Philipps Universität Marburg

Fachbereich Mathematik und Informatik

Inference and Application of Likelihood Based Methods for Hidden Markov Models

Kumulative Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Philipps-Universität Marburg

vorgelegt von

Florian Schwaiger
Dipl.-Math. oec.
aus Marburg

Erstgutachter: Prof. Dr. Hajo Holzmann
Zweitgutachter: Prof. Dr. Eyke Hüllermeier

Eingereicht: 27.06.2013
Tag der mündlichen Prüfung: 23.08.2013

Table of Contents

Acknowledgments

At first I would like to thank my supervisor Prof. Dr. Hajo Holzmann. I am very grateful that he gave me the opportunity to carry out this research. The discussions with him and his constant encouragement have been very important for me.

Furthermore, I thank Prof. Dr. Eyke Hüllermeier for taking the Koreferat.

I am also grateful to Dr. Florian Ketterer for helpful discussions at the beginning of my research activity and for reading parts of this thesis. Moreover, I would like to thank Dipl.-Math. oec. Daniel Hohmann for also reading parts of my thesis and for his valuable comments. Next, I thank Dipl.-Math. Matthias Eulert for many helpful discussions concerning computational issues, and I generally thank all of my office mates for providing a good working atmosphere.

I gratefully acknowledge financial support from the DFG, grant Ho 3260/3-1.

Finally, I thank Kristin Klingelhöfer and my parents Christa and Christoph for various forms of support.

1 Introduction

Mixture models are widely used to analyze datasets with independent observations showing heterogeneity in such a sense that there are multiple subpopulations. There are several applications in biology, pattern recognition and many other areas where the task of a model-based segmentation of a dataset is of interest. Hidden Markov models are an extension of mixture models because they additionally allow for a certain serial dependence between observations and are thus applied to deal with data that exhibit dependence over time. There is also a wide range of different areas of application as speech recognition, see e.g. Rabiner (1989), or financial economics, see e.g. Rydén, Teräsvirta and Åsbrink (1998).

Before presenting mixture and hidden Markov models in more detail, we introduce a general setting which covers the models of interest. To this end, let $(X_t, S_t)_{t\in\mathbb{T}}$ denote a bivariate stochastic process with $\mathbb{T} = \{1, \dots, T\}$ (from now on we omit the set $\mathbb{T}$ in the subscript). Hereby, for each $t \in \mathbb{T}$, S_t is a discrete random variable with $S_t \in \{1, \dots, k\}$ for some $k \in \mathbb{N}$. The process $(S_t)_t$ is named *state process* and its realizations are commonly not observable. The number of states k is a parameter of high interest for us and its choice in applications will be discussed below. The process $(X_t)_t$ is called *observable process* and is independent given $(S_t)_t$. Further, for $t = 1, \dots, T$, the conditional distribution of X_t given $(S_t)_t$ depends on S_t only. Throughout this thesis the observable process is real-valued and uni- or multivariate, depending on the considered scenario. Conditional on $S_t = j$, the distribution function of X_t is denoted by $F_j(\cdot)$, $j = 1, \dots, k$, and is referred to as *state-dependent distributions*. We assume that these distributions have densities w.r.t. Lebesgue measure. Summarizing, the parameters of the described model comprise the number of states $k \in \mathbb{N}$, the distribution of the state process $(S_t)_t$ and the state-dependent distributions.

1.1 Finite Mixture Models

A finite mixture model is obtained by letting the state process $(S_t)_t$ be an i.i.d. sequence with $P(S_t = j) = p_j$, $p_j \geq 0$, $j = 1, \dots, k$ and $\sum_{j=1}^k p_j = 1$. Hence, the observable process $(X_t)_t$ is also i.i.d. and the distribution function is given by $F(x) = \sum_{j=1}^k p_j F_j(x)$. Note that, due to the restriction $\sum_{j=1}^k p_j = 1$, p_k is already defined by $p_1, \dots, p_{k-1}$. If further the state-dependent distributions belong to a parametric family given by $F(\cdot;\theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, then X_t has distribution function

$$F(x; p_1, \dots, p_k, \theta_1, \dots, \theta_k) = \sum_{j=1}^k p_j F(x; \theta_j),$$

$\theta_j \in \Theta$, $j = 1, \ldots, k$. In the following we will concentrate on the case of parametric state-dependent distributions such that all belong to the same family and denote $\boldsymbol{p}(k) = (p_1, \ldots, p_k)$, $\boldsymbol{\theta}(k) = (\theta_1, \ldots, \theta_k)$, or simply $\boldsymbol{p}$ and $\boldsymbol{\theta}$ if the number of components is known, where in case of mixture models states are also referred to as components.

Identifiability of the parameters of a mixture model, that means that same distribution functions imply the same parameter, is an important issue e.g. for parameter estimation, see Frühwirth-Schnatter (2006, chap. 1) for a short discussion. This topic is well studied for many parametric families, see e.g. Teicher (1963) or Yakowitz and Spragins (1968) for the normal distribution.

The parameters of a k-state mixture can be estimated by maximum likelihood, i.e.

$$\big(\hat{\boldsymbol{p}}(k), \hat{\boldsymbol{\theta}}(k)\big) = \arg\max \big\{ l_T^{(k)}(\boldsymbol{p}, \boldsymbol{\theta}) \big| \textstyle\sum_{j=1}^k p_j = 1, p_j \geq 0, \theta_j \in \Theta, j = 1, \ldots, k \big\},$$

where the log-likelihood of a k-state mixture model is given by

$$l_T^{(k)}(\boldsymbol{p}, \boldsymbol{\theta}) = \sum_{t=1}^{T} \log \big(\sum_{j=1}^{k} p_j f(X_t; \theta_j) \big),$$

with $f(\cdot; \theta)$ denoting the density function of the parametric family w.r.t. Lebesgue measure. In applications we always calculate the MLEs (as well for hidden Markov models) by direct constrained numerical maximization of the log-likelihood. Another common way is to apply the EM algorithm which was introduced by Dempster, Laird and Rubin (1977).

The choice of the number of components k is an important aspect since it has crucial impact on the quality of statistical inferences made by using mixture models. One possible strategy is to apply model selection criteria such as the Akaike information criterion (AIC) or the Baysian information criterion (BIC). Keribin (2000) shows that BIC chooses asymptotically the right number of components. Another approach, which will be pursued by us, is to successively test the hypothesis

$$H_0 : k = k_0 \qquad \text{against} \qquad H_1 : k = k_1 > k_0,$$

using likelihood based tests, starting with $k_0 = 1$. Under H_0 the likelihood-ratio test (LRT) statistic

$$2 \cdot \Big(l_T^{(k_1)}\big(\hat{\boldsymbol{p}}(k_1), \hat{\boldsymbol{\theta}}(k_1)\big) - l_T^{(k_0)}\big(\hat{\boldsymbol{p}}(k_0), \hat{\boldsymbol{\theta}}(k_0)\big) \Big)$$

does asymptotically not exhibit the usual $\chi^2_{\cdot}$-distribution (even for $k_0 = 1$) since the true parameter is not identified in the alternative parameter space. Thus further theory and other approaches have been developed. Chen, Chen and

Kalbfleisch (2001), (2004) and Li and Chen (2010) propose modified likelihood-ratio tests for analyzing the hypothesis $H_0 : k = 1$, $H_0 : k = 2$ and generally $H_0 : k = k_0$ for state-dependent distributions with an one-dimensional parameter as e.g. the Poisson distribution or the univariate normal distribution with fixed mean. Basically, penalty functions on the weight parameters are applied to force the estimates away from zero. Especially for normal mixtures there are various tests, among which Chen, Li and Fu (2012) propose a test for assessing the general hypothesis of k_0 against $2k_0$ components for normal location- and scale-mixtures. In each paper the asymptotic distribution of the corresponding modified likelihood-ratio statistic is deduced and in finite-sample applications this distribution is used for testing. Generally, if the asymptotic distribution of the test statistic is not available or the finite-sample behavior of the test is not accurate, bootstrapping the (modified) likelihood-ratio statistic is a proper approach, see McLachlan (1987). In Vollmer, Holzmann and Schwaiger (2013) a parametric bootstrap technique is used to assess the number of components of a normal mixture for a GDP dataset.

Commonly, mixture models are applied for model-based clustering, as the unobserved components can be estimated and linked to groups. One aims to find the most likely component to each observation given the (estimated) mixture model. This can be done by maximum-a-posteriori estimation, i.e.

$$\hat{S}_t^{\text{MAP}} = \arg\max \big\{ \hat{p}_j f(X_t; \hat{\theta}_j) \big| j = 1, \dots, k \big\}, \quad t = 1, \dots, T,$$

where $\hat{p}_j, \hat{\theta}_j$, $j = 1, \dots, k$, are the estimated parameters. In case of mixture models having state-dependent distributions with distinctly different locations (relatively seen to scales) this leads to accurate decoding results, i.e. peaks in the density are linked to clusters. But components of a mixture model do not necessarily coincide with density based clusters: In case of scale mixtures, i.e. equal locations but different scales, an accurate decoding around the mode is nearly impossible. To overcome this difficulty Biernacki, Celeux and Govaert (2000) introduce a model selection criteria in order to find components which correspond to clusters, and Baudry, Raftery, Celeux, Lo and Gottardo (2010) merge components which represent density based clusters after estimation. A strategy of combining states in case of hidden Markov models has been investigated in Holzmann and Schwaiger (2013a).

1.2 Hidden Markov Models

A k-state hidden Markov model (HMM) is given by the process $(X_t, S_t)_t$ introduced above when the state process $(S_t)_t$ is considered as a (first order) Markov chain with state space $\{1, \dots, k\}$, i.e. for $t > 1$ satisfying

$$P(S_t = j_t | S_1 = j_1, \dots, S_{t-1} = j_{t-1}) = P(S_t = j_t | S_{t-1} = j_{t-1}),$$

with $j_\tau \in \{1, \dots, k\}$, $\tau = 2, \dots, t$. The name hidden Markov model is caused by the fact that only the process $(X_t)_t$ is observable while the Markov chain $(S_t)_t$ is hidden. Zucchini and MacDonald (2009) provide a practical introduction to hidden Markov models with applications and code examples.

A Markov chain is said to be homogeneous if for $i, j = 1, \dots, k$ the transition probabilities $P(S_t = j | S_{t-1} = i)$ are independent of the time t. The Markov chain is then characterized by its initial distribution $P(S_1 = j)$, $j = 1, \dots, k$, and the transition probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{i,j})_{i,j}$ with

$$\gamma_{i,j} = P(S_t = j | S_{t-1} = i), \quad t > 1.$$

Finite-state hidden Markov models are also called Markov-dependent finite mixtures suggesting their relation to finite-state mixture models. In detail, mixture models form a subclass of hidden Markov models since a t.p.m. with all rows being equal directly implies an independent state process. Assuming the Markov chain to be ergodic yields the unique existence of a stationary distribution which is denoted by $\boldsymbol{\pi} = \boldsymbol{\pi}(k) = (\pi_1, \dots, \pi_k)$ and satisfies $\boldsymbol{\pi} \cdot \mathbf{\Gamma} = \boldsymbol{\pi}$. Furthermore, for any starting distribution the probabilities $P(S_t = j)$ then converge towards π_j, $j = 1, \dots, k$, for $t \to \infty$. Hence, under ergodicity, the assumption to start the Markov chain in its stationary distribution is a mild restriction, especially for asymptotic considerations.

The observable process $(X_t)_t$ is dependent over time via the Markov chain. Thus, hidden Markov models are suitable to model serially-dependent data when the dependence is induced by a latent Markov dependent group belonging. When $F_j(\cdot)$ denotes again the conditional distribution function of X_t given $S_t = j$, the marginal distribution function of X_t is given by the mixture $\sum_{j=1}^{k} P(S_t = j) F_j(x)$. If the Markov chain is started in its stationary distribution the latter mixture is independent of t and the weights are given by $\boldsymbol{\pi}$.

A popular example of HMMs, which Rydén et al. (1998) use to describe log-returns of asset prices, is a normal HMM with fixed zero means and state-dependent standard deviations. Let therefore $(\varepsilon_t)_t$ be an i.i.d. sequence of standard normal random variables, $(S_t)_t$ a stationary k-state Markov chain and $\sigma_j \in \mathbb{R}$, $j = 1, \dots, k$, with $0 < \sigma_1 < \dots < \sigma_k$. Then by letting $X_t = \sigma_{S_t} \varepsilon_t$, the process $(X_t, S_t)_t$ is a k-state HMM. Its observable process has a scale-mixture of normals as stationary distribution, which covers the fat-tailed distribution of the log-returns. Further, each state of the Markov chain refers to a different market situation characterized by the level of variance of the observable process. The estimated transition behavior of the Markov chain is typically very persistent and Rydén et al. (1998) find by bootstrapping models with two or three states. In Holzmann and Schwaiger (2013b) normal and skew-normal HMMs are considered to model log-returns, and it is discussed that for time-periods containing the financial crisis in 2008 even four states are required to describe the data well, see below for further discussion on choosing the number of states of an HMM.

Analogously to mixture models we focus on the case that the state-dependent distributions are given by a parametric family with distribution function $F(\cdot;\theta)$ and density function $f(\cdot;\theta)$, $\theta \in \Theta \subset \mathbb{R}^d$, respectively. The parameters of a k-state HMM with ergodic Markov chain starting in its stationary distribution are then given by the transition probability matrix $\mathbf{\Gamma} \in \mathbb{R}^{k\times k}$ and the state-dependent parameters $\boldsymbol{\theta} = \boldsymbol{\theta}(k) = (\theta_1, \dots, \theta_k)$, such that $(X_t|S_t = j) \sim F(\cdot;\theta_j), j = 1, \dots, k$. Similarly to mixture models, identifiability of HMMs is well studied. One very important result in this context is given by Leroux (1992). If all state-dependent distributions belong to the same parametric family, then identifiability of the HMM holds if the corresponding finite mixture is identifiable. The log-likelihood of the observable part is given by

$$L_T(\mathbf{\Gamma}, \boldsymbol{\theta}) = \log\Big(\boldsymbol{\pi}\ \mathbf{\Gamma}\ \boldsymbol{f}(X_1;\boldsymbol{\theta})\ \mathbf{\Gamma}\ \boldsymbol{f}(X_2;\boldsymbol{\theta}) \cdot \ldots \cdot \mathbf{\Gamma}\ \boldsymbol{f}(X_T;\boldsymbol{\theta})\ \mathbf{1}^T\Big),$$

where $\boldsymbol{f}(X_t;\boldsymbol{\theta}) = \operatorname{diag}\big(f(X_t;\theta_1), \dots, f(X_t;\theta_k)\big)$ and $\mathbf{1} = (1, \dots, 1)$. The log-likelihood can be computed numerically stable by an algorithm given in chapter 3 of Zucchini and MacDonald (2009).

For a given number of states k, maximum likelihood estimation of a hidden Markov model can be done by direct numerical maximization of the log-likelihood or by the EM algorithm. Under regularity conditions, Leroux (1992) shows consistency of the MLE and Bickel, Ritov and Rydén (1998) its asymptotic normality. Lindgren (1978) introduces a quasi-likelihood based approach. He proposes to estimate the state-dependent parameters and the weights of the stationary mixture of an HMM by using the log-likelihood of a mixture model. Under regularity conditions these estimates are consistent. The transition probability matrix cannot be estimated by this approach since it is neither involved in the mixture log-likelihood nor is there a unique mapping from the stationary distribution back to the transition probability matrix.

In fact, besides selecting an appropriate parametric family, the choice of the number of states is the critical part in model estimation. Model selection criteria like AIC or BIC can be applied here as well, but a likelihood-ratio test for the number of states based on the likelihood $L_T(\mathbf{\Gamma}, \boldsymbol{\theta})$, as described for mixture models in the previous section, leads to problems. Even for testing $H_0 : k = 1$, the likelihood-ratio statistic does not converge towards a $\chi^2_\cdot$-distribution. Instead, it diverges to infinity as $T \to \infty$, see Gassiat and Keribin (2000). To overcome arising problems when using the full-model likelihood, Dannemann and Holzmann (2008) propose a quasi-likelihood based likelihood-ratio test for investigating the hypothesis $H_0 : k = 2$. Based on the marginal distribution of the HMM, they show that the asymptotic distribution of the modified likelihood-ratio statistic of Chen et al. (2004) still holds in case of a Markov dependent state process. In Holzmann and Schwaiger (2013b) this approach is followed and the work of Li and Chen (2010) and Chen et al. (2012) is extended for HMMs in order to test the hypothesis $H_0 : k = k_0$ for an arbitrary $k_0 \in \mathbb{N}$. However, likelihood ratio-tests using the full-model likelihood are in certain situations also possible for HMMs and the usual $\chi^2_{\cdot}$-distribution holds. Giudici et al. (2000) extend the theory of likelihood-ratio testing from the i.i.d. to the HMM setting. Their framework does not cover tests for the number of states but e.g. special restrictions on the state-dependent parameters can be tested. They test for equal entries in the state-dependent covariance matrices of a multivariate normal HMM. In Holzmann and Schwaiger (2013a) their result is used and it is shown that a full-model likelihood ratio test can be applied to test special restrictions on the transition probability matrix of an HMM.

The possibility of using an HMM for model based clustering is one of the reasons for the popularity of this model class. The serial dependence induced by a dataset or resp. an estimated model is a very important information for state decoding. In the introductory example of this section the stationary distribution is a scale mixture, but due to the persistent state transition an accurate decoding also around the common mean is possible. Given all realizations of the observable process, one can estimate the hidden states by e.g. calculating the most likely sequence of states, i.e.

$$\arg\max \big\{ P(S_1 = j_1, \dots, S_T = j_T | X_1 = x_1, \dots, X_T = x_T) \big| j_t = 1, \dots, k,\ t = 1, \dots, T \big\}.$$

The latter strategy is denoted as global decoding and can be done efficiently with the Viterbi algorithm, see Viterbi (1967). Further, states can be estimated using local decoding, i.e. by estimating the most likely state isolated given the whole observable process for each t, see Zucchini and MacDonald (2009) for details on both approaches. In Holzmann and Schwaiger (2013b) the Viterbi algorithm is used to identify volatility periods of financial assets.

2 Summary of Publications

2.1 English

Testing for the number of states in hidden Markov models

Hidden Markov models with state-dependent scale are used in the literature to model asset log-return time-series, see Section 1.2 or e.g. Rydén et al. (1998). Commonly, Markov chains with two or three states are used therefor. In "Testing for the number of states in hidden Markov models" we deal with quasi-likelihood based testing procedures, which enables us to investigate the question whether in light of the financial crisis of 2008 three volatility states are still sufficient or if a fourth crisis-state should be added.

In detail we generalize existing testing procedures for i.i.d. mixture models to hidden Markov models by considering penalized quasi-likelihood ratio tests. They can be applied in order to assess the number of states k of a hidden Markov model with univariate state-dependent distribution fulfilling certain regularity conditions. For two different models, which differ in the assumptions concerning the state-dependent distribution, we propose tests for the hypothesis

$$H_0 : k = k_0 \qquad \text{against} \qquad H_1 : k > k_0,$$

for fixed $k_0 \in \mathbb{N}$. Note that for $k_0 = 2$ a quasi-LRT for HMMs has been developed by Dannemann and Holzmann (2008).

In the setting of a k-state HMM with state-dependent distributions belonging to the same parametric family having density $f(\cdot, \theta)$ w.r.t. Lebesgue measure, the quasi-log-likelihood is given by $l_T^{(k)}(\boldsymbol{p}, \boldsymbol{\theta})$, see Section 1.1. Thus, the quasi-log-likelihood neglects the serial-dependence of the hidden Markov chain and replaces it by an i.i.d. state migration, see Lindgren (1978). The quasi-maximum-likelihood estimator (QMLE) is the parameter which maximizes the quasi-log-likelihood given the usual restrictions of mixture model estimation. We consider two different types of state-dependent distributions implying two different tests.

Normal HMM

The first test is based on the assumption of a normal state-dependent distribution where both the mean and standard deviation parameter depend on the state of the Markov chain, i.e. $\theta_j = (\mu_j, \sigma_j)$, $\mu_j \in \mathbb{R}$, $\sigma_j > 0$, $j = 1, \ldots, k_0$. The quasi-log-likelihood is unbounded in this case, but this problem can be solved by adding a penalty function which penalizes small values of σ_j. For estimation under the hypothesis of k_0 states we thus use the same penalty as in Chen et al. (2012). For estimation under the alternative of a mixture with $2k_0$ components a different penalty function on the scale parameters and more restrictions on the weights are applied. Essentially, we use the restricted parameter space for $2k_0$-dimensional weight vectors as used in Chen et al. (2012), which asymptotically bounds all weights away from zero under the null hypothesis. Our first main statement is the asymptotic distribution of the penalized quasi-LRT statistic Z_n, i.e.

$$Z_n \xrightarrow{d} \chi^2_{2k_0} \qquad \text{as } n \to \infty,$$

which holds for a k_0-state normal HMM with distinct means and an ergodic hidden Markov chain.

Univariate switching parameter

As a second model class we assume the parametric family of state-dependent distributions with density $f(\cdot, \theta)$ to be parametrized by a d-dimensional parameter $\boldsymbol{\nu} \in \Theta_1$ and a one dimensional parameter $\vartheta \in \Theta_2$, i.e. $\theta = (\boldsymbol{\nu}, \vartheta)$. The main difference to the previously discussed setting of a normal HMM is that $\boldsymbol{\nu}$ is assumed to be a nuisance parameter, i.e. $(X_t|S_t = j) \sim f(\cdot, \theta_j)$ with $\theta_j = (\boldsymbol{\nu}, \vartheta_j)$. Note that we extend the test of Li and Chen (2010) to additionally allow for a nuisance parameter and a Markov dependent state process.

In case of finite mixture models having a univariate state-dependent parameter, the asymptotic distribution of the LRT statistic is surprisingly more involved than in case of a normal state-dependent distribution with a bivariate parameter. The asymptotic distribution of the LRT statistic is, under a set of regularity conditions on the density and under a special estimation procedure, see Li and Chen (2010), given by a mixture of point mass in zero and k_0 $\chi^2_\cdot$-distributions. The weights of the asymptotic mixture depend on the true parameter value. This fact transfers to the case of a quasi LRT with structural parameter. Our second main statement is the asymptotic distribution of the quasi LRT statistic R_n under the hypothesis of a k_0-state HMM, i.e.

$$R_n \xrightarrow{d} \sum_{j=0}^{k_0} \alpha_j \chi^2_j \qquad \text{as } n \to \infty,$$

where χ^2_0 denotes the point mass at zero and the weights again depend on the true parameters. The asymptotic distribution holds e.g. for the Poisson distribution or (skew-)normal and t-distributions with state-dependent scale if a lower bound on the scale parameter is applied. Note that in applications the unknown weights can be replaced by estimates.

We provide an extensive simulation study for normal and skew-normal state dependent distributions, which indicate a good finite sample behavior of both tests.

Application to log-returns

To answer the question of how many volatility states are needed, we apply a normal and a skew-normal HMM and the corresponding tests to a 15-year dataset of daily log-returns of the S&P 500 index. In case of the skew-normal HMM the hypotheses of $k_0 = 1, 2$ and 3 can be clearly rejected using asymptotic critical values. In case of the normal HMM one and two states can also be rejected by this way, whereas three states can be rejected using simulated critical values. In both models the hypothesis of four states cannot be rejected, and estimating the maximum-a-posteriori state sequence using four states, in fact highlights a clear connection of the fourth state and the financial crisis of 2008.

Hidden Markov Models with state-dependent mixtures

Generally, hidden Markov models with state-dependent finite mixtures provide much more flexibility than a simple parametric family as e.g. the normal distribution. Using these HMMs for modeling or clustering serially-dependent data is thus a convenient way to benefit of good properties concerning adequate characterization of state-dependent distributions and of the simplicity of mixtures. In this paper we analyze the dependence structure of this model class. Our results have applications to model selection as well as to model-based clustering. We propose algorithms for both purposes.

The parametrization of such a model is not unique since one can parametrize an r-state HMM with mixtures as state-dependent distribution as a k-state HMM, where $r < k$, such that each component of the state-dependent mixture is interpreted as a single state of a k-state Markov chain. We therefore investigate the dependence structure of the hidden Markov chain and deduce a unique minimal representation of the HMM if the state-dependent densities belong to the same parametric family.

At first we analyze the dependence structure of Markov chains isolated and therefor define a function λ on the space of transition probability matrices with fixed number of states. In detail when the original Markov chain with t.p.m. $\mathbf{\Gamma}$ has k-states and $\mathcal{G} = \{G_1, \ldots, G_r\}$ denotes a partition of the state space $\{1, \ldots, k\}$, the mapped t.p.m. is given by

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

where $G(j) = l \Leftrightarrow j \in G_l$, $j = 1, \ldots, k$. As the sets G_l can be interpreted as groups of states, the mapped t.p.m. is thus given by the transition probabilities between groups and by the conditional probability within the attained group. Therefore, the mapping can be interpreted as a reduction of information concerning the state migration. The first main result is the existence of a unique, minimal partition $\mathcal{G}^*_{\mathbf{\Gamma}}$ such that $\lambda_{\mathcal{G}^*_{\mathbf{\Gamma}}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$. Note that minimal refers in this context to the number of sets in the partition.

We prove that an arbitrary HMM with Markov chain having t.p.m. $\mathbf{\Gamma}$ can be parametrized equivalently by an r-state HMM with state-dependent finite mixtures, whenever $\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ holds for a partition $\mathcal{G} = \{G_1, \ldots, G_r\}$ of the original state space $\{1, \ldots, k\}$. Equivalently here refers to identical distributions of the observable processes under both parameterizations. We attain the requested unique minimal representation of the HMM as a Corollary, if k-component mixtures of the used parametric family are identifiable. The resulting partition is denoted by *independence partition* and the sets contained therein by *independence clusters*. Note that we refer to the mentioned reparametrization as *merging*, since several states are interpreted as a new state having a finite mixture as state-dependent distribution.

For model selection purposes and in case that the state-dependent densities belong to a known parametric family, we propose a likelihood-ratio test. The test is based on the full-model log-likelihood $L_T(\mathbf{\Gamma}, \boldsymbol{\theta})$, see Section 1.2. In detail, for a given partition $\mathcal{G} = \{G_1, \ldots, G_r\}$ of the state space $\{1, \ldots, k\}$ we test

$$H_0 : \lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma} \text{ against } H_1 : \lambda_{\mathcal{G}}(\mathbf{\Gamma}) \neq \mathbf{\Gamma}.$$

Under the assumption of the true parameter to be an interior point of the null-hypothesis parameter space and under further regularity assumptions, we show that the LRT statistic is asymptotically $\chi^2_{\cdot}$-distributed with $k^2 - 2k - r^2 + 2r$ degrees of freedom. For the normal or respectively multivariate-normal distribution the regularity conditions are fulfilled if we impose lower bounds on the scale parameters or respectively on the determinants of the covariance matrices. In order to investigate the finite-sample behavior of the test we provide an extensive simulation study.

Since merging in general changes the distribution of the observable process, we provide an algorithm which uses the upper LRT in order to find the independence partition iteratively via backward selection. For a given dataset and a parametric family of densities the algorithm starts with the trivial partition $\mathcal{G}_0 = \{\{1\}, \ldots, \{k\}\}$. In each iteration step all partitions resulting from combining two sets are considered and for each of it the test is performed. The new partition is then given by the one associated with the highest p-value or the iteration is stopped if all tests can be rejected given a predefined level $\alpha > 0$. A simulation study indicates a level of $\alpha = 0.01$ to be a good choice.

In case of i.i.d. mixture models, component-distributions which refer to the same density based cluster are supposed to be interpreted as a single finite mixture component-distribution, since then a meaningful maximum-a-posteriori analysis is possible, see Baudry et al. (2010) or Hennig (2010). For HMMs the situation is more involved. If there is a strong serial-dependence in the dataset, then also states whose state-dependent distributions marginally strongly overlap can still be well separated. Only if the dependence structure justifies merging, i.e. states are in the same independence cluster, and if state-dependent distributions represent a density based cluster, states should be merged. Thus, we propose an entropy based algorithm which iteratively finds density based clusters within the independence partition. The entropy of the local-decoding probabilities are calculated for several candidate models and the one with minimal entropy is chosen. This procedure is iterated until all possible states w.r.t. the independence partition are merged. The final model is given by the model of the last iteration step, or if the plot of the entropy values shows an elbow the according model is selected.

Using normal state-dependent distributions, we apply our methodology to log-returns of daily gold prices covering a 15-year period. The information criteria AIC and resp. BIC choose six and resp. four states. We start with a six-state model and obtain a four-state HMM by applying iterative testing and entropy based merging. Two state-dependent distributions are given by normal distributions. The remaining two are both mixtures with two components.

Peaks vs Components

Quah (1996) finds a rich and a poor convergence club by relating peaks in the density of the GDP to welfare groups. In "Peaks vs Components"we illustrate at first that this approach can lead to ambiguous conclusions, since when associating peaks with groups the results are not invariant under changes of the scale. The number of peaks (modes of the density) can vary when e.g. the log-GDP data is considered.

In our paper we analyze welfare groups of countries all over the world by applying finite mixture models. The GDP per capita of 190 countries from 1970 to 2009 given by the "Penn World tables 7.0", see Heston, Summers and Aten (2011), is considered. Instead of peaks in the density we use components of a mixture model as decision criterion of the group membership of a country. The application of such models is not new to economic literature. Paap and Dijk (1998) apply a two-state mixture of a normal and Weibull distribution to model the GDP per capita. In the present paper we challenge the twin-peaks approach and suggest a finite mixture of normal distributions with state-dependent mean and standard deviation as an alternative.

At first we consider the mode-based approach an apply it to the kernel density estimate. In order to find an appropriate choice of the bandwidth we apply the Silvermantest, see Silverman (1981), and obtain a validated number of modes for each year. We thereby find evidence for two peaks at first (1970-1990) and for three peaks thereafter.

In a second step we investigate the panel dataset by estimating for each year a normal mixture. To this end we find and validate the number of components by iterative testing. Note that since the likelihood of the considered mixture model is unbounded and diverges as $\sigma_j \to \infty$, an usual LRT cannot be applied. Thus, we use a test based on a penalized version of the log-likelihood as proposed in Chen and Li (2009) which results by adding

$$-0.05 \sum_{j=1}^{m} \left(s_n^2/\sigma_j^2 + \log\left(\sigma_j^2/s_n^2\right)\right)$$

to the log-likelihood, where s_n denotes the empirical standard deviation of the dataset and σ_j the standard deviation of the jth component. The final test decision is then based on critical values which are the result of a parametric bootstrap.

The tests result in mixture models with at first three (1970-1995) and then two components (1996-2009). Since the component-means differ distinctly (relatively seen to the standard deviations) the three components can be interpreted as low-, middle- and high-income countries. Because of the same reason, the two components of the mixtures for years after 1995 also refer to low- and high-income countries. Additionally, the low-income component of the mixtures after 1995 can be seen as results of a union of two previous components. Finally, by computing a-posteriori probabilities we obtain a detailed insight into the group membership of all countries over the course of time.

2.2 German

Testing for the number of states in hidden Markov models

Hidden Markov Modelle (HMMs) mit zustandsabhängigem Skalenparameter sind in der Literatur weitverbreitet, um log-return Zeitreihen von Assetpreisen zu modellieren, siehe Abschnitt 1.2 oder z.B. Rydén et al. (1998). Gewöhnlich werden dazu Markov-Ketten mit zwei oder drei Zustände verwendet. In "Testing for the number of states in hidden Markov models" befassen wir uns mit quasi-Likelihood basierten Testverfahren, um unter anderem der Frage nachzugehen, ob angesichts der Finanzkrise von 2008 drei verschiedene Zustände der Volatilität ausreichen, oder ob ein vierter Krisen-Zustand hinzugefügt werden sollte.

Genauer verallgemeinern wir vorhandene Testverfahren von Mischungsmodellen auf HMMs, indem wir penalisierte quasi-Likelihood-Quotienten-Tests untersuchen. Falls die zustandsbedingte Verteilung des HMMs univariat ist und gewisse Regularitätsbedingungen erfüllt sind, können die vorgestellten Testverfahren dazu verwendet werden, die Anzahl der Zustände k eines HMMs festzustellen. Für zwei verschiedene Modelle, welche sich in den Annahmen bezüglich der zustandsbedingten Verteilungen unterscheiden, dient der Test dazu, die Hypothese

$$H_0 : k = k_0 \qquad \text{gegen} \qquad H_1 : k > k_0$$

für fixiertes $k_0 \in \mathbb{N}$ zu testen. Für $k_0 = 2$ wurde bereits von Dannemann and Holzmann (2008) ein quasi-Likelihood-Quotienten-Test vorgeschlagen.

Die quasi-log-Likelihood eines HMMs, welches über k Zustände und zustandsbedingte Verteilungen der selben parametrischen Familie mit Dichte $f(\cdot;\theta)$ bzgl. des Lebesgue Maßes verfügt, ist durch $l_T^{(k)}(\boldsymbol{p}, \boldsymbol{\theta})$ gegeben, siehe Abschnitt 1.1. Somit vernachlässigt die quasi-log-Likelihood die serielle Abhängigkeitsstruktur der Markov-Kette und ersetzt diese durch einen u.i.v. Zustandsübergang, siehe Lindgren (1978). Der quasi-Maximum-Likelihood-Schätzer ist definiert als der Parameter, welcher die quasi-log-Likelihood gegeben den üblichen Nebenbedingungen für Mischungsmodelle maximiert. Wir betrachten zwei verschiedenen zustandsbedingten Verteilungen, welche jeweils einen anderen Test zur Folge haben.

Normalverteilung

Dem ersten Test liegt die Annahme einer zustandsbedingten Normalverteilung zu Grunde, wobei sowohl Lokations- als auch Skalenparameter von der Markov-Kette abhängig sind, d.h. $\theta_j = (\mu_j, \sigma_j)$, $\mu_j \in \mathbb{R}$, $\sigma_j > 0$, $j = 1, \ldots, k_0$. Die quasi-log-Likelihood ist in diesem Fall unbeschränkt, jedoch kann dieses Problem durch hinzufügen einer Penaltyfunktion, welche kleine Werte für σ_j bestraft, gelöst werden. Für die Schätzung unter der Hypothese von k_0 Zuständen verwenden wir die Penaltyfunktion wie in Chen et al. (2012). Für die Schätzung unter der Alternative eines Mischungsmodells mit $2k_0$ Zuständen wenden wir eine weitere Penaltyfunktion bzgl. der Skalenparameter und zusätzliche Nebenbedingungen an die Gewichte der Mischung an. Im Grunde verwenden wir den eingeschränkten Parameterraum für $2k_0$-dimensionale Mischgewichte wie in Chen et al. (2012) eingeführt, welcher unter gültiger Nullhypothese asymptotisch alle Gewichte von Null weg beschränkt. Unser erstes Hauptresultat ist die asymptotische Verteilung der penalisierten quasi-Likelihood-Quotienten-Test Statistik Z_n,

$$Z_n \xrightarrow{d} \chi^2_{2k_0}, \qquad \text{für } n \to \infty$$

welche unter der Nullhypothese eines HMMs mit normalverteilten zustandsbedingten Verteilungen, k_0 Zuständen, verschiedenen Lokationsparametern und ergodischer Markov-Kette gilt.

Eindimensionaler zustandsabhängiger Parameter

Als zweite Modellklasse nehmen wir an, dass die parametrische Familie der zustandsbedingten Verteilungen mit Dichte $f(\cdot, \theta)$ durch einen d-dimensionalen Parameter $\boldsymbol{\nu} \in \Theta_1$ und einen eindimensionalen Parameter ϑ parametrisiert wird, d.h. $\theta = (\boldsymbol{\nu}, \vartheta)$. Der Hauptunterschied zu dem zuvor diskutierten Fall ist die Tatsache, dass $\boldsymbol{\nu}$ hier als Strukturparameter betrachtet wird, d.h. $(X_t|S_t = j) \sim f(\cdot, \theta_j)$ mit $\theta_j = (\boldsymbol{\nu}, \vartheta_j)$. Wir erweitern den Test von Li and Chen (2010), um einerseits einen Strukturparameter und andererseits einen Markov-abhängigen Zustandsprozess berücksichtigen zu können.

Im Falle eines endlichen Mischungsmodelles mit univariatem zustandsabhängigen Parameteren ist die asymptotische Verteilung der Likelihood-Quotienten-Test Statistik überraschenderweise komplizierter als bei zustandsabhängiger Normalverteilung mit bivariatem Parameter. Die asymptotische Verteilung ist unter gewissen Regularitätsbedingungen an die Dichte und unter Anwendung einer speziellen Schätzmethodik der Parameter, siehe Li and Chen (2010), gegeben durch ein Mischungsmodell aus der Punktmasse in Null und k_0 $\chi^2_\cdot$-Verteilungen. Die Gewichte der asymptotischen Mischung hängen dabei vom wahren Parameterwert ab. Unser zweites Hauptresultat ist die asymptotische Verteilung der quasi Likelihood-Quotienten-Test Statistik R_n unter der Hypothese eines HMMs mit k_0 Zuständen,

$$R_n \xrightarrow{d} \sum_{j=0}^{k_0} \alpha_j \chi_j^2 \qquad \text{für } n \to \infty,$$

wobei χ_0^2 die Punktmasse in Null bezeichnet und die Gewichte wieder vom wahren Parameterwert abhängen. Die asymptotische Verteilung gilt z.B. für die Poisson Verteilung oder für die (schiefe-)Normalverteilung bzw. t-Verteilung mit zustandsabhängigem Skalenparameter, falls eine untere Schranke für den Skalenparameter benutzt wird. In Anwendungen können die unbekannten Gewichte durch Schätzer ersetzt werden.

Eine umfangreiche Simulationsstudie für normale- und schief-normale zustandsbedingte Verteilungen belegt ein gutes Verhalten bei endlichen Stichprobengrößen.

Anwendung auf log-returns

Um der Frage nach der Anzahl benötigter Volatilitäts-Zustände auf den Grund zu gehen, wenden wir hidden Markov Modelle mit zustandsbedingter Normal- und schiefer Normalverteilung an. Wir betrachten die täglichen log-returns des S&P 500 über einen Zeitraum von 15 Jahren und wenden je nach Modell den zugehörigen Test an. Im Fall der schiefen Normalverteilung können die Hypothesen von $k_0 = 1, 2$ und 3 unter Berücksichtigung der asymptotischen kritischen Werte klar verworfen werden. Im Fall der Normalverteilung können sowohl ein als auch zwei Zustände auf diese Art verworfen werden. Drei Zustände können unter Verwendung simulierter kritischer Werte ebenso verworfen werden. Da vier Zustände für beide Modelle nicht verworfen werden können, wählen wir jeweils vier Zustände. Die Schätzung der maximum-a-posteriori Zustandsfolge stellt schließlich einen klaren Zusammenhang zwischen dem vierten Zustand und der Finanzkrise von 2008 her.

Hidden Markov Models with state-dependent mixtures

Hidden Markov Modelle mit zustandsbedingten endlichen Mischungsmodellen bieten generell deutlich mehr Flexibilität als einfache parametrische Familien wie z.B. die Normalverteilung. Bei der Anwendung solcher HMMs für die Modellierung oder das Clustering seriell abhängiger Daten profitiert man dabei von der Möglichkeit einer adäquaten Beschreibung der zustandsbedingten Verteilungen und der guten Handhabbarkeit von Mischungsmodellen. In dieser Arbeit analysieren wir die Abhängigkeitsstruktur dieser Modellklasse. Unsere Ergebnisse haben Anwendungen für die Modellwahl und das modellbasierte Clustering. Für beide Aufgaben schlagen wir je einen Algorithmus vor.

Die Parametrisierung eines solchen Modells ist nicht eindeutig, denn ein HMM mit r Zuständen und zustandsbedingten Mischungen lässt sich als HMM mit k Zuständen parametrisieren, wobei $r < k$ und jede Komponente der zustandsbedingten Mischungen als einzelner Zustand der Markov-Kette mit k Zuständen interpretiert wird. Wir untersuchen daher die Abhängigkeitsstruktur der latenten Markov-Kette und leiten daraus, falls die zustandsbedingten Dichten zur selben parametrischen Familie gehören, eine eindeutige, minimale Darstellung des HMMs ab.

Zunächst analysieren wir die Abhängigkeitsstruktur der Markov-Kette isoliert und definieren dazu eine Funktion λ auf dem Raum der Übergangsmatrizen mit fixierter Anzahl von Zuständen. Konkret bezeichne $\mathbf{\Gamma}$ die Übergangsmatrix einer Markov-Kette mit k Zuständen und $\mathcal{G} = \{G_1, \ldots, G_r\}$ eine Partition des Zustandsraums $\{1, \ldots, k\}$. Die abgebildete Übergangsmatrix ist dann gegeben durch

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

wobei $G(j) = l \Leftrightarrow j \in G_l$, $j = 1, \ldots, k$. Da die Mengen G_l als Gruppen von Zuständen interpretiert werden können, ist die abgebildete Übergangsmatrix gegeben durch die Übergangswahrscheinlichkeiten zwischen den einzelnen Gruppen und die bedingte Wahrscheinlichkeit in der eingetretenen Gruppe. Die Abbildung kann daher als Reduzierung der Information bzgl. des Zustandsübergangs der Markov-Kette angesehen werden. Das erste Hauptresultat ist die eindeutige Existenz einer minimalen Partition $\mathcal{G}^*_{\mathbf{\Gamma}}$ mit $\lambda_{\mathcal{G}^*_{\mathbf{\Gamma}}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$. Hierbei bezieht sich minimal auf die Anzahl der Mengen in der Partition.

Wir zeigen, dass ein beliebiges HMM mit Übergangsmatrix $\mathbf{\Gamma}$ und k Zuständen der Markov-Kette äquivalent parametrisiert werden kann durch ein HMM mit r Zuständen und zustandsbedingten endlichen Mischungen, falls $\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ gilt. Hierbei ist $\mathcal{G} = \{G_1, \ldots, G_r\}$ eine Partition des Zustandsraums $\{1, \ldots, k\}$ und äquivalent bezieht sich auf identische Verteilungen des beobachtbaren Prozesses unter beiden Parametrisierungen. Wir erhalten die gewünschte eindeutige, minimale Darstellung als Korollar, falls Mischungen mit k Komponenten der verwendeten parametrischen Familie identifizierbar sind. Die resultierende Partition bezeichnen wir als *independence partition* und die darin enthaltenen Mengen als *independence clusters*. Wir bezeichnen ferner die beschriebene Umparametrisierung als *merging*, denn mehrere Zustände werden als ein neuer Zustand interpretiert, welcher eine endliche Mischung als zustandsbedingte Verteilung hat.

Zum Zweck der Modellwahl und in dem Fall, dass die zustandsbedingten Dichten einer bekannten parametrischen Familie angehören, schlagen wir einen Likelihood-Quotienten-Test vor. Der Test basiert auf der vollen Likelihood des Modells $L_T(\mathbf{\Gamma}, \boldsymbol{\theta})$, siehe Abschnitt 1.2. Genauer sei $\mathcal{G} = \{G_1, \ldots, G_r\}$ erneut eine Partition des Zustandsraums $\{1, \ldots, k\}$. Wir testen

$$H_0 : \lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma} \text{ gegen } H_1 : \lambda_{\mathcal{G}}(\mathbf{\Gamma}) \neq \mathbf{\Gamma}.$$

Unter der Annahme, dass der wahre Parameter ein innerer Punkt des Nullhypothesen-Parameterraums ist, und unter weiteren Regularitätsannahmen zeigen wir, dass die Likelihood-Quotienten-Test Statistik asymptotisch $\chi^2_{\cdot}$-verteilt ist mit k^2-2k-r^2+2r Freiheitsgraden. Für die Normalverteilung bzw. die multivariate Normalverteilung sind die Regularitätsbedingungen erfüllt, falls man die Skalenparameter bzw. die Determinanten der Kovarianzmatrizen von Null weg beschränkt. Das Verhalten bei endlichen Stichprobengrößen untersuchen wir mittels einer umfangreichen Simulationsstudie.

Da merging im Allgemeinen die Verteilung des beobachtbaren Prozesses ändert, schlagen wir einen Algorithmus vor, welcher obigen Likelihood-Quotienten-Test verwendet und dazu dient, iterativ mittels backward selection die independence partition zu finden. Für einen gegebenen Datensatz und eine gegebene parametrische Familie von Dichten startet der Algorithmus mit der trivialen Partition $\mathcal{G}_0 = \{\{1\}, \ldots, \{k\}\}$. In jedem Schritt werden alle Partitionen betrachtet, die durch Vereinigung zweier Mengen entstehen und für jede Partition wir der Test durchgeführt. Die neue Partition ist dann gegeben durch diejenige, welche beim vorherigen Testen den größten P-Wert erzeugte oder die Iteration wird gestoppt, falls alle Tests bzgl. eines vorgegebenen Niveaus $\alpha > 0$ verworfen werden können. Eine Simulationsstudie zeigt, dass $\alpha = 0.01$ eine gute Wahl für das Niveau ist.

Im Fall von u.i.v. Mischungsmodellen sollten zustandsbedingte Verteilungen, die ein dichtebasiertes Cluster bilden, als eine einzelne Komponente mit einer Mischung als zustandsbedingte Verteilung betrachtet werden, denn unter dieser Voraussetzung ist eine aussagekräftige maximum-a-posteriori Analyse möglich, siehe Baudry et al. (2010) oder Hennig (2010). Für HMMs ist die Situation komplizierter. Falls die Daten stark seriell abhängig sind, lassen sich Zustände, deren zustandsbedinge Verteilungen sich marginal stark überlappen, dennoch gut trennen. Zustände sollten nur kombiniert werden, falls die Abhängigkeitsstruktur dies ermöglicht, d.h. wenn die Zustände im selben independence Cluster liegen und wenn die zustandsbedingten Verteilungen ein dichtebasiertes Cluster darstellen. Daher schlagen wir einen entropiebasierten Algorithmus vor, welcher iterativ dichtebasierte Cluster innerhalb der independence partition findet. Für verschiedene Kandidatenmodelle wird die Entropie der "local-decoding" Wahrscheinlichkeiten berechnet und das Modell mit dem minimalen Wert ausgewählt. Diese Prozedur wird solange iteriert, bis alle möglichen Zustände bzgl. der independence partition kombiniert wurden. Das finale Modell ist schließlich gegeben durch das Modell des letzten Iterationsschritts oder falls die Entropiewerte einen Ellenbogen aufweisen wird das zugehörige Modell ausgewählt.

Wir wenden unsere Methodik mit normalverteilten zustandsbedingten Verteilungen auf tägliche log-returns des Goldpreises an, wobei der Datensatz einen Zeitraum von 15 Jahren umfasst. Die Informationskriterien AIC bzw. BIC wählen sechs bzw. vier Zustände. Wir starten mit sechs Zuständen und erhalten durch iteratives Testen und entropiebasiertes merging ein HMM mit vier Zuständen. Zwei der zustandsbedingten Verteilungen sind gegeben durch Normalverteilungen. Die beiden übrigen bestehen jeweils aus einer Mischung mit zwei Komponenten.

Peaks vs Components

Quah (1996) unterteilt Länder der Welt in eine reiche und eine arme Gruppe, indem er Moden der Dichte des BIP mit Wohlfahrtsgruppen assoziiert. In "Peaks vs Components" zeigen wir zunächst, dass diese Vorgehensweise zu mehrdeutigen Schlussfolgerungen führen kann, da die so erzielten Ergebnisse nicht invariant unter Veränderung der Skala sind. Konkret kann die Anzahl der Moden der Dichteschätzung variieren, wenn z.B. die log-BIP Daten betrachtet werden.

In unserer Arbeit wenden wir endliche Mischungsmodelle an, um Wohlfahrtsgruppen von Ländern der gesamten Welt zu analysieren. Wir betrachten das BIP pro Kopf von 190 Ländern in den Jahren 1970 bis 2009, wie es in den "Penn World tables 7.0"veröffentlicht wurde, siehe Heston, Summers and Aten (2011). Anstatt Moden der Dichte ziehen wir Komponenten des Mischungsmodells als Entscheidungskriterium der Gruppenzugehörigkeit heran. Die Anwendung solcher Modelle ist nicht neu in der volkswirtschaftlichen Literatur. Paap and Dijk (1998) verwenden ein Mischungsmodell mit zwei Komponenten bestehend aus einer Normal- und einer Weibullverteilung um das BIP zu modellieren. In unserer Arbeit hinterfragen wir den Ansatz zweier Komponenten kritisch und schlagen die Verwendung endlicher Mischungsmodelle bestehend aus Normalverteilungen mit variablem Lokations- und Skalenparameter als Alternative vor.

Zunächst betrachten wir den modenbasierten Ansatz und wenden diesen auf die Kerndichteschätzung an. Um eine geeignete Bandbreite zu finden verwenden wir den Silvermantest, siehe Silverman (1981), und erhalten somit für jedes

Jahr eine validierte Anzahl von Moden. Wir finden dadurch Belege für zunächst zwei (1970-1990) und anschließend für drei Moden.

Im zweiten Schritt untersuchen wir den Paneldatensatz, indem wir für jedes Jahr ein Mischungsmodell mit zustandsabhängigen Normalverteilungen schätzen. Wir finden und validieren die Anzahl der Komponenten durch iteratives Testen. Da die Likelihood des hier betrachteten Mischungsmodells unbeschränkt ist und für $\sigma_j \to \infty$ divergiert, kann ein gewöhnlicher Likelihood-Quotienten-Test nicht angewendet werden. Daher verwenden wir einen Test, der auf einer penalisierten Version der log-Likelihood, wie in in Chen and Li (2009) vorgeschlagen, basiert. Diese penalisierte Version der log-Likelihood resultiert durch Addieren von

$$-0.05 \sum_{j=1}^{m} \left(s_n^2/\sigma_j^2 + \log\left(\sigma_j^2/s_n^2\right)\right)$$

zur log-Likelihood. Hierbei bezeichnet s_n die empirische Standardabweichung des Datensatzes und σ_j die Standardabweichung der j-ten Komponente. Die finale Testentscheidung wird schließlich auf Grund kritischer Werte, die durch ein parametrisches Bootstrap-Verfahren ermittelt wurden, gefällt.

Die durchgeführten Tests führen zu Mischungsmodellen mit zunächst drei (1970-1995) und anschließend zwei Komponenten (1996-2009). Auf Grund deutlicher Unterschiede der Lokationsparameter, relativ gesehen zu den Skalanparametern, können die drei Komponenten als Länder mit niedrigem, mittlerem und hohem Einkommen interpretiert werden. Analog lassen sich die zwei Komponenten der Mischungen der Jahre nach 1995 als Länder mit geringerem und höherem Einkommen interpretieren. Zusätzlich kann die ärmere Komponente der Mischungen nach 1995 als Vereinigung zweier vorheriger Komponenten angesehen werden. Indem wir abschließend die a-posteriori Wahrscheinlichkeiten berechnen, erhalten wir detaillierte Erkenntnisse über die Gruppenzugehörigkeit aller Länder im Zeitablauf.

References

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.

Bickel, P. J., Ritov, Y. and Rydén, T. (1998). Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models. *The Annals of Statistics*, **26** 1614–1635.

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.

Chen, H., Chen, J. and Kalbfleisch, J. (2004). Testing for a Finite Mixture Model with Two Components. *Journal of the Royal Statistical Society: Series B*, **66** 95–115.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A Modified Likelihood Ratio Test for Homogeneity in Finite Mixture Models. *Journal of the Royal Statistical Society: Series B*, **63** 19–29.

Chen, J. and Li, P. (2009). Hypothesis Test for Normal Mixture Models: The EM Approach. *Annals of Statistics*, **37** 2523–2542.

Chen, J., Li, P. and Fu, Y. (2012). Inference on the Order of a Normal Mixture. *Journal of the American Statistical Association*, **107** 1096–1105.

Dannemann, J. and Holzmann, H. (2008). Testing for two states in a hidden Markov model. *Canadian Journal of Statistics*, **36** 505–520.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, **39** 1–38.

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

Gassiat, E. and Keribin, C. (2000). The likelihood ratio test for the number of components in a mixture with markov regime. *ESAIM: Probability and Statistics*, **4** 25–52.

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.

HESTON, A., SUMMERS, R. and ATEN, B. (2011). Penn World Tables Version 7.0. *Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.*

HOLZMANN, H. and SCHWAIGER, F. (2013a). Hidden Markov models with state-dependent mixtures: Minimal representation, model testing and applications to clustering. *Preprint.*

HOLZMANN, H. and SCHWAIGER, F. (2013b). Testing for the number of states in hidden Markov models. *Preprint.*

KERIBIN, C. (2000). Consistent Estimation of the Order of Mixture Models. *Sankhyā: The Indian Journal of Statistics, Series A* 49–66.

LEROUX, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, **40** 127–143.

LI, P. and CHEN, J. (2010). Testing the Order of a Finite Mixture. *Journal of the American Statistical Association*, **105** 1084–1092.

LINDGREN, G. (1978). Markov Regime Models for Mixed Distributions and Switching Regressions. *Scandinavian Journal of Statistics* 81–91.

MCLACHLAN, G. (1987). On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Applied Statistics*, **36** 318–324.

PAAP, R. and DIJK, H. K. (1998). Distribution and Mobility of Wealth of Nations. *European Economic Review*, **42** 1269–1293.

QUAH, D. (1996). Twin peaks: Growth and convergence in models of distribution dynamics. *Economic Journal*, **106** 1045–1055.

RABINER, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, **77** 257–286.

RYDÉN, T., TERÄSVIRTA, T. and ÅSBRINK, S. (1998). Stylized Facts of Daily Return Series and the Hidden Markov Model. *Journal of Applied Econometrics*, **13** 217–244.

SILVERMAN, B. W. (1981). Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society: Series B*, **43** 97–99.

TEICHER, H. (1963). Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, **34** 1265–1269.

VITERBI, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13** 260–269.

Vollmer, S., Holzmann, H. and Schwaiger, F. (2013). Peaks vs components. *Review of Development Economics*, **17** 352–364.

Yakowitz, S. J. and Spragins, J. D. (1968). On the Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, **39** 209–214.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series*. Chapman & Hall.

Testing for the number of states in hidden Markov models

Hajo Holzmann and Florian Schwaiger

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

Under the mixture of distributions hypothesis asset returns marginally follow a scale mixture of a certain, e.g. the normal, distribution, a simple specification being a three component scale mixture with high, intermediate and low volatility states. We propose tests for the number of states in hidden Markov models, and use these to assess whether in view of recent financial turbulences, three volatility states are still sufficient. Based on a quasi-likelihood which neglects the dependence structure of the regime, our tests extend existing tests for independent finite mixtures. Here, our main theoretical insight is the surprising fact that the asymptotic distribution of the proposed tests for HMMs is the same as for independent mixtures with corresponding weights. Thus, our results also state that existing tests for independent mixtures are indeed robust against Markov-dependence in the regime. As application we determine the number of volatility states for logarithmic returns of the S&P 500 index in two HMMs, one with state-dependent normal distributions and switching mean and scale, and the other with state-dependent skew-normal distributions with switching scale and structural mean and skewness parameters. It turns out that in both models, four states are indeed required, and a maximum-a-posteriori analysis shows that the highest volatility state mainly corresponds to the recent financial crisis. Finally, simulations indicate the good finite sample performance of the proposed testing methodology.

Keywords: hypothesis testing, hidden Markov models, volatility states, finite mixtures

1. Introduction

The mixture of distributions hypothesis for asset returns refers to specifications for which the marginal distribution of the returns follows a scale-mixture of a certain, e.g. the normal, distribution (Shephard and Andersen 2009), thus generating heteroscedastic return volatility. A simple version is a finite scale-mixture of normals, as proposed in Kon (1984), typically with three states corresponding to high, intermediate and low volatility. In order to induce volatility clustering one additionally requires positive serial correlation of the latent scale process, e.g. via a stationary finite-state Markov chain with high diagonal entries (Rydén et al. 1998).

For the resulting class of processes, called hidden Markov models (HMMs), we shall propose tests with a tractable asymptotic distribution for the number of states of the underlying unobserved regime, and investigate whether in view of recent financial turbulences, three volatility states are still sufficient.

More precisely, an HMM is a bivariate process $(S_t, X_t)_{t\geq 1}$, where $(S_t)_{t\geq 1}$ is an unobservable, finite-state Markov chain and $(X_t)_{t\geq 1}$ is the observable process with values in some Borel-subset of a Eucledian space, which are related as follows. Given $(S_t)_{t\geq 1}$, the $(X_t)_{t\geq 1}$ are conditionally independent, and for each $t \geq 1$, the conditional distribution of X_t depends on S_t only. The unobservable Markov chain is also called the regime or the latent process of the HMM. We shall assume that (S_t) is stationary and ergodic with state space $\mathcal{M} = \{1, \ldots, k\}$, so that the stationary distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ of the associated transition matrix $\gamma_{lm} = P(S_{t+1} = m | S_t = l)$, $l, m \in \mathcal{M}$ is uniquely determined.

The conditional distributions of X_t given $S_t = l$, $l = 1, \ldots, k$, called the state-dependent distributions, are assumed to have densities $f(\cdot, \boldsymbol{\nu}, \boldsymbol{\vartheta}_l)$ from some parametric family w.r.t. some σ-finite measure. Thus, $\boldsymbol{\nu} \in \Theta_1 \subset \mathbb{R}^{d_1}$ is a structural parameter and the $\boldsymbol{\vartheta}_l \in \Theta_2 \subset \mathbb{R}^{d_2}$ are actually state-dependent.

HMMs provide a flexible and very widely used class of models for dependent data, in particular in the presence of overdispersion (for series of count data) or unobserved heterogeneity, see the monographs by MacDonald and Zucchini (1997) for further examples of applications, and by Cappé *et al.* (2005) for a state-of-the-art overview of theoretical developments for HMMs.

In statistical applications of HMMs, selection of the number of states k of the latent process is a task of major importance. To this end, in certain models for fixed $k_0 \in \mathbb{N}$ we shall propose tests for the hypothesis

$$H_0 : k = k_0 \qquad \text{against} \qquad H_1 : k > k_0.$$

Since Gassiat and Keribin (2000) show that the LRT statistic for testing $k = 1$ against $k \geq 2$ for an HMM diverges to ∞, we shall follow the quasi-likelihood based approach in Lindgren (1978) and in Dannemann and Holzmann (2008) and proceed via the marginal finite mixture.

Specifically, we use the testing approaches for the number of states in a finite mixture by Chen, Li and Fu (2012) for normal state-dependent distributions with switching means and scales, as well as that by Li and Chen (2010) for a univariate switching parameter, extended to allow for nuisance parameters. Our main theoretical insight is the surprising fact that the asymptotic distribution of these tests for HMMs is the same as for independent mixtures with corresponding weights. Thus, our results also state that existing tests for independent mixtures are indeed robust against Markov-dependence in the regime.

The structure of the paper is as follows. In Section 2 we develop the relevant testing methodology. Section 3 contains results of an extensive simulation study.

As application, in Section 4 we determine the number of volatility states for logarithmic returns of the S&P 500 index in two HMMs, one with states dependent normal distributions and switching mean and scale, and the other with state-dependent skew-normal distributions with switching scale and structural mean and skewness parameters. It turns out that in both models, four states are indeed required, and a maximum-a-posteriori analysis shows that the highest volatility state mainly corresponds to the recent financial crisis.

The Appendix contains a proof of the main insight that the asymptotic distribution of the test by Chen et al. (2012) remains the same for HMMs as for independent finite mixtures. The supplementary Appendix B contains technical details for the asymptotic distribution theory, while Appendix C contains details on the finite-sample tuning of the tests, additional simulation results as well as results of an application to oil price logarithmic returns.

2. Quasi-likelihood based estimation and testing

2.1 Quasi-likelihood estimation

Following Lindgren (1978) and Dannemann and Holzmann (2008), we consider a quasi log-likelihood which neglects the dependence in the regime. For given number of states k, set $\boldsymbol{\theta} = \boldsymbol{\theta}(k) = (\boldsymbol{\nu}^T, \boldsymbol{\vartheta}_1^T, \ldots, \boldsymbol{\vartheta}_k^T)^T \in \Theta^{(k)} = \Theta_1 \times \Theta_2^k$,

$$f_{mix}^{(k)}(X_t; \boldsymbol{\theta}, \boldsymbol{\pi}) := \sum_{j=1}^{k} \pi_j \; f\big(X_t | S_t = j; \boldsymbol{\theta}\big) = \sum_{j=1}^{k} \pi_j \; f(X_t; \boldsymbol{\nu}, \boldsymbol{\vartheta}_j),$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ with $\pi_j \geq 0$, $\pi_1 + \ldots + \pi_k = 1$, and

$$l_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{t=1}^{n} \log\big(f_{mix}^{(k)}(X_t; \boldsymbol{\theta}, \boldsymbol{\pi})\big).$$

The quasi maximum-likelihood estimator (QMLE) is then given by

$$(\widehat{\boldsymbol{\theta}}(k), \widehat{\boldsymbol{\pi}}(k)) := (\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}) := \arg\max\{l_n^{(k)}(\boldsymbol{\theta}, \boldsymbol{\pi}) : \boldsymbol{\theta} \in \Theta^{(k)}, \sum_{j=1}^{k-1} \pi_j \leq 1,\ \pi_j \geq 0\}.$$

We are mainly interested in two specific situations, for which we intend to extend the testing methodology for mixtures to the case of HMMs.

Example 1 (*Normal HMMs)*. One of the most important classes of HMMs are those with normal state-dependent distributions. If both mean μ and variance σ^2 are allowed to switch, we have that

$$f_{X_t|S_t=j}(x) = \phi(x; \mu_j, \sigma_j), \qquad j = 1, \dots, k,$$

where ϕ denotes the normal density. We let $\phi_{mix}^{(k)}(x; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$ denote the corresponding k-component normal mixture. Without compactness assumption on the parameter space, and thus in particular a lower bound for the standard deviations σ_j, it is well known that the quasi log-likelihood $l_n^{(k)}$ is unbounded. However, there are ways of dealing with this issue, in particular by adding penalty terms, see e.g. Chen, Tan and Zhang (2008). We shall follow the approach by Chen et al. (2012), and use the penalized quasi log-likelihood

$$pl_n^{(k)}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \sum_{t=1}^{n} \log\left(\phi_{mix}^{(k)}(X_t; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})\right) \ + \ p^{(k)}(X_1, \dots, X_n, \boldsymbol{\sigma}), \tag{1}$$

where

$$p^{(k)}(X_1, \dots, X_n, \boldsymbol{\sigma}) = -\frac{1}{n} \sum_{j=1}^{k} \left\{\hat{s}_n^2/\sigma_j^2 + \log(\sigma_j^2/\hat{s}_n^2)\right\}$$

and $\hat{s}_n^2$ is the empirical variance of $X_1, \dots, X_n$.

Example 2 (*Univariate switching parameters)*. Here the parameter space Θ_2 for the switching parameter is one-dimensional. Important special cases are Poisson HMMs (for which there are no nuisance parameters), as well as HMMs with normal or more generally skew-normal state-dependent distributions, where only a single parameter is allowed to switch.

For the normal distribution, the asymptotic theory below does not apply to the case of a switching mean with a structural variance. However, it does apply in case of a switching scale parameter, if the other parameters are taken as structural. In this case, a lower bound or a penalty on the scale parameters is again required to avoid unbounded quasi-likelihoods and achieve consistency.

We shall present our general asymptotic theory for the case of a one-dimensional switching parameter without penalty function (which is not required for Poisson HMMs), under conditions which guarantee consistency of MLE of the mixing distribution in case of i.i.d. observations. Thus, lower bounds on the scale parameter for (skew-) normal HMMs with switching scale parameter are required. Nevertheless, we also investigate penalty functions on the scale parameter in our simulation study.

2.2 Testing for the number of states: Normal HMMs

First, we consider normal HMMs with switching means and standard deviations.

The testing procedure is a simplified version of that proposed by Chen et al. (2012) in case of independent finite mixtures, see below for further comments.

Our main contribution is to show that the asymptotic distribution remains unchanged if we pass from an independent mixture to an HMM with same the marginal mixture. This is quite surprising since, for example, the asymptotic distribution of the QMLE under an HMM is quite different from that of the MLE in case of independent mixtures. Thus, our results show that the test by Chen et al. (2012) is robust w.r.t. Markov dependence in the regime.

Let

$$\big(\widehat{\boldsymbol{\mu}}(k_0), \widehat{\boldsymbol{\sigma}}(k_0), \widehat{\boldsymbol{\pi}}(k_0)\big) = \big(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}}, \widehat{\boldsymbol{\pi}}\big) := \arg\max_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}} pl_n^{(k_0)}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$$

denote the (penalized quasi-likelihood) estimates under k_0-states, where $pl_n^{(k_0)}$ is defined in (1). We assume that the estimated means $\hat{\mu}_1 < \ldots < \hat{\mu}_{k_0}$ are increasingly ordered.

The test by Chen et al. (2012) is against a specific class of mixtures with $2k_0$ components. To define it, consider the set of $2k_0$-dimensional weight vectors

$$\begin{aligned}\Omega_{2k_0}(\mathcal{J}) :=& \Big\{\big(\pi_1\beta_1,\ \pi_1(1-\beta_1),\ \ldots,\ \pi_{k_0}\beta_{k_0},\ \pi_{k_0}(1-\beta_{k_0})\big) : \\ & \beta_j \in \mathcal{J},\ \sum_{j=1}^{k_0} \pi_j = 1,\ \pi_j \geq 0\Big\},\end{aligned} \tag{2}$$

where $\mathcal{J} \subset (0, 0.5]$ with $0.5 \in \mathcal{J}$ is a finite set (we shall use $\mathcal{J} = \{0.1, 0.3, 0.5\}$).

Further, partition the real line into k_0 subsets by $I_j := (\eta_{j-1}, \eta_j]$, where $\eta_0 = -\infty, \eta_{k_0} = \infty$ and $\eta_j = (\widehat{\mu}_j(k_0) + \widehat{\mu}_{j+1}(k_0))/2$ for $j = 1, \ldots, k_0 - 1$.

Then estimate the specific alternative $2k_0$-state normal mixture model with weights in $\Omega_{2k_0}(\mathcal{J})$ and two successive μ's in each set I_j as follows

$$\begin{aligned}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{\boldsymbol{\pi}}) = \arg\max \big\{\widetilde{pl_n}^{(2k_0)}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) :\ \boldsymbol{\pi} \in \Omega_{2k_0}(\mathcal{J}), \mu_{2j-1}, \mu_{2j} \in I_j,\\ j = 1, \dots, k_0, \boldsymbol{\sigma} \in \mathbb{R}_+^{2k_0}\big\},\\ \widetilde{pl_n}^{(2k_0)}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}) = \sum_{t=1}^{n} \log\big(\phi_{mix}^{(2\,k_0)}(X_t; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})\big) - a_n \sum_{j=1}^{k_0} w\big(\sigma_{2j-1}, \sigma_{2j}, \hat\sigma_j(k_0)\big), \quad (3)\\ w\big(\sigma_{2j-1}, \sigma_{2j}, \hat\sigma_j(k_0)\big) = \hat\sigma_j^2(k_0)/\sigma_{2j-1}^2 - 1 + \log(\sigma_{2j-1}^2/\hat\sigma_j^2(k_0))\\ + \hat\sigma_j^2(k_0)/\sigma_{2j}^2 - 1 + \log(\sigma_{2j}^2/\hat\sigma_j^2(k_0)),\end{aligned}$$

where $a_n > 0$ is a tuning constant, the choice of which is further discussed below. Finally, the quasi-likelihood ratio test statistic is then given by

$$Z_n = 2\Big(\widetilde{pl_n}^{(2k_0)}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{\boldsymbol{\pi}}) - l_n^{(k_0)}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}}, \widehat{\boldsymbol{\pi}})\Big). \quad (4)$$

Assumption A1. $(X_t, S_t)_t$ is a hidden Markov model. The Markov chain $(S_t)_t$ is finite-state, stationary, irreducible and aperiodic. $\diamond$

Theorem 1. *If $(X_t, S_t)_t$ is a k_0-state normal HMM fulfilling assumption A1 with distinct means μ_i in each state, we have for the quasi-likelihood test statistic that*

$$Z_n \xrightarrow{d} \chi^2_{2k_0}. \quad (5)$$

Remarks a. *Fixed proportions and EM iterations.* We test against fixed proportions β_j only, and do not perform EM-iterations since this does not seem to increase the power substantially, and requires an additional penalty function on the proportions. However, the EM-version is possible as well, and the asymptotic theory also applies.

b. *Tuning parameters.* The test statistic requires the choice of a tuning parameter a_n in $\widetilde{pl_n}^{(2k_0)}$, and in fact, since the penalty term involving this tuning constant is not only used for estimation, but is also included in the test statistic in (4), the finite-sample performance crucially depends on its choice. For a range of possible values of the true mixture, Chen et al. (2012) give recommendations based on simulations. Since our test statistic is slightly different (no penalties on the proportions), we also need to calibrate a_n distinctly. See the simulations for details. Further, care is needed if the true underlying parameter constellation is far from those used for calibration of a_n, see Section 4.

2.3 Testing for the number of states: Univariate switching parameter

Next we consider the general case, but with a univariate switching parameter $\vartheta \in \Theta_2 \subset \mathbb{R}$. Here, we extend the test by Li and Chen (2010) to include nuisance parameters, and as above show that its asymptotic distribution remains unchanged if we pass from an independent mixture to an HMM with the same marginal mixture.

We shall write $\widehat{\boldsymbol{\theta}}(k_0) = \widehat{\boldsymbol{\theta}} = \big(\widehat{\boldsymbol{\nu}}, \widehat{\vartheta}_1, \dots, \widehat{\vartheta}_{k_0}\big)$ for the QMLE, where we assume that the entries of $\widehat{\boldsymbol{\vartheta}}$ are ordered: $\hat{\vartheta}_1 \leq \dots \leq \hat{\vartheta}_{k_0}$. As above, the test is against a specific class of mixtures with $2k_0$ components. The set of weights $\Omega_{2k_0}(\mathcal{J})$ is defined as in (2).

Partition Θ_2 into k_0 subsets by $I_j := (\eta_{j-1}, \eta_j]$, where $\eta_0 = \inf \Theta_2, \eta_{k_0} = \sup \Theta_2$ and $\eta_j = (\widehat{\vartheta}_j + \widehat{\vartheta}_{j+1})/2$ for $j = 1, \dots, k_0 - 1$. We further restrict the parameter space of the state-dependent parameters, such that each parameter in $\widehat{\boldsymbol{\vartheta}}$ is possibly split into two components within the interval I_j. Thus, we set

$$\begin{aligned} R_n &= 2\Big(l_n^{(2k_0)}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\pi}}) - l_n^{(k_0)}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}})\Big), \\ (\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\pi}}) &= \arg\max \big\{ l_n^{(2k_0)}(\boldsymbol{\theta}, \boldsymbol{\pi}) : \boldsymbol{\pi} \in \Omega_{2k_0}(\mathcal{J}), \boldsymbol{\theta} = (\boldsymbol{\nu}, \vartheta_1, \dots, \vartheta_{2k_0}), \\ &\qquad \boldsymbol{\nu} \in \Theta_1,\ \vartheta_{2j-1}, \vartheta_{2j} \in I_j,\ j = 1, \dots, k_0 \big\}. \end{aligned} \tag{6}$$

We now present the asymptotic distribution of the test statistic under the hypothesis of k_0 states, which is somewhat more involved and requires additional notation.

Let $\pi_j^* := P(S_t = j)$ for $j \in \{1, \dots, k_0\}$ denote the true stationary probability of the Markov chain for state j and $\boldsymbol{\pi}^* := \big(\pi_1^*, \dots, \pi_{k_0}^*\big)$, $\boldsymbol{\theta}^* = \big(\boldsymbol{\nu}^*, \vartheta_1^*, \dots, \vartheta_{k_0}^*\big)$ with $\vartheta_1^* < \dots < \vartheta_{k_0}^*$ and $\boldsymbol{\nu}^* := (\nu_1^*, \dots, \nu_{d_1}^*)$. The assumption of irreducibility, see A1, implies $\pi_j^* > 0$.

Assumption A2. The support of $f(x; \boldsymbol{\nu}, \vartheta)$ does not depend on the parameter $(\boldsymbol{\nu}, \vartheta) \in \Theta_1 \times \Theta_2$. Further, the derivatives

$$\frac{\partial^{i_1}}{\partial \vartheta^{i_1}} f(x; \boldsymbol{\nu}, \vartheta) \quad \text{and} \quad \frac{\partial^{i_2+i_3+i_4+i_5}}{\partial \nu_l^{i_2} \partial \nu_i^{i_3} \partial \nu_h^{i_4} \partial \vartheta^{i_5}} f(x; \boldsymbol{\nu}, \vartheta)$$

where $i_1 = 1, 2, 3, 4$ and $i_2, i_3, i_4, i_5 = 0, 1, 2, 3$ with $i_2 + i_3 + i_4 + i_5 \leq 3$ exist for $l, i, h = 1, \dots, d_1$. $\diamond$

For $t = 1, \dots, n$, $j = 1, \dots, k_0$ and $l = 1, \dots, d_1$ let

$$Y_{tj}' := \frac{f_\vartheta(X_t; \boldsymbol{\nu}*, \vartheta_j^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}, \qquad Y_{tj}'' := \frac{f_{\vartheta\vartheta}(X_t; \boldsymbol{\nu}*, \vartheta_j^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)},$$

where f_ϑ denotes the partial derivative of $f(x; \boldsymbol{\nu}, \vartheta)$ w.r.t. ϑ. Note that Y_{tj}' is (except for the constant π_j^*) the partial derivative of $\log(f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}, \boldsymbol{\pi}))$ w.r.t. the state-dependent parameter ϑ_j evaluated at the true values of the parameters. Further, set

$$U_t^{\{l\}} := \frac{f_{mix,\nu_l}^{(k_0)}(X_t; \boldsymbol{\nu}^*, \vartheta_1^*, \dots, \vartheta_k^*, \boldsymbol{\pi}^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)},$$

where $f^{(k_0)}_{mix,\nu_l}$ is the partial derivative of $f^{(k_0)}_{mix}$ w.r.t. ν_l, and in addition, let

$$\Delta_{tj} := \frac{f(X_t;\boldsymbol{\nu}^*,\vartheta_j^*) - f(X_t;\boldsymbol{\nu}^*,\vartheta_{k_0}^*)}{f^{(k_0)}_{mix}(X_t;\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)}.$$

Set $\boldsymbol{b}_t^T := \big(\boldsymbol{b}_{1t}^T, \boldsymbol{b}_{2t}^T\big)$ where

$$\begin{aligned}\boldsymbol{b}_{1t}^T &= \Big(U_t^{\{1\}},\dots,U_t^{\{d\}},\Delta_{t1},\dots,\Delta_{tk_0-1},Y'_{t1},\dots,Y'_{tk_0}\Big),\\ \boldsymbol{b}_{2t}^T &= \big(Y''_{t1},\dots,Y''_{tk_0}\big).\end{aligned}$$

Further, let $\boldsymbol{\Sigma} := \mathrm{COV}(\boldsymbol{b}_1)$ with submatrices $\boldsymbol{\Sigma}_{ij} := E([b_{i1} - E(b_{i1})][b_{j1} - E(b_{j1})]^T)$, $i,j = 1,2$, (the moments exist by Assumption A4 below), and define

$$\tilde{\boldsymbol{b}}_{2t} := \boldsymbol{b}_{2t} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{b}_{1t}, \qquad \widetilde{\boldsymbol{\Sigma}}_{22} := \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \mathrm{COV}(\tilde{\boldsymbol{b}}_{2t}).$$

All expected values are computed w.r.t. the true parameters $\boldsymbol{\theta}^* = (\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*)$ with $\vartheta_1^* < \dots < \vartheta_{k_0}^*$ and $\boldsymbol{\Gamma}^*$, the true underlying transition matrix. The next assumption guarantees consistency of the marginal mixing distribution.

Assumption A3 (*Consistency of the mixing distribution*)**.** Assume that $\Theta_1 \subset \mathbb{R}^{d_1}$, $\Theta_2 \subset \mathbb{R}$ are closed, and that

a. $E\big|\log f^{(k_0)}_{mix}(X_1;\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)\big| < \infty$,

b. $\lim_{\|(\boldsymbol{\nu},\vartheta)\|\to\infty} f(x;\boldsymbol{\nu},\vartheta) = 0$ for all x,

c. The density $f(x;\boldsymbol{\nu},\vartheta)$ is uniformly bounded in $\boldsymbol{\nu} \in \Theta_1$, $\vartheta \in \Theta_2$ and in x.

d. Finite mixtures in $f(x;\boldsymbol{\nu},\vartheta)$ with structural $\boldsymbol{\nu}$ are identifiable.

More refined conditions are possible, see e.g. Leroux (1992). However, mixtures of Poisson distributions, (skew-)normal distributions and t-distributions with lower bound on the scale parameter, which are of main interest here, satisfy the assumption. For details on part d. see e.g. Holzmann, Munk and Gneiting (2006).

The next assumption requires that the components in the score are, locally around the true parameter values, uniformly dominated by an integrable function. It is essential for the asymptotic expansion of the quasi likelihood test statistic R_n in (6). In the supplementary Appendix B, we show that the following two assumptions are satisfied for the skew-normal distribution with fixed skewness parameter, structural location and switching scale as well as for the t-distribution with fixed degrees of freedom, and either fixed scale and switching location or fixed location and switching scale.

Assumption A4. Given $\epsilon > 0$ let $E_1(\boldsymbol{\nu}, \epsilon) := \{\boldsymbol{\nu}' \in \Theta_1 : ||\boldsymbol{\nu}' - \boldsymbol{\nu}|| \leq \epsilon\}$ for $\epsilon > 0$ and $E_2(\vartheta, \epsilon) := \{\vartheta' \in \Theta_2 : |\vartheta' - \vartheta| \leq \epsilon\}$, and set

$$Y_t'(\boldsymbol{\nu}, \vartheta) := \frac{f_\vartheta(X_t; \boldsymbol{\nu}, \vartheta)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}, \ldots, Y_t''''(\boldsymbol{\nu}, \vartheta) := \frac{f_{\vartheta\vartheta\vartheta\vartheta}(X_t; \boldsymbol{\nu}, \vartheta)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)},$$
$$\Delta_{tj}(\boldsymbol{\nu}) := \frac{f(X_t; \boldsymbol{\nu}, \vartheta_j^*) - f(X_t; \boldsymbol{\nu}, \vartheta_{k_0}^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}.$$

There exists an integrable function g, i.e. $E\big|g(X_t)\big| < \infty$, and an $\epsilon_0 > 0$, such that for $\boldsymbol{\nu} \in E_1(\boldsymbol{\nu}^*, \epsilon_0)$ and $\vartheta \in \bigcup_{j=1}^{k_0} E_2(\vartheta_j^*, \epsilon_0)$, we have that

$$\big|\Delta_{tj}(\boldsymbol{\nu})\big|^3 \leq g(X_t), \ \big|Y_t'(\boldsymbol{\nu}, \vartheta)\big|^3 \leq g(X_t), \ \ldots \ , \ \big|Y_t''''(\boldsymbol{\nu}, \vartheta)\big|^3 \leq g(X_t),$$

$$\Big|\frac{\partial^{i_1+i_2+i_3+i_4}/(\partial\nu_l^{i_1}\partial\nu_i^{i_2}\partial\nu_h^{i_3}\partial\vartheta^{i_4})f(X_t; \boldsymbol{\nu}, \vartheta)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}\Big|^3 \leq g(X_t),$$

for $i_1 + i_2 + i_3 + i_4 \leq 3, \ i_m \geq 0$ and $l, i, h = 1, \ldots, d_1$. $\diamond$

Finally, we need the following assumption, which guarantees that an expansion in terms of second derivatives of the switching parameters suffices. For the validity, see the supplementary Appendix B.

Assumption A5. The covariance matrix $\boldsymbol{\Sigma} = \mathrm{COV}(\boldsymbol{b}_1)$ is positive definite. $\diamond$

Theorem 2. *Under the hypothesis of k_0 states, for the test statistic R_n we have under the Assumptions A1-A5 that*

$$R_n \xrightarrow{d} \sum_{j=0}^{k_0} \alpha_j \chi_j^2, \tag{7}$$

where

$$\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_{22}), \qquad \hat{\boldsymbol{v}} := \arg\sup_{\boldsymbol{v} \geq 0} \Big(2\boldsymbol{v}'\boldsymbol{w} - \boldsymbol{v}'\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{v}\Big), \qquad \alpha_j = P\Big(\sum_{h=1}^{k_0} I\left(\hat{v}_h > 0\right) = j\Big)$$

and $\{\boldsymbol{v} \geq 0\} := \{(v_1, \ldots, v_{k_0}) : v_j \geq 0, \ j = 1, \ldots, k_0\}$.

Remark. *(Estimating parameters in the asymptotic distribution)* The asymptotic distribution in (7) depends on parameters through $\widetilde{\boldsymbol{\Sigma}}_{22}$ and has to be estimated. To this end, the true parameters in the vectors $\boldsymbol{b}_t$ are replaced by estimators, leading to $\hat{\boldsymbol{b}}_t$, say. Then $\boldsymbol{\Sigma}$ is estimated as the empirical covariance matrix of the $\hat{\boldsymbol{b}}_t$. A more explicit form of the asymptotic distribution in case of $k_0 = 3$ states can be found in Li and Chen (2010).

3. Simulations

Here we present some of the results of an extensive simulation study of the proposed tests. Further simulation results as well as some technical details are provided in Appendix C. We investigate hidden Markov models with switching variance and skew-normal innovations as well as normal HMMs with both parameters switching, focusing on the hypotheses of $k^* = 2$ and $k^* = 3$ states. Throughout, we denote by $\Gamma = (\gamma_{lm})_{l,m=1,\dots,k_0}$ the transition probability matrix (t.p.m.) of the hidden Markov chain. For the simulations we apply the software R in the version 2.15 and compute 10000 repetitions for simulated sizes, and 2500 for power simulations respectively. For the calculation of the quasi maximum likelihood estimators, we use the function `constrOptim`.

	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6
$\gamma_{1,1}$	85	25	50	95	76.25	80
$\gamma_{2,2}$	85	25	50	80	5	20

(a) $k^* = 2$

	Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
$\gamma_{1,1}$	90	20	100/3	90	15	25
$\gamma_{1,2}$	5	40	100/3	10	85	50
$\gamma_{2,1}$	5	40	100/3	5	42.5	25
$\gamma_{2,2}$	90	20	100/3	80	32.5	50
$\gamma_{3,1}$	5	40	100/3	0	0	25
$\gamma_{3,3}$	90	20	100/3	70	50	25

(b) $k^* = 3$

Table 1: Transition probability matrices with $k^* = 2, 3$ states (in percent).

3.1 Normal HMMs with switching means and variances

The choice of the tuning parameter a_n in $\widetilde{pl}_n^{(2k_0)}$ is quite crucial for the finite-sample performance since the penalty term involving this tuning constant is not only used for estimation, but is also included in the test statistic in (4). Therefore we follow Chen et al. (2012) and Chen and Li (2011) by reproducing their tuning of this constant with exactly the same design as in the first mentioned paper, see Appendix C.1 for further details.

For analyzing the finite sample behavior of the test we simulate rejection rates for normal HMMs under the true hypotheses of two and three states. The two state HMM, denoted by Nor_1, has parameters $(\mu_1, \mu_2) = (-1.75, 1.75)$ and $(\sigma_1, \sigma_2) = (1, 1)$, the three state HMM, denoted by Nor_2, has parameters $(\mu_1, \mu_2, \mu_3) =$
$(-3.5, 0, 4.5)$ and $(\sigma_1, \sigma_2, \sigma_3) = (0.6, 1.2, 0.6)$. For each scenario the simulations are performed with six different t.p.m.s, i.e. for Nor_1 with t.p.m.s $\Gamma_1, \dots, \Gamma_6$ and for Nor_2 with $\Gamma_7, \dots, \Gamma_{12}$, see Tables 1a and 1b. Note that Γ_1, Γ_2 and Γ_3 have the same stationary distribution and Γ_3 is the i.i.d. benchmark. The same holds for the triples $(\Gamma_4, \Gamma_5, \Gamma_6)$, $(\Gamma_7, \Gamma_8, \Gamma_9)$ and $(\Gamma_{10}, \Gamma_{11}, \Gamma_{12})$. The results are listed in Table 2. One directly can observe, that the dependence structure has

little effect on the finite sample behavior. Further, due to tuning the penalty constants, the levels are quite accurate.

Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6
9.96	10.15	10.4	8.15	8.57	8.46
4.92	5.12	5.23	3.71	3.81	3.89
1.01	1.08	0.86	0.65	0.61	0.67

(a) $n = 200$ and Nor_1

Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6
10.91	10.93	10.67	9.15	9.24	9.26
5.72	5.52	5.67	4.45	4.85	4.22
1.16	1.32	1.09	0.75	0.91	0.74

(b) $n = 400$ and Nor_1

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
9.25	9.94	9.62	9.89	9.24	10.18
4.27	4.72	4.80	5.06	4.30	5.01
0.79	0.83	0.95	0.97	0.69	0.93

(c) $n = 200$ and Nor_2

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
10.65	10.65	11.42	11.26	11.59	11.18
5.51	5.41	5.78	5.41	5.65	5.76
1.22	0.91	1.15	0.98	1.17	1.11

(d) $n = 400$ and Nor_2

Table 2: Simulated levels in percent for normal hidden Markov models under the hypothesis of two and three regimes with different sample sizes n. Each table lists line by line the rejection rates on levels $10\%, 5\%$ and 1%.

3.2 Skew-Normal HMM with switching variances

We consider a location-scale family of the standardized skew-normal family, parametrized in the skewness ξ, see Appendix C.1 for further details. We refer to this family by $\mathcal{SN}(\mu, \sigma, \xi)$.

Now, we simulate from two- and three-state skew-normal HMMs with parameters listed in Table 3. When using this parametric family one has to face two challenges: The likelihood is unbounded for scale parameters converging to zero and for finite sample sizes there is a non-negligible probability for the shape parameter to diverge towards the boundary of its parameter space, see Azzalini and Arellano-Valle (2012).

First, the likelihood is unbounded for $\sigma_j \to 0$. Using appropriate fixed lower bounds theoretically solves this problem. However, in the simulations we use the penalties from the normal case in the estimation process. The test statistic itself remains the same as in (6), the penalty is not included here, thus, the penalty does not play such an important role in this context. The penalization of the scale parameters depends on two tuning constants. Under the hypothesis we again use the inverse of the sample size and under the alternative we always use the constant 0.05, as it turned out that its choice is not that crucial.

Second, one obtains anti-conservative tests for moderate sample sizes (e.g. $n = 500$ for a three state HMM), if the skewness parameter is distinctly away from zero (e.g. $\xi = 0.3$), because estimates of the skewness parameter may diverge. Thus, we also penalize ξ in a data dependent way, see Appendix C.1.

Table 3 list the null distributions considered in the simulation study, the results for SN_2 and SN_4 are provided in Appendix C.2. The scenarios $SN_1 - SN_4$ correspond to cases under the hypothesis, the scenarios SP_1 and SP_2 are alternative scenarios.

For $SN_1 - SN_4$, each simulation is performed for sample sizes $n = 500$, $n = 1000$ and the same transition probability matrices as given in table 1. For the investigation of the finite sample behavior under the true hypothesis of two states we performed 10.000 repetitions. In case of three states we simulated 5.000 repetitions due to the higher computational complexity. The resulting sizes are displayed in Tables 4 and 8: The test performs well for all models under consideration.

Concerning power under alternatives SP_1 (when testing for two states) and SP_2 (testing for three states), for proper estimation of the power we use simulated critical values. Precisely, for given alternative, we generate a single large sample ($n = 25.000$) from this alternative and fit a corresponding null model to this sample by (full-model) maximum likelihood. Note that this will approximate the null model with minimal Kullback-Leibler distance to the given alternative. From this null model, we generate 2.500 samples of sizes 500 or 1000 (depending on the scenario), and in each case compute the test statistics. Finally, the finite sample critical values in the actual simulation are calculated as empirical quantiles of the previously performed simulation.

When testing for two states under a three state HMM, the results in Table 5a and 5b indicate that the test has good finite sample power properties, and that the power depends on the t.p.m. mainly through its stationary distribution.

For analysing the power under the false hypothesis of three states we simulated w.r.t. parameters given by SP_2 in table 3 and transition probability matrices Γ_{13} - Γ_{18}, given in the simulation appendix C.2. Again, $(\Gamma_{13}, \Gamma_{14}, \Gamma_{15})$ have the same stationary distribution, and so have $(\Gamma_{16}, \Gamma_{17}, \Gamma_{18})$. As before the first t.p.m. has persistent structure, the second is non-persistent and the third the i.i.d. benchmark. The simulated power is somewhat lower than in the above scenario when testing for two states (see tables 5c and 5d), but remains reasonably high, and also mainly depends on the t.p.m. through its stationary distribution.

	σ_1	σ_2	σ_3	σ_4	μ	ξ
SN_1	1	3			10	0.2
SN_2	2	7			0	−0.15
SN_3	1	10	30		−5	0.1
SN_4	1	3	9		5	−0.3
SP_1	1	5	10		−3	−0.25
SP_2	0.5	5	15	35	3	1/3

Table 3: Parameters of the skew-normal hidden Markov model.

Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6
9.37	9.47	9.28	8.77	8.53	9.24
4.80	4.94	4.51	4.57	4.58	5.18
1.02	1.09	0.95	1.04	1.22	1.18

(a) $n = 500$ and SN_1

Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6
9.44	9.51	9.91	8.76	9.13	8.90
4.80	4.99	4.94	4.37	4.57	4.50
0.86	1.05	1.08	0.87	1.03	1.00

(b) $n = 1000$ and SN_1

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
8.88	8.46	9.60	10.06	10.22	10.34
4.66	3.90	4.94	5.52	5.34	5.32
0.98	0.76	0.70	1.26	1.30	1.54

(c) $n = 500$ and SN_3

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
9.66	10.50	11.02	10.48	10.72	11.62
5.00	5.28	5.74	5.52	5.24	5.90
0.88	1.04	1.40	1.14	1.16	1.34

(d) $n = 1000$ and SN_3

Table 4: Simulated levels (10.000 and resp. 5.000 repetitions) in percent for skew-normal hidden Markov models under the hypothesis of two and three states with different sample sizes n and different parameters. Each table lists line by line the rejection rates on levels $10\%, 5\%$ and 1%.

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
81.16	85.68	86.16	92.04	94.64	94.28
71.08	78.08	77.44	86.80	89.88	88.64
47.48	53.96	56.44	71.84	74.36	71.12

(a) $n = 500$ and SP_1

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
97.40	97.76	98.00	99.88	99.92	99.92
94.40	96.12	95.88	99.64	99.56	99.64
83.72	89.72	87.68	97.52	98.28	97.96

(b) $n = 1000$ and SP_1

Γ_{13}	Γ_{14}	Γ_{15}	Γ_{16}	Γ_{17}	Γ_{18}
57.84	61.04	60.08	63.76	66.52	62.56
44.40	47.76	45.08	50.32	53.88	47.92
23.12	22.08	18.84	25.08	28.80	22.20

(c) $n = 500$ and SP_2

Γ_{13}	Γ_{14}	Γ_{15}	Γ_{16}	Γ_{17}	Γ_{18}
85.28	87.88	87.96	92.68	92.40	89.36
76.84	77.84	79.36	85.52	85.88	80.52
52.88	53.88	61.60	71.48	71.32	57.96

(d) $n = 1000$ and SP_2

Table 5: Simulated power (2.500 repetitions) in percent for skew-normal hidden Markov models for the hypothesis of two, resp. three regimes under alternatives with three resp. four states, with different sample sizes n and different parameters. Each table lists line by line the rejection rates on levels $10\%, 5\%$ and 1%.

4. Volatility states in series of asset returns

In this section we examine the number of volatility states in a time series of logarithmic returns of daily closing prices of the S&P 500 price return index (GSPC) for the last 15 years from August 31st 2012 backwards, using two distinct classes of HMMs: one with skew-normal state-dependent distributions with switching scale and structural mean and skewness, and another with normal state-dependent distributions with switching means and variances. See the simulations for further details on the implementation of the tests for these models, both of which cover skewness in the marginal distribution of the process.

For each model we apply our appropriate test methodology and in addition compare the results to those of model selection criteria such as AIC or BIC based on the full-model likelihood of the HMM. All computations were performed using our R Package *qLRT*[1]. A further applications to oil price logarithmic returns is given in Appendix C.3.

For the skew-normal HMM we reject the hypotheses of $k_0 = 1, 2$ and $k_0 = 3$ volatility states but could not reject four states. Since AIC and BIC select five and resp. four states we decide for the HMM with four states.

In case of the normal HMM, we reject the hypotheses of $k_0 = 1$ and $k_0 = 2$ states but could not reject three states based on the asymptotic distribution of the test statistic, i.e. χ^2_6. Again AIC selects five and BIC four states, therefore, we investigated more closely the finite sample distribution of the test statistic in case of strongly switching scales but only moderately switching means. To this end we simulated the finite-sample distribution of the test statistic given the MLE of the normal HMM with three states as null model (2.500 repetitions, each dataset has the same length as the original dataset). It turned out that the test using the asymptotic distribution is strongly conservative and that using the simulated critical values we can also reject the hypothesis of $k_0 = 3$ in case of the normal HMM (p-value = 1.04%). Four states cannot be rejected. Note that the simulation section focused on scenarios where the means switch substiantially, compared to the standard deviations.

We also estimated the full models with four states, which turn out to be quite similar:

$$\begin{aligned}
\hat{\boldsymbol{\sigma}}_{\boldsymbol{SN}} &= (0.58, 1.01, 1.59, 3.55), & \hat{\mu}_{SN} &= 0.06, \quad \hat{\xi}_{SN} = -0.05 \\
\hat{\boldsymbol{\sigma}}_{\boldsymbol{Nor}} &= (0.58, 1.01, 1.59, 3.53), & \hat{\boldsymbol{\mu}}_{Nor} &= (0.10, 0.03, -0.04, -0.33),
\end{aligned}$$

$$\hat{\mathbf{\Gamma}}_{SN} = \begin{pmatrix} 98.05 & 1.95 & 0.00 & 0.00 \\ 1.40 & 97.39 & 1.05 & 0.17 \\ 0.00 & 1.96 & 97.36 & 0.68 \\ 0.00 & 0.00 & 4.97 & 95.03 \end{pmatrix}, \qquad \hat{\mathbf{\Gamma}}_{Nor} = \begin{pmatrix} 97.93 & 2.07 & 0.00 & 0.00 \\ 1.49 & 97.29 & 1.06 & 0.16 \\ 0.00 & 1.95 & 97.35 & 0.70 \\ 0.00 & 0.00 & 4.99 & 95.01 \end{pmatrix},$$

[1] available on http://www.uni-marburg.de/fb12/stoch/research/rpackage

with resulting stationary distributions

$$\hat{\boldsymbol{\pi}}_{SN} = (29.19, 40.66, 25.28, 4.87), \qquad \hat{\boldsymbol{\pi}}_{Nor} = (29.27, 40.52, 25.34, 4.88).$$

Finally, we estimate the maximum-a-posteriori paths of the underlying four-state Markov chains, given the observations, using the Viterbi algorithm. Fig. 1 shows the time series of log-returns together with the estimated states of both models, where the lowest value corresponds to state one and the highest to state four.

Both paths are rather similar: They switch into the most volatility state on September 15, 2008, which was the day when Lehman Brothers filed for chapter 11. For both HMMs the period linked to the financial crisis holds until March 30th 2009. We observe four further high-volatility periods: In October 1997 (two days for both HMMs: 27th and 28th), which corresponds to the Asian financial crisis in 1997, in August/September 1998 (11 days: August 27th till September 11th), in July/August 2002 (16 days for the normal HMM: July 18th till August 8th; 20 days for the skew-normal HMM: July 18th till August 14th) and in August 2011 (13 days for the normal HMM: 2nd till 18th; 8 days for the skew-normal HMM: 2nd till 11th), which corresponds to the European sovereign debt crisis in 2011. Note, that on August 4th 2011 the European Central Bank announced to buy government bonds of the countries involved in this crisis (Portugal, Italy, Ireland, Greece and Spain). The highest (fourth) volatility state thus may be interpreted as a crisis state. In contrast, the time from October 2003 until June 2007 was dominated by the smallest (first) volatility state, which was followed by a few months in the second and third state, before in fall 2008 the financial crisis started and the volatility switched to the highest level.

	normal HMM			skew-normal HMM		
	p-value	AIC	BIC	p-value	AIC	BIC
$k_0 = 1$	$\leq 10^{-4}$	12946.73	12961.20	$\leq 10^{-4}$	12917.17	12935.88
$k_0 = 2$	$\leq 10^{-4}$	11740.28	11777.70	$\leq 10^{-4}$	11735.89	11773.31
$k_0 = 3$	0.01	11436.83	11511.67	0.003	11443.62	11512.22
$k_0 = 4$	**0.051**	11375.17	**11499.90**	**0.628**	11381.23	**11493.49**
$k_0 = 5$	–	**11366.58**	11553.67	–	**11370.45**	11538.84
$k_0 = 6$	–	11372.07	11634.00	–	11382.64	11619.62

Table 6: Selecting the number of states of S&P 500 log-returns.

Acknowledgements

The authors would like to thank Florian Ketterer and Jörn Dannemann for helpful comments. Financial support from the Deutsche Forschungsgemeinschaft is

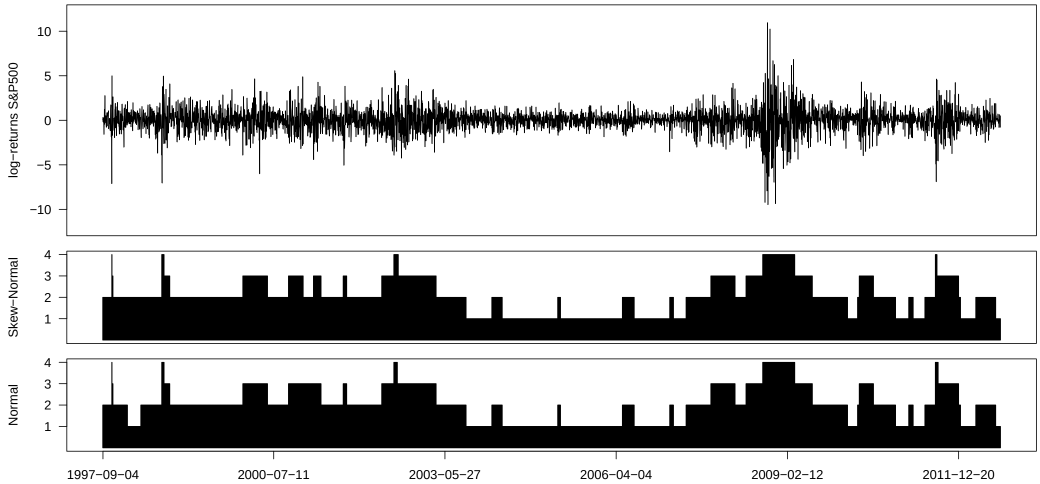


Figure 1: Daily logarithmic returns of the S&P 500 price return index from September 2nd 1997 until August 31st 2012 with estimated regimes of the hidden Markov chains of both models.

gratefully acknowledged.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171-178.

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

Chen, J., Tan, X. and Zhang, R. (2008). Consistency of penalized MLE for normal mixtures in mean and variance. Statistica Sinica. 18, 443-465.

Chen, J. and Li, P. (2011). Tuning the EM-test for the order of finite mixture models. The Canadian Journal of Statistics. 39, 389-404.

Chen, J., Li, P. and Fu, Y. (2012). Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107, 1096-1105.

Dannemann, J. and Holzmann, H. (2008). Testing for two states in a hidden Markov model. *Canadian Journal of Statistics*, **36**, 505-520.

Gassiat, E. and Keribin, C. (2000) The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM, Probab. Stat.* **4**, 25-52.

Holzmann, H., Munk, A. and Gneiting, T. (2006) Identifiability of finite mixtures of elliptical distributions. *Scand. J. Statist.* **33**, 753-764.

Kon, S.J. (1984). Models of Stock Returns - A Comparison. *The Journal of Finance*, **39**, 147-165.

Leroux, B. G. (1992). Consistent estimation of the mixing distribution. *Ann. Statist.* **20**, 1350-1360.

Li, P. and Chen, H. (2010). Testing the order of a finite mixture model. *Journal of the American Statistical Association*, **105**, 1084-1092.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, **5**, 81-91.

MacDonald, I. L. and Zucchini, W. (1997) *Hidden Markov and other models for discrete-valued time series*, London: Chapman & Hall.

Rydén, T., Teräsvirta, T. and Åsbrink, S. (1998). Stylized Facts of Daily Return Series and the Hidden Markov Model. *Journal of Applied Economics*, **13**, 217-244.

Shephard, N. and Andersen, T.G. (2009). Stochastic Volatility: Origins and Overview. In: Mikosch et al. *Handbook of Financial Time Series*, Springer Berlin Heidelberg, **1**, 233-254.

A. Appendix

A.1. Proof of Theorem 1

Proof. Set

$$\begin{aligned} D_{tj} &:= \phi(X_t;\mu_j^*,\sigma_j^*) - \phi(X_t;\mu_{k_0}^*,\sigma_{k_0}^*), && j=1,\dots,k_0-1 \\ A_{tj}^{(l)} &:= \partial^l \phi(X_t;\mu_j^*,\sigma_j^*)/\partial^l \mu, && j=1,\dots,k_0,\ l=1,\dots,4. \end{aligned}$$

and let $\boldsymbol{c}_t^T = (\boldsymbol{c}_{1t}^T, \boldsymbol{c}_{2t}^T)$ where

$$\begin{aligned} \boldsymbol{c}_{1t}^T &:= \left(D_{t1},\dots,D_{t(k_0-1)}, A'_{t1},\dots,A'_{tk_0}, A''_{t1},\dots,A''_{tk_0}\right) / \phi_{mix}^{(k)}(X_t;\boldsymbol{\mu}^*,\boldsymbol{\sigma}^*,\boldsymbol{\pi}^*) \\ \boldsymbol{c}_{2t}^T &:= \left(A'''_{t1},\dots,A'''_{tk_0}, A''''_{t1},\dots,A''''_{tk_0}\right) / \phi_{mix}^{(k)}(X_t;\boldsymbol{\mu}^*,\boldsymbol{\sigma}^*,\boldsymbol{\pi}^*) \end{aligned}$$

Then set $\boldsymbol{\Psi} := \mathrm{COV}(\boldsymbol{c}_t)$ with submatrices $\boldsymbol{\Psi}_{ji} := E([\boldsymbol{c}_{j1} - E(\boldsymbol{c}_{j1})][\boldsymbol{c}_{i1} - E(\boldsymbol{c}_{i1})]^T)$, $i,j=1,2$, and define

$$\tilde{\boldsymbol{c}}_{2t} := \boldsymbol{c}_{2t} - \boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{c}_{1t}, \qquad \widetilde{\boldsymbol{\Psi}}_{22} := \boldsymbol{\Psi}_{22} - \boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12} = \mathrm{COV}(\tilde{\boldsymbol{c}}_{2t}).$$

Due to the mixing properties of the HMM, the arguments in Chen, Li and Fu (2012) apply (the order assessments remain valid) to obtain the expansion

$$Z_n = \left(n^{-1/2}\sum_{t=1}^n \tilde{\boldsymbol{c}}_{2t}^T\right) \widetilde{\boldsymbol{\Psi}}_{22}^{-1} \left(n^{-1/2}\sum_{t=1}^n \tilde{\boldsymbol{c}}_{2t}\right) + o_P(1).$$

Now the sequence $\tilde{\boldsymbol{c}}_{2t}$ is centered, and $n^{-1/2}\sum_{t=1}^n \tilde{\boldsymbol{c}}_{2t}$ asymptotically normal by the central limit theorem for stationary sequences, with long-run covariance matrix

$$\widetilde{\boldsymbol{\Psi}}_{22} + 2\sum_{t=2}^{\infty} E\big(\tilde{\boldsymbol{c}}_{2t}\tilde{\boldsymbol{c}}_{21}^T + \tilde{\boldsymbol{c}}_{21}\tilde{\boldsymbol{c}}_{2t}^T\big).$$

If this reduces to $\widetilde{\boldsymbol{\Psi}}_{22}$, the conclusion of the theorem is evident. To show this, we show that $(\tilde{\boldsymbol{c}}_{2t})_t$ is actually a martingale difference sequence.

To this end, we consider the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ with

$$\mathcal{F}_t := \sigma\left(S_j, \boldsymbol{c}_j; j \le t\right) \text{ for } t \in \mathbb{N}.$$

Then $\mathcal{L}\left(\boldsymbol{c}_t \mid \mathcal{F}_{t-1}\right) = \mathcal{L}\left(\boldsymbol{c}_t \mid S_{t-1}\right)$, where $\mathcal{L}$ denotes the (conditional) distribution of a random variable, and therefore also $\mathcal{L}(\tilde{\boldsymbol{c}}_{2t}|\mathcal{F}_{t-1}) = \mathcal{L}(\tilde{\boldsymbol{c}}_{2t}|S_{t-1})$. Thus, it remains to show that

$$E\left(\tilde{\boldsymbol{c}}_{2t} \mid S_{t-1} = j\right) = 0, \qquad j = 1,\dots,k_0. \tag{8}$$

Let

$$\boldsymbol{\lambda}_h := E\left(\boldsymbol{c}_1 \mid S_1 = h\right),$$

and recall that $\gamma_{jh} := P\left(S_t = h \mid S_{t-1} = j\right)$ for $h = 1, \ldots, k_0$. As the Markov chain can adopt k_0 states under the hypothesis, it follows that

$$E\left(\boldsymbol{c}_t \mid S_{t-1} = j\right) = \sum_{h=1}^{k_0} \gamma_{jh} \boldsymbol{\lambda}_h \quad \text{and} \quad E\left(\boldsymbol{c}_{lt} \mid S_{t-1} = j\right) = \sum_{h=1}^{k_0} \gamma_{jh} \boldsymbol{\lambda}_{hl} \text{ for } l = 1, 2,$$

where we partition $\boldsymbol{\lambda}_h^T = \left(\boldsymbol{\lambda}_{h1}^T, \boldsymbol{\lambda}_{h2}^T\right)$ with $\boldsymbol{\lambda}_{h1} \in \mathbb{R}^{3k_0-1}$. We get

$$E\left(\tilde{\boldsymbol{c}}_{2t} \mid S_{t-1} = j\right)^T = \sum_{h=1}^{k_0} \gamma_{jh} \boldsymbol{\lambda}_{h2}^T - \left(\sum_{h=1}^{k_0} \gamma_{jh} \boldsymbol{\lambda}_{h1}^T\right) \boldsymbol{\Psi}_{11}^{-1} \boldsymbol{\Psi}_{12}. \tag{9}$$

Since $0 = E\left(\boldsymbol{c}_1\right) = \sum_{h=1}^{k_0} \pi_h^* \boldsymbol{\lambda}_h$, we obtain

$$\boldsymbol{\lambda}_{k_0} = \sum_{h=1}^{k_0-1} \alpha_h \boldsymbol{\lambda}_h, \text{ with } \alpha_h := -\pi_h^*/\pi_{k_0}^*, \tag{10}$$

and setting $d_h := \gamma_{jh} + \gamma_{jk_0} \alpha_h$ for $h = 1, \ldots, k_0 - 1$ and inserting (10) in (9) gives

$$E\left(\tilde{\boldsymbol{c}}_{2t} \mid S_{t-1} = j\right)^T = \sum_{h=1}^{k_0-1} d_h \boldsymbol{\lambda}_{h2}^T - \left(\sum_{h=1}^{k_0-1} d_h \boldsymbol{\lambda}_{h1}^T\right) \boldsymbol{\Psi}_{11}^{-1} \boldsymbol{\Psi}_{12}. \tag{11}$$

Now observe that

$$E\left(D_{1h} \boldsymbol{c}_1\right) = \boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{k_0}, \quad h = 1, \ldots, k_0 - 1.$$

Let

$$\boldsymbol{S} := \begin{pmatrix} \boldsymbol{I}_{k_0-1} \\ \boldsymbol{0}_{(4k_0)\times(k_0-1)} \end{pmatrix}, \qquad \boldsymbol{T} := \begin{pmatrix} \boldsymbol{I}_{k_0-1} \\ \boldsymbol{0}_{2k_0\times(k_0-1)} \end{pmatrix},$$

then from the definition of $\boldsymbol{\Psi}$ and (10) we get

$$\begin{aligned} \boldsymbol{\Psi S} &= \left[E\left(D_{11} \boldsymbol{c}_1\right), \ldots,\ E\left(D_{1k_0-1} \boldsymbol{c}_1\right)\right] \\ &= \left[\boldsymbol{\lambda}_1 - \sum_{h=1}^{k_0-1} \alpha_h \boldsymbol{\lambda}_h, \ldots,\ \boldsymbol{\lambda}_{k_0-1} - \sum_{h=1}^{k_0-1} \alpha_h \boldsymbol{\lambda}_h\right] =: \boldsymbol{\Lambda}, \end{aligned} \tag{12}$$

where $\boldsymbol{0}_{.}$ denotes matrices of zeros and $\boldsymbol{I}_{.}$ are identity matrices, all with the appropriate dimensions. This result also holds for the partitioned $\boldsymbol{\lambda}$ vectors, i.e.

$$\boldsymbol{\Psi}_{l1} \boldsymbol{T} = \left(\boldsymbol{\lambda}_{1l} - \textstyle\sum_{h=1}^{k_0-1} \alpha_h \boldsymbol{\lambda}_{hl}\ ,\quad \cdots \quad,\ \boldsymbol{\lambda}_{(k_0-1)l} - \textstyle\sum_{h=1}^{k_0-1} \alpha_h \boldsymbol{\lambda}_{hl}\right), \quad l = 1, 2.$$

Now, we show below that

$$\operatorname{span}\left(\boldsymbol{\Lambda}\right) = \operatorname{span}\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{k_0-1}\}, \tag{13}$$

where span$\,(\boldsymbol{\Lambda})$ denotes the space spanned by the columns of $\boldsymbol{\Lambda}$. Therefore, there is a matrix $\boldsymbol{M} \in \mathbb{R}^{(k_0-1)\times(k_0-1)}$ such that $\boldsymbol{\Lambda}\,\boldsymbol{M} = \big[d_1\boldsymbol{\lambda}_1\ ,\dots,\ d_{k_0-1}\boldsymbol{\lambda}_{k_0-1}\big]$ and thus from (12)

$$\boldsymbol{\Psi}\,\boldsymbol{S}\boldsymbol{M} = \big[d_1\boldsymbol{\lambda}_1,\dots,\ d_{k_0-1}\boldsymbol{\lambda}_{k_0-1}\big]$$

and hence for the submatrices of $\boldsymbol{\Psi}$

$$\boldsymbol{\Psi}_{l1}\,\boldsymbol{T}\boldsymbol{M} = \big[d_1\boldsymbol{\lambda}_{1l},\dots,d_{k_0-1}\boldsymbol{\lambda}_{(k_0-1)l}\big], \qquad l=1,2.$$

This implies

$$\big(1\ ,\cdots,\ 1\big)\,\boldsymbol{M}^T\boldsymbol{T}^T\boldsymbol{\Psi}_{1l} = \sum_{h=1}^{k_0-1} d_h\boldsymbol{\lambda}_{hl}^T, \qquad l=1,2.$$

Using this subsequently for $l=1$ and $l=2$ we get

$$\left(\sum_{h=1}^{k_0-1} d_h\boldsymbol{\lambda}_{h1}^T\right)\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12} = \big(1\ ,\cdots,\ 1\big)\,\boldsymbol{M}^T\boldsymbol{T}^T\boldsymbol{\Psi}_{11}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12} = \sum_{h=1}^{k_0-1} d_h\boldsymbol{\lambda}_{h2}^T$$

which due to (11) implies (8).

It remains to show (13). Here, we evidently only need to show that for any $\boldsymbol{\omega} = (\omega_1,\dots,\omega_{k_0-1})^T \in \mathbb{R}^{k_0-1}$ the vector $\sum_{h=1}^{k_0-1}\omega_h\boldsymbol{\lambda}_h$ is in span$\,(\boldsymbol{\Lambda})$. This is the case since the matrix

$$\begin{pmatrix} 1-\alpha_1 & -\alpha_1 & \cdots & -\alpha_1 \\ -\alpha_2 & 1-\alpha_2 & \cdots & -\alpha_2 \\ \vdots & & \ddots & \vdots \\ -\alpha_{k_0-1} & \cdots & & 1-\alpha_{k_0-1} \end{pmatrix}$$

has determinant $1-\left(\sum_{h=1}^{k_0-1}\alpha_h\right) = (\pi_{k_0}^*)^{-1} \neq 0$ and hence is invertible.

□

B. Supplement: Technical details

B.1. Concerning the Assumptions A4 and A5

We start by discussing the validity of Assumption A5 in some relevant models. For a one-parameter family (and hence no structural parameters), under the assumption that the corresponding moments exist this follows from linear independence of the families of densities and their first and second derivatives (essentially the notion of strong identifiability by Chen 1995). It is satisfied e.g. by the Poisson distribution.

Now suppose that the $f > 0$ is a twice-continuously differentiable density w.r.t. Lebesgue measure, and let $f(x;\mu,\sigma) = f\big((x-\mu)/\sigma\big)/\sigma$ denote the corresponding location-scale family.

Lemma 3. *a. Suppose that for any $\mu \in \mathbb{R}$, $0 < \sigma_1 < \ldots < \sigma_{k_0}$ and $a_j, b_j, c_j, d_j \in \mathbb{R}$, $j = 1, \ldots, k_0$, the condition*

$$\sum_{j=1}^{k_0} \big(a_j f(x;\mu,\sigma_j) + b_j \frac{\partial f(x;\mu,\sigma_j)}{\partial\mu} + c_j \frac{\partial f(x;\mu,\sigma_j)}{\partial\sigma} + d_j \frac{\partial^2 f(x;\mu,\sigma_j)}{\partial^2\sigma}\big) = 0 \quad (14)$$

$\forall x \in \mathbb{R}$ implies that $a_j = b_j = c_j = d_j = 0$, $j = 1, \ldots, k_0$. Then, under the assumption that the moments exist, Assumption A5 is satisfied in the HMM with state-dependend densities from $f(\cdot;\mu,\sigma)$ with switching scale σ and structural location μ.
b. Similarly, suppose that for any $\mu_1 < \ldots < \mu_{k_0} \in \mathbb{R}$, $\sigma > 0$ and $a_j, b_j, c_j, d_j \in \mathbb{R}$, $j = 1, \ldots, k_0$, the condition

$$\sum_{j=1}^{k_0} \big(a_j f(x;\mu_j,\sigma) + b_j \frac{\partial f(x;\mu_j,\sigma)}{\partial\mu} + c_j \frac{\partial^2 f(x;\mu_j,\sigma)}{\partial^2\mu} + d_j \frac{\partial f(x;\mu_j,\sigma)}{\partial\sigma}\big) = 0 \quad (15)$$

$\forall x \in \mathbb{R}$ implies that $a_j = b_j = c_j = d_j = 0$, $j = 1, \ldots, k_0$. Then, under the assumption that the moments exist, Assumption A5 is satisfied in the HMM with state-dependent densities from $f(\cdot;\mu,\sigma)$ with switching location μ and structural scale σ.

Proof of Lemma 3. We only show part a., part b. is completely analogous. Under the assumption that the corresponding moments exist, Assumption A5 is satisfied if and only if the corresponding random variables are linearly independent in L_2. From the definitions, this is the case if for all $a_j, b_j, c_j, d \in \mathbb{R}$,

$$\sum_{j=1}^{k_0} \Big(a_j f(X_1;\mu,\sigma_j) + b_j \frac{\partial f(X_1;\mu,\sigma_j)}{\partial\sigma} + c_j \frac{\partial^2 f(X_1;\mu,\sigma_j)}{\partial^2\sigma}\Big) + d \frac{\partial f_{mix}^{(k_0)}(X_1;\mu,\sigma_1,\ldots,\sigma_{k_0})}{\partial\mu} = 0 \quad \text{a.s.}$$

implies $a_j = b_j = c_j = d = 0$. Since the distribution of X_1 is equivalent to Lebesgue measure, this is equivalent to

$$\sum_{j=1}^{k_0} \Big(a_j f(x; \mu, \sigma_j) + b_j \frac{\partial f(x; \mu, \sigma_j)}{\partial \sigma} + c_j \frac{\partial^2 f(x; \mu, \sigma_j)}{\partial^2 \sigma} \Big) + d \sum_{j=1}^{k_0} \pi_j \frac{\partial f(x; \mu, \sigma_j)}{\partial \mu} = 0$$

for Lebesgue-almost all $x \in \mathbb{R}$, which, by continuity, then holds for all $x \in \mathbb{R}$. Then $a_j = b_j = c_j = d = 0$ follows immediately from (14), as required. □

We consider two special cases. The skew-normal family $\mathcal{SN}(\alpha)$ is defined by

$$\tilde{f}_0(x; \alpha) = 2\, \phi(x)\, \Phi(\alpha x), \qquad \alpha \in \mathbb{R} \tag{16}$$

where Φ is the distribution function of the standard normal, see Azzalini (1985). For a fixed α, we show below that condition a. of the lemma is satisfied for the corresponding location-scale family. However, it is well-known that condition b. is not satisfied for the normal distribution, see e.g. Chen and Li (2009), and indeed, Assumption A5 is not satisfied. For the t-distribution with density

$$f(x) = \Gamma\left(\frac{\nu+1}{2}\right) \left(\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi\nu} \left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2} \right)^{-1}, \qquad \nu > 0, \tag{17}$$

we show below that for fixed degrees of freedom ν, both conditions are satisfied for the corresponding location-scale family.

Lemma 4. *1. For the location-scale family of the skew-normal distribution with fixed skewness parameter α, condition a. of Lemma 3 is satisfied.*
2. For the location-scale family of the t-distribution with fixed degrees of freedom α, conditions a. and b. of Lemma 3 are satisfied.

Proof. Ad. 1.: Let

$$\psi(t, \mu, \sigma) = \exp(\mu i t - 0.5\, \sigma^2 t^2) \cdot (1 + i\, \mathrm{erf}(\sigma \delta t / \sqrt{2}))$$

be the characteristic function of the skew-normal distribution with parameters μ, σ, where

$$\delta = \alpha / \sqrt{1 + \alpha^2} \qquad \text{and} \qquad \mathrm{erf}(x) = 2/\sqrt{\pi} \int_0^x \exp(-t^2)\, dt.$$

Then, taking Fourier transforms and interchanging derivative and integrals in (14) gives

$$\sum_{j=1}^{k_0} \Big(a_j \psi(t, \mu, \sigma_j) + b_j \psi_\sigma(t, \mu, \sigma_j) + c_j \psi_{\sigma\sigma}(t, \mu, \sigma_j) + d_j \psi_\mu(t, \mu, \sigma_j) \Big) = 0, \tag{18}$$

$\forall t$ where $\psi_\sigma, \psi_{\sigma\sigma}$ and ψ_μ are the derivatives of the characteristic function w.r.t. σ, twice σ and μ. Now

$$\begin{aligned}
\psi_\mu(t,\mu,\sigma) =& it \cdot \psi(t;\mu,\sigma)\\
\psi_\sigma(t,\mu,\sigma) =& -\sigma t^2 \cdot \psi(t;\mu,\sigma)\,(1+i\,\mathrm{erf}(\sigma\delta t/\sqrt{2})) + i\,\frac{\delta t\sqrt{2}}{\sqrt{\pi}} \cdot \exp\big(-\sigma^2 t^2 \delta^2/2\big)\,\exp(\mu i t - 0.5\,\sigma^2 t^2),\\
\psi_{\sigma\sigma}(t,\mu,\sigma) =& \psi(t;\mu,\sigma)\,\big(\sigma^2 t^4 - t^2\big) - i\exp(\mu i t - 0.5\,\sigma^2 t^2)\,\exp\big(-\sigma^2 t^2 \delta^2/2\big)\,t^3 \sigma \frac{\sqrt{2}}{\sqrt{\pi}}\,\big(2\delta + \delta^3\big).
\end{aligned}$$

Plugging this into (18), multiplying by $e^{-i\mu t}$ and taking the real part gives

$$\sum_{j=1}^{k_0} \exp(-1/2\,\sigma_j^2\,t^2)\big(a_j - b_j\,\sigma_j\,t^2 + c_j\,\big(\sigma_j^2\,t^4 - t^2\big) - d_j\,t\,\,\mathrm{erf}(\sigma_j \delta t/\sqrt{2})\big) = 0\ \forall t. \tag{19}$$

Suppose that $\sigma_1^2 < \ldots < \sigma_{k_0}^2$. Now in (19),

1. multiply by $\exp(\sigma_1^2 t^2/2)/t^4$, let $t \to \infty$ to conclude $c_1 = 0$,

2. multiply by $\exp(\sigma_1^2 t^2/2)/t^2$, let $t \to \infty$ to conclude $b_1 = 0$,

3. multiply by $\exp(\sigma_1^2 t^2/2)/t$, let $t \to \infty$ to conclude $d_1 = 0$,

4. multiply by $\exp(\sigma_1^2 t^2/2)$, let $t \to \infty$ to conclude $a_1 = 0$.

Now proceed by induction over k_0.
Ad 2.: The characteristic function of the location-scale family of the t-distribution is given by (cf. Hurst 1995)

$$\varphi(t;\mu,\sigma) = e^{i\mu t}\frac{K_m\left(\sqrt{\nu}\sigma|t|\right)\left(\sqrt{\nu}\sigma|t|\right)^m}{\Gamma\left(m\right)2^{m-1}}, \qquad m = \frac{1}{2}\nu, \tag{20}$$

where $\Gamma(\cdot)$ is the Gamma function and $K_p(\cdot)$ is the modified Bessel function of the second kind and order p (cf. Andrews 1986, chapter 6). The partial derivatives are given by

$$\begin{aligned}
\varphi_\mu(t;\mu,\sigma) &= it\ e^{i\mu t}\ \frac{K_m\left(\sqrt{\nu}\sigma|t|\right)\left(\sqrt{\nu}\sigma|t|\right)^m}{\Gamma\left(m\right)2^{m-1}},\\
\varphi(t;\mu,\sigma)_{\mu\mu} &= -t^2\ e^{i\mu t}\ \frac{K_m\left(\sqrt{\nu}\sigma|t|\right)\left(\sqrt{\nu}\sigma|t|\right)^m}{\Gamma\left(m\right)2^{m-1}}\\
\varphi_\sigma(t;\mu,\sigma) &= -|t|\ e^{i\mu t}\frac{K_{m-1}(\sqrt{\nu}\sigma|t|)\sqrt{\nu}\left(\sqrt{\nu}\sigma|t|\right)^m}{\Gamma(m)2^{m-1}},\\
\varphi_{\sigma\sigma}(t;\mu,\sigma) &= \frac{|t|\,\sqrt{\nu}e^{i\mu t}}{\Gamma(m)2^{m-1}}\ \left(\sqrt{\nu}|t|\right)^m\ \sigma^{m-1}\left(\sqrt{\nu}\sigma|t|K_{m-2}(\sqrt{\nu}\sigma|t|) - K_{m-1}(\sqrt{\nu}\sigma|t|)\right),
\end{aligned}$$

cf. Andrews (1986).

(a)*: Switching* σ. As above, taking the Fourier transform and interchanging integral and derivative, and dividing by $(\sqrt{\nu}|t|)^m e^{i\mu t}/\Gamma(m)2^{m-1}$ gives

$$\sum_{j=1}^{k_0} K_m(\sigma_j\sqrt{\nu}|t|)\sigma_j^m\big(a_j - b_j|t|\sqrt{nu}\,\frac{K_{m-1}(\sigma_j\sqrt{\nu}|t|)}{K_m(\sigma_j\sqrt{\nu}|t|)} \\ + c_j\big(\nu t^2\frac{K_{m-2}(\sigma_j\sqrt{\nu}|t|)}{K_m(\sigma_j\sqrt{\nu}|t|)} - \sqrt{\nu}\frac{K_{m-1}(\sigma_j\sqrt{\nu}|t|)}{K_m(\sigma_j\sqrt{\nu}|t|)\sigma_j}\big) + itd_j\big) = 0\ \forall t.$$

Now use $K_{m-1}(x)/K_m(x) \to 1$ and $K_m(x) \sim \sqrt{2/(\pi x)}\,e^{-x}$ as $x \to \infty$. Consider the real part in the above display, and

1. multiply by $e^{\sigma_1 x}/t^{3/2}$, let $t \to \infty$ to conclude $c_1 = 0$,

2. multiply by $e^{\sigma_1 x}/t^{1/2}$, let $t \to \infty$ to conclude $b_1 = 0$,

3. multiply by $e^{\sigma_1 x}\,t^{1/2}$, let $t \to \infty$ to conclude $a_1 = 0$,

4. Finally, consider the imaginary part, multiply by $e^{\sigma_1 x}/t^{1/2}$, let $t \to \infty$ to conclude $d_1 = 0$.

Now proceed by induction.

(b). *Switching* μ. As above, taking the Fourier transform and interchanging integral and derivative, and dividing by $(\sqrt{\nu}\sigma|t|)^m K_m\left(\sqrt{\nu}\sigma|t|\right)/\Gamma(m)2^{m-1}$ gives

$$\sum_{j=1}^{k_0} e^{i\mu_j t}\left(a_j + b_j it - c_j t^2 - d_j|t|\frac{K_{m-1}(\sigma\sqrt{\nu}|t|)}{K_m(\sigma\sqrt{\nu}|t|)}\right) = 0\ \forall t.$$

Now, multiply by $e^{-i\mu_1 t}t^{-2}$, and average the resulting equation over $t = t_0, 2t_0, \ldots, nt_0$ to obtain

$$c_1 + \sum_{j=2}^{k_0} c_j \frac{1}{n}\sum_{l=1}^{n} e^{ilt_0(\mu_j-\mu_1)} = o(1), \qquad n \to \infty, \tag{21}$$

where $t_0 > 0$ is chosen such that $t_0(\mu_j - \mu_1) \notin 2\pi\mathbb{Z}$. Now since for $j = 2, \ldots, k_0$,

$$\frac{1}{n}\sum_{l=1}^{n} e^{ilt_0(\mu_j-\mu_1)} = \frac{1}{n}\,\frac{e^{i(n+1)t_0(\mu_j-\mu_1)} - 1}{e^{it_0(\mu_j-\mu_1)} - 1} \to 0, \quad n \to \infty,$$

we obtain $c_1 = 0$ by letting $n \to \infty$ in (21). Repeat this argument to obtain $c_2 = \ldots = c_{k_0} = 0$.

Next, proceed similarly when multiply by $e^{-i\mu_1 t}t^{-1}$ to conclude $-d_1 + ib_1 = 0$, that is, $d_1 = b_1 = 0$, and similarly for $j = 2, \ldots, k_0$. Finally, multiply by $e^{-i\mu_j t}$ to deal with the a_j. □

Next, consider Assumption 4.

Lemma 5. *Assumption A4 is satisfied*
a. for the skew-normal distribution with fixed skewness parameter, structural location and switching scale
b. for the t-distribution with fixed degrees of freedom, and either fixed scale and switching location or fixed location and switching scale.

Proof. a. We start by bounding the partial derivatives of the skew-normal density (see (16)) which arise in the assumption. First note that for fixed $\alpha \in \mathbb{R}$, $F(x;\mu,\sigma) := \Phi(\alpha\,(x-\mu)/\sigma)$ as well as any finite number of partial derivatives w.r.t μ and σ are uniformly bounded in $x \in \mathbb{R}$ and μ, σ varying over compact sets. Therefore, for compact K and $\epsilon, \sigma_0 > 0$,

$$\sup_{\mu \in K}\ \sup_{\sigma_0-\epsilon \le \sigma \le \sigma_0+\epsilon} \big|\partial_\sigma^j\, \partial_\mu^l f(x;\alpha,\mu,\sigma)\big| \le C(1+x^8)\exp\big(-\frac{x^2}{2(\sigma_0+\epsilon)}\big), \quad x \in \mathbb{R},\ j+l \le 4,$$

where $f(x;\alpha,\mu,\sigma)$ is the skew-normal density with skewness parameter α, location μ and scale σ. Now, suppose that $\sigma_1 < \ldots < \sigma_{k_0}$ are ordered. Then

$$\frac{1}{|f_{mix}^{(k_0)}(x;\mu^*,\sigma_1^*,\ldots,\sigma_{k_0}^*,\boldsymbol{\pi}^*)|^2} \le C_1 \exp\big(x^2/\,\sigma_{k_0}^2\big), \qquad x \in \mathbb{R}.$$

Choose

$$\epsilon < \sigma_{k_0}^*(\sqrt{3/2}-1)$$

and consider parameters (μ,σ) such that μ is in a compact neighborhood of μ^* and

$$\sigma \in U := \cup_{j=1}^{k_0} U_\epsilon(\sigma_j^*), \qquad U_\epsilon(\sigma_j^*) := \{\sigma : |\sigma-\sigma_j^*| \le \epsilon\}.$$

Since X_t has density $f_{mix}^{(k_0)}(x;\mu^*,\sigma_1^*,\ldots,\sigma_{k_0}^*,\boldsymbol{\pi}^*)$, for $j+l \le 4$ we obtain

$$\begin{aligned} &E \sup_{\mu\in K}\sup_{\sigma \in U} \frac{\big|\partial_\sigma^j\,\partial_\mu^l f(X_t;\alpha,\mu,\sigma)\big|^3}{f_{mix}^{(k_0)}(X_t;\mu^*,\sigma_1^*,\ldots,\sigma_{k_0}^*,\boldsymbol{\pi}^*)^3} \\ &\le \int C(1+x^8)^3 \exp\big(-x^2\big(3/(2(\sigma_{k_0}+\epsilon)^2) - 1/\sigma_{k_0}^2\big)\big)\,dx < \infty \end{aligned}$$

by the choice of ϵ.

b. For the location-scale family $f(x;\nu,\mu,\sigma)$ of the t-distribution (see (17)), we have for compact sets K_1, K_2 that

$$\sup_{\mu\in K_1}\ \sup_{\sigma \in K_2} \big|\partial_\mu^j \partial_\sigma^k f(x;\nu,\mu,\sigma)\big| \le C(1+|x|)^{-(\nu+1)}, \quad j+k \le 4.$$

Since also

$$\frac{1}{f_{mix}^{(k_0)}(x;\mu^*,\sigma_1^*,\ldots,\sigma_{k_0}^*,\boldsymbol{\pi}^*)} \le C_1|x|^{1+\nu}, \quad x \in \mathbb{R},$$

the integrability assumption is obvious. □

B.2. Proof of Theorem 2

The proof proceeds in several steps. First we show the following tightness statement, which is required for the expansion of the quasi-likelihood statistic. For $i, l = 1, \dots, d_1$, let

$$R_t^{\{l\}}(\boldsymbol{\nu}, \vartheta) := \frac{\frac{\partial}{\partial \nu_l} f(X_t; \boldsymbol{\nu}, \vartheta)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}, \; T_t^{\{l\}}(\boldsymbol{\nu}, \vartheta) := \frac{\frac{\partial^2}{\partial \nu_l \partial \vartheta} f(X_t; \boldsymbol{\nu}, \vartheta)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}, \\ V_t^{\{l,i\}}(\boldsymbol{\nu}, \boldsymbol{\vartheta}) := \frac{\frac{\partial^2}{\partial \nu_l \partial \nu_i} f_{mix}^{(k_0)}(X_t; (\boldsymbol{\nu}, \boldsymbol{\vartheta}), \boldsymbol{\pi}^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)^{-1}} \tag{22}$$

Lemma 6. *Under A2 and A4, for fixed $j = 1, \dots, k_0$ and $(\boldsymbol{\nu}, \vartheta) \in E_1(\boldsymbol{\nu}^*, \epsilon_0) \times E_2(\vartheta_j^*, \epsilon_0)$ the following processes are tight*

$$n^{-1/2} \sum R_t^{\{l\}}(\boldsymbol{\nu}, \vartheta), \qquad n^{-1/2} \sum V_t^{\{l,i\}}(\boldsymbol{\nu}, \boldsymbol{\vartheta}), \qquad n^{-1/2} \sum T_t^{\{l\}}(\boldsymbol{\nu}, \vartheta) \\ n^{-1/2} \sum Y_t'(\boldsymbol{\nu}, \vartheta), \qquad n^{-1/2} \sum Y_t''(\boldsymbol{\nu}, \vartheta), \qquad n^{-1/2} \sum Y_t'''(\boldsymbol{\nu}, \vartheta),$$

where $i, l = 1, \dots, d_1$.

Proof of Lemma 6. Let $\boldsymbol{\zeta} := (\boldsymbol{\nu}, \vartheta) \in E_1(\boldsymbol{\nu}^*, \epsilon_0) \times E_2(\vartheta_j^*, \epsilon_0)$. We show the tightness of

$$\widetilde{T}_n^{\{l\}}(\boldsymbol{\zeta}) := n^{-1/2} \sum T_t^{\{l\}}(\boldsymbol{\zeta}).$$

Using Billingsley (1968, p. 95) or Klicnarova (2007, prop. 1) for the multivariate case, it suffices to show that for some $C > 0$,

$$E\big(\widetilde{T}_n^{\{l\}}(\boldsymbol{\zeta}_1) - \widetilde{T}_n^{\{l\}}(\boldsymbol{\zeta}_2)\big)^2 \leq C \; ||\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2||_1^2, \tag{23}$$

for all n. Now,

$$E\big(\widetilde{T}_n^{\{l\}}(\boldsymbol{\zeta}_1) - \widetilde{T}_n^{\{l\}}(\boldsymbol{\zeta}_2)\big)^2 = E\big(T_1^{\{l\}}(\boldsymbol{\zeta}_1) - T_1^{\{l\}}(\boldsymbol{\zeta}_2)\big)^2 \\ + \frac{2}{n} \sum_{t=2}^{n} (n + 1 - t) \; E\Big(\big(T_1^{\{l\}}(\boldsymbol{\zeta}_1) - T_1^{\{l\}}(\boldsymbol{\zeta}_2)\big)\big(T_t^{\{l\}}(\boldsymbol{\zeta}_1) - T_t^{\{l\}}(\boldsymbol{\zeta}_2)\big)\Big). \tag{24}$$

Using the multivariate mean-value theorem and assumption A4 gives

$$\big|T_1^{\{l\}}(\boldsymbol{\zeta}_1) - T_1^{\{l\}}(\boldsymbol{\zeta}_2)\big| = \big|\nabla_{\boldsymbol{\zeta}}^T T_1^{\{l\}}(\boldsymbol{\zeta}_0) \cdot (\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2)\big| \leq \big(g(X_1)^{1/3}\big) \; ||\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2||_1. \tag{25}$$

This immediately bounds the first term on the right side of (24).

As for the second, we let

$$\lambda_m = E\big(T_1^{\{l\}}(\boldsymbol{\zeta}_1) - T_1^{\{l\}}(\boldsymbol{\zeta}_2) | S_1 = m\big), \qquad \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{k_0})^T.$$

Then for $t = 2, \dots, n$, we have that

$$E\Big(\big(T_1^{\{l\}}(\boldsymbol{\zeta}_1) - T_1^{\{l\}}(\boldsymbol{\zeta}_2)\big)\big(T_t^{\{l\}}(\boldsymbol{\zeta}_1) - T_t^{\{l\}}(\boldsymbol{\zeta}_2)\big)\Big) = \boldsymbol{\lambda}^T \Gamma^{t-1} \boldsymbol{\lambda}.$$

Now

$$0 = E\big(T_1^{\{l\}}(\boldsymbol{\zeta}_1) - T_1^{\{l\}}(\boldsymbol{\zeta}_2)\big) = \boldsymbol{\lambda}^T \boldsymbol{\pi},$$

and therefore for some $c > 0$,

$$\big|\boldsymbol{\lambda}^T \Gamma^{t-1} \boldsymbol{\lambda}\big| \leq c r^{t-1} \|\boldsymbol{\lambda}\|^2,$$

where $0 \leq r < 1$ can be chosen slightly larger than the second-largest eigenvalue of Γ, see e.g. Seneta (2006, theorem 1.2). By (25), we get for some $c_1 > 0$

$$\|\boldsymbol{\lambda}\|^2 \leq c_1 ||\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2||_1^2.$$

Therefore

$$|\frac{2}{n}\sum_{t=2}^{n}(n+1-t)\, E\Big(\big(T_1^{\{l\}}(\boldsymbol{\zeta}_1) - T_1^{\{l\}}(\boldsymbol{\zeta}_2)\big)\big(T_t^{\{l\}}(\boldsymbol{\zeta}_1) - T_t^{\{l\}}(\boldsymbol{\zeta}_2)\big)\Big)| \leq 2cc_1 \sum_{t=2}^{\infty} r^{t-1} ||\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2||_1^2$$

which concludes the proof. □

Let us next turn to consistency of the QMLE.

Recall that we assume the entries of $\big(\vartheta_1^*, \dots, \vartheta_{k_0}^*\big)$ to be distinct and ordered $\vartheta_1^* < \cdots < \vartheta_{k_0}^*$, $\boldsymbol{\nu}^* := (\nu_1^*, \dots, \nu_{d_1}^*)$, $\boldsymbol{\theta}^* = (\boldsymbol{\nu}^*, \vartheta_1^*, \dots, \vartheta_{k_0}^*)$. Let $\pi_j^* := P(S_t = j)$ for $j \in \{1, \dots, k_0\}$ denote the true stationary probability of the Markov chain for state j and $\boldsymbol{\pi}^* := \big(\pi_1^*, \dots, \pi_{k_0}^*\big)$. The assumption of irreducibility, see A1, implies $\pi_j^* > 0$.

For the QMLE under the hypothesis we write $\widehat{\boldsymbol{\theta}}(k_0) = \widehat{\boldsymbol{\theta}} = \big(\widehat{\boldsymbol{\nu}}, \widehat{\vartheta}_1, \dots, \widehat{\vartheta}_{k_0}\big)$, where $\hat{\vartheta}_1 \leq \ldots \leq \hat{\vartheta}_{k_0}$. For the QMLE $(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\pi}})$ under our specific alternative with $2k_0$ states, see (6), we write

$$\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\nu}}, \tilde{\vartheta}_1, \dots, \tilde{\vartheta}_{2k_0}), \qquad \widetilde{\boldsymbol{\pi}} = \big(\tilde{\beta}_1\tilde{\pi}_1, (1-\tilde{\beta}_1)\tilde{\pi}_1, \dots, \tilde{\beta}_{k_0}\tilde{\pi}_{k_0}, (1-\tilde{\beta}_{k_0})\tilde{\pi}_{k_0}\big),$$

where each $\tilde{\beta}_j \in \mathcal{J}$.

Lemma 7. *Under Assumptions A1, A2 and A3, we have that*

a. $\widehat{\boldsymbol{\nu}} \to \boldsymbol{\nu}^*$, $\hat{\pi}_j \to \pi_j^*$ *and* $\hat{\vartheta}_j \to \vartheta_j^*$, $j = 1, \dots, k_0$, *in probability,*

b. $\widetilde{\boldsymbol{\nu}} \to \boldsymbol{\nu}^*$, $\tilde{\pi}_j \to \pi_j^*$ *and* $\tilde{\vartheta}_{2j-1}, \tilde{\vartheta}_{2j} \to \vartheta_j^*$, $j = 1, \dots, k_0$, *in probability.*

Proof of Lemma 7. a. Let $\bar{\Theta}_1$ be the closure of Θ_1 in $\bar{\mathbb{R}}^{d_1}$, where $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$, and similarly for $\bar{\Theta}_2$. For parameters $\boldsymbol{\theta} = (\boldsymbol{\nu}, \vartheta_1, \dots, \vartheta_{k_0})$, $\boldsymbol{\nu} \in \bar{\mathbb{R}}^{d_1}$, $\vartheta_j \in \bar{\Theta}_2$ and k_0 weights $\boldsymbol{\pi}$ let

$$G_{\boldsymbol{\theta},\boldsymbol{\pi}}(t, \boldsymbol{s}) = \sum_{j=1}^{k_0} \pi_j \, I(\vartheta_j \le t, \nu_1 \le s_1, \dots, \nu_{d_1} \le s_{d_1}), \qquad t \in \bar{\Theta}_2, \boldsymbol{s} \in \bar{\Theta}_1$$

denote the corresponding mixing distribution with at most k_0 components. Let $d_w()$ denote a metric which metrizes weak convergence of probabilities on $\bar{\Theta}_2 \times \bar{\Theta}_1$. Our claim follows from the weak convergence

$$d_w\big(G_{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\pi}}}, G_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}\big) \to 0 \qquad \text{in probability,} \tag{26}$$

since by assumption $G_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}$ has k_0 distinct support points, so that the (ordered) support points and weights of $G_{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\pi}}}$ must converge as well.

To show (26), we apply the classical consistency result by Wald (1949), in the version of theorem 5.14 in van der Vaart (1998) for general M-estimators. Since the result only relies on a law of large numbers of an integrable function in the observations, the theorem also applies in case of stationary, ergodic observations (the P in van der Vaart is the univariate marginal distribution). In our case, the parameter is the mixing distribution $G_{\boldsymbol{\theta},\boldsymbol{\pi}}$, $\boldsymbol{\nu} \in \bar{\Theta}_2$, $\vartheta_j \in \bar{\Theta}_1$, $j = 1, \dots, k_0$, which ranges through a compact set by compactness of $\bar{\Theta}_2$ and $\bar{\Theta}_1$, and the criterion function is the mixing density for $\boldsymbol{\nu} \in \Theta_2$, $\vartheta_j \in \Theta_1$, $j = 1, \dots, k_0$,

$$m_{G_{\boldsymbol{\theta},\boldsymbol{\pi}}}(x) := \int_{\Theta_2 \times \Theta_1} f(x; \boldsymbol{s}, t) \, dG_{\boldsymbol{\theta},\boldsymbol{\pi}}(t, \boldsymbol{s}) = f_{mix}^{(k)}(x; \boldsymbol{\theta}, \boldsymbol{\pi}),$$

and $m_{G_{\boldsymbol{\theta},\boldsymbol{\pi}}}(x) = -\infty$, if the parameters are not all contained in Θ_1 and Θ_2. The quasi log-likelihood of section 2, expressed in terms of the mixing distribution, is thus given by

$$l_n\left(G_{\boldsymbol{\theta},\boldsymbol{\pi}}\right) = \sum_{t=1}^{n} m_{G_{\boldsymbol{\theta},\boldsymbol{\pi}}}(X_t).$$

It remains to check the assumptions for Theorem 5.14 in van der Vaart (1998). First, by identifiability of finite mixtures and the existence of the Kulback divergence, Assumption A3 a. and d., from the definiteness of the Kulback-Leibler divergence and the boundary condition $m_{G_{\boldsymbol{\theta},\boldsymbol{\pi}}}(x) = -\infty$ the set of maximizers of $Em_{G_{\boldsymbol{\theta},\boldsymbol{\pi}}}(X_1)$ in $G_{\boldsymbol{\theta},\boldsymbol{\pi}}$ is the singleton $G_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}$, and as noted above the space of mixing distributions is compact.

Now, condition (5.13) in van der Vaart (1998) is immediate from the uniform boundedness condition Assumption A3 c., For condition (5.12), if $d_w(G_{\boldsymbol{\theta}_l,\boldsymbol{\pi}_l}, G_{\boldsymbol{\theta},\boldsymbol{\pi}}) \to 0$, $l \to \infty$, where all mixing distributions as above have at most k_0 support points, then support points of the $G_{\boldsymbol{\theta}_l,\boldsymbol{\pi}_l}$ must converge to some support point of $G_{\boldsymbol{\theta},\boldsymbol{\pi}}$, or their weight converges to 0. Further, the sum of the weights of the support points converging to a specific support of $G_{\boldsymbol{\theta},\boldsymbol{\pi}}$ converges to the weight of that support point. Therefore, (5.12) in van der Vaart (1998) follows by the continuity and limit properties of the densities, Assumption A2 and Assumption A3 b.

Finally, by definition of $G_{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\pi}}}$ we have that

$$l_n\big(G_{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\pi}}}\big) \geq l_n\big(G_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}\big), \tag{27}$$

so that (26) finally follows from theorem 5.14 in van der Vaart (1998).

b. Now consider mixing distributions $G_{\boldsymbol{\theta},\boldsymbol{\pi}}$ with up to $2k_0$ states for parameters $\boldsymbol{\theta} = (\boldsymbol{\nu}, \vartheta_1, \dots, \vartheta_{2k_0})$ and $2k_0$-dimensional weights $\boldsymbol{\pi}$ (potentially with zero entries). We shall show that

$$d_w\big(G_{\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\pi}}}, G_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}\big) \to 0 \qquad \text{in probability,}$$

then by the specific forms of the parameter vector $\widetilde{\boldsymbol{\theta}}$ and the weight vector $\widetilde{\boldsymbol{\pi}}$, the claim in part b. follows. In order to apply theorem 5.14 in van der Vaart (1998), by the arguments in part a. we only need to check that

$$l_n\big(G_{\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\pi}}}\big) \geq l_n(G_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}) + o_P(1). \tag{28}$$

Now $G_{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\pi}}}$ can apparently be written as an element in

$$\{G_{\boldsymbol{\theta},\boldsymbol{\pi}} \ : \ \boldsymbol{\pi} \in \Omega_{2k_0}(\mathcal{J}), \boldsymbol{\nu} \in \Theta_1, \vartheta_{2j-1}, \vartheta_{2j} \in I_j, j = 1, \dots, k_0\}.$$

Since $G_{\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\pi}}}$ is by definition the maximizer of l_n over this class, we have

$$l_n\big(G_{\widetilde{\boldsymbol{\theta}},\widetilde{\boldsymbol{\pi}}}\big) \geq l_n\big(G_{\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\pi}}}\big),$$

which together with (27) implies (28). □

Setting

$$\widetilde{\boldsymbol{s}}_n := \sum_{t=1}^{n} \tilde{\boldsymbol{b}}_{2t},$$

we have the following quadratic approximation to the test statistics.

Lemma 8. *For the test statistic we have under the assumptions A1-A5 that*

$$\Big|R_n - \sup_{\boldsymbol{z}\geq 0}\Big(2\boldsymbol{z}'\widetilde{\boldsymbol{s}}_n - n\boldsymbol{z}'\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{z}\Big)\Big| = o_P(1), \tag{29}$$

where $\{\boldsymbol{z} \geq 0\} := \{(z_1, \dots, z_{k_0}) : z_j \geq 0,\ j = 1, \dots, k_0\}$.

Proof of Lemma 8. The proof is quite similar to those in Chen, Chen and Kalbfleisch (2004) and Li and Chen (2010).

Decompose

$$\begin{aligned} R_n &= 2(l_n^{(2k_0)}(\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\pi}}) - l_n^{(k_0)}(\boldsymbol{\theta}^*, \boldsymbol{\pi}^*)) - 2(l_n^{(k_0)}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}) - l_n^{(k_0)}(\boldsymbol{\theta}^*, \boldsymbol{\pi}^*)) \\ &=: R_n^{(1)} - R_n^{(0)}. \end{aligned}$$

Consider $R_n^{(1)}$: We have $R_n^{(1)} = 2\sum_{t=1}^n \log(1+\delta_t)$, where

$$\delta_t := \frac{f_{mix}^{(2k_0)}(X_t; \widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\pi}}) - f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}.$$

First we derive an upper bound on $R_n^{(1)}$. Since, $\log(1+x) \leq x - x^2/2 + x^3/3$ we shall consider $\sum_{t=1}^n \delta_t^l$ for $l = 1, 2, 3$.

For $t = 1, \ldots, n$ we have

$$\begin{aligned} \sum_{j=1}^{k_0-1} (\tilde{\pi}_j - \pi_j^*)\Delta_{tj}(\widetilde{\boldsymbol{\nu}}, \boldsymbol{\vartheta}^*) &= \sum_{j=1}^{k_0-1} \frac{(\tilde{\pi}_j - \pi_j^*) f(X_t; \widetilde{\boldsymbol{\nu}}, \vartheta_j^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)} - \frac{f(X_t; \widetilde{\boldsymbol{\nu}}, \vartheta_{k_0}^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)} \sum_{j=1}^{k_0-1} (\tilde{\pi}_j - \pi_j^*) \\ &= \sum_{j=1}^{k_0} \frac{(\tilde{\pi}_j - \pi_j^*) f(X_t; \widetilde{\boldsymbol{\nu}}, \vartheta_j^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}, \end{aligned}$$

since

$$\sum_{j=1}^{k_0-1} (\tilde{\pi}_j - \pi_j^*) = (1 - \tilde{\pi}_{k_0}) - (1 - \pi_{k_0}^*) = -\tilde{\pi}_{k_0} + \pi_{k_0}^* = -(\tilde{\pi}_{k_0} - \pi_{k_0}^*).$$

Now, we subtract the right side of the previous equation and add the resulting zero to δ_t. This gives

$$\begin{aligned} \delta_t = &\sum_{j=1}^{k_0-1} (\tilde{\pi}_j - \pi_j^*)\Delta_{tj}(\widetilde{\boldsymbol{\nu}}, \boldsymbol{\vartheta}^*) + \frac{f_{mix}^{(k_0)}\big(X_t; (\widetilde{\boldsymbol{\nu}}, \boldsymbol{\vartheta}^*), \boldsymbol{\pi}^*\big) - f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)} \\ &+ \sum_{j=1}^{k_0} \frac{\tilde{\pi}_j \tilde{\beta}_j \big(f(X_t; \widetilde{\boldsymbol{\nu}}, \tilde{\vartheta}_{2j-1}) - f(X_t; \widetilde{\boldsymbol{\nu}}, \vartheta_j^*)\big)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)} \qquad (30) \\ &+ \sum_{j=1}^{k_0} \frac{\tilde{\pi}_j (1 - \tilde{\beta}_j) \big(f(X_t; \widetilde{\boldsymbol{\nu}}, \tilde{\vartheta}_{2j}) - f(X_t; \widetilde{\boldsymbol{\nu}}, \vartheta_j^*)\big)}{f_{mix}^{(k_0)}(X_t; \boldsymbol{\theta}^*, \boldsymbol{\pi}^*)}. \end{aligned}$$

Now we expand each of the terms in (30). To start, for $t = 1, \ldots, n$ and $j = 1, \ldots, k_0 - 1$

$$\begin{aligned} \Delta_{tj}(\widetilde{\boldsymbol{\nu}}, \boldsymbol{\vartheta}^*) &= \Delta_{tj}(\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*) + \big(\Delta_{tj}(\widetilde{\boldsymbol{\nu}}, \boldsymbol{\vartheta}^*) - \Delta_{tj}(\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*)\big) \\ &= \Delta_{tj}(\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*) + (\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}^*)^T \, \nabla_{\boldsymbol{\nu}}\big(\Delta_{tj}(\overline{\boldsymbol{\nu}}, \boldsymbol{\vartheta}^*)\big) \\ &= \Delta_{tj}(\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*) + \sum_{l=1}^{d} \big((\tilde{\nu}_l - \nu_l^*) \; [R_t^{\{l\}}(\overline{\boldsymbol{\nu}}, \vartheta_j^*) - R_t^{\{l\}}(\overline{\boldsymbol{\nu}}, \vartheta_{k_0}^*)]\big) \end{aligned}$$

for some $\overline{\boldsymbol{\nu}}$ between $\widetilde{\boldsymbol{\nu}}$ and $\boldsymbol{\nu}^*$, and where $R_t^{\{l\}}$ is defined in (22). Therefore, we obtain

$$\sum_{j=1}^{k_0-1}(\tilde{\pi}_j - \pi_j^*)\Delta_{tj}(\widetilde{\boldsymbol{\nu}}, \boldsymbol{\vartheta}^*) = \sum_{j=1}^{k_0-1}(\tilde{\pi}_j - \pi_j^*)\Delta_{tj}(\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*) + \varepsilon_{tn}^{(A)},$$

where

$$\varepsilon_{tn}^{(A)} := \sum_{j=1}^{k_0-1}\Big((\tilde{\pi}_j - \pi_j^*) \cdot \sum_{l=1}^{d}\Big[\big(\tilde{\nu}_l - \nu_l^*\big)\ \big(R_t^{\{l\}}(\widetilde{\boldsymbol{\nu}}, \vartheta_j^*) - R_t^{\{l\}}(\widetilde{\boldsymbol{\nu}}, \vartheta_{k_0}^*)\big)\Big]\Big).$$

Therefore,

$$\sum_{t=1}^{n}\sum_{j=1}^{k_0-1}(\tilde{\pi}_j - \pi_j^*)\Delta_{tj}(\widetilde{\boldsymbol{\nu}}, \boldsymbol{\vartheta}^*) = \sum_{t=1}^{n}\sum_{j=1}^{k_0-1}\big((\tilde{\pi}_j - \pi_j^*)\Delta_{tj}(\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*)\big) + \varepsilon_n^{(A)}, \tag{31}$$

where due to Lemmas 6 and 7,

$$\varepsilon_n^{(A)} = \sum_{t=1}^{n}\varepsilon_{tn}^{(A)} = o_P(n^{1/2})\sum_{j=1}^{k_0-1}\big(\tilde{\pi}_j - \pi_j^*\big).$$

The second part of (30) can be expanded similarly. Here, for brevity we omit $X_t, \boldsymbol{\pi}^*, \boldsymbol{\vartheta}^*, k_0$ in the marginal mixture, i.e. writing $f_{t,mix}(\boldsymbol{\nu})$ for $f_{mix}^{(k_0)}\big(X_t; (\boldsymbol{\nu}, \boldsymbol{\vartheta}^*), \boldsymbol{\pi}^*\big)$. We obtain

$$\begin{aligned}\frac{f_{t,mix}(\widetilde{\boldsymbol{\nu}}) - f_{t,mix}(\boldsymbol{\nu}^*)}{f_{t,mix}(\boldsymbol{\nu}^*)} &= \frac{\big(\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\big)^T\ \nabla_{\boldsymbol{\nu}}\ f_{t,mix}(\boldsymbol{\nu}^*) + 1/2\ \big(\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\big)^T\ \nabla_{\boldsymbol{\nu}\boldsymbol{\nu}}\ f_{t,mix}(\overline{\boldsymbol{\nu}})\ \big(\widetilde{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\big)}{f_{t,mix}(\boldsymbol{\nu}^*)}\\ &= \sum_{l=1}^{d}(\tilde{\nu}_l - \nu_l^*)\ U_t^{\{l\}}(\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*) + 1/2\ \sum_{l,i=1}^{d}\big\{(\tilde{\nu}_l - \nu_l^*)(\widetilde{\nu}_{1i} - \nu_i^*)\ V_t^{\{l,i\}}(\overline{\boldsymbol{\nu}}, \boldsymbol{\vartheta}^*)\\ &=: \sum_{l=1}^{d}(\tilde{\nu}_l - \nu_l^*)\ U_t^{\{l\}}(\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*) + \varepsilon_{tn}^{(B)}\end{aligned}$$

where and $\overline{\boldsymbol{\nu}}$ is again between $\widetilde{\boldsymbol{\nu}}$ and $\boldsymbol{\nu}^*$, and $V_t^{\{l,i\}}(\boldsymbol{\nu}, \boldsymbol{\vartheta})$ is defined in (22). By Lemmas 6 and 7 it follows that

$$\begin{aligned}&\sum_{i=1}^{n}\frac{f_{mix}^{(k_0)}\big(X_t; (\boldsymbol{\nu}, \boldsymbol{\vartheta}^*), \boldsymbol{\pi}^*\big) - f_{mix}^{(k_0)}\big(X_t; (\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*), \boldsymbol{\pi}^*\big)}{f_{mix}^{(k_0)}\big(X_t; (\boldsymbol{\nu}^*, \boldsymbol{\vartheta}^*), \boldsymbol{\pi}^*\big)}\\ =&\sum_{i=1}^{n}\sum_{l=1}^{d}\big(U_t^{\{l\}}(\boldsymbol{\nu}^*)\ (\tilde{\nu}_l - \nu_l^*)\big) + \varepsilon_n^{(B)},\end{aligned} \tag{32}$$

where

$$\varepsilon_n^{(B)} = \sum_{t=1}^{n}\varepsilon_{tn}^{(B)} = o_P(n^{1/2})\ \sum_{l=1}^{d}(\tilde{\nu}_l - \nu_l^*).$$

To expand the remaining term in (30), we now consider

$$\big(f(X_t;\widetilde{\boldsymbol{\nu}},\tilde{\vartheta}_{2j-i}) - f(X_t;\widetilde{\boldsymbol{\nu}},\vartheta_j^*)\big)/f_{mix}^{(k_0)}(X_t;\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)$$

for $t = 1,\dots,n$, $j = 1,\dots,k_0$ and $i = 0,1$. We have

$$\frac{f(X_t;\widetilde{\boldsymbol{\nu}},\tilde{\vartheta}_{2j-i}) - f(X_t;\widetilde{\boldsymbol{\nu}},\vartheta_j^*)}{f_{mix}^{(k_0)}(X_t;\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)} = \frac{f(X_t;\boldsymbol{\nu}^*,\tilde{\vartheta}_{2j-i}) - f(X_t;\boldsymbol{\nu}^*,\vartheta_j^*)}{f_{mix}^{(k_0)}(X_t;\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)} + \varepsilon_{tjin}^{(C_1)}$$

where

$$\begin{aligned}\varepsilon_{tjin}^{(C_1)} &:= \sum_{l=1}^{d}(\tilde{\nu}_l - \nu_l^*)\frac{f_{\nu_l}(X_t;\overline{\boldsymbol{\nu}}_j,\tilde{\vartheta}_{2j-i}) - f_{\nu_l}(X_t;\overline{\boldsymbol{\nu}}_j,\vartheta_j^*)}{f_{mix}^{(k_0)}(X_t;\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)}\\ &= \sum_{l=1}^{d}(\tilde{\nu}_l - \nu_l^*)\ (\tilde{\vartheta}_{2j-i} - \vartheta_j^*)\frac{f_{\nu_l\vartheta}(X_t;\overline{\boldsymbol{\nu}}_j,\overline{\vartheta}_{2j-i})}{f_{mix}^{(k_0)}(X_t;\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)},\end{aligned}$$

and $\overline{\vartheta}_{2j-i}$ and $\overline{\boldsymbol{\nu}}_j$ lie between the appropriate parameters.

Moreover,

$$\begin{aligned}&\frac{f(X_t;\boldsymbol{\nu}^*,\tilde{\vartheta}_{2j-i}) - f(X_t;\boldsymbol{\nu}^*,\vartheta_j^*)}{f_{mix}^{(k_0)}(X_t;\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)}\\ &= Y_{tj}'(\tilde{\vartheta}_{2j-i} - \vartheta_j^*) + 1/2\ Y_{tj}''(\tilde{\vartheta}_{2j-i} - \vartheta_j^*)^2 + 1/6\ Y_{tj}'''(\boldsymbol{\nu}^*,\overline{\vartheta}_{2j-i}')(\tilde{\vartheta}_{2j-i} - \vartheta_j^*)^3.\end{aligned}$$

Therefore, setting $\varepsilon_{tjin}^{(C_2)} := 1/6\ Y_{tj}'''(\boldsymbol{\nu}^*,\tilde{\vartheta}_{2j-i})(\tilde{\vartheta}_{2j-i} - \vartheta_j^*)^3$ let us define the error term by

$$\varepsilon_{tn}^{(C)} := \sum_{j=1}^{k_0}\Big[\tilde{\pi}_j\tilde{\beta}_j\Big(\varepsilon_{tj1n}^{(C_1)} + \varepsilon_{tj1n}^{(C_2)}\Big) + \tilde{\pi}_j(1-\tilde{\beta}_j)\Big(\varepsilon_{tj0n}^{(C_1)} + \varepsilon_{tj0n}^{(C_2)}\Big)\Big],$$

We obtain that

$$\begin{aligned}&\sum_{t=1}^{n}\sum_{j=1}^{k_0}\frac{\tilde{\pi}_j\tilde{\beta}_j\big(f(X_t;\widetilde{\boldsymbol{\nu}},\tilde{\vartheta}_{2j-1}) - f(X_t;\widetilde{\boldsymbol{\nu}},\vartheta_j^*)\big) + \tilde{\pi}_j(1-\tilde{\beta}_j)\big(f(X_t;\widetilde{\boldsymbol{\nu}},\tilde{\vartheta}_{2j}) - f(X_t;\widetilde{\boldsymbol{\nu}},\vartheta_j^*)\big)}{f_{mix}^{(k_0)}(X_t;\boldsymbol{\theta}^*,\boldsymbol{\pi}^*)}\\ =&\sum_{t=1}^{n}\sum_{j=1}^{k_0}\big(\tilde{\pi}_j\widehat{m}_{1j}Y_{tj}' + \tilde{\pi}_j\widehat{m}_{2j}Y_{tj}''\big) + \varepsilon_n^{(C)},\end{aligned} \tag{33}$$

where

$$\widehat{m}_{hj} := \tilde{\beta}_j(\tilde{\vartheta}_{2j-1} - \vartheta_j^*)^h + (1-\tilde{\beta}_j)(\tilde{\vartheta}_{2j} - \vartheta_j^*)^h, \qquad \text{for} \quad h = 1,2$$

and

$$\varepsilon_n^{(C)} := \sum_{t=1}^{n}\varepsilon_{tn}^{(C)} = o_P(n^{1/2})\ \sum_{j=1}\tilde{\pi}_j(\widehat{m}_{1j} + \widehat{m}_{2j})$$

by Lemmas 6 and 7. Due to equations (31), (32) and (33) we may write

$$\sum_{t=1}^{n} \delta_t = \sum_{t=1}^{n} \boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}} + \varepsilon_n,$$

where $\varepsilon_n = \varepsilon_n^{(A)} + \varepsilon_n^{(B)} + \varepsilon_n^{(C)}$ and

$$\begin{aligned}\widehat{\boldsymbol{\tau}} = \big(&\tilde{\pi}_1 - \pi_1^*, \ldots, \tilde{\pi}_{k_0-1} - \pi_{k_0-1}^*, \tilde{\nu}_1 - \nu_1^*, \ldots, \tilde{\nu}_{d_1} - \nu_{d_1}^*, \\ &\tilde{\pi}_1 \widehat{m}_{11}, \ldots, \tilde{\pi}_{k_0} \widehat{m}_{1k_0}, \tilde{\pi}_1 \widehat{m}_{21}, \ldots, \tilde{\pi}_{k_0} \widehat{m}_{2k_0}\big)^T.\end{aligned} \tag{34}$$

Using $|x| \leq 1 + x^2$ we further see that

$$|\varepsilon_n| \leq o_P(1) \sum_{j=1}^{3k_0-1+d_1} n^{1/2} |\tau_j| \leq o_P(1) \sum_{j=1}^{3k_0-1+d_1} (n\ \tau_j^2 + 1) = o_P(n)\ \widehat{\boldsymbol{\tau}}^T \widehat{\boldsymbol{\tau}} + o_P(1).$$

Turning to $\sum_{t=1}^n \delta_i^2$ we have

$$\sum_{t=1}^{n} \delta_t^2 = \sum_{t=1}^{n} (\boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}})^2 + \varepsilon_n^{(Q)}$$

where

$$\varepsilon_n^{(Q)} := \sum_{t=1}^{n} \big(\varepsilon_{tn}^{(A)} + \varepsilon_{tn}^{(B)} + \varepsilon_{tn}^{(C)}\big)^2 + 2 \sum_{t=1}^{n} \big[\boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}}\ (\varepsilon_{tn}^{(A)} + \varepsilon_{tn}^{(B)} + \varepsilon_{tn}^{(C)})\big].$$

Now

$$\begin{aligned}|\varepsilon_{tn}^{(A)}| &\leq o_P(1)\ g(X_t)^{1/3} \sum_{j=1}^{k_0-1} (\tilde{\pi}_j - \pi_j^*), \\ |\varepsilon_{tn}^{(B)}| &\leq o_P(1)\ g(X_t)^{1/3} \sum_{l=1}^{d} (\tilde{\nu}_l - \nu_l^*), \\ |\varepsilon_{tn}^{(C)}| &\leq o_P(1)\ g(X_t)^{1/3} \sum_{j=1}^{k_0} \tilde{\pi}_j\ (\widehat{m}_{1j} + \widehat{m}_{2j}).\end{aligned}$$

By integrability of $g(X_t)$, we get from the ergodic theorem

$$\begin{aligned}&\sum_{t=1}^{n} \big(\varepsilon_{tn}^{(A)} + \varepsilon_{tn}^{(B)} + \varepsilon_{tn}^{(C)}\big)^2 \leq 4 \sum_{t=1}^{n} \big((\varepsilon_{tn}^{(A)})^2 + (\varepsilon_{tn}^{(B)})^2 + (\varepsilon_{tn}^{(C)})^2\big) \\ \leq &o_P(n)\ \widehat{\boldsymbol{\tau}}^T \widehat{\boldsymbol{\tau}} + o_P(1) = O_P(\varepsilon_n) + o_P(1).\end{aligned}$$

As in Li and Chen (2010), by the Cauchy inequality the second error term of the expansion of $\sum_{t=1}^n \delta_t^2$ results in no higher order. Since the remainder term of the expansion of $\sum_{t=1}^n \delta_t^3$ is also $O_P(\varepsilon_n)$, we obtain the following bound for $R_n^{(1)}$

$$R_n^{(1)} \leq 2 \sum_{t=1}^n \boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}} - \sum_{t=1}^n (\boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}})^2 + 2/3 \sum_{t=1}^n (\boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}})^3 + O_P(\varepsilon_n). \tag{35}$$

In order to estimate the cubic term, from

$$n^{-1} \sum_{t=1}^n \boldsymbol{b}_t \boldsymbol{b}_t^T \overset{a.s.}{\to} E(\boldsymbol{b}_1 \boldsymbol{b}_1^T)$$

we obtain

$$\sum_{t=1}^n (\boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}})^2 = n\ \widehat{\boldsymbol{\tau}}^T \boldsymbol{\Sigma} \widehat{\boldsymbol{\tau}}\ (1 + o_P(1)).$$

Because of the positive definiteness of $\boldsymbol{\Sigma}$, we further get

$$\sum_{t=1}^n (\boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}})^2 + O_P(\varepsilon_n) = n\ \widehat{\boldsymbol{\tau}}^T \boldsymbol{\Sigma} \widehat{\boldsymbol{\tau}}\ (1 + o_P(1)) + o_P(1)$$

and

$$\frac{\sum_{t=1}^n (\boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}})^3}{\sum_{t=1}^n (\boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}})^2} \leq \max(|\widehat{\boldsymbol{\tau}}|) = o_P(1).$$

Thus, (35) reduces to the following bound

$$R_n^{(1)} \leq 2 \sum_{t=1}^n \boldsymbol{b}_t^T \widehat{\boldsymbol{\tau}} - n\ \widehat{\boldsymbol{\tau}}^T \boldsymbol{\Sigma} \widehat{\boldsymbol{\tau}}\ (1 + o_P(1)) + o_P(1).$$

Now, analogously to Li and Chen (2010) the just established upper bound for $R_n^{(1)}$ is bounded by $O_P(1)$ and thus we deduce $\widehat{\boldsymbol{\tau}} = O_P(n^{-1/2})$. As for $R_n^{(0)}$, the classic expansion is

$$R_n^{(0)} = \sum_{t=1}^n \boldsymbol{b}_{1t}^T \big(n\ \boldsymbol{\Sigma}_{11}\big)^{-1} \sum_{t=1}^n \boldsymbol{b}_{1t}^T + o_P(1).$$

Therefore,

$$\begin{aligned} R_n^{(1)} - R_n^{(0)} \leq & \sup_{\boldsymbol{\tau} \in \mathbb{R}^{3k_0 - 1 + d_1}} \Big(2\ (\sum_{t=1}^n \boldsymbol{b}_t^T) \boldsymbol{\tau} - n \boldsymbol{\tau}^T \boldsymbol{\Sigma} \boldsymbol{\tau}\Big) - \sum_{t=1}^n \boldsymbol{b}_{1t}^T \big(n\ \boldsymbol{\Sigma}_{11}\big)^{-1} \sum_{t=1}^n \boldsymbol{b}_{1t}^T + o_P(1) \\ = & \sup_{\boldsymbol{\tau}_1 \in \mathbb{R}^{2k_0 - 1 + d_1}} \Big(2 (\sum_{t=1}^n \boldsymbol{b}_{1t}^T) \boldsymbol{\tau}_1 - n \boldsymbol{\tau}_1^T \boldsymbol{\Sigma}_{11} \boldsymbol{\tau}_1\Big) + \sup_{\{\boldsymbol{\tau}_2 \geq 0\}} \Big(2 \boldsymbol{\tau}_2^T (\sum_{t=1}^n \tilde{\boldsymbol{b}}_{2t}) - n \boldsymbol{\tau}_2^T \widetilde{\boldsymbol{\Sigma}}_{22} \boldsymbol{\tau}_2\Big) \\ & - \sum_{t=1}^n \boldsymbol{b}_{1t}^T \big(n\ \boldsymbol{\Sigma}_{11}\big)^{-1} \sum_{t=1}^n \boldsymbol{b}_{1t}^T + o_P(1) \\ = & \sup_{\{\boldsymbol{\tau}_2 \geq 0\}} \Big(2 \boldsymbol{\tau}_2^T (\sum_{t=1}^n \tilde{\boldsymbol{b}}_{2t}) - n \boldsymbol{\tau}_2^T \widetilde{\boldsymbol{\Sigma}}_{22} \boldsymbol{\tau}_2\Big) + o_P(1), \end{aligned}$$

where $\{\boldsymbol{z} \geq 0\} := \{(z_1, \ldots, z_{k_0}) : z_j \geq 0,\ j = 1, \ldots, k_0\}$.

The reasoning why this upper bound is attained in our setting is analogous to the i.i.d. case without structural parameters, i.e. to this in Li and Chen (2010). Let $\boldsymbol{\tau}^* := (\tilde{\boldsymbol{\tau}}_1^*, \boldsymbol{\tau}_2^*)$, with

$$\begin{aligned}\tilde{\boldsymbol{\tau}}_1^* &= \arg \sup_{\boldsymbol{\tau}_1 \in \mathbb{R}^{2k_0-1+d_1}} \Big(2\big(\sum_{t=1}^n \boldsymbol{b}_{1t}^T\big)\boldsymbol{\tau}_1 - n\boldsymbol{\tau}_1^T \boldsymbol{\Sigma}_{11} \boldsymbol{\tau}_1\Big) = n^{-1}\boldsymbol{\Sigma}_{11}^{-1} \sum_{t=1}^n b_{1t} = O_P(n^{-1/2}),\\ \boldsymbol{\tau}_2^* &= \arg \sup_{\boldsymbol{\tau}_2 \geq 0} \Big(2\boldsymbol{\tau}_2^T\big(\sum_{t=1}^n \tilde{\boldsymbol{b}}_{2t}\big) - n\boldsymbol{\tau}_2^T \widetilde{\boldsymbol{\Sigma}}_{22} \boldsymbol{\tau}_2\Big),\end{aligned} \tag{36}$$

denote the vector attaining the upper bound of the previous display, where the order assessment of $\tilde{\boldsymbol{\tau}}_1^*$ is due to the CLT for stationary weak dependent processes.

The unrestricted optimal point of the second function in (36) is $n^{-1}\widetilde{\boldsymbol{\Sigma}}_{22}^{-1}\widetilde{\boldsymbol{s}}_n = O_P(n^{-1/2})$, since $n^{-1/2}\widetilde{\boldsymbol{s}}_n$ is asymptotically normal. This implies that the unrestricted and hence the restricted optimum of the second function are bounded by $n^{-1}\widetilde{\boldsymbol{s}}_n^T\widetilde{\boldsymbol{\Sigma}}_{22}^{-1}\widetilde{\boldsymbol{s}}_n = O_P(1)$. Therefore, we also have $\boldsymbol{\tau}_2^* = O_P(n^{-1/2})$, because otherwise we would get a contradiction to the $O_P(1)$ upper bound.

Denote by $\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\pi}}$ the parameter leading, under the same mapping as in (34), to $\boldsymbol{\tau}^*$. Due to the non-negativity restriction in (36) and $\boldsymbol{\tau}^* = O_P(n^{-1/2})$ its existence is obvious. Further, since $\boldsymbol{\tau}^* = O_P(n^{-1/2})$

$$\begin{gathered}\bar{\boldsymbol{\pi}} - \boldsymbol{\pi}^* = O_P(n^{-1/2}), \qquad \bar{\boldsymbol{\nu}} - \boldsymbol{\nu}^* = O_P(n^{-1/2}),\\ \bar{\vartheta}_{2j-1} - \vartheta_j^* = O_P(n^{-1/4}), \qquad \bar{\vartheta}_{2j} - \vartheta_j^* = O_P(n^{-1/4}), \qquad j = 1, \ldots, k_0.\end{gathered}$$

Now, due to the previous order assessment and a further expansion, see Chen, Chen and Kalbfleisch (2004, proof of Lemma 2) for a similar argument, we obtain

$$\begin{aligned}\bar{R}_n^{(1)} :&= 2(l_n^{(2k_0)}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\pi}}) - l_n^{(k_0)}(\boldsymbol{\theta}^*, \boldsymbol{\pi}^*)) = 2\ \big(\sum_{t=1}^n \boldsymbol{b}_t^T\big)\boldsymbol{\tau}^* - n(\boldsymbol{\tau}^*)^T\boldsymbol{\Sigma}\boldsymbol{\tau}^* + o_P(1)\\ &= \sup_{\boldsymbol{\tau} \in \mathbb{R}^{3k_0-1+d_1}} \Big(2\ \big(\sum_{t=1}^n \boldsymbol{b}_t^T\big)\boldsymbol{\tau} - n\boldsymbol{\tau}^T\boldsymbol{\Sigma}\boldsymbol{\tau}\Big) + o_P(1)\end{aligned}$$

and thus

$$\bar{R}_n^{(1)} - R_n^{(0)} = \sup_{\{\boldsymbol{\tau}_2 \geq 0\}} \Big(2\boldsymbol{\tau}_2^T\big(\sum_{t=1}^n \tilde{\boldsymbol{b}}_{2t}\big) - n\boldsymbol{\tau}_2^T\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{\tau}_2\Big) + o_P(1).$$

Since $R_n^{(1)} \geq \bar{R}_n^{(1)}$ due to the maximizing property of the QMLE under the alternative, it holds

$$R_n^{(1)} - R_n^{(0)} \geq \bar{R}_n^{(1)} - R_n^{(0)} = \sup_{\{\boldsymbol{\tau}_2 \geq 0\}} \Big(2\boldsymbol{\tau}_2^T\big(\sum_{t=1}^n \tilde{\boldsymbol{b}}_{2t}\big) - n\boldsymbol{\tau}_2^T\widetilde{\boldsymbol{\Sigma}}_{22}\boldsymbol{\tau}_2\Big) + o_P(1).$$

This ends the proof of Lemma 8.

□

To conclude the proof of Theorem 2 we show that $(\tilde{\boldsymbol{b}}_{2t})_t$ is a martingale difference sequence, which is quite analogous to the case in Appendix 1. Then (7) follows as in the i.i.d. setting of Li and Chen (2010).

Consider the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ with

$$\mathcal{F}_t := \sigma\left(S_j, \boldsymbol{b}_j; j \leq t\right) \text{ for } t \in \mathbb{N}.$$

Then $\mathcal{L}\left(\boldsymbol{b}_t \mid \mathcal{F}_{t-1}\right) = \mathcal{L}\left(\boldsymbol{b}_t \mid S_{t-1}\right)$, and therefore also $\mathcal{L}(\tilde{\boldsymbol{b}}_{2t}|\mathcal{F}_{t-1}) = \mathcal{L}(\tilde{\boldsymbol{b}}_{2t}|S_{t-1})$. Thus, it remains to show that

$$E\left(\tilde{\boldsymbol{b}}_{2t} \mid S_{t-1} = j\right) = 0, \qquad j = 1, \ldots, k_0. \tag{37}$$

Let

$$\boldsymbol{\lambda}_h := E\left(\boldsymbol{b}_1 \mid S_1 = h\right) \text{ and } \gamma_{jh} := P\left(S_t = h \mid S_{t-1} = j\right) \text{ for } h = 1, \ldots, k_0.$$

As the Markov chain can adopt k_0 states under the hypothesis, it follows that

$$E\left(\boldsymbol{b}_t \mid S_{t-1} = j\right) = \sum_{h=1}^{k_0} \gamma_{jh}\boldsymbol{\lambda}_h \quad \text{ and } \quad E\left(\boldsymbol{b}_{lt} \mid S_{t-1} = j\right) = \sum_{h=1}^{k_0} \gamma_{jh}\boldsymbol{\lambda}_{hl} \text{ for } l = 1, 2,$$

where we partition $\boldsymbol{\lambda}_h^T = \left(\boldsymbol{\lambda}_{h1}^T, \boldsymbol{\lambda}_{h2}^T\right)$ with $\boldsymbol{\lambda}_{h1} \in \mathbb{R}^{2k_0-1+d}$. We get

$$E\left(\tilde{\boldsymbol{b}}_{2t} \mid S_{t-1} = j\right)^T = \sum_{h=1}^{k_0} \gamma_{jh}\boldsymbol{\lambda}_{h2}^T - \left(\sum_{h=1}^{k_0} \gamma_{jh}\boldsymbol{\lambda}_{h1}^T\right) \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{38}$$

Since $0 = E\left(\boldsymbol{b}_1\right) = \sum_{h=1}^{k_0} \pi_h^*\boldsymbol{\lambda}_h$, we obtain

$$\boldsymbol{\lambda}_{k_0} = \sum_{h=1}^{k_0-1} c_h\boldsymbol{\lambda}_h, \text{ with } c_h := -\pi_h^*/\pi_{k_0}^*, \tag{39}$$

and inserting (39) in (38) gives setting $d_h := \gamma_{jh} + \gamma_{jk_0}c_h$ for $h = 1, \ldots, k_0 - 1$,

$$E\left(\tilde{\boldsymbol{b}}_{2t} \mid S_{t-1} = j\right)^T = \sum_{h=1}^{k_0-1} d_h\boldsymbol{\lambda}_{h2}^T - \left(\sum_{h=1}^{k_0-1} d_h\boldsymbol{\lambda}_{h1}^T\right) \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{40}$$

Now observe that

$$E\left(\Delta_{1h}\boldsymbol{b}_1\right) = \boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{k_0}, \quad h = 1, \ldots, k_0 - 1.$$

Let

$$\boldsymbol{S} := \begin{pmatrix} \mathbf{0}_{d\times(k_0-1)} \\ \boldsymbol{I}_{k_0-1} \\ \mathbf{0}_{(2k_0)\times(k_0-1)} \end{pmatrix}, \qquad \boldsymbol{T} := \begin{pmatrix} \mathbf{0}_{d\times(k_0-1)} \\ \boldsymbol{I}_{k_0-1} \\ \mathbf{0}_{k_0\times(k_0-1)} \end{pmatrix},$$

then from the definition of $\boldsymbol{\Sigma}$ and (39) we get

$$\begin{aligned} \boldsymbol{\Sigma}\boldsymbol{S} &= \left[E\left(\Delta_{11}\boldsymbol{b}_1\right), \ldots,\ E\left(\Delta_{1k_0-1}\boldsymbol{b}_1\right)\right] \\ &= \left[\boldsymbol{\lambda}_1 - \sum_{h=1}^{k_0-1} c_h\boldsymbol{\lambda}_h, \ldots,\ \boldsymbol{\lambda}_{k_0-1} - \sum_{h=1}^{k_0-1} c_h\boldsymbol{\lambda}_h\right] =: \boldsymbol{\Lambda}, \end{aligned} \tag{41}$$

where $\mathbf{0}.$ denotes matrices of zeros and $\boldsymbol{I}.$ are identity matrices, all with the appropriate dimensions. This result also holds for the partitioned $\boldsymbol{\lambda}$ vectors, i.e.

$$\boldsymbol{\Sigma}_{l1}\boldsymbol{T} = \left(\boldsymbol{\lambda}_{1l} - \textstyle\sum_{h=1}^{k_0-1} c_h\boldsymbol{\lambda}_{hl}\ ,\ \ \cdots\ \ ,\ \boldsymbol{\lambda}_{(k_0-1)l} - \textstyle\sum_{h=1}^{k_0-1} c_h\boldsymbol{\lambda}_{hl}\right), \quad l = 1, 2.$$

As in Appendix 1, one shows that

$$\operatorname{span}(\boldsymbol{\Lambda}) = \operatorname{span}\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{k_0-1}\},$$

where $\operatorname{span}(\boldsymbol{\Lambda})$ denotes the space spanned by the columns of $\boldsymbol{\Lambda}$. Therefore, there is a matrix $\boldsymbol{M} \in \mathbb{R}^{(k_0-1)\times(k_0-1)}$ such that $\boldsymbol{\Lambda}\,\boldsymbol{M} = \left[d_1\boldsymbol{\lambda}_1\ , \ldots,\ d_{k_0-1}\boldsymbol{\lambda}_{k_0-1}\right]$ and thus from (41)

$$\boldsymbol{\Sigma}\,\boldsymbol{S}\boldsymbol{M} = \left[d_1\boldsymbol{\lambda}_1, \ldots,\ d_{k_0-1}\boldsymbol{\lambda}_{k_0-1}\right]$$

and hence for the submatrices of $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma}_{l1}\,\boldsymbol{T}\boldsymbol{M} = \left[d_1\boldsymbol{\lambda}_{1l}, \ldots, d_{k_0-1}\boldsymbol{\lambda}_{(k_0-1)l}\right], \qquad l = 1, 2.$$

This implies

$$\left(1\ , \cdots\ ,\ 1\right)\boldsymbol{M}^T\boldsymbol{T}^T\boldsymbol{\Sigma}_{1l} = \sum_{h=1}^{k_0-1} d_h\boldsymbol{\lambda}_{hl}^T, \qquad l = 1, 2.$$

Using this subsequently for $l = 1$ and $l = 2$ we get

$$\left(\sum_{h=1}^{k_0-1} d_h\boldsymbol{\lambda}_{h1}^T\right)\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \left(1\ , \cdots\ ,\ 1\right)\boldsymbol{M}^T\boldsymbol{T}^T\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \sum_{h=1}^{k_0-1} d_h\boldsymbol{\lambda}_{h2}^T$$

which due to (40) implies (37). This ends the proof of Theorem 2.

C. Supplement: Details on Simulations, additional results, further applications

C.1. Details on Simulation

Normal HMMs with switching means and variances

The constant in the penalty function in $\widetilde{pl_n}^{(2k_0)}$ in (3) is a crucial choice for the finite sample behavior of the test, since it appears directly in the quasi-likelihood ratio test statistic. Therefore we follow Chen, Li and Fu (2012) and Chen and Li (2011) by reproducing their tuning of this constant using the design from the first mentioned paper: For different null models, sample sizes and penalty constants (of the alternative fit) the actual rejection rate given a base tuning level (e.g. 5%) is simulated. Then a function reflecting the difference of these levels is linearly regressed by the misclassification rate of the model (see Maitra and Melnykov, 2010), the inverse of the sample size and a function of the penalty constant. The root of the regressed function finally gives the tuned penalty constant as a function of the sample size and the misclassification rate of a given model.

We obtained the following tuning formulas $a_n^{(2)}$ and $a_n^{(3)}$ when testing for two and three states

$$
\begin{aligned}
a_n^{(2)} &= \frac{0.35\ \exp(-1.168 - 0.772\ \bar{w}_{12} - 153.712\ n^{-1})}{1 + \exp(-1.168 - 0.772\ \bar{w}_{12} - 153.712\ n^{-1})},\\
a_n^{(3)} &= \frac{0.35\ \exp(-1.237 - 0.34\ \bar{w}_{123} - 186.733\ n^{-1})}{1 + \exp(-1.237 - 0.34\ \bar{w}_{123} - 186.733\ n^{-1})},
\end{aligned}
$$

where n denotes the sample size, $\bar{w}_{12} = \log(w_{12}/(1-w_{12}))$ and $\bar{w}_{123} = \log(w_{12}w_{23} /(1-w_{12})(1-w_{23}))$, where $w_{ij} = P(S_t = i)\ w_{j|i} + P(S_t = j)\ w_{i|j}$ is the average misclassification rate between state i and j, i.e. $w_{i|j}$ denotes the probability to assign an observation to state i although it comes from state j.

Skew-Normal HMM with switching variances

Consider the skew-normal family $\mathcal{SN}(\alpha)$ defined by the density

$$
2\,\phi(x)\,\Phi(\alpha x), \qquad \alpha \in \mathbb{R}
$$

where Φ is the distribution function of the standard normal and ϕ its density function, see Azzalini (1985). Let $\delta = \alpha/(1+\alpha^2)$, and let $\mu(\alpha) = \delta\sqrt{2/\pi}$, $\sigma^2(\alpha) = 1 - 2\,\delta^2/\pi$, denote the mean and variance of $\mathcal{SN}(\alpha)$. We shall use the standardized version of the skew-normal distribution with density

$$
f_0(x,\alpha) = \sigma(\alpha)\ 2\ \phi\big(\sigma(\alpha)x + \mu(\alpha)\big)\ \Phi\big(\alpha(\sigma(\alpha)x + \mu(\alpha))\big).
$$

As proposed by Azzalini (1985), we now consider a parametrization of the family of skew-normals, which is given by the location-scale-shape family generated by the upper distribution and a different parametrization of the skewness. Indeed, if Z_α has density $f_0(x, \alpha)$, we parametrize in $\xi(\alpha) = E(Z_\alpha^3)$ instead of α, and further consider the location-scale-shape family generated by r.v.'s $\mu + \sigma Z_\alpha$ with $\mu \in \mathbb{R}$ and $\sigma > 0$. We refer to this family by $\mathcal{SN}(\mu, \sigma, \xi)$ or simply as the family of skew-normal distributions (although this is not the common parametrization). Using this parametrization has two major advantages: The parameters can directly be interpreted as the mean, the standard deviation and the skewness. Further, the Fisher information of this family is not singular, see Azzalini (1985).

Penalty on ξ: We require a penalty function that converges to zero for $n \to \infty$. Since we don't want to push the estimated skewness towards zero, we use a penalty function $pen_\xi : (-0.996, 0.996) \to (-\infty, 0]$ which is very flat in the middle and drops rapidly at the boundary of the domain,

$$pen_\xi(y) = -5000/n \cdot \Big(\exp\big([h(y) - h(0)]^4\big) - 1 \Big), \quad h(x) := \phi^{-1}\big((x + 0.996)/1.992\big),$$

where ϕ^{-1} is the quantile function of a normal distribution with standard deviation 1/3 and mean zero. Nevertheless, the penalty function has it maximum zero for $\xi = 0$, but due to the flat structure it mainly prevents the divergence of $\hat{\xi}$.

C.2. Additional Simulations

Poisson HMMs: Levels

We simulate from a three-state Poisson HMM with parameters $1, 5$ and 15 for the state-dependent Poisson distributions. We use the transition probability matrices $\Gamma_7, \dots, \Gamma_{12}$ as defined in Table 1b. The simulated sizes, using the test according to Theorem 2, are listed in Table 7: For both sample sizes ($n = 250$ and $n = 500$) the tests are slightly conservative, but improve with rising sample size.

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
6.91	7.26	7.36	7.27	7.48	7.05
3.34	3.40	3.43	3.33	3.26	3.20
0.37	0.46	0.47	0.60	0.47	0.53

(a) Sample size $n = 250$

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
8.83	8.60	8.74	8.14	8.22	8.72
4.24	4.19	4.33	3.97	3.97	4.31
0.70	0.85	0.53	0.62	0.69	0.76

(b) Sample size $n = 500$

Table 7: Simulated rejection rates in percent for Poisson HMMs under the true hypothesis of three regimes and different sample sizes n. Each table lists line by line the rejection rates on levels $10\%, 5\%$ and 1%.

Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6
9.11	8.84	9.30	9.37	9.48	9.23
4.61	4.28	4.49	4.80	5.21	5.07
0.94	0.95	1.01	1.15	1.02	1.16

(a) $n = 500$ and SN_2

Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6
9.91	4.72	9.36	9.88	8.49	9.70
5.39	4.90	4.89	4.60	4.13	4.89
1.02	1.02	0.99	1.01	0.88	1.03

(b) $n = 1000$ and SN_2

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
10.14	9.86	10.02	11.84	11.60	11.84
5.92	5.86	5.44	7.02	7.08	7.02
1.56	1.74	1.42	2.08	1.90	2.04

(c) $n = 500$ and SN_4

Γ_7	Γ_8	Γ_9	Γ_{10}	Γ_{11}	Γ_{12}
9.06	7.98	8.92	10.50	10.54	10.94
4.14	4.02	4.76	6.28	5.88	6.74
1.02	0.76	1.20	1.76	1.58	1.62

(d) $n = 1000$ and SN_4

Table 8: Simulated rejection rates (10.000 and resp. 5.000 repetitions) in percent for skew-normal hidden Markov models under the true hypothesis of two and three states with different sample sizes n and different parameters. Each table lists line by line the rejection rates on levels $10\%, 5\%$ and 1%.

Further results for skew-normal HMMs

Transition probability matrices with four states

For simulating the finite sample power properties under the false hypothesis of three states, we apply the following transition probability matrices

$$\Gamma_{13} = \begin{pmatrix} 0.91 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.91 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.91 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.91 \end{pmatrix}, \qquad \Gamma_{14} = \begin{pmatrix} 0.1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.1 \end{pmatrix},$$

$$\Gamma_{16} = \begin{pmatrix} 0.95 & 0.05 & 0 & 0 \\ 0.0\overline{6} & 0.92 & 0.01\overline{3} & 0 \\ 0 & 0.02 & 0.905 & 0.075 \\ 0 & 0 & 0.15 & 0.850 \end{pmatrix}, \qquad \Gamma_{17} = \begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.8 & 0.1 & 0.1 & 0 \\ 0 & 0.15 & 0.45 & 0.4 \\ 0 & 0 & 0.8 & 0.2 \end{pmatrix}$$

and Γ_{15} and Γ_{18} being transition probability matrices of the i.i.d. case corresponding to $(\Gamma_{13}, \Gamma_{14})$ and $(\Gamma_{16}, \Gamma_{17})$, i.e. are given by probabilities $1/4\ (1, 1, 1, 1)$ and $(0.4, 0.3, 0.2, 0.1)$.

C.3. Application to oil price logarithmic returns

We further investigate the log-returns of the daily WTI oil spot prices (source US Department of Energy via wikiposit.org; September 2nd 1997 until August 31st 2012). We again use the normal and skew-normal hidden Markov models and reject one and two states with p-values of $\leq 10^{-4}, 0.001$ for the skew-normal HMM and $\leq 10^{-4}, 0.017$ respectively for the normal one. Three states cannot be rejected (even using simulated critical values for the normal HMM; p-value of 0.497 for the normal HMM and of 1 for the skew-normal one). Since BIC indicates three states for both models we decide nevertheless for the three state model (AIC selects five states for the normal and four states for the skew-normal HMM). The parameter estimates are given by $\hat{\boldsymbol{\sigma}}_{SN} = (1.71, 2.39, 5.79)$, $\hat{\mu}_{SN} = 0.08$, $\hat{\xi} = -0.12$ for the skew-normal HMM and by $\hat{\boldsymbol{\sigma}}_{Nor} = (1.70, 2.32, 5.62)$, $\hat{\boldsymbol{\mu}}_{Nor} = (0.09, 0.13, -0.63)$ for the normal one with t.p.m's

$$\hat{\mathbf{\Gamma}}_{SN} = \begin{pmatrix} 98.31 & 0.79 & 0.90 \\ 1.56 & 97.04 & 1.40 \\ 1.26 & 11.47 & 87.27 \end{pmatrix}, \qquad \hat{\mathbf{\Gamma}}_{Nor} = \begin{pmatrix} 98.34 & 0.65 & 1.00 \\ 1.61 & 96.48 & 1.91 \\ 0.31 & 14.24 & 85.46 \end{pmatrix}.$$

The estimated models are again similar, where the dependence structure (of both) is slightly different to the fit of the S&P 500, since there is a higher probability to remain in the most volatility state.

Additional References

Andrews, L. C. (1986). *Special Functions for Engineers and Applied Mathematicians*. Macmillan Publishing Company, New York.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Annals of Statistics*, **23**, 221-233.

Chen, H., Chen J. and Kalbfleisch J.D. (2004). Testing for a Finite Mixture Model with Two Components. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **66**, 95-115.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Annals of Statistics*, **37**, 2523-2542.

Hurst, S. (1995). The characteristic function of student t distribution. Financial Mathematics Research Report No. FMRR 006-95, Statistical Research Report No. SRR044-95.

Klicnarová, J. (2007). Central limit theorem for Hölder processes on $\mathbb{R}^m$-unit cube. *Commentationes Mathematicae Universitatis Carolinae*, **48**, 83-91.

Maitra, R. and Melnykov, V. (2010). Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms. *Journal of Computational and Graphical Statistics*, **19**, 354-376.

Seneta, E. (2006). *Non-negative Matrices and Markov Chains.* Springer Series in Statistics.

van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge University Press.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **20**, 595-601.

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 analyze the structure of the resulting Markov chains in detail, and in particular 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 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 testing, and in the second step we merge components in the resulting state-dependent finite mixtures by using a local entropy 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, based on a local entropy criterion, similar to Baudry et al. (2010) in the context of independent finite mixtures. 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 methodology, Section 3 presents our clustering algorithm. 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. Some technical arguments and some further numerical results are provided in an appendix.

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

Let $(S_t)_t$ be a k-state Markov chain with ergodic transition probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{i,j})_{i,j=1,\dots,k}$ having the stationary distribution $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$. In the following we always assume $\pi_j > 0$ for $j = 1, \dots, k$.
For a (disjoint) partition $\mathcal{G} = \{G_1, \dots, G_r\}$ of the state space (into non-empty sets), we let $G(j)$ be the function which maps a state $j \in \{1, \dots, k\}$ 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

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

Lemma 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*
$$P\big(S_t \in G(j)\big|S_{t-1} = i\big) = P\big(S_t \in G(j)\big|S_{t-1} \in G(i)\big)$$
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 $a(g) = l :\Leftrightarrow g \in G_l$.

The elementary proof is provided in the appendix.

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 1. *There exits 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 2.

2.2. Representations of HMMs with state-dependent mixtures

Let $(X_t, S_t)_t$ be a 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

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

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

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 2. *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 $(X_t)_t \overset{(d)}{=} \big(X_t^{(\mathcal{G})}\big)_t$.

The proof is given in the appendix.

Corollary 1. *Let $(X_t, S_t)_t$ be a 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 $\{1,\dots,k\}$ 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 $\mathcal{G}^*_{\mathbf{\Gamma}} = \{G_1,\dots,G_r\}$ of the HMM its *independence clusters*. The corollary follows from Theorems 1 and 2, since identifiability of k-component mixtures garantuees identifiability of the parameters of the HMM.

2.3. Testing the validity of reducing states to mixture components

Suppose that the state-dependent densities $f_1(\cdot),\dots,f_k(\cdot)$ belong to a known parametric family, i.e. $f_j(x) = f(x;\theta(j))$, $\theta(j) \in M \subset \mathbb{R}^p$, $j = 1,\dots,k$. We denote the complete parameter vector by $\boldsymbol{\eta} = \big((\gamma_{i,j})_{i,j=1,\dots,k}, \theta(1),\dots,\theta(k)\big) \in \Theta \subset \mathbb{R}^d$. 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

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

where $p_{\boldsymbol{\eta}}$ denotes the density function of $(X_1,\dots,X_T)$ 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 fulfils the restriction introduced in section 2.1. Specifically, for a given partition $\mathcal{G} = \{G_1,\dots,G_r\}$ of the state space consider

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

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 3. *Assume the Markov chain $(S_t)_t$ to be ergodic, the MLE $\hat{\boldsymbol{\eta}}_T$ to be strongly consistent, assumptions A2 - A4 of Giudici et al. (2000) concerning the parametric family $f(\cdot, \theta)$ to hold and the Fisher information $\mathcal{J}(\boldsymbol{\eta}_0)$ of the HMM to be nonsingular. If $\boldsymbol{\eta}_0 \in \Theta_{0,\mathcal{G}}$ and $P_{\boldsymbol{\eta}_0}(S_t \in G_l | S_{t-1} \in G_m) > 0$ for $l, m = 1, \ldots, r$ then*

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

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

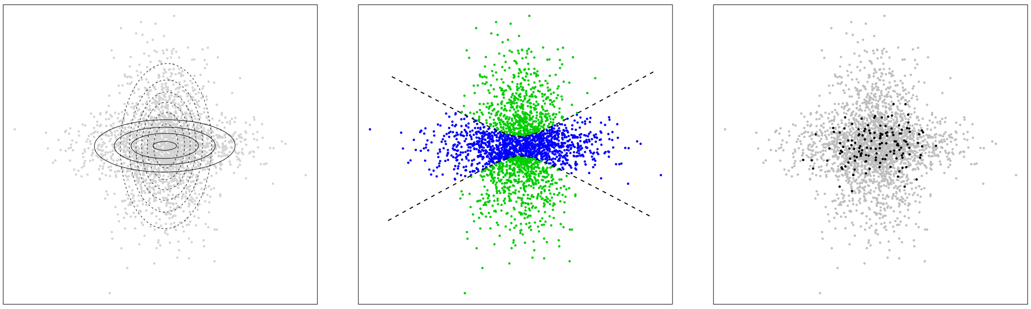
Clusters for independent data usually correspond to peaks of the density, the aim is to determine density-based clusters. In the following example we illustrate the well-known fact that for serially dependent data, groups which marginally strongly overlap, thus forming only a single density-based cluster, can still be very well separated (decoded) when taking advantage of the serial dependence.

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

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

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. Marginally, we only see a single (density-based) cluster. When fitting a two-component normal mixture by ML and determining states by maximum-a-posteriori, the cluster assignment does not reflect the true overlapping group structure and 647 of 2.500 observations are wrongly classified (25.88% of the data). Figure 1b illustrates this result, where data assigned to the first (resp. second) component are colored blue (resp. green), and the dashed line depicts the border of the maximum a posteriori clustering.

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.



(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.

3.2. Merging states in HMMs

When clustering data by using (independent) finite mixtures based on a maximum-a-posteriori analysis, the components need to correspond well to cluster shapes. Otherwise, a single cluster will be modeled by several mixture components and a direct maximum-a-posteriori analysis will lead to too many clusters.

A popular alternative to using more flexible (but more involved) state-dependent distributions is to stick with relatively simple ones (like normals), but to merge the components corresponding 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) compares distinct methods in a simulation study.

For HMMs, the situation apparently is more involved. As seen above, strongly overlapping components, which marginally (density-based) might be merged into a single component, can still be well-separated by strong serial dependence. Therefore, we only allow to merge states within the same elements of the independence partition, for which we shall use the local-decoding entropy of the HMM.

Let us describe our merging and clustering algorithm in detail.

Input The observed series $x_1, \dots, x_T$ 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 $\mathcal{G}^* = \{G_1, \dots, G_r\}$ of the selected HMM with k states using a backward selection based on the p-values of the test in Theorem 3, 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 $\hat{\theta}_1, \dots, \hat{\theta}_k$ 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 $\mathcal{H}_0 = \{\{1\}, \dots, \{k\}\}$, $i = 0$. 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

$$LDE(0) := -\sum_{t=1}^{T} \sum_{j=1}^{k} \phi_{t,j}(\mathcal{H}_0) \cdot \log\big(\phi_{t,j}(\mathcal{H}_0)\big),$$

$$\phi_{t,j}(\mathcal{H}_0) := P\big(S_t^{(\mathcal{H}_0)} = j \big| X_1^{(\mathcal{H}_0)} = x_1, \dots, X_T^{(\mathcal{H}_0)} = x_T\big),$$

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

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 $(X_t^{(\mathcal{H})}, S_t^{(\mathcal{H})})_t$

$$LDE(\mathcal{H}) := -\sum_{t=1}^{T} \sum_{j=1}^{k-(i+1)} \phi_{t,j}(\mathcal{H}) \cdot \log\big(\phi_{t,j}(\mathcal{H})\big),$$

$$\phi_{t,j}(\mathcal{H}) := P\big(S_t^{(\mathcal{H})} = j \big| X_1^{(\mathcal{H})} = x_1, \dots, X_T^{(\mathcal{H})} = x_T\big),$$

for $j = 1, \dots, k - (i + 1),\ t = 1, \dots, T$. 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

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

together with the local decoding entropies

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

and choose $0 \leq i^* \leq k - r$ appropriately, e.g. as an elbow in the entropy plot or if the relative reduction in the entropy exeeds a certain threshold.

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

3.3. Numerical Illustrations

We present two numerical illustrations of the above algorithm.

1. Five-state normal HMM with two independence clusters

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

$$\begin{aligned}
&\boldsymbol{\mu}_1 = \begin{pmatrix}2.5\\1.5\end{pmatrix}^T \quad \boldsymbol{\mu}_2 = \begin{pmatrix}3.5\\2\end{pmatrix}^T \quad \boldsymbol{\mu}_3 = \begin{pmatrix}2\\7\end{pmatrix}^T \quad \boldsymbol{\mu}_4 = \begin{pmatrix}3\\0.5\end{pmatrix}^T \quad \boldsymbol{\mu}_5 = \begin{pmatrix}2.5\\6\end{pmatrix}^T\\
&\boldsymbol{\Sigma}_1 = \begin{pmatrix}0.30 & 0.18\\0.18 & 0.30\end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix}0.30 & -0.18\\-0.18 & 0.30\end{pmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix}0.48 & -0.42\\-0.42 & 0.48\end{pmatrix}\\
&\boldsymbol{\Sigma}_4 = \begin{pmatrix}1.20 & 0.27\\0.27 & 1.20\end{pmatrix} \quad \boldsymbol{\Sigma}_5 = \begin{pmatrix}0.5 & 0.4\\0.4 & 0.5\end{pmatrix}\\
&\boldsymbol{\Gamma} = \begin{pmatrix}25.50 & 17.00 & 42.50 & 10.00 & 5.00\\25.50 & 17.00 & 42.50 & 10.00 & 5.00\\25.50 & 17.00 & 42.50 & 10.00 & 5.00\\6.00 & 4.00 & 10.00 & 70.00 & 10.00\\4.50 & 3.00 & 7.50 & 10.00 & 75.00\end{pmatrix}
\end{aligned} \tag{4}$$

Its independence partition is given by $\mathcal{G}^* = \big\{\{1,2,3\},\{4\},\{5\}\big\}$, but within $\{1,2,3\}$, only states $\{1,2\}$ form a density-based cluster. 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:

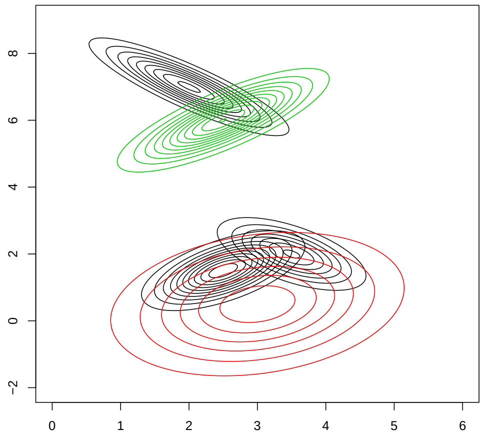


Figure 2: Contour lines of the state dependent bivariate normals. States one to three are depicted by (black) solid lines, state four by (red) dashed lines and state five by (green) dashed lines.

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 appendix. 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%	$\{\{1,3\},\{2\},\{4\},\{5\}\}$
2	72.78%	$\{\{1,2,3\},\{4\},\{5\}\}$
3	$\leq 10^{-4}$	$\{\{1,2,3,4\},\{5\}\}$

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

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

for $\boldsymbol{\Gamma}$. 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.

2. Two-state skew-normal HMM

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

$$\begin{aligned} &\boldsymbol{\Sigma}_1 = \begin{pmatrix} 4.80 & -0.48 \\ -0.48 & 1.20 \end{pmatrix}, \; \boldsymbol{\Sigma}_2 = \begin{pmatrix} 4.0 & -0.4 \\ -0.4 & 1.0 \end{pmatrix}, \qquad \boldsymbol{\Gamma} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \\ &\boldsymbol{\alpha}_1 = \begin{pmatrix} 14 & -6 \end{pmatrix}, \; \boldsymbol{\alpha}_2 = \begin{pmatrix} 14 & 0 \end{pmatrix}, \\ &\boldsymbol{\mu}_1 = \begin{pmatrix} -5.0 & 3.3 \end{pmatrix}, \; \boldsymbol{\mu}_2 = \begin{pmatrix} -1.5 & 6.0 \end{pmatrix}. \end{aligned} \tag{5}$$

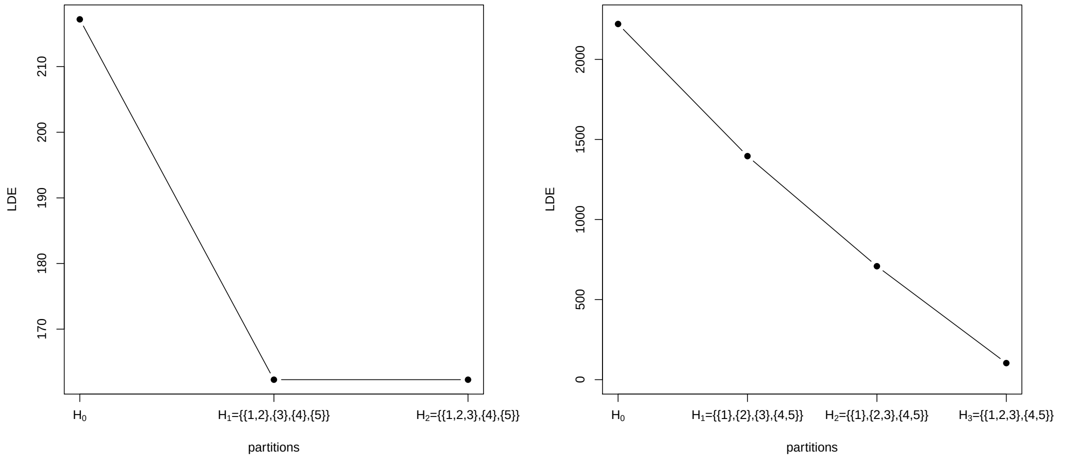
Specifically, the two-dimensional skew-normal density is given by

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

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

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

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



(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.

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

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

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:

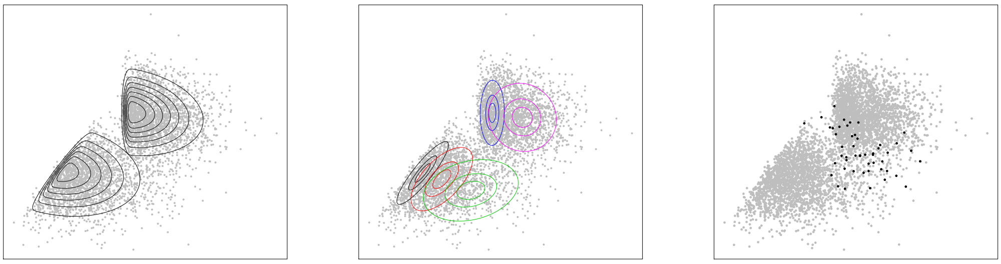
Step i	Max. p-value of $\lambda_{\mathcal{G}_i}(\mathbf{\Gamma}) = (\mathbf{\Gamma})$	Partition $\mathcal{G}_i$ with max. p-value
1	62.99%	$\big\{\{1,3\},\{2\},\{4\},\{5\}\big\}$
2	22.45%	$\big\{\{1,3\},\{2\},\{4,5\}\big\}$
3	11.94%	$\big\{\{1,2,3\},\{4,5\}\big\}$
4	$\leq 10^{-4}$	$\big\{\{1,2,3,4,5\}\big\}$

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

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

and fitted values for the state-dependent parameters are listed in the appendix, see Figures 4a and 4b for contour plots of the true densities and the fitted normal state-dependent densities.

When applying our merging algorithm, 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. The misclassifications from a clustering using global decoding, together with the true group assigment, is plotted in Figure 4c.



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

Figure 4: 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.

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 (4). The regularity conditions of Theorem 3 are satisfied if we impose lower bounds on the determinants of the state-dependent covariance matrices.

For the three partitions $\mathcal{G}_1 = \big\{\{1,2\},\{3\},\{4\},\{5\}\big\}$, $\mathcal{G}_2 = \big\{\{1\},\{2,3\},\{4\},\{5\}\big\}$ and $\mathcal{G}_3 = \big\{\{1,2,3\},\{4\},\{5\}\big\}$ 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%	16.30	12.46	11.32
5%	9.30	6.90	5.80
1%	2.38	1.44	1.16

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

Table 1: Simulated rejection rates in percent (series lenghts 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) χ^2_7 - approximation, (c) 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 3 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 $H_0 : \lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ 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 $\tilde{\mathcal{G}} = \{\{1, \dots, k\}\}$ is the trivial partition (in step $i = k-1$), we let $\mathcal{G}^* = \{\{1, \dots, k\}\}$ 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 (4), where we always start with a HMM with five states. The results are given in Table 2. 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 at the given t.p.m. is quite high.

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 selected sets in independence cluster according to used level. $M = 1.000$ repetitions for lenghts 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 (5), but fit normal HMMs. We generate $M = 1.000$ repetitions for lenghts $T = 1.000$ 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. clus.	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 form 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 the estimated t.p.m. is given by

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

the other parameter estimates are given in the appendix. 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%	$\big\{\{1\},\{2\},\{3\},\{4\},\{5,6\}\big\}$
2	45.80%	$\big\{\{1,2\},\{3\},\{4\},\{5,6\}\big\}$
3	0.47%	$\big\{\{1,2,3\},\{4\},\{5,6\}\big\}$

giving $\mathcal{G}^* = \big\{\{1,2\},\{3\},\{4\},\{5,6\}\big\}$ as independence partition.

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

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

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. $\lambda_{\mathcal{G}^*}(\hat{\boldsymbol{\Gamma}})$, 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 5 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 second to the fourth state. The fourth state, which arises as a scale mixture of two normals, has mean almost $= 0$ but a heavier tail than an ordinary normal distribution. Finally, the third state has distinctly the highest volatility. Both states 3 and 4 are highly persistent.

Using the merged model and the Viterbi algorithm we estimated the most likely series of states, see Figure 6. The third state e.g. occurred from October 26th 2007 until April 6th 2009, a time period containing the financial crisis.

Acknowledgements

The authors gratefully acknowledge financial support from the DFG, grant Ho 3260/3-1.

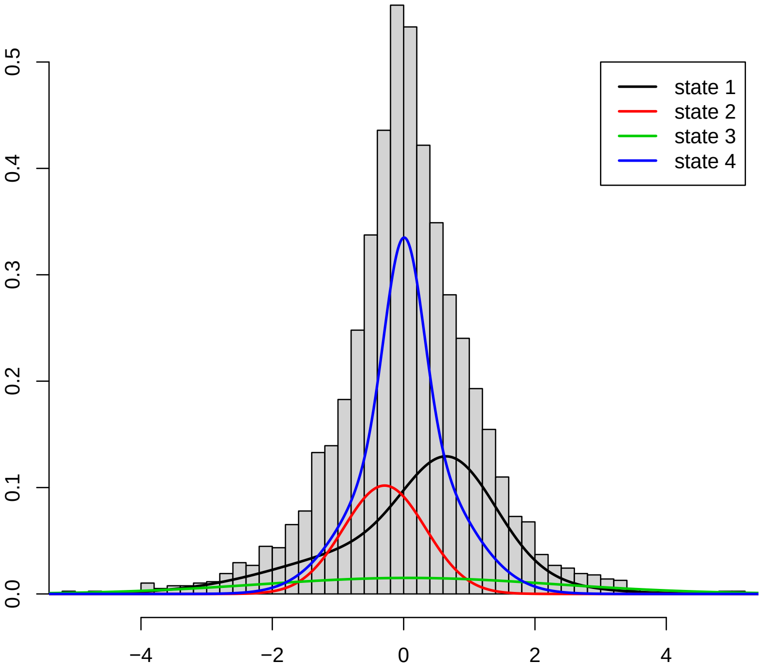


Figure 5: Histogram of logarithmic returns of gold prices in percent (September 2nd 1997 until August 31st 2012) and estimated state dependent densities of the merged four-state hidden Markov model (densities of states one to four are colored in black, red, green and blue).

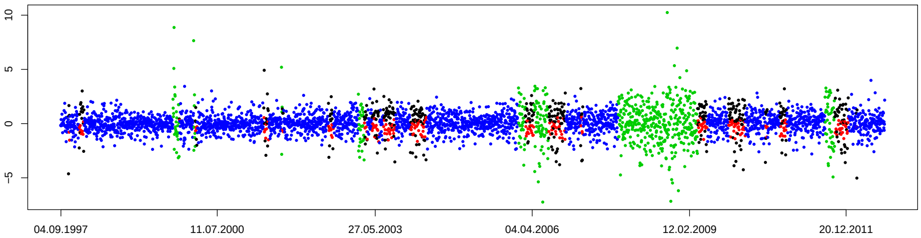


Figure 6: Time-series of logarithmic returns of gold prices in percent (September 2nd 1997 until August 31st 2012) and estimated series of states using the merged model and the Viterbi algorithm (states one to four are given by colors black, red, green and blue).

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.

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

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.

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

Appendix

A. Proofs

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

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

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

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

implies $\gamma_{i,j} = \gamma_{h,j}$ for all $h \in G(i)$, i.e. under 1. $\boldsymbol{\Gamma}$ has equal rows with indices in the same group of 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)

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

and hence also the second claim of 2. follows.

Now, assume 2. to hold, then for $l, m = 1,\dots,r$ 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 $\boldsymbol{p} = (p_1,\dots,p_k)$ has the desired property of 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. Hence, for $i, j = 1, \dots, k$

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

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 2. *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 2. 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,

$$\gamma_{i,j} = P\big(S_t \in I(j) \big| S_{t-1} \in I(i)\big) \cdot P\big(S_t = j \big| S_t \in I(j)\big) \qquad 1 \leq i, j \leq k. \tag{6}$$

Note that (6) 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 $i, j \in I_1$, $i \in G_l$, $i' \in G_{l'}$, 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,

$$\gamma_{i,g} = \gamma_{h,g}, \quad i,h \in I \in \mathcal{I}, \quad 1 \le g \le k. \tag{7}$$

Now, due to (7),

$$\begin{aligned}
\big(\lambda_{\mathcal{I}}(\mathbf{\Gamma})\big)_{i,j} &= P\big(S_t = j \big| S_t \in I(j)\big) \cdot P\big(S_t \in I(j) \big| S_{t-1} \in I(i)\big) \\
&= \frac{\pi_j}{\sum_{g \in I(j)} \pi_g} \cdot \sum_{g \in I(j)} \sum_{h \in I(i)} \left(\frac{\pi_h}{\sum_{l \in I(i)} \pi_l} \cdot \gamma_{h,g} \right) \\
&= \frac{\pi_j}{\sum_{g \in I(j)} \pi_g} \cdot \sum_{g \in I(j)} \gamma_{i,g} \sum_{h \in I(i)} \frac{\pi_h}{\sum_{l \in I(i)} \pi_l} \\
&= \frac{\pi_j}{\sum_{g \in I(j)} \pi_g} \cdot \sum_{g \in I(j)} \gamma_{i,g},
\end{aligned} \tag{8}$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$ denotes the stationary distribution of $(S_t)_t$. Therefore, in order to show (6), it sufficies to show that

$$\frac{\gamma_{i,j}}{\pi_j} = \frac{\sum_{g \in I(j)} \gamma_{i,g}}{\sum_{g \in I(j)} \pi_g}, \qquad 1 \le i, j \le k. \tag{9}$$

which is equivalent to

$$\frac{\gamma_{i,j}}{\pi_j} = \frac{\gamma_{i,a}}{\pi_a}, \qquad 1 \le i, j \le k, \; a \in I(j). \tag{10}$$

Indeed, (9) evidently implies (10), while using (10) 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, (9).

Now, in order to show (10), we note that the corresponding property holds for the partitions $\mathcal{G}$ and $\mathcal{H}$, so that (10) evidently holds if $I(j) \ne I_1$. To cover this case, suppose that $j, a \in I_1$, so that $j \in G_l$ and $a \in G_{l'}$ for some $1 \le l, l' \le r$. 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 $1 \le i \le k$:

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

□

Proof of Theorem 1. Let

$$\mathcal{G}_{\mathbf{\Gamma}}^* \in \operatorname{argmin}\big\{\operatorname{card} \mathcal{G} : \; \mathcal{G} \text{ is partition of } \{1, \dots, k\} \text{ with } \lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}\big\}. \tag{11}$$

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 2, 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 (11). □

Proof of Theorem 2. 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 $f_j(x)$, $j = 1,\ldots,k$.

Proving $(Y_t^{(\mathcal{G})})_t \overset{(d)}{=} (X_t^{(\mathcal{G})})_t$ yields the claim, since then under $\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ directly $(X_t)_t \overset{(d)}{=} (Y_t^{(\mathcal{G})})_t$ (t.p.m.'s and state dependent densities coincide) and thus $(X_t)_t \overset{(d)}{=} \big(X_t^{(\mathcal{G})}\big)_t$.

The remainder is to prove $(Y_t^{(\mathcal{G})})_t \overset{(d)}{=} (X_t^{(\mathcal{G})})_t$: At first, reducing dependence information does not change the stationary distribution of the Markov chain, i.e. for $j \in \{1,\ldots k\}$, we have

$$P\big(T_t^{(\mathcal{G})} = j\big) = P\big(S_t = j\big).$$

To show this, let $\boldsymbol{\pi}$ denote the stationary distribution of $(S_t)_t$, we have

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

and thus

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

Further, as mentioned before, for the MC $(S_t^{(\mathcal{G})})_t$ it holds

$$P\big(S_t^{(\mathcal{G})} = l\big) = P\big(S_t \in G_l\big), \qquad l = 1,\ldots,r$$

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

Further, since $f_j(x) = f_{X_t|S_t=j}(x)$ denotes the state dependent density of the original HMM, the state dependent densities of the transformed HMMs are given for $x \in \mathbb{R}^d$ by

$$f_l^{(\mathcal{G})}(x) = f_{X_t^{(\mathcal{G})}|S_t^{(\mathcal{G})}=l}(x) = \sum_{g \in G_l} P\big(S_t = g|S_t \in G_l\big) \cdot f_g(x), \qquad l = 1, \dots, r$$

and

$$f_j(x) = f_{Y_t^{(\mathcal{G})}|T_t^{(\mathcal{G})}=j}(x), \qquad j = 1, \dots, k.$$

Assuming the Markov chain to start in its stationary distribution given by the t.p.m., the density of the observable process $(Y_t^{(\mathcal{G})})_{t=1,\dots,T}$ of the reduced model is given by

$$\begin{aligned}
&f_{(Y_1^{(\mathcal{G})},\dots,Y_T^{(\mathcal{G})})}(x_1, \dots, x_T) \\
&= \sum_{j_1,\dots,j_T=1}^{k} \Big(P\big(T_1^{(\mathcal{G})} = j_1\big) f_{j_1}(x_1) \prod_{t=2}^{T} P\big(T_t^{(\mathcal{G})} = j_t|T_{t-1}^{(\mathcal{G})} = j_{t-1}\big) f_{j_t}(x_t)\Big) \\
&= \sum_{l_1,\dots,l_T=1}^{r} \sum_{g_1 \in G_{l_1}} \cdots \sum_{g_T \in G_{l_T}} \Big(P\big(S_1 = g_1\big) f_{g_1}(x_1) \prod_{t=2}^{T} \big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{g_{t-1},g_t} f_{g_t}(x_t)\Big)
\end{aligned} \tag{12}$$

Since here $g_{t-1} \in G_{l_{t-1}}$ and $g_t \in G_{l_t}$,

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

Therefore

$$\begin{aligned}
&f_{(Y_1^{(\mathcal{G})},\dots,Y_T^{(\mathcal{G})})}(x_1, \dots, x_T) \\
&= \sum_{l_1,\dots,l_T=1}^{r} \sum_{g_1 \in G_{l_1}} \cdots \sum_{g_T \in G_{l_T}} \Big(P\big(S_1 = g_1\big) f_{g_1}(x_1) \prod_{t=2}^{T} \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t} P\big(S_t = g_t|S_t \in G_{l_t}\big) f_{g_t}(x_t)\Big)
\end{aligned}$$

and hence in the latter sum only $P(S_T = g_T|S_T \in G_{l_T}) \cdot f_{g_T}(x)$ depends on g_T, i.e. everything else can be factorized. Iterating this procedure over g_t gives

$$\begin{aligned}
&f_{(Y_1^{(\mathcal{G})},\dots,Y_T^{(\mathcal{G})})}(x_1, \dots, x_T) \\
&= \sum_{l_1,\dots,l_T=1}^{r} \Big(\sum_{g_1 \in G_{l_1}} P(S_1 = g_1) f_{g_1}(x_1) \prod_{t=2}^{T} \Big\{ \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t} \sum_{g_t \in G_{l_t}} P(S_t = g_t|S_t \in G_{l_t}) f_{g_t}(x_t) \Big\} \Big) \\
&= \sum_{l_1,\dots,l_T=1}^{r} \Big(P\big(S_1^{(\mathcal{G})} = l_1\big) f_{l_1}^{(\mathcal{G})}(x_1) \prod_{t=2}^{T} \Big\{ \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t} f_{l_t}^{(\mathcal{G})}(x_t) \Big\} \Big) \\
&= f_{(X_1^{(\mathcal{G})},\dots,X_T^{(\mathcal{G})})}(x_1, \dots, x_T).
\end{aligned}$$

□

Proof of Theorem 3. 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 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

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

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

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 $G_1 = \{1, \dots, n_1\}, G_2 = \{n_1 + 1, \dots, n_2\}, \dots, G_r = \{n_{r-1} + 1, \dots, n_r\}$. Thus, the Markov chain can be parametrized via

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

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

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

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

Therefore, H_0 is equivalent to

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

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

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, $\mathbf{\Gamma}_{\boldsymbol{\eta}}$ 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. □

B. Additional estimation results

B.1. Results concerning the simulations

All transition probability matrices are given in percent.

1. Five-state normal HMM with two independence clusters

The estimated BIC-optimal unrestricted five-state HMM of the first simulated dataset is given by

$$\tilde{\boldsymbol{\mu}}_1 = \begin{pmatrix} 2.51 \\ 1.51 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_2 = \begin{pmatrix} 3.57 \\ 2.01 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_3 = \begin{pmatrix} 2.03 \\ 6.98 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_4 = \begin{pmatrix} 2.94 \\ 0.40 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_5 = \begin{pmatrix} 2.48 \\ 6.02 \end{pmatrix}^T$$

$$\tilde{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.20 & 0.14 \\ 0.14 & 0.32 \end{pmatrix} \tilde{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.29 & -0.23 \\ -0.23 & 0.32 \end{pmatrix} \tilde{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 0.43 & -0.37 \\ -0.37 & 0.42 \end{pmatrix}$$

$$\tilde{\boldsymbol{\Sigma}}_4 = \begin{pmatrix} 1.09 & 0.13 \\ 0.13 & 1.04 \end{pmatrix} \tilde{\boldsymbol{\Sigma}}_5 = \begin{pmatrix} 0.56 & 0.42 \\ 0.42 & 0.49 \end{pmatrix}$$

$$\tilde{\boldsymbol{\Gamma}} = \begin{pmatrix} 25.01 & 20.21 & 40.10 & 12.42 & 2.26 \\ 21.95 & 19.54 & 46.37 & 9.72 & 2.43 \\ 29.31 & 13.69 & 40.86 & 13.98 & 2.16 \\ 6.04 & 3.14 & 10.62 & 70.27 & 9.94 \\ 9.02 & 0.00 & 7.11 & 8.93 & 74.93 \end{pmatrix}.$$

Under the independence restriction $\mathcal{G}^* = \big\{\{1,2,3\},\{4\},\{5\}\big\}$, which has been found by backward selection, the estimated HMM is given by

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 2.48 \\ 1.47 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 3.53 \\ 2.03 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_3 = \begin{pmatrix} 2.03 \\ 6.98 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_4 = \begin{pmatrix} 2.94 \\ 0.39 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_5 = \begin{pmatrix} 2.48 \\ 6.02 \end{pmatrix}^T$$

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.19 & 0.13 \\ 0.13 & 0.30 \end{pmatrix} \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.30 & -0.22 \\ -0.22 & 0.31 \end{pmatrix} \hat{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 0.43 & -0.37 \\ -0.37 & 0.42 \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_4 = \begin{pmatrix} 1.09 & 0.12 \\ 0.12 & 1.03 \end{pmatrix} \hat{\boldsymbol{\Sigma}}_5 = \begin{pmatrix} 0.56 & 0.43 \\ 0.43 & 0.49 \end{pmatrix}$$

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

2. Two-state skew-normal HMM

The estimated BIC-optimal unrestricted five-state HMM of the second simulated dataset is given by

$$
\begin{aligned}
&\tilde{\boldsymbol{\mu}}_1 = \begin{pmatrix} -4.52 \\ 3.16 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_2 = \begin{pmatrix} -3.52 \\ 2.87 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_3 = \begin{pmatrix} -2.00 \\ 2.35 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_4 = \begin{pmatrix} -0.88 \\ 5.94 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_5 = \begin{pmatrix} 0.69 \\ 5.73 \end{pmatrix}^T \\
&\tilde{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.76 & 0.73 \\ 0.73 & 0.88 \end{pmatrix} \tilde{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1.08 & 0.62 \\ 0.62 & 0.89 \end{pmatrix} \tilde{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 2.57 & 0.35 \\ 0.35 & 0.84 \end{pmatrix} \\
&\tilde{\boldsymbol{\Sigma}}_4 = \begin{pmatrix} 0.16 & 0.01 \\ 0.01 & 0.94 \end{pmatrix} \tilde{\boldsymbol{\Sigma}}_5 = \begin{pmatrix} 1.33 & -0.08 \\ -0.08 & 1.03 \end{pmatrix} \\
&\hat{\boldsymbol{\Gamma}}_{Nor} = \begin{pmatrix} 25.92 & 39.45 & 23.92 & 7.37 & 3.35 \\ 23.35 & 44.16 & 21.30 & 2.43 & 8.75 \\ 20.72 & 44.38 & 25.37 & 4.12 & 5.41 \\ 2.89 & 4.32 & 2.50 & 33.76 & 56.53 \\ 3.75 & 2.85 & 3.45 & 36.20 & 53.75 \end{pmatrix}.
\end{aligned}
$$

Under the independence restriction $\mathcal{G}^* = \big\{\{1,2,3\},\{4,5\}\big\}$, which has been found by backward selection, the estimated HMM is given by

$$
\begin{aligned}
&\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} -4.58 \\ 3.13 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} -3.54 \\ 2.89 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_3 = \begin{pmatrix} -2.02 \\ 2.38 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_4 = \begin{pmatrix} -0.88 \\ 5.93 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_5 = \begin{pmatrix} 0.70 \\ 5.73 \end{pmatrix}^T \\
&\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.74 & 0.72 \\ 0.72 & 0.88 \end{pmatrix} \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1.03 & 0.61 \\ 0.61 & 0.91 \end{pmatrix} \hat{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 2.53 & 0.34 \\ 0.34 & 0.85 \end{pmatrix} \\
&\hat{\boldsymbol{\Sigma}}_4 = \begin{pmatrix} 0.17 & 0.01 \\ 0.01 & 0.95 \end{pmatrix} \hat{\boldsymbol{\Sigma}}_5 = \begin{pmatrix} 1.33 & -0.08 \\ -0.08 & 1.02 \end{pmatrix} \\
&\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 89.37 \cdot \begin{pmatrix} 0.26 & 0.47 & 0.27 \\ 0.26 & 0.47 & 0.27 \\ 0.26 & 0.47 & 0.27 \end{pmatrix} & 10.63 \cdot \begin{pmatrix} 0.39 & 0.61 \\ 0.39 & 0.61 \\ 0.39 & 0.61 \end{pmatrix} \\ 09.92 \cdot \begin{pmatrix} 0.26 & 0.47 & 0.27 \\ 0.26 & 0.47 & 0.27 \end{pmatrix} & 90.08 \cdot \begin{pmatrix} 0.39 & 0.61 \\ 0.39 & 0.61 \end{pmatrix} \end{pmatrix}.
\end{aligned}
$$

B.2. Results concerning the application

The estimated AIC-optimal unrestricted six-state HMM is given by

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

When estimating under the independence restrictions implied by $\mathcal{G}^* = \big\{\{1,2\},\{3\},\{4\},\{5,6\}\big\}$, we obtain

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} -0.259 & 0.717 & -0.290 & 0.072 & 0.026 & 0.006 \end{pmatrix}$$

$$\hat{\boldsymbol{\sigma}} = \begin{pmatrix} 1.531 & 0.695 & 0.625 & 2.244 & 0.805 & 0.296 \end{pmatrix}$$

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

Peaks vs Components

Sebastian Vollmer[1]
Hajo Holzmann[2]
Florian Schwaiger[3]

We analyze the cross-national distribution of gross domestic product (GDP) per capita and its evolution from 1970 to 2009. We argue that peaks are not a suitable measure for distinct convergence clubs/equilibria in the cross-country distribution of GDP per capita, because the number of peaks is not invariant under non-linear strictly monotonic transformations of the data such as the logarithmic transformation. Instead, we model the distribution as a finite mixture, and determine its number of components via statistical testing. We find that the number of components in the cross-country distribution changes from three to two in the mid 1990s.

JEL classification: C12, O11, O47, F01
Keywords: twin peaks, economic growth, convergence.
Published in: Review of Development Economics, **17** 352-364.

[1] University of Göttingen & Harvard University
[2] University of Marburg
[3] University of Marburg

1 Introduction

The notion of twin peaks in the cross-country distribution of gross domestic product (GDP) per capita was introduced by Quah (1993, 1996, 1997). He interpreted the emergence of twin peaks as polarization of the distribution of per capita income into a rich and a poor convergence club. Bianchi (1997) confirmed Quah's observation of twin peaks via rigorous statistical testing. The contributions of Quah are part of a larger literature on convergence (e.g. Barro, 1991; Barro and Sala-i-Martin, 1992; Mankiw et al., 1992; Sala-i-Martin, 1996; Galor, 1996; Jones, 1997; Graham and Temple, 2006). It is controversial in this literature whether the twin peaks represent locally stable equilibria/convergence clubs (Quah, 1996) or whether they are only a temporary phenomenon caused by a high frequency of growth miracles (Jones, 1997).

The unified growth theory (c.f. Galor, 2010 for an overview) provides another explanation for multiple regimes in the cross-country distribution of GDP per capita which also uncovers the forces that have lead to the emergence of these regimes. The theory suggests that growth segments economies into three fundamental regimes: a Malthusian regime with slow growing economies, fast growing economies in a sustained growth regime, and a third group in the transition from one regime to the other. One important difference to models with multiple equilibria is that this segmentation does not represent the long-run steady state of these economies. Variations in the levels of income only reflect country-specific characteristics and not the actual stage of development. Thus, there are no critical levels that permit economies to switch from one regime to the other, but rather critical rates of progress.

Recent theoretical work by Schumacher (2009) and Strulik (2012) provides alternative explanations for the emergence of multiple equilibria in the cross-country distribution of GDP per capita. Schumacher (2009) endogenizes discounting via wealth in a neoclassical growth model and shows that this can generate multiple equlibria. Strulik (2012) formulates an endogenous growth theory with endogenous patience, which can explain the take-off from stagnation to modern growth. He concludes that either all countries adjust to the same balanced growth path or that lagging countries will never catch up.

In this paper we challenge Quah's twin peaks result. We show that the number of peaks of a distribution is not preserved under strictly monotonic transformations of the data: a simple log transformation may change the number of peaks in the crosscountry distribution of GDP per capita. This fact casts doubts on the economic interpretation of twin peaks: it does not make much sense to call a country high income on the original scale of the GDP per capita data and middle income on a log scale of the same data. A suitable measure of convergence clubs or growth regimes should not be affected by a simple log transformation of the data.

We therefore propose another method to identify different regimes within a distribution which does not have this problem. We will use mixture models to estimate the cross-country distribution of GDP per capita and to statistically assign countries to different convergence clubs. Mixture models are not new to the economic literature and have been used in quite a few articles to model income distributions. Most prominently, Paap and Dijk (1998) used a two-component mixture, consisting of a truncated normal distribution and a Weibull distribution, to model the cross-country distribution of GDP per capita. Nevertheless, we are not aware of any article that challenges the twin peaks approach and suggests mixture models as an alternative.

2 Data

We use the Penn World Tables 7.0 (PWT) data for the period from 1970 to 2009 (Heston et al. 2011). The PWT is a panel dataset containing 190 countries and 38 variables. We use the variable rgdpch, which is PPP converted GDP per capita (chain series) at 2005 constant prices. We consider the mentioned GDP per capita variable on its original scale (US$1000) and on a logarithmic scale with base 10.

We exclude a few small countries whose economies heavily depend on oil export from the analysis: Bahrain, Brunei, Equatorial Guinea, Gabon, Kuwait, Qatar, Suriname, Timor-Leste and Trinidad and Tobago. The reason for this choice is, that these countries show large fluctuations in GDP per capita, which are mostly driven by fluctuations of the oil price. Arguably, these countries are not essential for understanding multiple equilibria in the world's cross-country distribution of GDP per capita. The PWT dataset contains two versions of China, we thus exclude the second version (CH2) from the analysis. We believe that using a balanced panel is most appropriate for analyzing the cross-country distribution of GDP per capita over time, because a balanced panel is not affected by changes in the sample composition. This leaves 151 countries in the dataset, for which we have GDP per capita data for all years from 1970 to 2009.

3 Peaks

Figure 1 shows simple kernel density estimates of the cross-country distribution of GDP per capita in 1985 on the original scale (US$1000) and on a logarithmic scale with base 10. The density of the data on the original scale has two peaks and the density of the data on a log scale has three peaks. This simple picture illustrates that the number of peaks is not preserved under a simple log transformation: Quah's twin peaks become triple peaks on the log scale.

However, the different numbers of peaks in the plots could be a simple artifact of the nonparametric curve estimates, e.g. from inaccurate choice of the tuning parameter. It is therefore necessary to validate the statistical significance of the peaks via rigorous statistical testing. To this end we utilize Silverman's test. Formally, a *peak* of a density f (and similarly of the kernel estimator $\hat{f}$) is a local maximum of f (or $\hat{f}$). Silverman (1981) showed that the number of modes of $\hat{f}$ is a right-continuous, monotonically decreasing function of the bandwidth h if the normal kernel $K(x) = (2\pi)^{-1}\exp(-x^2/2)$ is employed. This allowed him to define the k-critical bandwidth $h_c(k)$ as the minimal bandwidth h for which $\hat{f}$ still just has k peaks and not yet $k+1$ peaks. Based on the notion of the k-critical bandwidth, Silverman (1981) proposed a bootstrap test for the hypotheses

$$\tilde{H}_k : f \text{ has at most } k \text{ modes} \qquad \text{against} \qquad \tilde{K}_k : f \text{ has more than } k \text{ peaks.}$$

This test is known to be slightly conservative (even asymptotically), for $\tilde{H}_1$ we therefore use the adjustment proposed by Hall and York (2001). The tests were performed using our R-package silvermantest (available online at http://www.uni-marburg.de/fb12/stoch/research/rpackage). We apply Silverman's test to the distributions of GDP per capita and log-GDP per capita for all years from 1970 to 2009. We report the p-values in Tables 1 and 2.

For the distribution of GDP per capita we can reject the null hypothesis of a single peak from 1970 to 1990, but we cannot reject the null hypothesis of two peaks in favor of three or more peaks. This is basically the period that Quah studied in his influential papers, and our results confirm his findings. From 1991 onwards we can also reject the null hypothesis of two peaks in favor of three peaks (but not more). Thus, we find evidence for two peaks from 1970 to 1990 and for three peaks thereafter. For the distribution of log-GDP per capita we can reject the null hypothesis of two peaks in favor of three peaks (but not more) from 1970 to 1990, but we fail to reject the null hypothesis of a single peak. Note that this result does not mean that the null hypothesis of a single peak is correct, it just means that there is not enough evidence to reject it at a level of 5%. Thus, there is evidence of three peaks, but none of only two peaks from 1970 to 1990. From 1991 onwards we cannot reject any of the null hypotheses (with the exception of a few transition years from 1992 to 1994 where the distribution appears to have four peaks) and thus find evidence for only a single peak.

What do we learn from this analysis? The number of peaks is relevant information for the proper visualization of data. However, our results show clearly that peaks should neither be used for economic interpretation of the cross-country distribution of GDP per capita nor for assigning countries to convergence clubs, growth regimes and the like. It does not make sense to conclude that the distribution of the GDP per capita consists of two convergence clubs between 1970 and 1990, while the distribution of log GDP per capita consists of three convergence clubs over the same period.

4 Components

Methods

We now turn to mixture models to estimate the cross-country distribution of GDP per capita. Let f_X denote the density of the cross-country distribution of GDP per capita X for a given year. We model

$$f_X(x) = \alpha_1 g(x;\phi_1) + \ldots + \alpha_m g(x,\phi_m), \qquad x > 0,$$

where $g(x;\phi)$ is a parametric family of densities and the weights $\alpha_i \geq 0$ sum up to one. There is no general simple connection between the number of modes of f and the number of components m. Typically, for single-peaked g, the number of peaks of f will be at most m, but often will be less than m. The number of components is preserved if the data are transformed via a strictly monotonic transformation (if densities are correspondingly transformed). We let $Y = \log X$ and model the density of log-income f_Y by

$$f_Y(y) = \alpha_1\varphi(y;\mu_1,\sigma_1) + \ldots + \alpha_m\varphi(y;\mu_m,\sigma_m),$$

where $\varphi(\cdot,\mu,\sigma)$ is the density of the normal distribution with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$. Then $g(\cdot;\phi)$ in the representation of f_X is the log-normal distribution. The number of components is determined via statistical inference: We aim to test successively for ascending m in $\mathbb{N}$ the hypotheses

$$H_{m_0} : m_0 = m \quad \text{against} \quad K_{m_0} : m_0 \geq m + 1,$$

where $m_0 \in \mathbb{N}$ is the true, unknown number of components. Testing in parametric models is often accomplished by using the likelihood ratio test (LRT). However, the standard theory of the LRT does not apply for the number of components in finite mixture models (Dacunha-Castelle and Gassiat, 1999). Recently, Chen et al. (2001, 2004) and Chen and Kalbfleisch (2005) suggested modified LRTs, which retain comparatively simple limit theory as well as the good power properties of the LRT. Unfortunately, these tests are only valid if the switching parameter is one-dimensional and hence we cannot apply them for selecting the number of components.

In our setting with switching μ and σ only an asymptotic test for homogeneity, i.e. for H_1, is available, see Chen and Li (2009). Therefore, in order to test all hypotheses under investigation with the same methodology, we apply the commonly used parametric bootstrap.[4] As is well known, since μ and σ both switch

[4] We used 1000 bootstrap replications.

the likelihood function is unbounded if small values of the standard deviation are allowed. Therefore, we use a penalized log-likelihood as proposed in Chen and Li (2009) as follows:

$$l_n(X_1, \ldots, X_n; \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \log \Big(\sum_{j=1}^{m} \alpha_j \varphi(X_i; \mu_j, \sigma_j) \Big) + p_n(X_1, \ldots, X_n, \boldsymbol{\sigma}),$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{m-1})$ with $\alpha_m := 1 - \sum_{j=1}^{m-1} \alpha_j$ for $m > 1$ and $\boldsymbol{\alpha} = 1$ for $m = 1$, and

$$p_n(X_1, \ldots, X_n, \boldsymbol{\sigma}) = -\frac{1}{20} \sum_{j=1}^{m} \left(\frac{s_n^2}{\sigma_j^2} + \log \left(\frac{\sigma_j^2}{s_n^2} \right) \right),$$

where $s_n^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ with $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. The function p_n penalizes small values of a σ_j, and guarantees a bounded (penalized) likelihood.

After fitting the model and selecting the number of components, we can use the mixture model for cluster analysis, see e.g. Fraley and Raftery (2002). Each observation can be assigned *a-posteriori* probabilities to belong to each of the components in the mixture model. Specifically, the *a-posteriori* probability of an observation y to belong to group j is equal to

$$p_j(y) = \frac{\hat{\alpha}_j \varphi(y; \hat{\mu}_j, \hat{\sigma}_j)}{\hat{\alpha}_1 \varphi(y; \hat{\mu}_1, \hat{\sigma}_1) + \ldots + \hat{\alpha}_m \varphi(y; \hat{\mu}_m, \hat{\sigma}_m)},$$

for $m = 2$ or $m = 3$ in case of a two- or a three-component mixture. Therefore, we do not merely assign an income level to each country, but rather a probability distribution, which makes transitions from one group to the other much more transparent. One may then assign an observation y to one of the components by using the maximum *a-posteriori* estimate (MPE), which assigns the j to the country i for which $p_j(y)$ is maximal. One can also determine the threshold $t_{j,j+1}$, $j = 1, \ldots, m-1$, for the values of the log-GDP per capita at which the MPE changes between group j and $j+1$, by solving the equations

$$p_j(t_{j,j+1}) = p_{j+1}(t_{j,j+1}),$$

restricted to the interval $[\hat{\mu}_j, \hat{\mu}_{j+1}]$. This yields three levels of income which we label poor, intermediate, and rich, with indices 1, 2, 3 when a three-component mixture is fitted or two levels of income which we label poor and rich when a two-component mixture is fitted.

Results

Table 3 displays the results of the parametric bootstrap test based on 1000 bootstrap samples. We can always reject the hypothesis of homogeneity, i.e. of a single normal distribution. Further, we cannot reject the null hypothesis of two components in 1970, 1971 and 1972 at the 5 percent significance level, however, the p-values are already quite. From 1973 to 1995 we can reject the null hypothesis of two components with p-values at the 5% level. From 1996 to 2001 the p-values are still quite low, but we cannot reject the null hypothesis at the 5 % level anymore. After 2002 the p-values are rather large and the null hypothesis cannot be rejected. Overall, we observe a three component mixture that evolves into a two component mixture. We thus model the cross country distribution of GDP per capita with three components from 1970 to 1995 and with two components from 1996 to 2009.

In Figure 2 we show the fitted three-component mixtures for 1975 and 1985 and compare it to the corresponding kernel density estimators based on the smallest bandwidths which produce three peaks. Further, Figure 3 shows the fitted two-component mixtures for 1996 and 2005 with the corresponding kernel density estimators based on the smallest bandwidths which produce two peaks. We also provide quantile-quantile (qq) plots of the data against the fitted mixture models, see figure 4 and 5. The qq-plots show that the three respectively the two component mixtures describe the data well.

Figure 6 shows the development of the different component means over time as well as the thresholds where the maximum *a-posteriori* estimate changes from one component to the other. The component means are also shown in Table 4. The mean of the low-income and the middle-income component hardly changes between 1970 and 1990, but both component means show substantial increases from 1991 to 1995. The mean of the high-income component steadily grows from 1970 to 1995 (by roughly 50 percent over the entire period).

After 1995 the three components merge into two components. The new higher-income component basically continues on the growth path of the high-income component from the previous model, whereas, the low-income and middle-income components from the previous model merge into a new lower-income component. Both component means steadily grow between 1996 and 2009 (both roughly by one third over the entire period).

The observation that the low-income and middle-income components of the three-component model merge into a new lower-income component in the two-component model is also supported by the component weights which are displayed in Table 5. In 1970 the low-income component constitutes about 50 percent of the countries, whereas the middle-income and high-income components represent 33 and 17 percent respectively. Over time, this picture reverses: Between 1970 and 1990, the size of the low-income component decreases to roughly 31 % and the size of the middle income component increases to 50 %. After 1996 the lower-income component is about as large as the low-income and middle-income components were jointly, which again supports the observation that those two components merged into a new lower-income component. Between 1991 and 1995 there is some variation in the component sizes, but before and after this picture is remarkably stable.

It is also important to keep the relative component sizes in mind when we interpret the component means. Even though the means of both the low-income and middle-income components stagnated between 1970 and 1990, there was still quite a bit of growth, because many countries made transition from the low-income component to the middle income component.

5 Concluding Remarks

In this paper we challenge the long standing twin peaks finding in the cross-country distribution of GPD per capita. We show that the number of peaks of a distribution depends on the scale (e.g. original or logarithmic) and argue that this feature is highly undesirable for economic interpretations. As an alternative approach to peaks, we use finite mixture models to investigate the cross-country distribution of GDP per capita, since (1) the number of components does not depend on the scale, (2) components in the mixture arguably correspond better to income clubs in the distribution than peaks, and (3) finite mixture models allow for an accurate analysis of the intra-distributional dynamics by using *a-posteriori* probability estimates.

Interestingly, our conclusions are not so different from Quah's, however, this might well be a coincidence. For the period that Quah studied, we find that the cross-country distribution of GDP per capita consisted of three components, which seem more like transition regimes rather than convergence clubs. Only for more recent years did we find that the cross-country distribution of GDP per capita consists of two groups which are quite stable and follow their own growth paths, and thus could potentially be interpreted as convergence clubs. In any case, we wanted to make the point that in our opinion, components should take the place of peaks in the literature on economic growth, because they do not suffer from the inherent shortcomings that peaks have and thus can lead to more meaningful economic interpretations.

References

Barro, R. (1991) Economic Growth in a Cross Section of Countries. *Quarterly Journal of Economics* **106**: 407–443.

Barro, R. and X. Sala-i-Martin (1992) Convergence. *Journal of Political Economy* **100**: 223–251.

Bianchi, M. (1997) Testing for Convergence: Evidence from Non-Parametric Multimodality Tests. *Journal of Applied Econometrics* **12**: 393–409.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001) A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society Series B* **63**: 19–29.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2004) Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society Series B* **66**: 95–115.

Chen, J. and Kalbfleisch,J. D. (2005) Modified likelihood ratio test in finite mixture models with a structural parameter. *Journal of Statistical Planning and Inference* **129**: 93–107.

Chen, J. and Li, P. (2009) Hypothesis test for normal mixture models: the EM approach. *Annals of Statistics* **37**: 2523–2542.

Dacunha-Castelle, D. and Gassiat, E. (1999) Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Annals of Statistics* **27**: 1178–1209

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**: 611–631.

Galor, O. (1996) Convergence? Inference from Theoretical Models. *Economic Journal* **106**: 1056–1069.

Galor, O. (2010) The 2008 Klein Lecture: Comparative Economic Development: Insights from Unified Growth Theory. *International Economic Review* **51**: 1–44.

Graham, B. S. and Temple, J. R. W. (2006) Rich nations, poor nations: how much can multiple equilibria explain? *Journal of Economic Growth* **11**: 5–41.

Hall, P. and York, M. (2001) On the calibration of Silverman's test for multimodality. *Statistica Sinica* **11**: 515–536.

Heston, A.; Summers, R. and Aten, B. (2011) Penn World Tables Version 7.0. *Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.*

Jones, C. (1997) On the Evolution of the World Income Distribution. *Journal of Economic Perspectives* **11**: 19–36.

Mankiw, G., D. Romer and D. Weil (1992) A Contribution to the Empirics of Economic Growth. *Quarterly Journal of Economics* **107**: 407–437.

Paap, R. and Dieck, H. K. (1998) Distribution and mobility of wealth of nations. *European Economic Review* **42**: 1269–1293.

Quah, D. T. (1993) Galton's Fallacy and Tests of the Convergence Hypothesis. *Scandinavian Journal of Economics* **95**: 427–443.

Quah, D. T. (1996) Twin Peaks: Growth and Convergence in Models of Distribution Dynamics. *Economic Journal* **106**: 1045–1055.

Quah, D. T. (1997) Empirics for growth and distribution: Stratification, polarisation and convergence clubs. *Journal of Economic Growth* **2**: 27–59.

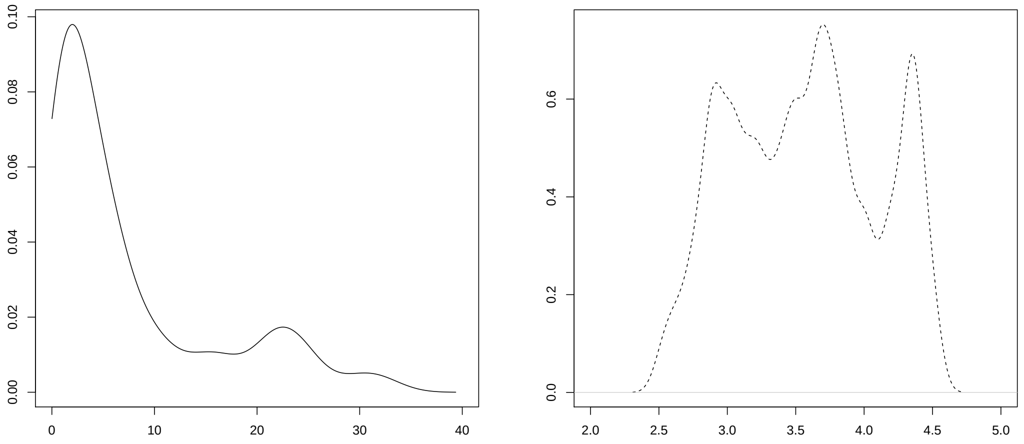
Sala-i-Martin, X. (1996) The Classic Approach to Convergence Analysis. *Economic Journal* **106**: 1019–1036.

Schumacher, I. (2009) Endogenous discounting via wealth, twin-peaks and the role of technology, *Economics Letters*, **103** : 78–80

Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society Series B* **43**: 97–99.

Strulik H. (2012) Patience and Prosperity *Journal of Economic Theory* **147**: 336–352.

Tables and Figures



(a) GDP per capita in US$1000

(b) log GDP per capita

Figure 1: Kernel density estimate for GDP per capita in US$1000 (a) and log GDP per capita (b) for 1985. We use the logarithm to the base 10.

	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
at most 1	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.030	0.023	0.008
at most 2	0.267	0.245	0.286	0.170	0.154	0.429	0.932	0.814	0.954	0.916
at most 3	0.585	0.644	0.620	0.768	0.874	0.812	0.858	0.528	0.740	0.854
at most 4	0.170	0.232	0.238	0.624	0.775	0.678	0.728	0.770	0.352	0.596
	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
at most 1	0.001	0.002	0.001	0.000	0.000	0.000	0.007	0.000	0.001	0.000
at most 2	0.650	0.302	0.672	0.806	0.430	0.364	0.390	0.784	0.833	0.508
at most 3	0.569	0.251	0.814	0.598	0.316	0.038	0.042	0.476	0.361	0.523
at most 4	0.546	0.954	0.836	0.520	0.784	0.750	0.488	0.152	0.184	0.175
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
at most 1	0.003	0.011	0.021	0.044	0.041	0.016	0.009	0.011	0.021	0.061
at most 2	0.251	0.024	0.017	0.002	0.008	0.031	0.142	0.138	0.042	0.002
at most 3	0.434	0.460	0.038	0.022	0.009	0.065	0.178	0.247	0.407	0.296
at most 4	0.326	0.124	0.134	0.106	0.222	0.254	0.245	0.116	0.188	0.220
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
at most 1	0.031	0.029	0.026	0.036	0.029	0.025	0.011	0.006	0.005	0.009
at most 2	0.000	0.000	0.002	0.000	0.000	0.002	0.008	0.017	0.042	0.034
at most 3	0.300	0.136	0.552	0.580	0.638	0.676	0.608	0.310	0.221	0.154
at most 4	0.221	0.101	0.536	0.235	0.278	0.378	0.416	0.144	0.022	0.204

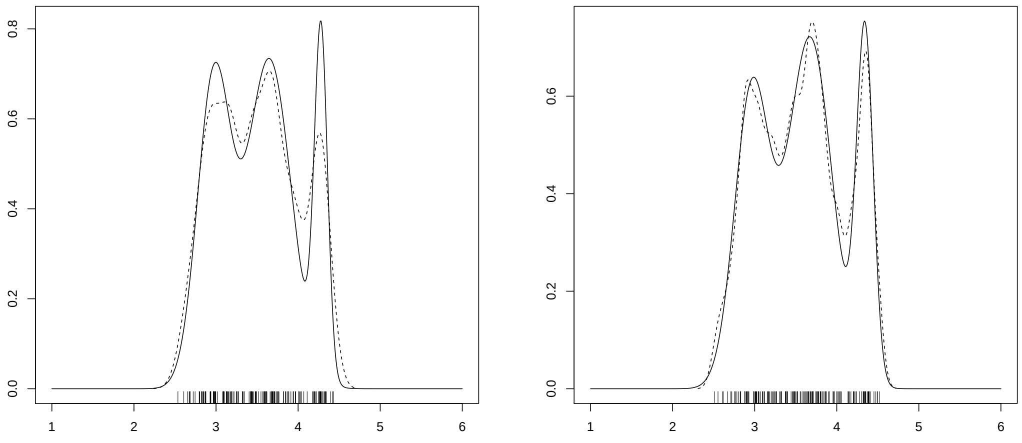
Table 1: P-values for testing the number of peaks in the cross-country distribution of GDP per capita with Silverman's test.

	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
at most 1	0.215	0.212	0.129	0.011	0.033	0.104	0.308	0.551	0.393	0.529
at most 2	0.084	0.051	0.003	0.006	0.013	0.024	0.034	0.130	0.075	0.118
at most 3	0.202	0.352	0.413	0.858	0.824	0.360	0.144	0.054	0.048	0.011
at most 4	0.729	0.596	0.656	0.654	0.367	0.290	0.711	0.276	0.822	0.850
	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
at most 1	0.533	0.163	0.249	0.167	0.197	0.221	0.180	0.246	0.264	0.264
at most 2	0.128	0.048	0.039	0.034	0.018	0.011	0.013	0.020	0.036	0.012
at most 3	0.002	0.032	0.063	0.699	0.859	0.956	0.607	0.850	0.734	0.404
at most 4	0.811	0.822	0.414	0.770	0.780	0.816	0.744	0.642	0.764	0.348
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
at most 1	0.220	0.220	0.228	0.275	0.246	0.227	0.173	0.168	0.139	0.103
at most 2	0.008	0.166	0.118	0.180	0.128	0.352	0.386	0.235	0.202	0.154
at most 3	0.878	0.004	0.013	0.013	0.120	0.214	0.408	0.166	0.389	0.388
at most 4	0.708	0.840	0.457	0.605	0.199	0.070	0.174	0.848	0.594	0.091
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
at most 1	0.136	0.105	0.102	0.109	0.133	0.126	0.157	0.171	0.245	0.262
at most 2	0.208	0.140	0.068	0.122	0.173	0.112	0.112	0.158	0.117	0.104
at most 3	0.048	0.326	0.259	0.475	0.544	0.667	0.566	0.492	0.563	0.722
at most 4	0.126	0.284	0.444	0.472	0.138	0.250	0.242	0.258	0.261	0.436

Table 2: P-values for testing the number of peaks in the cross-country distribution of log GDP per capita with Silverman's test.

	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
1 vs. 2	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
2 vs. 3	0.07	0.16	0.16	0.01	0.01	0.02	0.01	0.02	0.07	0.02
	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
1 vs. 2	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
2 vs. 3	0.02	0.01	0.00	0.01	0.02	0.01	0.01	0.02	0.04	0.02
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
1 vs. 2	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
2 vs. 3	0.01	0.04	0.41	0.07	0.07	0.04	0.06	0.19	0.06	0.05
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1 vs. 2	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
2 vs. 3	0.16	0.10	0.05	0.27	0.40	0.23	0.33	0.38	0.41	0.29

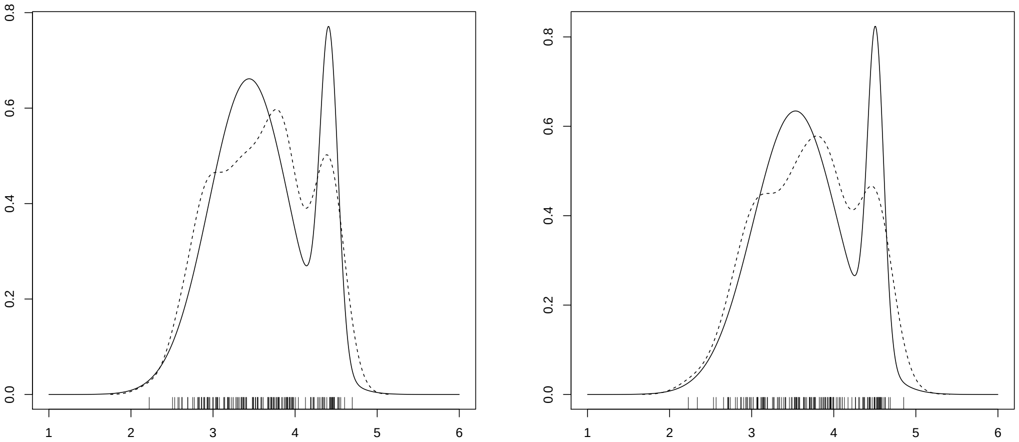
Table 3: Bootstrap p-values for testing the hypotheses of one and two components in the cross-country distribution of GDP per capita.



(a) 1975

(b) 1985

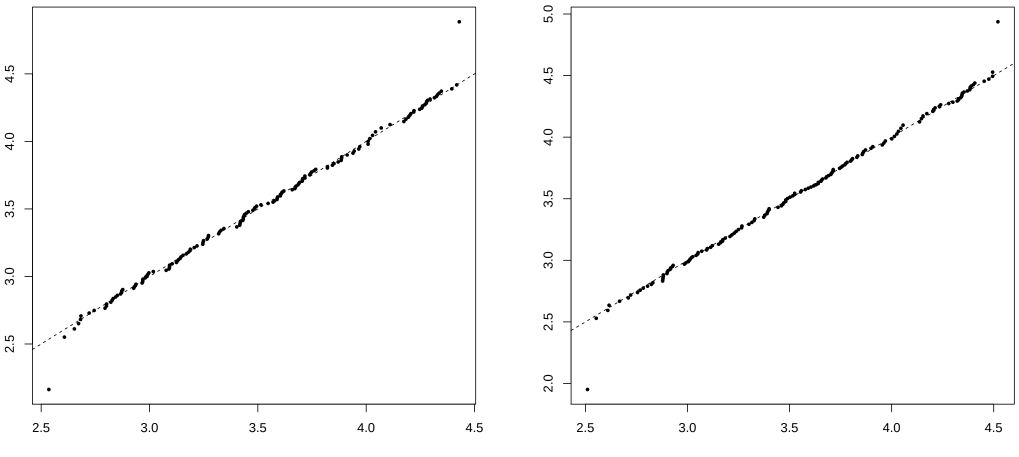
Figure 2: Fitted three-component mixture densities (solid line) and kernel density estimate based on $h_c(3)$ (dashed line) for the log-data.



(a) 1996

(b) 2005

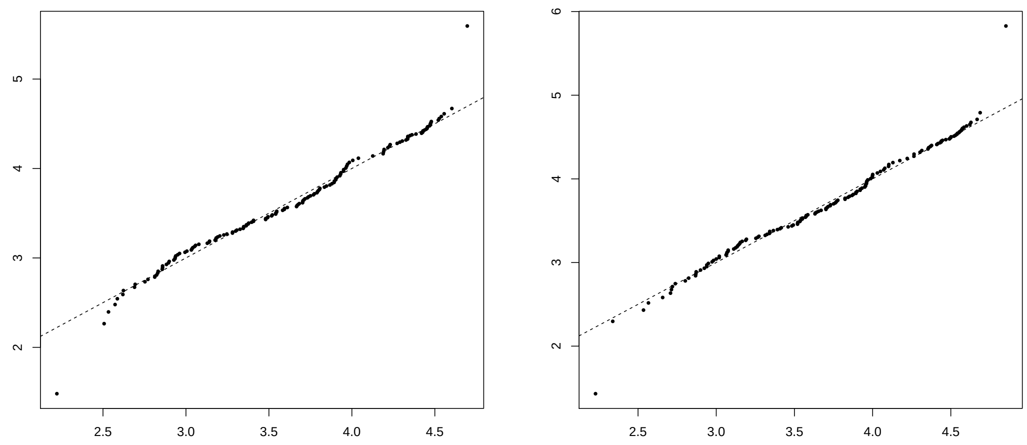
Figure 3: Fitted two-component mixture densities (solid line) and kernel density estimate based on $h_c(2)$ (dashed line) for the log-data.



(a) 1975

(b) 1985

Figure 4: QQ plot of the log-data with three components.



(a) 1996

(b) 2005

Figure 5: QQ plot of the log-data with two components.

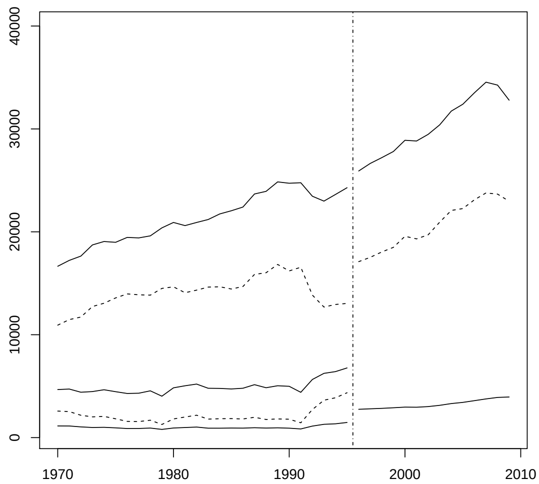


Figure 6: Estimated means (based on the log GDP data, but displayed on the original scale) of the three or respectively two distinct groups (solid lines). Income levels where the maximum a-posteriori estimates switch from one group to the other (dashed lines).

	low	middle	high
1970	1136	4669	16651
1971	1129	4719	17225
1972	1042	4411	17641
1973	983	4475	18726
1974	1000	4652	19058
1975	948	4463	18980
1976	881	4290	19457
1977	886	4314	19408
1978	928	4547	19607
1979	798	4025	20385
1980	933	4838	20910
1981	979	5038	20604
1982	1025	5202	20908
1983	916	4796	21196
1984	915	4782	21739
1985	930	4728	22048
1986	922	4798	22407
1987	962	5144	23678
1988	927	4847	23933
1989	952	5040	24853
1990	916	4985	24721
1991	842	4401	24771
1992	1120	5649	23460
1993	1292	6245	22985
1994	1342	6416	23634
1995	1472	6773	24290

(a) Component means for the years 1970 to 1995 (balanced dataset).

	low	high
1996	2752	25910
1997	2799	26646
1998	2842	27211
1999	2901	27807
2000	2964	28897
2001	2953	28827
2002	3016	29470
2003	3137	30396
2004	3308	31732
2005	3423	32402
2006	3591	33512
2007	3760	34544
2008	3905	34259
2009	3946	32791

(b) Component means for the years 1996 to 2009 (balanced dataset).

Table 4: Estimated means (based on the log GDP data, but displayed on the original scale) of the three or respectively two distinct groups.

	low	middle	high
1970	0.50	0.33	0.17
1971	0.48	0.35	0.17
1972	0.43	0.40	0.17
1973	0.40	0.44	0.17
1974	0.39	0.44	0.17
1975	0.35	0.49	0.15
1976	0.30	0.56	0.14
1977	0.29	0.56	0.15
1978	0.31	0.54	0.16
1979	0.23	0.62	0.15
1980	0.33	0.51	0.16
1981	0.35	0.47	0.17
1982	0.37	0.46	0.17
1983	0.33	0.51	0.17
1984	0.33	0.49	0.18
1985	0.33	0.49	0.18
1986	0.32	0.50	0.18
1987	0.34	0.48	0.18
1988	0.30	0.51	0.18
1989	0.31	0.52	0.17
1990	0.31	0.50	0.19
1991	0.25	0.57	0.18
1992	0.42	0.36	0.22
1993	0.49	0.27	0.24
1994	0.50	0.25	0.24
1995	0.53	0.23	0.24

(a) Component weights for the years 1970 to 1995 (balanced dataset).

	low	high
1996	0.81	0.19
1997	0.81	0.19
1998	0.81	0.19
1999	0.81	0.19
2000	0.82	0.18
2001	0.81	0.19
2002	0.81	0.19
2003	0.82	0.18
2004	0.83	0.17
2005	0.82	0.18
2006	0.82	0.18
2007	0.83	0.17
2008	0.83	0.17
2009	0.83	0.17

(b) Component weights for the years 1996 to 2009 (balanced dataset).

Table 5: Estimated weights (based on the log GDP data) of the three or respectively two distinct groups.