`[2]. This package provides functions to perform the backward selection based on the LRT (see Section 4.2), to calculate iteratively local decoding entropies given an estimated model, to estimate HMMs via MLE under dependence-structure restrictions, to simulate datasets and to perform a maximum-a-posteriori analysis with the Viterbi algorithm.

1. Five-state normal HMM with two independence clusters

First, we consider the following five-state bivariate normal HMM.

$$\begin{gathered}\boldsymbol{\mu}_1 = \begin{pmatrix}2.5\\1.5\end{pmatrix}^T \quad \boldsymbol{\mu}_2 = \begin{pmatrix}3.5\\2\end{pmatrix}^T \quad \boldsymbol{\mu}_3 = \begin{pmatrix}2\\7\end{pmatrix}^T \quad \boldsymbol{\mu}_4 = \begin{pmatrix}3\\0.5\end{pmatrix}^T \quad \boldsymbol{\mu}_5 = \begin{pmatrix}2.5\\6\end{pmatrix}^T \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix}0.30 & 0.18\\0.18 & 0.30\end{pmatrix}\\ \boldsymbol{\Sigma}_2 = \begin{pmatrix}0.30 & -0.18\\-0.18 & 0.30\end{pmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix}0.48 & -0.42\\-0.42 & 0.48\end{pmatrix} \quad \boldsymbol{\Sigma}_4 = \begin{pmatrix}1.20 & 0.27\\0.27 & 1.20\end{pmatrix} \quad \boldsymbol{\Sigma}_5 = \begin{pmatrix}0.5 & 0.4\\0.4 & 0.5\end{pmatrix},\\ \boldsymbol{\Gamma} = \begin{pmatrix}25.50 & 17.00 & 42.50 & 10.00 & 5.00\\25.50 & 17.00 & 42.50 & 10.00 & 5.00\\25.50 & 17.00 & 42.50 & 10.00 & 5.00\\6.00 & 4.00 & 10.00 & 70.00 & 10.00\\4.50 & 3.00 & 7.50 & 10.00 & 75.00\end{pmatrix}\end{gathered} \tag{5}$$

Its independence partition is given by $\mathcal{G}^* = \big\{\{1,2,3\},\{4\},\{5\}\big\}$, but within $\{1,2,3\}$, only the densities of states $\{1,2\}$ overlap. See Figure 2 for the contour lines of the state-dependent densities. We generate a sequence of 1000 observations, for a (correctly specified) normal HMM the BIC indeed selects five states:

no. of states	2	3	4	5	6	7
BIC	6492.975	6000.316	5894.382	**5891.908**	5979.094	6073.929

the parameter estimates for five states are listed in the supplementary material. The backward selection then leads to the independence partition $\mathcal{G}^*$, as follows.

Step i	Max. P-value of $\lambda_{\mathcal{G}_i}(\boldsymbol{\Gamma}) = (\boldsymbol{\Gamma})$	Partition $\mathcal{G}_i$ with max. p-value
1	60.61%	$\big\{\{1,3\},\{2\},\{4\},\{5\}\big\}$
2	72.78%	$\big\{\{1,2,3\},\{4\},\{5\}\big\}$
3	$\leq 10^{-4}$	$\big\{\{1,2,3,4\},\{5\}\big\}$

[2] http://www.unimarburg.de/fb12/stoch/research/rpackage

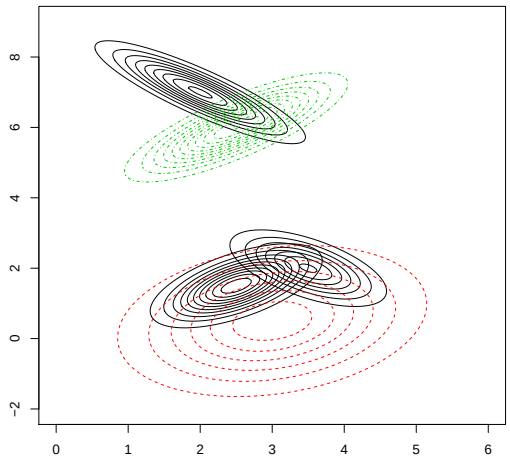


Figure 2: Contour lines of the state dependent bivariate normals. States one to three are depicted by (black) solid lines, state four by (red) dashed lines and state five by (green) dot-dashed lines.

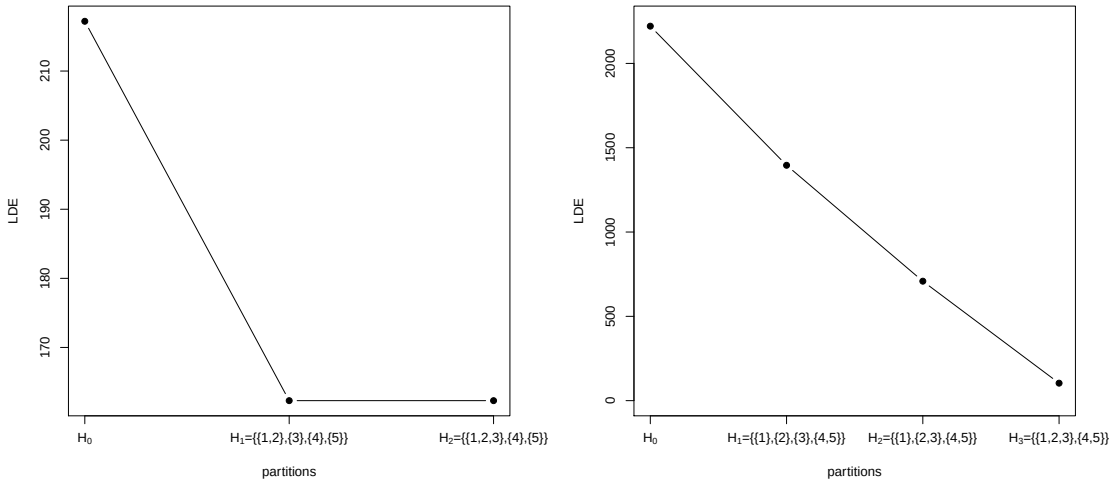
Under the independence restrictions implied by $\mathcal{G}^*$ we obtain the estimate

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 26.00 & 17.25 & 41.71 & 12.82 & 2.22 \\ 26.00 & 17.25 & 41.71 & 12.82 & 2.22 \\ 26.00 & 17.25 & 41.71 & 12.82 & 2.22 \\ 6.36 & 4.22 & 10.20 & 69.28 & 9.94 \\ 4.72 & 3.13 & 7.57 & 9.67 & 74.91 \end{pmatrix}$$

for $\mathbf{\Gamma}$, see the supplement for the remaining parameter estimates. First, the local decoding entropies, together with the corresponding partitions, are plotted in Figure 3a. There is a distinctive elbow after the first merge, so that the four elements of the partition $\mathcal{H}_{1^*} = \{\{1,2\},\{3\},\{4\},\{5\}\}$ correspond to the clusters, and only states 1 and 2 are merged. The two-component normal mixture formed from states 1 and 2 (the weights being $(0.255, 0.17)/(0.255 + 0.17)$) is actually slightly bimodal, but with a ridgeline-ratio (see Hennig 2010) of 0.95 very close to 1. See Figure 4 for a ridgline plot, i.e. the value of the density along the parameter α which parametrizes the ridgeline.

The algorithm of Volant et al. (2013) merges all states except state 5 into a single cluster, see the supplementary material Holzmann and Schwaiger (2014) for the numerical output. This is contrary to our clustering philosophy in to respects: First, state 4 does not belong to the same element of the independence partition as states 1-3, second, state 3 can be well discriminated from states 1,2,4 based on its marginal density.

2. Two-state skew-normal HMM



(a) normal (b) skew-normal

Figure 3: Local decoding entropies of estimated HMMs(a): series according to five-state normal HMM, (b) series according to two-state skew-normal HMM, local decoding entropies based on fitted five-state normal HMM.

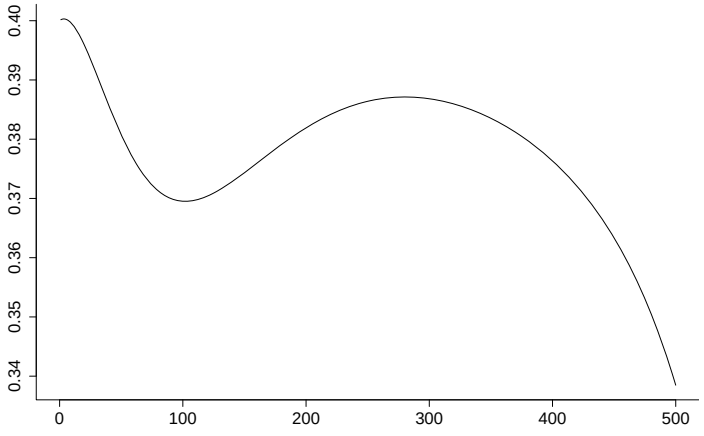


Figure 4: Plot of density along parameter of the ridgline for first two states of five-state normal HMM

Second, we consider the following two-state bivariate skew-normal HMM :

$$\begin{aligned}
&\boldsymbol{\Sigma}_1 = \begin{pmatrix} 4.80 & -0.48 \\ -0.48 & 1.20 \end{pmatrix},\ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 4.0 & -0.4 \\ -0.4 & 1.0 \end{pmatrix}, \qquad \boldsymbol{\Gamma} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \\
&\boldsymbol{\alpha}_1 = \begin{pmatrix} 14 & -6 \end{pmatrix},\ \boldsymbol{\alpha}_2 = \begin{pmatrix} 14 & 0 \end{pmatrix}, \\
&\boldsymbol{\mu}_1 = \begin{pmatrix} -5.0 & 3.3 \end{pmatrix},\ \boldsymbol{\mu}_2 = \begin{pmatrix} -1.5 & 6.0 \end{pmatrix}.
\end{aligned} \tag{6}$$

Specifically, the two-dimensional skew-normal density is given by

$$2\varphi(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \Phi_1(\boldsymbol{\alpha}\boldsymbol{\omega}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})^T),$$

where $\Phi_1(\cdot)$ is the distribution function of univariate standard-normal, and

$$\boldsymbol{\omega}^{-1} = \operatorname{diag}\left(\boldsymbol{\Sigma}_{11}^{-1/2}, \boldsymbol{\Sigma}_{22}^{-1/2}\right).$$

We consider a series of length 5000, and fit a (misspecified) normal HMM. In order to fit strongly skewed state-dependent densities, the BIC selects 5 states, the first three corresponding to the first component, the other two to the second component:

no. of states	2	3	4	5	6	7
BIC	34274.92	33814.44	33364.45	**33324.97**	33330.39	33419.65

The estimated five-state transition matrix in the normal HMM is

$$\hat{\boldsymbol{\Gamma}}_{Nor} = \begin{pmatrix} 25.92 & 39.45 & 23.92 & 7.37 & 3.35 \\ 23.35 & 44.16 & 21.30 & 2.43 & 8.75 \\ 20.72 & 44.38 & 25.37 & 4.12 & 5.41 \\ 2.89 & 4.32 & 2.50 & 33.76 & 56.53 \\ 3.75 & 2.85 & 3.45 & 36.20 & 53.75 \end{pmatrix},$$

which has approximate independence restrictions. If we apply the backward selection procedure in this misspecified situation, we obtain $\mathcal{G}^* = \big\{\{1,2,3\},\{4,5\}\big\}$ as independence partition:

Step i	Max. P-value of $\lambda_{\mathcal{G}_i}(\boldsymbol{\Gamma}) = (\boldsymbol{\Gamma})$	Partition $\mathcal{G}_i$ with max. p-value
1	62.99%	$\big\{\{1,3\},\{2\},\{4\},\{5\}\big\}$
2	22.45%	$\big\{\{1,3\},\{2\},\{4,5\}\big\}$
3	11.94%	$\big\{\{1,2,3\},\{4,5\}\big\}$
4	$\leq 10^{-4}$	$\big\{\{1,2,3,4,5\}\big\}$

Under the independence restrictions implied by $\mathcal{G}^*$, the fitted transition matrix is given by

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 89.37 \cdot \begin{pmatrix} 0.26 & 0.47 & 0.27 \\ 0.26 & 0.47 & 0.27 \\ 0.26 & 0.47 & 0.27 \end{pmatrix} & 10.63 \cdot \begin{pmatrix} 0.39 & 0.61 \\ 0.39 & 0.61 \\ 0.39 & 0.61 \end{pmatrix} \\ 09.92 \cdot \begin{pmatrix} 0.26 & 0.47 & 0.27 \\ 0.26 & 0.47 & 0.27 \end{pmatrix} & 90.08 \cdot \begin{pmatrix} 0.39 & 0.61 \\ 0.39 & 0.61 \end{pmatrix} \end{pmatrix},$$

and fitted values for the state-dependent parameters are listed in the supplementary material, see Figures 5a and 5b for contour plots of the true densities and the fitted normal state-dependent densities.

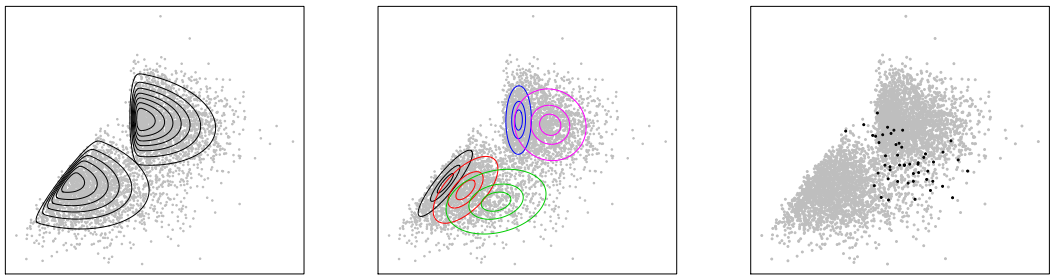
When applying our merging algorithms, we obtain the LDEs with corresponding partitions as plotted in Figure 3b. There is no elbow, so that we ought to perform all possible merges, leading to $\mathcal{H}_{3^*} = \mathcal{G}^*$, the elements of which correspond to the two clusters.

Next, consider the modality-based merging methods. First, consider the states 4 and 5. The resulting two-component normal mixture is bimodal, with a ridgeline ratio of 0.82. For the states $\{1, 2, 3\}$, the two-component mixture of states 2 and 3 (with weights $\pi = 0.27/(0.27 + 0.47)$ for states 2 and $1 - \pi$ for state 3) is unimodal (the others being bimodal), so that these two states are merged in the first step. Next, we form the mean vector and the covariance matrix of the resulting two-component normal mixture of states 2 and 3, which are given by (vectors are taken as column vectors)

$$\mu_{mix} = \pi\,\mu_2 + (1-\pi)\,\mu_3, \qquad \Sigma_{mix} = \pi\big(\Sigma_2 + \mu_2^T\,\mu_2\big) + (1-\pi)\,\big(\Sigma_3 + \mu_3^T\,\mu_3\big) - \mu_{mix}^T\mu_{mix}.$$

Then the ridgeline is formed for the parameters (μ_1, Σ_1), $(\mu_{mix}, \Sigma_{mix})$ and weights $(0.26, 0.74)$, yielding a unimodal density. Thus, the ridgeline ratio method with tuning parameter 0.8 yields two clusters.

The misclassifications from a clustering using global decoding is plotted in Figure 5c.



(a) true densities (b) normal densities (c) merged normal HMM

Figure 5: State dependent densities of (a) true skew-normal HMM and (b) estimated normal HMM; (c) wrongly estimated states using the Viterbi algorithm and the estimated, merged normal HMM (49 of 5.000 observations). The states of the normal fit are ordered ascending by mean of the x-coordinate.

In this example, the algorithm of Volant et al. (2013) gives the same result with two remaining clusters, see the supplementary material. Neglecting dependence, we also applied the algorithm of Baudry et al. (2010). It also selects two clusters, but the resulting clustering is slightly different, since in the independent case there is a hard threshold line between the clusters, these cannot "overlap". The maximum a-posteriori clustering of the mixture-model classified 92 of 5.000 observations incorrectly, in comparison to 49 for the merged HMM. Thus, both methods perform well, but as expected the merged HMM can recover more observations which lie on the boundary of the two clusters. See the supplement for the numerical results.

Further simulation results in the above two settings are presented in Section 4.

4 Further simulation results

4.1 Simulated sizes

We simulate the levels of the likelihood-ratio test for the five-state normal HMM with three independence clusters as specified in (5). The regularity conditions of Theorem 5 are satisfied if we impose lower bounds on the determinants of the state-dependent covariance matrices.

For the four partitions $\mathcal{G}_1 = \big\{\{1,2\},\{3\},\{4\},\{5\}\big\}$, $\mathcal{G}_2 = \big\{\{1\},\{2,3\},\{4\},\{5\}\big\}$, $\mathcal{G}_3 = \big\{\{1,3\},\{2\},\{4\},\{5\}\big\}$ and $\mathcal{G}_4 = \big\{\{1,2,3\},\{4\},\{5\}\big\}$ for which $\lambda_{\mathcal{G}_i}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ is satisfied, we simulate the levels of the corresponding tests for three different sample sizes ($T = 1.000, 2.500, 5.000$), with $M = 5.000$ simulations each. The sizes corresponding to asymptotic levels of $\alpha = 10\%, 5\%, 1\%$ are listed in Table 1. The tests are somewhat anti-conservative for the smaller sample sizes, but quite accurate for higher ones. Note that states 2 and 3 are much better separated than states 1 and 2, which also leads to somewhat more accurate levels of the test. The simulations were conducted on the MaRC2 supercomputer of the university of Marburg, and their duration was a few days.

level/T	1.000	2.500	5.000
10%	16.62	12.84	11.80
5%	9.22	7.36	6.20
1%	2.20	1.50	1.68

(a) $\mathcal{G}_1 = \big\{\{1,2\},\{3\},\{4\},\{5\}\big\}$

level/T	1.000	2.500	5.000
10%	15.36	10.92	10.76
5%	8.36	5.94	5.60
1%	2.18	1.50	1.32

(b) $\mathcal{G}_2 = \big\{\{1\},\{2,3\},\{4\},\{5\}\big\}$

level/T	1.000	2.500	5.000
10%	14.96	12.30	10.72
5%	7.52	6.42	5.68
1%	1.66	1.52	1.24

(c) $\mathcal{G}_3 = \big\{\{1,3\},\{2\},\{4\},\{5\}\big\}$

level/T	1.000	2.500	5.000
10%	16.30	12.46	11.32
5%	9.30	6.90	5.80
1%	2.38	1.44	1.16

(d) $\mathcal{G}_4 = \big\{\{1,2,3\},\{4\},\{5\}\big\}$

Table 1: Simulated rejection rates in percent (series lengths 1.000, 2.500 and 5.000) for accessing finite sample behavior of $\chi^2_\cdot$ - approximation in case of a normal HMM, row-wise to levels $10\%, 5\%$ and 1%. (a), (b), (c) χ^2_7 - approximation, (d) χ^2_{12} - approximation.

4.2 Backward selection

We start by spelling out the backward selection algorithm for determining the independence partition based on Theorem 5 in detail.

Input: The observed series $x_1, \ldots, x_T$ and the parametric family of the state dependent densities $f(\cdot, \theta)$, and the test level $\alpha > 0$.

Step 1 Select and fit an appropriate finite-state HMM with state dependent densities from $f(\cdot, \theta)$, e.g. by using the BIC (or possibly the AIC). Denote the number of states of the selected HMM by k.

Step 2 Initialize $\mathcal{G}_0 = \{\{1\}, \ldots, \{k\}\}$, $i = 1$.

Iteration For each partition $\mathcal{G}$ which is a coarsening of $\mathcal{G}_{i-1}$ with one element less than $\mathcal{G}_{i-1}$, compute the p-value of the likelihood-ratio test of $H : \lambda_{\mathcal{G}}(\Gamma) = \Gamma$ based on the asymptotic χ^2-distribution with $2i(k-1) - i^2$ degrees of freedom.

If the maximal p-value of these tests is $< \alpha$, we set $\mathcal{G}^* = \mathcal{G}_{i-1}$ and stop.

Otherwise we choose the partition $\tilde{\mathcal{G}}$ with maximal p-value. If $\tilde{\mathcal{G}} = \{\{1, \ldots, k\}\}$ is the trivial partition (in step $i = k - 1$), we let $\mathcal{G}^* = \{\{1, \ldots, k\}\}$ and stop,

otherwise we let $\mathcal{G}_i = \tilde{\mathcal{G}}$ and continue the iteration with $i + 1$.

We continue by simulating the performance of the backward selection in two examples.

Five-state normal HMM with two independence clusters

We apply the backward selection algorithm to the five-state normal HMM with three independence clusters as specified in (5), where we always start with a HMM with five states. The results are given in Table 2, corresponding simulation results for series of length 500 are provided in the supplement. The backward selection most often selects the independence partition with three elements. Since no partition with less states is selected, the power of the test for the given t.p.m. is quite high. The situation is changed if we consider instead the t.p.m.

$$\mathbf{\Gamma} = \begin{pmatrix} 3.00 & 2.00 & 5.00 & 85.00 & 5.00 \\ 3.00 & 2.00 & 5.00 & 85.00 & 5.00 \\ 3.00 & 2.00 & 5.00 & 85.00 & 5.00 \\ 6.00 & 4.00 & 10.00 & 70.00 & 10.00 \\ 4.50 & 3.00 & 7.50 & 10.00 & 75.00 \end{pmatrix}$$

Here, states 1-3 and 4 are sometimes merged, in particular for the shorter series, see the supplement for the simulation results.

length T	1.000			2.500			5.000		
level/ind. clusters	3	4	5	3	4	5	3	4	5
10%	830	127	43	881	101	18	448	44	8
5%	905	79	16	934	59	7	471	23	6
1%	977	21	2	991	8	1	496	4	-

Table 2: Simulation results of backward selection under a normal HMM: Absolute frequency of number of elements in independence partition according to used level. $M = 1.000$ repetitions for lengths of $T = 1.000$ and 2.500; $M = 500$ repetitions for lenght $T = 5.000$.

Two-state skew-normal HMM

Finally, we apply the backward selection algorithm in the misspecified situation where we simulate series from the two-state skew-normal HMM in (6), but fit normal HMMs. We generate $M = 1.000$ repetitions for lengths $T = 1.000$ and 2.500, as well as $M = 500$ repetitions for length $T = 5.000$.

We start with a BIC-choice for the number of states, the results are as follows.

length / states	3	4	5	6
1.000	138	862	0	0
2.500	0	815	185	0
5.000	0	2	158	340

	length T	1.000			2.500				5.000				
BIC choice	level/ind. clusters	2	3	4	2	3	4	5	2	3	4	5	6
3 states	10%	128	10	-	-	-	-	-	-	-	-	-	-
	5%	132	6	-	-	-	-	-	-	-	-	-	-
	1%	136	2	-	-	-	-	-	-	-	-	-	-
4 states	10%	730	76	56	702	71	42	-	2	-	-	-	-
	5%	788	51	23	755	43	17	-	2	-	-	-	-
	1%	844	14	4	809	5	1	-	2	-	-	-	-
5 states	10%	-	-	-	153	16	15	1	136	16	6	-	-
	5%	-	-	-	162	13	10	-	148	7	3	-	-
	1%	-	-	-	176	5	4	-	155	2	1	-	-
6 states	10%	-	-	-	-	-	-	-	280	34	22	2	2
	5%	-	-	-	-	-	-	-	311	20	8	-	1
	1%	-	-	-	-	-	-	-	344	4	1	1	-

Table 3: Simulation results of the backward selection of normal HMMs under a true skew-normal HMM: Absolute frequency of number of elements in independence partition according to used level.

Finally, the results of the backward selection for the number of states in the independence partition, split according to the initial BIC-choice, are given in Table 3.

A two-element independence partition is chosen most often in all settings.

5 Model selection: An application to logarithmic returns of daily gold prices

We conclude with an application which illustrates how our methodology can be used for model selection and fine-tuning.

We consider a series of logarithmic returns of the daily gold prices in London in U.S. dollar from September 2nd 1997 until August 31st 2012. When fitting normal HMMs, the AIC selects six states, while the BIC selects only four:

no. of states	2	3	4	5	6	7	8
AIC	11250	11125	11060	11035	**11023**	11034	11052
BIC	11294	11213	**11204**	11248	11318	11423	11547

We therefore start with the six-state HMM, for which we obtain the estimates

$$\tilde{\boldsymbol{\mu}} = \begin{pmatrix} -0.252 & 0.821 & -0.202 & 0.119 & 0.018 & 0.008 \end{pmatrix}$$
$$\tilde{\boldsymbol{\sigma}} = \begin{pmatrix} 1.454 & 0.652 & 0.561 & 2.281 & 0.796 & 0.293 \end{pmatrix}$$
$$\tilde{\boldsymbol{\Gamma}} = \begin{pmatrix} 47.75 & 28.74 & 23.42 & 0.08 & 0.01 & 0.00 \\ 11.86 & 15.53 & 71.96 & 0.65 & 0.00 & 0.00 \\ 54.48 & 33.26 & 8.01 & 0.00 & 4.26 & 0.00 \\ 2.53 & 0.00 & 0.00 & 97.47 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.24 & 60.31 & 39.45 \\ 3.22 & 1.05 & 0.00 & 0.18 & 69.36 & 26.19 \end{pmatrix}.$$

Next we apply the backward-selection algorithm to find the independence partition of the six-state HMM, which yields

Step i	Max. P-value of $\lambda_{\mathcal{G}_i}(\boldsymbol{\Gamma}) = (\boldsymbol{\Gamma})$	Partition $\mathcal{G}_i$ with max. p-value
1	94.04%	$\{\{1\},\{2\},\{3\},\{4\},\{5,6\}\}$
2	45.80%	$\{\{1,2\},\{3\},\{4\},\{5,6\}\}$
3	0.47%	$\{\{1,2,3\},\{4\},\{5,6\}\}$

giving $\mathcal{G}^* = \{\{1,2\},\{3\},\{4\},\{5,6\}\}$ as independence partition. Here, in the second step of the algorithm all except two of the eight p-values are $\leq 10^{-3}$, and the second highest being 2.63% for the partition $\{\{1,3\},\{2\},\{4\},\{5,6\}\}$, which is much lower than the $45,80\%$ for the selected partition $\{\{1,2\},\{3\},\{4\},\{5,6\}\}$. Although transitions from states 1 and 3 look similar, a heuristic reason might be that states 4 and 5 can only be reached from state 3, but not from states 1 (and 2). Therefore, state 3 does not belong to the same element of the independence partition as state 1.

When estimating under the independence restrictions implied by $\mathcal{G}^*$, we obtain

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} -0.259 & 0.717 & -0.290 & 0.072 & 0.026 & 0.006 \end{pmatrix}$$
$$\hat{\boldsymbol{\sigma}} = \begin{pmatrix} 1.531 & 0.695 & 0.626 & 2.244 & 0.805 & 0.296 \end{pmatrix}$$

and

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 51.04 \cdot \begin{pmatrix} 0.5043 & 0.4957 \\ 0.5043 & 0.4957 \end{pmatrix} & 48.58 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} & 0.37 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} & 0.01 \cdot \begin{pmatrix} 0.661 & 0.339 \\ 0.661 & 0.339 \end{pmatrix} \\ 95.64 \cdot \begin{pmatrix} 0.5043 & 0.4957 \end{pmatrix} & 0.00 & 0.00 & 4.36 \cdot \begin{pmatrix} 0.661 & 0.339 \end{pmatrix} \\ 2.57 \cdot \begin{pmatrix} 0.5043 & 0.4957 \end{pmatrix} & 0.00 & 97.43 & 0.00 \cdot \begin{pmatrix} 0.661 & 0.339 \end{pmatrix} \\ 1.42 \cdot \begin{pmatrix} 0.5043 & 0.4957 \\ 0.5043 & 0.4957 \end{pmatrix} & 0.00 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} & 0.22 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} & 98.36 \cdot \begin{pmatrix} 0.661 & 0.339 \\ 0.661 & 0.339 \end{pmatrix} \end{pmatrix}$$

The local decoding entropy for the initial model is given by 3300.245, after the first merging step (states 5 and 6) by 2521.548, and after the second step by 1896.366.

Therefore, also for clustering purposes it is reasonable to consider the reduced representation with four states, t.p.m. $\lambda_{\mathcal{G}^*}(\hat{\mathbf{\Gamma}})$, and state-dependent densities

$$
\begin{aligned}
f_1(x) &= \boldsymbol{p}_1^{(1)} \cdot \varphi(x; \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\sigma}}_1) + \boldsymbol{p}_2^{(1)} \cdot \varphi(x; \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\sigma}}_2), \quad f_2(x) = \varphi(x; \hat{\boldsymbol{\mu}}_3, \hat{\boldsymbol{\sigma}}_3) \\
f_4(x) &= \boldsymbol{p}_1^{(2)} \cdot \varphi(x; \hat{\boldsymbol{\mu}}_5, \hat{\boldsymbol{\sigma}}_5) + \boldsymbol{p}_2^{(2)} \cdot \varphi(x; \hat{\boldsymbol{\mu}}_6, \hat{\boldsymbol{\sigma}}_6), \quad f_3(x) = \varphi(x; \hat{\boldsymbol{\mu}}_4, \hat{\boldsymbol{\sigma}}_4) \\
\boldsymbol{p}^{(1)} &= \big(0.5043; 0.4957\big)^T, \qquad \boldsymbol{p}^{(2)} = \big(0.661; 0.339\big)^T.
\end{aligned}
$$

Figure 6 illustrates the estimated state dependent distributions.

Let us briefly describe and comment on the resulting four-state model.

State 1 has a positive-mean, comparatively high volatility and is left skewed. State 2 has a negative mean and small volatility, these two states form a kind of cycle, out of which transition is (almost) only possible from the State 2 to 4. State 4, which arises as a scale mixture of two normals, has mean almost $= 0$ but a heavier tail than an ordinary normal distribution. This state is highly persistent, which implies that we observe long periods of moderate growth corresponding to that state.

State 3 (green) can be interpreted as a turbulent or crisis state since its state-dependent distribution is a scale mixture with a very high variance and almost zero mean.

Using the merged model and the Viterbi algorithm we estimated the most likely series of states, see Figure 7 for visualizations of the results.

State 3 (crisis) occurs e.g. from October 26th 2007 until April 6th 2009, a time period containing the financial crisis, in 2011, which contained the US debt ceiling crisis, in March 2003 at the beginning of the Iraq war and at the time following the September 11, 2001 attacks. In contrast, state 4 occurs over the long stable periods of moderate growth.

The volatility in the cycle consisting of states 1 and is between that of states 4 and 3, it can therefore be interpreted as an intermediate phase between moderate growth (state 4) and crisis (state 3).

Acknowledgements

The authors gratefully acknowledge financial support from the DFG, grants Ho 3260/3-1 and Ho 3260/3-2. Further, we would like to thank the associate editor as well as two anonymous referees for helpful comments and for pointing out several relevant references.

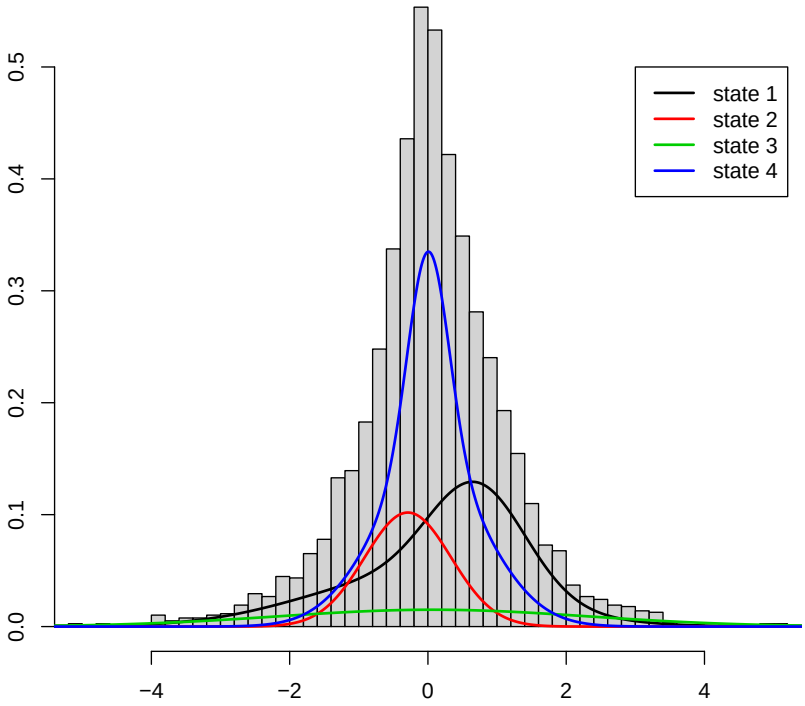
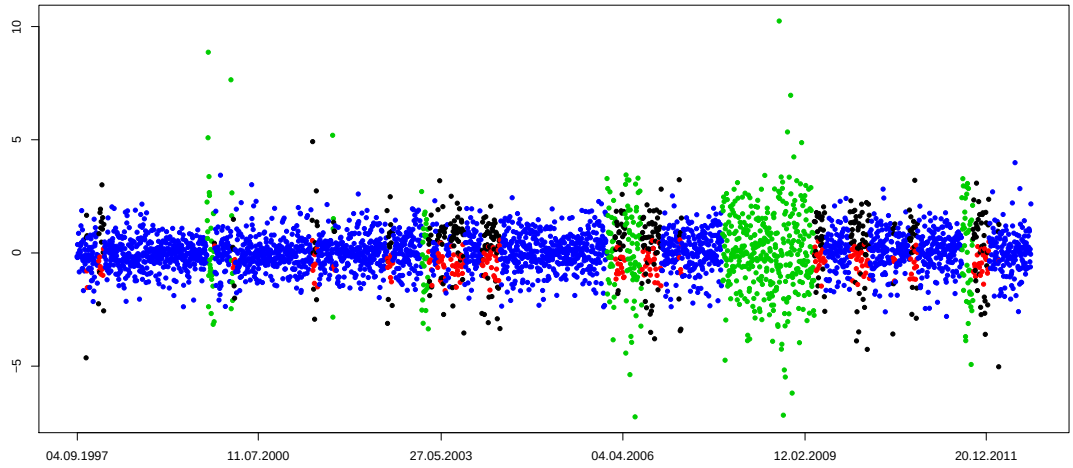
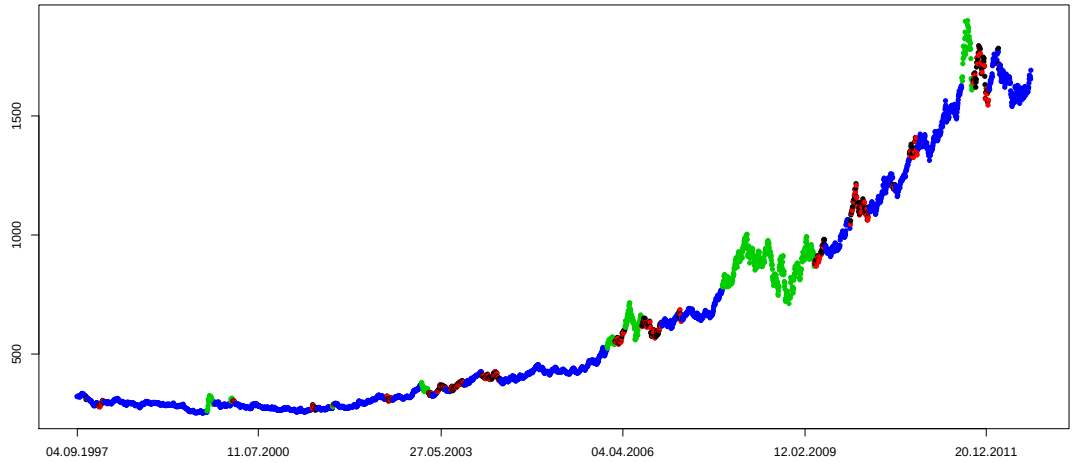


Figure 6: Histogram of logarithmic returns of gold prices in percent (September 2nd 1997 until August 31st 2012) and estimated state dependent densities of the merged four-state hidden Markov model (densities of states one to four are colored in black, red, green and blue).



(a) log-returns of gold prices



(b) gold prices

Figure 7: Viterbi Clustering of gold prices and log-returns. Coloring is chosen as in Figure 6

References

Ajmera, J. and Wooters, C. (2003). *A robust speaker clustering algorithm.* In: Automatic Speech Recognition and Understanding, 411-416.

Baudry, J.-P., Raftery, A.E., Celeux, G., Lo, K. and Gottardo, R. (2010). *Combining Mixture Components for Clustering*, Journal of Computational and Graphical Statistics, 19, 332-353.

Biernacki, C., Celeux, G. and Govaert, G. (2000). *Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 719-725.

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in hidden Markov models.* New York: Springer.

Chiu, A. WL., Derchansky, M., Cotic, M., Carlen, P. L., O Turner, S. and Bardakjian, B. L. (2011). *Wavelet-based Gaussian-mixture hidden Markov model for the detection of multistage seizure dynamics: A proof-of-concept study*, BioMedical Engineering OnLine 10-29.

Geweke, J. and Amisano, G. (2011). *Hierarchical Markov normal mixture models with applications to financial asset returns*, Journal of Applied Econometrics, 26, 1-29.

Giudici, P., Rydén, T., and Vandekerkhove P. (2000). *Likelihood-Ratio Tests for Hidden Markov Models*, Biometrics, 56, 742-747.

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, 4, 3-34.

Holzmann, H. and Schwaiger, F. (2014). Supplementary Material to:
Hidden Markov Models with state-dependent mixtures: Minimal representation, model testing and applications to clustering. Working paper, Marburg University.

Leroux, B.G. (1992). *Maximum-likelihood estimation for hidden Markov models*, Stochastic Processes and their Applications, 40, 127-143.

Kemeny, J. G. and Snell, J. L. (1960). *Finite Markov Chains.* Springer, New York.

Volant S., Bérard, C., Martin-Magniette, M.-L. and Robin, S. (2013). *Hidden Markov Models with mixtures as emission distributions*, Statistics and Computing. DOI 10.1007/s11222-013-9383-7.

White, L. B., Mahony, R. and Brushe, G. D. (2000). Lumpable Hidden Markov Models: Model Reduction and Reduced Complexity Filtering. *IEEE Transactions on Automatic Control*, 45, 2297-2306.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*, London: Chapman & Hall.

Appendix: Proofs

Proof of Proposition 1. $\lambda_{\mathcal{G}}(\mathbf{\Gamma})$ is a t.p.m. since

$$\sum_{j=1}^{k} \big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{i,j} = \sum_{l=1}^{r} \sum_{g \in G_l} P\big(S_t \in G_l \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = g \big| S_t \in G_l\big) = 1.$$

In order to validate that 1. implies 2., note that

$$\gamma_{i,j} = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big), \qquad i,j = 1,\ldots,k,$$

implies $\gamma_{i,j} = \gamma_{h,j}$ for all $h \in G(i)$, i.e. that under 1. $\mathbf{\Gamma}$ has equal rows with indices in the same set of the partition $\mathcal{G}$. Thus,

$$P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) = \sum_{g \in G(j)} \sum_{h \in G(i)} \left(\frac{P\big(S_{t-1} = h\big)}{P\big(S_{t-1} \in G(i)\big)} \cdot \gamma_{h,g} \right)$$

$$= \sum_{g \in G(j)} \gamma_{i,g} \sum_{h \in G(i)} \left(\frac{P\big(S_{t-1} = h\big)}{P\big(S_{t-1} \in G(i)\big)} \right) = \sum_{g \in G(j)} \gamma_{i,g} = P\big(S_t \in G(j) \big| S_{t-1} = i\big),$$

which gives the first claim of 2. If further $P\big(S_t \in G(j), S_{t-1} = i\big) > 0$ then $\gamma_{i,g} > 0$ for at least one $g \in G(j)$ and thus $P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) > 0$. Further, due to 1. and (1)

$$\begin{aligned} &P\big(S_t \in G(j) \big| S_{t-1} = i\big) \cdot P\big(S_t = j \big| S_{t-1} = i,\ S_t \in G(j)\big) \\ &= P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big) \end{aligned}$$

and hence also the second claim of 2. follows.

Now, assume 2. to hold, then for $l, m = 1, \ldots, r$ set $\nu_{l,m} = P(S_t \in G_m | S_{t-1} \in G_l)$ and $p_j = P(S_t = j | S_t \in G(j))$. Note $a(\cdot)$ is constant on each group of the partition, $\boldsymbol{N} = (\nu_{l,m})_{l,m} \in \mathbb{R}^{r \times r}$ defines a t.p.m. and $\boldsymbol{p} = (p_1, \ldots, p_k)$ is normalized as in 3. At first, if $P\big(S_{t-1} = i,\ S_t \in G(j)\big) = 0$ we have $\gamma_{i,j} = 0$ and also $P\big(S_t \in G(j) \big| S_{t-1} = i\big) = 0$. Thus, due to 2. also $P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) = 0$ and hence 3. holds in this case. If otherwise $P\big(S_{t-1} = i,\ S_t \in G(j)\big) > 0$, due to (1) and the validity of both statements in 2.,

$$\gamma_{i,j} = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big) = \nu_{a(i),a(j)} \cdot p_j.$$

Finally, assume 3. to hold. Then for $i, j = 1, \ldots, k$

$$P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) = \sum_{g \in G(j)} \sum_{h \in G(i)} \left(\frac{P\big(S_{t-1} = h\big)}{P\big(S_{t-1} \in G(i)\big)} \cdot \nu_{a(h),a(g)} \cdot p_g \right) = \nu_{a(i),a(j)},$$

$$P\big(S_t \in G(j) \big| S_{t-1} = i\big) = \sum_{g \in G(j)} \big(\nu_{a(h),a(g)} \cdot p_g\big) = \nu_{a(i),a(j)}.$$

If $P\big(S_{t-1} = i,\ S_t \in G(j)\big) = 0$, again $\gamma_{i,g} = 0$ for all $g \in G(j)$. Due to 3. we further have for $h \in G(i)$ and $g \in G(j)$ $\gamma_{h,g} = \nu_{a(i),a(g)} \cdot p_g = \gamma_{i,g} = 0$. Thus also $\nu_{a(i),a(j)} = 0$ and hence $\gamma_{i,j} = \big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{i,j} = 0$. If otherwise $P\big(S_{t-1} = i,\ S_t \in G(j)\big) > 0$ due to (1) we directly get $p_j = P\big(S_t = j \big| S_{t-1} = i,\ S_t \in G(j)\big)$, and since p_j is independent of i, $p_j = P\big(S_t = j \big| S_t \in G(j)\big)$, which finally gives

$$\gamma_{i,j} = \nu_{a(i),a(j)} \cdot p_j = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big) = \big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{i,j}.$$

□

Lemma 6. *Let $\mathbf{\Gamma} = (\gamma_{i,j})_{i,j=1,\dots,k}$ denote the (ergodic) t.p.m. of the stationary Markov chain $(S_t)_t$. Suppose that $\mathcal{G} = \{G_1, \dots, G_r\}$ and $\mathcal{H} = \{H_1, \dots H_q\}$ are two distinct partitions of the state space for which $\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \lambda_{\mathcal{H}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ and $\mathcal{H}$ is not a refinement of $\mathcal{G}$. Then there exists a partition $\mathcal{I}$ which is a strict coarsening of $\mathcal{G}$ and for which $\lambda_{\mathcal{I}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$.*

Proof of Lemma 6. Since $\mathcal{H}$ is not a refinement of $\mathcal{G}$, there exist $H \in \mathcal{H}$, $G \in \mathcal{G}$ with $H \cap G \neq \emptyset$ and $H \setminus G \neq \emptyset$, w.l.o.g. this is true for G_1 and H_1. Define the partition $\mathcal{I}$ by

$$\begin{aligned} I_1 &= G_1 \cup \bigcup_{\{l\,:\,G_l \cap H_1 \neq \varnothing\}} G_l, \\ I_l &= G_l, \text{ for } l \in \{1, \dots, r\} \text{ with } G_l \cap H_1 = \varnothing. \end{aligned}$$

Evidently $\mathcal{I}$ is a coarsening of $\mathcal{G}$ and has at least one element less than $\mathcal{G}$. We shall prove that $\lambda_{\mathcal{I}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$, that is,

$$\gamma_{i,j} = P\big(S_t \in I(j) \big| S_{t-1} \in I(i)\big) \cdot P\big(S_t = j \big| S_t \in I(j)\big) \qquad 1 \leq i, j \leq k. \tag{7}$$

Note that (7) in particular requires that rows of $\mathbf{\Gamma}$ with indices in the same element of the partition $\mathcal{I}$ be equal. Since $\mathbf{\Gamma} = \lambda_{\mathcal{G}}(\mathbf{\Gamma})$, this is true for $\mathcal{G}$, and hence evidently for all elements of the partition $\mathcal{I}$ exept for I_1. Suppose that $i, i' \in I_1$, $i \in G_l$, $i' \in G_{l'}$, we need to show that the i^{th} and the i'^{th} row of $\mathbf{\Gamma}$ be equal. By definition of I_1, there exist $j \in G_l \cap H_1$ and $j' \in G_{l'} \cap H_1$, and hence the i^{th} and the j^{th} row as well as the i'^{th} and the j'^{th} row of $\mathbf{\Gamma}$ are equal. But since also $\mathbf{\Gamma} = \lambda_{\mathcal{H}}(\mathbf{\Gamma})$, the j^{th} and the j'^{th} row of $\mathbf{\Gamma}$ are also equal, and the conclusion of equal rows for indices in the elements of $\mathcal{I}$ follows, formally,

$$\gamma_{i,g} = \gamma_{h,g}, \quad i, h \in I \in \mathcal{I}, \quad 1 \leq g \leq k. \tag{8}$$

Now, due to (8),

$$\begin{aligned} \big(\lambda_{\mathcal{I}}(\mathbf{\Gamma})\big)_{i,j} &= P\big(S_t = j \big| S_t \in I(j)\big) \cdot P\big(S_t \in I(j) \big| S_{t-1} \in I(i)\big) \\ &= \frac{\pi_j}{\sum_{g \in I(j)} \pi_g} \cdot \sum_{g \in I(j)} \sum_{h \in I(i)} \left(\frac{\pi_h}{\sum_{l \in I(i)} \pi_l} \cdot \gamma_{h,g} \right) \\ &= \frac{\pi_j}{\sum_{g \in I(j)} \pi_g} \cdot \sum_{g \in I(j)} \gamma_{i,g} \sum_{h \in I(i)} \frac{\pi_h}{\sum_{l \in I(i)} \pi_l} \\ &= \frac{\pi_j}{\sum_{g \in I(j)} \pi_g} \cdot \sum_{g \in I(j)} \gamma_{i,g}, \end{aligned} \tag{9}$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$ denotes the stationary distribution of $(S_t)_t$. Therefore, in order to show (7), it sufficies to show that

$$\frac{\gamma_{i,j}}{\pi_j} = \frac{\sum_{g \in I(j)} \gamma_{i,g}}{\sum_{g \in I(j)} \pi_g}, \qquad 1 \leq i, j \leq k. \tag{10}$$

which is equivalent to

$$\frac{\gamma_{i,j}}{\pi_j} = \frac{\gamma_{i,a}}{\pi_a}, \qquad 1 \leq i, j \leq k,\ a \in I(j). \tag{11}$$

Indeed, (10) evidently implies (11), while using (11) one computes

$$\gamma_{i,j} = \frac{1}{\sum_{g \in I(j)} \pi_g} \cdot \sum_{g \in I(j)} \pi_g\, \gamma_{i,j} = \frac{1}{\sum_{g \in I(j)} \pi_g} \cdot \sum_{g \in I(j)} \pi_j\, \gamma_{i,g} = \frac{\pi_j}{\sum_{g \in I(j)} \pi_g} \cdot \sum_{g \in I(j)} \gamma_{i,g},$$

that is, (10).

Now, in order to show (11), we note that the corresponding property holds for the partitions $\mathcal{G}$

and $\mathcal{H}$, so that (11) evidently holds if $I(j) \neq I_1$. To cover this case, suppose that $j, a \in I_1$, so that $j \in G_l$ and $a \in G_{l'}$ for some $1 \leq l, l' \leq r$. By definition of I_1, there exist $j' \in G_l \cap H_1$ and $a' \in G_{l'} \cap H_1$, and therefore for $1 \leq i \leq k$:

$$\frac{\gamma_{i,j}}{\pi_j} = \frac{\gamma_{i,j'}}{\pi_{j'}} = \frac{\gamma_{i,a'}}{\pi_{a'}} = \frac{\gamma_{i,a}}{\pi_a}.$$

□

Proof of Theorem 2. Let

$$\mathcal{G}^*_{\mathbf{\Gamma}} \in \operatorname{argmin}\big\{\operatorname{card}\mathcal{G} :\ \mathcal{G} \text{ is partition of} \{1,\ldots,k\} \text{ with } \lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}\big\}. \tag{12}$$

Note that $\lambda_{\{\{1\},\ldots,\{k\}\}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ is always satisfied. Suppose that $\mathcal{H}$ is a partition of $\{1,\ldots,k\}$ with $\mathbf{\Gamma} = \lambda_{\mathcal{H}}(\mathbf{\Gamma})$, then $\mathcal{H}$ must be a refinement of $\mathcal{G}^*_{\mathbf{\Gamma}}$, since otherwise by Lemma 6, there would exist a strict coarsening $\mathcal{I}$ of $\mathcal{G}^*_{\mathbf{\Gamma}}$ satisfying $\mathbf{\Gamma} = \lambda_{\mathcal{I}}(\mathbf{\Gamma})$, thus contradicting the choice of $\mathcal{G}^*_{\mathbf{\Gamma}}$. Therefore $\mathcal{G}^*_{\mathbf{\Gamma}}$ is the unique minimizer in (12). □

Proof of Theorem 3. Denote by $(Y_t^{(\mathcal{G})}, T_t^{(\mathcal{G})})_t$ the HMM with Markov chain $(T_t^{(\mathcal{G})})_t$ having t.p.m. $\lambda_{\mathcal{G}}(\mathbf{\Gamma})$ and observable process $(Y_t^{(\mathcal{G})})_t$ with state dependent densities $f_j(x)$, $j = 1,\ldots,k$.

First, we show that $\boldsymbol{\pi}$ is also the stationary distribution of the Markov chain $(T_t)^{\mathcal{G}}$, i.e. $P\big(T_t^{(\mathcal{G})} = j\big) = P\big(S_t = j\big)$ for $j \in \{1,\ldots k\}$. Since

$$\begin{aligned}
&\sum_{i=1}^{k} \pi_i\, P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) = \sum_{l=1}^{r} \sum_{g \in G_l} \pi_g\, P\big(S_t \in G(j) \big| S_{t-1} \in G_l\big) \\
=&\sum_{l=1}^{r} \Big\{ P\big(S_t \in G(j) \big| S_{t-1} \in G_l\big)\, P\big(S_t \in G_l\big)\Big\} = \sum_{l=1}^{r} \Big\{ P\big(S_t \in G(j),\ S_{t-1} \in G_l\big)\Big\} = P\big(S_t \in G(j)\big),
\end{aligned}$$

we get that

$$\begin{aligned}
\boldsymbol{\pi} \cdot \big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{\cdot,j} &= \sum_{i=1}^{k} \Big\{ \pi_i\, P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big)\, P\big(S_t = j \big| S_t \in G(j)\big)\Big\} \\
=&P\big(S_t = j \big| S_t \in G(j)\big)\ \sum_{i=1}^{k} \Big\{ \pi_i\, P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big)\Big\} = P(S_t = j).
\end{aligned}$$

The density of the observable process $(Y_t^{(\mathcal{G})})_{t=1,\ldots,n}$ of the reduced model is given by

$$\begin{aligned}
&f_{(Y_1^{(\mathcal{G})},\ldots,Y_n^{(\mathcal{G})})}(x_1,\ldots,x_n) \\
&= \sum_{j_1,\ldots,j_n=1}^{k} \Big(P\big(T_1^{(\mathcal{G})} = j_1\big) \cdot f_{j_1}(x_1) \cdot \prod_{t=2}^{n} P\big(T_t^{(\mathcal{G})} = j_t \big| T_{t-1}^{(\mathcal{G})} = j_{t-1}\big) \cdot f_{j_t}(x_t)\Big) \\
&= \sum_{l_1,\ldots,l_n=1}^{r} \sum_{g_1 \in G_{l_1}} \cdots \sum_{g_n \in G_{l_n}} \Big(P\big(S_1 = g_1\big) \cdot f_{g_1}(x_1) \cdot \prod_{t=2}^{n} \big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{g_{t-1},g_t} \cdot f_{g_t}(x_t)\Big) \\
&= \sum_{l_1,\ldots,l_n=1}^{r} \sum_{g_1 \in G_{l_1}} \cdots \sum_{g_n \in G_{l_n}} \Big(P\big(S_1 = g_1\big) \cdot f_{g_1}(x_1) \cdot \prod_{t=2}^{n} \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t} \cdot P\big(S_t = g_t | S_t \in G_{l_t}\big) \cdot f_{g_t}(x_t)\Big)
\end{aligned}$$

since for $g_{t-1} \in G_{l_{t-1}}$ and $g_t \in G_{l_t}$,

$$\big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{g_{t-1},g_t} = \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t} \cdot P\big(S_t = g_t | S_t \in G_{l_t}\big).$$

Only $P(S_n = g_n|S_n \in G_{l_n}) \cdot f_{g_n}(x)$ depend on g_n in the sum $\sum_{g_n \in G_{l_n}}$, everything else can be factorized. Iterating this procedure over g_t gives

$$
\begin{aligned}
&f_{(Y_1^{(\mathcal{G})},\ldots,Y_n^{(\mathcal{G})})}(x_1,\ldots,x_n) \\
&= \sum_{l_1,\ldots,l_n=1}^{r} \Big(\sum_{g_1 \in G_{l_1}} P\big(S_1 = g_1\big) \cdot f_{g_1}(x_1) \cdot \prod_{t=2}^{n} \Big\{ \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t} \cdot \sum_{g_t \in G_{l_t}} P(S_t = g_t|S_t \in G_{l_t}) \cdot f_{g_t}(x_t) \Big\} \Big) \\
&= \sum_{l_1,\ldots,l_n=1}^{r} \Big(P\big(S_1^{(\mathcal{G})} = l_1\big) \cdot f_{l_1}^{(\mathcal{G})}(x_1) \cdot \prod_{t=2}^{n} \Big\{ \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t} \cdot f_{l_t}^{(\mathcal{G})}(x_t) \Big\} \Big) \\
&= f_{(X_1^{(\mathcal{G})},\ldots,X_n^{(\mathcal{G})})}(x_1,\ldots,x_n).
\end{aligned}
$$

□

In the following we denote by $\mathcal{X}$ the observational space of the HMM, where the (X_t) take their values, and by ν the dominating measure for the densities $f(x,\theta)$, where $\theta = (\theta_1,\ldots,\theta_p)^T \in M \subset \mathbb{R}^p$.

Assumption 1. The set M is open, and for for every $x \in \mathcal{X}$ the map $\theta \mapsto f(x,\theta)$ is two-times continuously partially differentiable.

Assumption 2. The support of $f(\cdot,\theta)$ does not depend on $\theta \in M$.

Assumption 3. There is a $\delta > 0$ such that for all $1 \le i \le k$, $1 \le a,b \le p$

$$
E_{\boldsymbol{\eta}_0}\Big(\sup_{\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|<\delta} \big|\partial_{\theta_a} f(X_1,\theta_{\boldsymbol{\eta}}(i))\big|^2 \Big) < \infty, \qquad E_{\boldsymbol{\eta}_0}\Big(\sup_{\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|<\delta} \big|\partial_{\theta_a}\,\partial_{\theta_b} f(X_1,\theta_{\boldsymbol{\eta}}(i))\big| \Big) < \infty,
$$

$$
\int_{\mathcal{X}} \sup_{\|\theta-\theta_0\|<\delta} \big|\partial_{\theta_a} f(y,\theta)\big|\, d\nu(y) < \infty, \qquad \int_{\mathcal{X}} \sup_{\|\theta-\theta_0\|<\delta} \big|\partial_{\theta_a}\,\partial_{\theta_b} f(y,\theta)\big|\, d\nu(y) < \infty.
$$

Proof of Theorem 5. We intend to apply theorem 2 from Giudici et al. (2000). To this end, we note that their assumptions A1-A6 are satisfied and in particular that η_0 lies in the interior of the parameter set.

It remains to study the restriction on $\mathbf{\Gamma}$ and determine the degrees of freedom.

Note that in $\mathbf{\Gamma}_{\boldsymbol{\eta}}$ one arbitrary column is redundant, leading to $k^2 - k$ parameters. To prove the claim it is sufficient to show that $\mathbf{\Gamma}_{\boldsymbol{\eta}}$ can be smoothly parametrized in dependence of $\mathcal{G}$ via (k^2-k) parameters such that $(k^2 - 2k - r^2 + 2r)$ of them are zero if and only if $\boldsymbol{\eta} \in \Theta_{0,\mathcal{G}}$.

Since we assume that $P_{\mathbf{\Gamma}_{\boldsymbol{\eta}}}(S_t \in G_l|S_{t-1} \in G_m) > 0$, we may set

$$
\boldsymbol{A} = (\alpha_{i,l})_{\substack{i=1,\ldots,k \\ l=1,\ldots,r}}, \quad (\alpha_{i,l}) = \sum_{g \in G_l} (\mathbf{\Gamma}_{\boldsymbol{\eta}})_{i,g},
$$

$$
\boldsymbol{B} = (\beta_{i,j})_{i,j=1,\ldots,k}, \quad \beta_{i,j} = \alpha_{i,a(j)}^{-1} \cdot (\mathbf{\Gamma}_{\boldsymbol{\eta}})_{i,j}.
$$

Obviously, $(\mathbf{\Gamma}_{\boldsymbol{\eta}})_{i,j} = \alpha_{i,a(j)} \cdot \beta_{i,j}$. In the parametrization via $\boldsymbol{A},\boldsymbol{B}$ also one column in $\boldsymbol{A}$ (since all rows of $\boldsymbol{A}$ have to sum up to one), and r columns in $\boldsymbol{B}$ (since all columns of $\boldsymbol{B}$ with indices in the same group have to sum up to one) are redundant. In order to access the non-redundant parameters in a convenient way, consider a label switching in the Markov chain, such that $G_1 = \{1,\ldots,n_1\}, G_2 = \{n_1+1,\ldots,n_2\},\ldots,G_r = \{n_{r-1}+1,\ldots,n_r\}$. Thus, the Markov chain can be parametrized via

$$
\boldsymbol{A} = (\alpha_{i,l})_{\substack{i=1,\ldots,k \\ l=1,\ldots,r-1}}, \quad \boldsymbol{B} = (\beta_{i,j})_{\substack{i=1,\ldots,k \\ j=1,\ldots,k,\ j \neq n_1,\ldots,n_r}}.
$$

Note,

$$
\alpha_{i,l} = P_{\mathbf{\Gamma}_{\boldsymbol{\eta}}}\big(S_t \in G_l|S_{t-1} = i\big), \quad \beta_{i,j} = P_{\mathbf{\Gamma}_{\boldsymbol{\eta}}}\big(S_t = j|S_{t-1} = i, S_t \in G_l\big).
$$

Due to Proposition 1, $\lambda_{\mathcal{G}}(\boldsymbol{\Gamma}_{\boldsymbol{\eta}}) = \boldsymbol{\Gamma}_{\boldsymbol{\eta}}$, i.e. H_0, is equivalent to

$$\alpha_{i,l} = P_{\boldsymbol{\Gamma}_{\boldsymbol{\eta}}}\big(S_t \in G_l | S_{t-1} \in G_{a(i)}\big), \quad i = 1, \ldots, k, l = 1, \ldots, r-1,$$

$$\beta_{i,j} = P_{\boldsymbol{\Gamma}_{\boldsymbol{\eta}}}\big(S_t = j | S_t \in G_l\big), \quad i = 1, \ldots, k,\ j = 1, \ldots, k,\ j \neq n_1, \ldots, n_r.$$

Therefore, H_0 is equivalent to

$$\alpha_{1+n_{(m-1)},l} = \cdots = \alpha_{n_m,l}, \quad m = 1, \ldots, r,\ l = 1, \ldots, r-1,$$

$$\beta_{1,j} = \cdots = \beta_{k,j}, \quad j = 1, \ldots, k,\ j \neq n_1, \ldots, n_r$$

where $n_0 = 0$, which yields $(r-1) \cdot (k-r)$ restrictions to $\boldsymbol{A}$ and $(k-r) \cdot (k-1)$ restrictions to $\boldsymbol{B}$. Altogether, $\boldsymbol{\Gamma}_{\boldsymbol{\eta}}$ can be parametrized via matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and H_0 can be formulated via equality restrictions according to the new parameters. Thus, doing a second re-parametrization, where for each group of parameters that should be equal under H_0, all these parameters are expressed as the difference to one base parameter, yields the requested parametrization. Finally, $k^2 - 2k - r^2 + 2r$ parameters in the re-parametrized version being zero is equivalent to H_0, which concludes the proof. □