PHILIPPS-UNIVERSITÄT MARBURG

Hidden Markov models: Estimation theory and economic applications

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

vorgelegt von

Anna Maria Leister

Master of Science
aus Kassel

Marburg, 2016

Betreuer: Prof. Dr. Hajo Holzmann
Zweitgutachter: Prof. Dr. Jürgen Franke
Eingereicht: 18. Februar 2016
Tag der Disputation: 02. Mai 2016
Erscheinungsort: Marburg
Hochschulkennziffer: 1180

Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation

Hidden Markov models: Estimation theory and economic applications

selbst und ohne fremde Hilfe verfasst habe. Ich habe keine weiteren Quellen oder Hilfsmittel benutzt als angegeben und habe vollständig oder sinngemäße Zitate als solche gekennzeichnet.

Dies ist mein erster Versuch einer Promotion. Die Dissertation wurde bisher weder in der vorliegenden, noch in ähnlicher Form bei einer anderen in- oder ausländischen Hochschule anlässlich eines Promotionsgesuchs oder zu anderen Prüfungszwecken eingereicht.

Anna Leister
Marburg, 18. Februar 2016

“Die großen Sprünge schafft man nur in vielen kleinen Schritten”

Blumentopf - Keine Zeit

Acknowledgements

During my PhD study, I was supported by many people. Most importantly, I would like to express gratitude to my supervisor. Hajo, thank you for suggesting the problem, mentoring this thesis and bringing substantial ideas to my work. I am grateful for your constant support and encouragement, as well as your patient openness to all of my questions.

I thank Prof. Dr. Jürgen Franke for taking the Koreferat.

I would like to thank my fellow working group members and colleagues from the math-department for motivating, encouraging and entertaining talks and the good working atmosphere. Special thanks to Viktor Bengs for giving valuable comments and helpful corrections when reading this thesis.

Thanks to the Evangelisches Studienwerk Villigst for the financial support as well as many opportunities of academic and social exchange.

I gratefully acknowledge financial support from the DFG, grant HO 3260/3-2.

Thanks to my parents for supporting my studies and my family for providing a stable background, which is fundamental for the completion of this challenging project.

Contents

Introduction

Hidden Markov models are a common statistical instrument for modelling time series data. They were introduced in the 1960s by Baum and Petrie (1966) and became very popular for a wide range of applications including speech recognition (Rabiner et al., 1993), biology (Zucchini et al., 2008), signal processing (Cappé et al., 2005) and financial economics (Bhar and Hamori, 2010, Rydén et al., 1996) ever since.
The idea of the model is relatively simple. Considering a bivariate process $(X_t, S_t)_{t\in\mathcal{T}}$, the first component $(X_t)_{t\in\mathcal{T}}$ is assumed to represent a series of observations, while $(S_t)_{t\in\mathcal{T}}$ is unobserved. The sequential dependence which characterizes time series data is then modelled by assuming $(S_t)_{t\in\mathcal{T}}$ to be a Markov chain. It is further assumed that the distribution which generates an observation is determined by the corresponding state of the Markov chain. Detailed introductions can be found in Zucchini and MacDonald (2009), Cappé et al. (2005) or Elliott et al. (1995).

In this thesis the focus is on finite state space models in discrete time. The standard model is to consider a homogeneous Markov chain and state-dependent distributions from a parametric family like Gaussians. In this setting, statistical estimation theory is well developed, see Leroux (1992a) and Bickel et al. (1998) for results on consistency and asymptotic normality of the maximum likelihood estimator and Gassiat and Rousseau (2014) for asymptotic results in a Bayesian context. Computation of the maximum likelihood estimate is also very convenient in the standard model because the EM-algorithm gives closed-form estimates for many choices of the parametric state-dependent distributions.
In three chapters of this thesis, the focus is on maximum likelihood estimation in hidden Markov models where some of the standard assumptions are relaxed. Several choices of nonparametric densities which yield more flexibility of the state-dependent densities are considered, penalized estimation in certain types of models in order to allow for sparsity is discussed, and to enable time-dependent model parameters, inhomogeneous Markov chains are investigated. These adjustments illustrate the flexibility of this class of models and open it for a broader type of data sets. Next to some theoretical questions concerning maximum likelihood estimation in those models, the corresponding adjustments of the EM-algorithm are developed and numerical examples for particular models are presented.

The first chapter is an introduction to hidden Markov models and the related class of mixture models, which are relevant throughout the thesis. In addition, the general idea of the EM-algorithm, next to its formulation in context of mixture models and hidden Markov models are presented.
In Chapter 2, nonparametric maximum likelihood estimation for hidden Markov models and recent developments in this context are considered. Since identifiability of nonparametric hidden Markov models has been studied only recently, in Section 2.1 the most general statements on that issue together with required assumptions are summarized. In Sections 2.2 and 2.3, classes of state-dependent densities are specified and the corresponding maximum likelihood estimators are investigated. For the class of general mixture models, existence and consistency of a nonparametric maximum likelihood estimator are proven. In addition, the EM-algorithm is adapted and a simulation study to illustrate the theoretical results is given. For state-dependent log-concave densities, existence of the maximum likelihood estimator is proven and its shape is specified. Based on these results, numerical examples for simulated and real data are given.
Chapter 3 is about penalized estimation for hidden Markov models. Its main objective is to investigate sparsity of certain parameters in a parametric Gaussian hidden Markov model. In particular, conditional independence of random variables is considered by exploring zero entries in state-dependent precision matrices. Three penalty functions are introduced. Next to the popular l_1-penalty, hard thresholding and the SCAD-penalty function introduced by Fan and Li (2001) are considered. Known results from penalized estimation theory are transferred to the presented model, in order to prove consistency and asymptotic normality of the penalized maximum likelihood estimator. Additionally, a simulation study and a real data example compare the finite sample performance of the three penalty functions.
Chapter 4 is an application of hidden Markov models to a set of panel data containing the GDP of several countries over a certain period of time. Four models differing in dependency structure and the inclusion of covariables, which are all based on mixture models or hidden Markov models, are compared. In each model the parameters are estimated, followed by a-posteriori analysis in order to examine different income groups next to advances and decline of countries' income over time. The inclusion of covariables is an attempt to explain those developments. Since the considered models are time-inhomogeneous, the required adjustments of the EM-algorithm are deduced.
Additional material for Chapters 3 and 4 can be found in the appendix.

1. Introductory theory on hidden Markov models

Hidden Markov models are a comprehensive class of flexible statistical models to handle data possessing dependence over time. The basic idea is to model serial dependence between observations using an underlying, unobserved Markov chain. Extensive introductions to hidden Markov models can be found in Zucchini and MacDonald (2009), Cappé et al. (2005), Bhar and Hamori (2010) and Elliott et al. (1995).
In this chapter we first introduce mixture models which are closely related to hidden Markov models but designed to deal with independent data, before presenting hidden Markov models in more detail to give a theoretical background for the following chapters of this thesis.

1.1. Mixture models

Mixture models are a common statistical model for independent data which arise from a heterogeneous population that consists of various homogeneous subpopulations. Comprehensive introductions on their theory and applications can be found, for example, in Lindsay (1995), McLachlan and Peel (2004) and Frühwirth-Schnatter (2006).

A random vector X has a mixture distribution, if its probability density function admits the form

$$f_\pi(x) = \int_\Theta g_\vartheta(x)\pi(d\vartheta). \tag{1.1}$$

In that representation, $(g_\vartheta)_{\vartheta\in\Theta}$ is a parametric family of densities with respect to a σ-finite measure ν, Θ is the corresponding parameter space and π is a Borel probability measure on Θ, called the mixing distribution.
If the mixing distribution is discrete with finite support, the model is called finite mixture model and has a very illustrative presentation: For $i = 1, \ldots, n$, (X_i, S_i) is an independent random sample of the mixture model, where X_i describes the ith observation of the sample and S_i is a latent variable, indicating which subpopulation X_i belongs

to. Since π has finite support $\{\vartheta_1, \ldots, \vartheta_m\}$, the mixture density (1.1) can be written as

$$f_\pi(x) = \sum_{j=1}^{m} \pi_j g_{\vartheta_j}(x), \tag{1.2}$$

thus there are m homogeneous subpopulations. The corresponding component weights π_j ($\pi_j \geq 0$ and $\sum_{j=1}^{m} \pi_j = 1$) denote the proportion of the jth component regarding the total population. In addition, given $S_i = j$, we know that X_i is drawn from the component density g_{ϑ_j}. The support points of the mixing distribution are called the component parameters for the component densities.

The aspect identifiability of mixtures covers the question of a unique characterization of the mixture model. In our context, a mixture model is called identifiable, if for any probability measures π, π^*, the fact $\int_\Theta g_\vartheta(x) d\pi(\vartheta) = \int_\Theta g_\vartheta(x) d\pi^*(\vartheta)$ implies that $\pi = \pi^*$. This problem was discussed for example in Teicher (1960, 1961) and was extensively studied in the context of finite mixtures, see Teicher (1963) and Yakowitz and Spragins (1968).

In the following we will address the problem of estimating the mixing distribution or the parameters of the mixture model, when given a random sample $x_1, \ldots, x_n$ from the mixture density f_π. A classical estimator is a maximizer of the likelihood function

$$\mathcal{L}_n(\pi) = \prod_{i=1}^{n} f_\pi(x_i).$$

Using methods from convex analysis, Lindsay (1983) showed that a maximum likelihood estimator of the mixing distribution exists and has finite support size $m \leq n$. For a detailed introduction see Lindsay (1995). We will use parts of this method when investigating hidden Markov models with state dependent mixtures in Section 2.2.
Results on consistency of the maximum likelihood estimator can be found for example in Kiefer and Wolfowitz (1956), Pfanzagl (1988) and Leroux (1992b).

A common algorithm for computing the maximum likelihood estimator in the context of finite mixture models is the EM-algorithm introduced by Dempster et al. (1977) and Redner and Walker (1984). Laird (1978) and DerSimonian (1986) gave advice on the computation of the nonparametric maximum likelihood estimator of the mixing distribution. We will revisit the methodology of the EM-algorithm in Section 1.3.

After estimating the parameters of a finite mixture model, a common task is to match the observations to the estimated components of the model. This can be realized by assigning observation x_i ($i = 1, \ldots, n$) to the component which has the highest estimated

posterior probability,

$$\underset{k=1,\ldots,m}{\operatorname{argmax}} \frac{\hat{\pi}_k g_{\hat{\vartheta}_k}(x)}{\sum_{j=1}^{m} \hat{\pi}_j g_{\hat{\vartheta}_j}(x)}.$$

Detailed illustrations of clustering and classification methods in the context of mixture models can be found in Fraley and Raftery (2002) and Ritter (2014).

1.2. Hidden Markov models

In this section we introduce the setting of hidden Markov models (HMMs) with finite state space and in discrete time, as well as some aspects of inference in this context, which will be revisited in the following chapters of this thesis.
A hidden Markov model is a bivariate process $(X_t, S_t)_{t\in\mathbb{N}}$, where $(X_t)_{t\in\mathbb{N}}$ represents the process of observations and $(S_t)_{t\in\mathbb{N}}$ is a latent, unobserved process. We assume X_t to take values in a subset of Euclidean space $\mathcal{X} \subset \mathbb{R}^p$ and $(S_t)_{t\in\mathbb{N}}$ to be a K-state first order time-homogeneous Markov chain, i.e. for $t \in \mathbb{N}$,

$$P(S_t = k_t | S_1 = k_1, \ldots, S_{t-1} = k_{t-1}) = P(S_t = k_t | S_{t-1} = k_{t-1}) =: \gamma_{k_{t-1},k_t},$$

$k_t \in \{1, \ldots, K\}$. The transition probabilities are summarized in the transition probability matrix (t.p.m.) $\Gamma = (\gamma_{k,l})_{k,l=1,\ldots,K}$. The Markov chain is thus characterized by Γ and its initial distribution $\alpha_k = P(S_1 = k)$, $k = 1, \ldots, K$.
The observable process $(X_t)_{t\in\mathbb{N}}$ is assumed to be independent given $(S_t)_{t\in\mathbb{N}}$ and the conditional distribution of X_u given $(S_t)_{t\in\mathbb{N}}$ depends on S_u only and is called state-dependent distribution. We denote the distribution functions of the state-dependent distributions for X_t given $S_t = k$ by F_k $(k = 1, \ldots, K)$ and assume that they have densities $f_1, \ldots, f_K \in \mathcal{F}$, where $\mathcal{F}$ denotes a class of densities on $\mathcal{X}$ with respect to a σ-finite measure ν.
If the Markov chain is stationary and thus has a stationary starting distribution δ satisfying $\delta\Gamma = \delta$, the marginal distribution function of X_t is given by a finite mixture of the state-dependent distribution functions, where the component weights are given by the stationary distribution. Due to this relation, hidden Markov models are also called Markov-dependent mixtures.

An important issue when working with hidden Markov models is the question of identifiability. Leroux (1992a) showed that identifiability of parametric models is strongly connected to results on identifiability of mixtures (see Teicher, 1963). Namely it follows if identifiability holds for the marginal finite mixture of the hidden Markov model.
For semi- or nonparametric settings, using the dependence structure of a hidden Markov

model allows for identifiability results in greater generality than for mixture models. Results on this aspect were developed in Allman et al. (2009), Gassiat et al. (2016) and Gassiat and Rousseau (2016), finally Alexandrovich et al. (2016) showed that if the transition probability matrix of the Markov chain is ergodic, has full rank, and if the state-dependent distributions are all distinct, the parameters of a nonparametric hidden Markov model are identified. An overview on identifiability in semi-parametric hidden Markov models can be found in Dannemann et al. (2014).

For a given number of states K, the parameters of a hidden Markov model are commonly estimated using the maximum likelihood method. Let $\lambda = (\alpha, \Gamma, f_1, \ldots, f_K) \in \Lambda$ denote the parameter vector, where Λ denotes the parameter space of the model. For a sample $x = (x_1, \ldots, x_T)$, the likelihood function is given by

$$\mathcal{L}_T(\lambda) = \sum_{k_1=1}^{K} \ldots \sum_{k_T=1}^{K} \alpha_{k_1} f_{k_1}(x_1) \prod_{t=2}^{T} \gamma_{k_{t-1},k_t} f_{k_t}(x_t). \tag{1.3}$$

Because of the nonlinearity of the likelihood function, there is no analytic solution to calculate the maximum likelihood estimator (MLE). Thus, estimation is usually done performing numerical optimization of the (log-)likelihood function or using the EM-algorithm (in the HMM-context also called Baum–Welch algorithm), which will be introduced in Section 1.3.

For parametric hidden Markov models, the parameter vector has the form $\lambda = (\alpha, \Gamma, \vartheta_1, \ldots, \vartheta_K)$, where $\vartheta_k \in \Theta$ $(k = 1, \ldots, K)$ are the state-dependent parameters of the specified parametric class of densities and Θ is the corresponding parameter space. Popular examples are Poisson– or Gaussian hidden Markov models (see e.g. Zucchini and MacDonald, 2009).

In parametric settings, theoretical properties of the maximum likelihood estimator based on identifiability results are well studied. Under certain regularity conditions which will be further discussed in Chapter 3, Leroux (1992a) proved consistency of the maximum likelihood estimator, while Bickel et al. (1998) established its asymptotic normality.

From a theoretical point of view, nonparametric maximum likelihood estimation for hidden Markov models is not very well developed, while numerical approaches are available, see e.g. Dannemann (2012) for a nonparametric EM-algorithm. In Chapter 2, based on the identifiability result of Alexandrovich et al. (2016), we investigate existence and consistency of a nonparametric maximum likelihood estimator, when assuming the state-dependent densities of the model to be general mixtures of a parametric family. In addition, we consider state-dependent log-concave densities and prove that in this case a maximum likelihood estimator exists and its shape can be specified.

In many applications an important issue is to detect the sequence of states $\hat{s}_1, \ldots, \hat{s}_T$

of the Markov chain, which is most likely given the parameter estimates and the observations. Analogously to classification in mixture models, the most likely state of the Markov chain at time $t \in \{1, \dots, T\}$ can be derived by calculating the conditional distribution of S_t given the observations. To derive the most likely sequence of states, the joint probabilities of $X_1, \dots, X_T$ and $S_1 = s_1, \dots, S_T = s_T$, $\alpha_{s_1} f_{s_1}(x_1) \prod_{t=2}^{T} \gamma_{s_{t-1}, s_t} f_{s_t}(x_t)$ must be maximized over all possible sequences $s_1, \dots, s_T$, $s_t \in \{1, \dots, K\}$ for $t \in \{1, \dots, T\}$. Viterbi (1967) introduced an algorithm to efficiently solve this problem. We sketch his idea shortly as described in Rabiner et al. (1993):
Set

$$\xi_{1,k} = \delta_k f_k(x_1),$$

the joint probability for x_1 and $S_1 = k$ and

$$\xi_{t,k} = \max_{s_1, \dots, s_{t-1}, s_t = k} P(s_1, \dots, s_t, x_1, \dots, x_t),$$

the highest probability at time t along a path, which captures the first t observations and ends in state k. It satisfies the recursion

$$\xi_{t,k} = (\max_{j=1,\dots,K} (\xi_{t-1,j} \gamma_{j,k})) f_k(x_t), \qquad k = 1, \dots, K,\ t = 2, \dots, T.$$

The most likely sequence of states can be determined from $\hat{s}_T = \text{argmax}_{k=1,\dots,K} \xi_{T,k}$, $\hat{s}_t = \text{argmax}_{k=1,\dots,K} (\xi_{t,k} \gamma_{k, \hat{s}_{t+1}})$, $t = 1, \dots, T-1$.

Hidden Markov models can serve as statistical model in several areas of application. Rabiner et al. (1993) described how they can be applied to characterize speech in the context of speech recognition. Zucchini et al. (2008) provided a hidden Markov model for time series data on animal behaviour. Bhar and Hamori (2010) introduced applications of hidden Markov models to financial economics.
A popular practice is to use hidden Markov models to model volatility states of the financial market, see Holzmann and Schwaiger (2014), Zucchini and MacDonald (2009), Fiecas et al. (2012). The idea is to model daily log-returns of a number of shares using a hidden Markov model, assuming state-dependent Gaussian distributions. The states of the Markov chain then correspond to different volatility states of the market. We will revisit this example in Chapter 3 to illustrate our methods on penalized maximum likelihood estimation for hidden Markov models. Another similar approach in this context is the integration of a hidden Markov chain to autoregressive models for times series data from economics, in order to allow regime switches. See for example Hamilton (1989) for a parametric model or Franke et al. (2012) for a nonparametric approach and examples for financial time series.

1.3. The EM-algorithm

Since a major issue in this thesis is parameter estimation in hidden Markov models, which we will - next to theoretical results - consider in simulation studies as well as real data examples, we need to choose a stable tool to perform calculations. As we focus on estimation using the maximum likelihood method, it seems natural to apply the Expectation-Maximization (EM)-algorithm introduced by Dempster et al. (1977) as a "broadly applicable algorithm for computing maximum likelihood estimates from incomplete data". The advantage of this approach is that in a hidden Markov model, the unobserved sequence of states can be treated as missing data. Due to the work of Baum et al. (1970) in the context of hidden Markov models the EM-algorithm is also known as Baum–Welch algorithm.
In this section we state the general idea of the EM-algorithm and describe the application to mixture models as well as to hidden Markov models, comprehensive material can be found in McLachlan and Krishnan (2007).

The idea of the EM-algorithm is to maximize the complete-data likelihood function, based on the observed and missing data instead of maximizing the likelihood function of the observed data. Based on initial values for the parameters, the iterating procedure of Expectation (E)-step and Maximization (M)-step is progressed until a convergence criterion is satisfied.
In the E-step, given the observations and the current estimate of the parameter, the conditional expectation of the complete-data log-likelihood function is computed.
In the M-step, the conditional expectation of the complete-data log-likelihood function given the observations is maximized with respect to the parameters.
The resulting parameter is a (possibly local) maximum of the likelihood function.

Let us sketch the algorithm for an observable random variable X with density function $f(x;\theta)$ with parameter θ and an unobserved random variable S. The observed data are denoted by $x = (x_1, \ldots, x_n)$ and the missing data are $s = (s_1, \ldots, s_n)$. The density function of the random vector (X, S) is denoted by $f^c(x, s; \theta)$, thus the complete-data log-likelihood function, which assumes s to be observable is $\ell_n^c(\theta) = \sum_{i=1}^n \log(f^c(x_i, s_i; \theta))$.
Let $\hat{\theta}$ denote a current estimate or starting value of the parameter θ. The iterating procedure is as follows:

E-Step: Calculate the conditional expectation $E_{\hat{\theta}}(\ell_n^c(\theta) \mid x)$

M-Step: Maximize the conditional expectation with respect to θ: $\max_\theta E_{\hat{\theta}}(\ell_n^c(\theta) \mid x)$.

The EM-algorithm for finite mixture models

Assume that the density of a random vector X has the form (1.2). The vector of unknown parameters is thus $\theta = (\pi_1, \ldots, \pi_m, \vartheta_1, \ldots, \vartheta_m)$. Let $x = (x_1, \ldots, x_n)$ denote an observed random sample of (1.2). The log-likelihood function is then given by

$$\ell_n(\theta) = \sum_{i=1}^{n} \log(\sum_{j=1}^{m} \pi_j g_{\vartheta_j}(x_i)).$$

As described in Section 1.1, the sample $(s_1, \ldots, s_n)$ of the parameter process is not observed. We introduce indicator variables $u_i \in \{0,1\}^m$, where $u_{ij} = \mathbb{1}_{\{s_i=j\}}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.
Thus, the complete-data log-likelihood function has the form

$$\ell_n^c(\theta) = \sum_{j=1}^{m}\sum_{i=1}^{n} u_{ij} \log(\pi_j g_{\vartheta_j}(x_i)) = \sum_{j=1}^{m}\sum_{i=1}^{n} u_{ij} \log(\pi_j) + \sum_{j=1}^{m}\sum_{i=1}^{n} u_{ij} \log(g_{\vartheta_j}(x_i)). \quad (1.4)$$

For calculating the conditional expectations $\hat{u}_{ij}$ given x and the current estimate of the parameter θ, we apply the Bayes rule to obtain

$$\hat{u}_{ij} = P_\theta(S_i = j|x) = \frac{\pi_j g_{\vartheta_j}(x_i)}{\sum_{k=1}^{m} \pi_k g_{\vartheta_k}(x_i)}, \qquad i = 1, \ldots, n,\ j = 1, \ldots, m. \quad (1.5)$$

In the M-step we replace the u_{ij} in (1.4) by $\hat{u}_{ij}$. Maximization of the conditional expectation of the complete-data log-likelihood function can be performed considering two separate tasks. By maximizing the function with respect to π we observe

$$\hat{\pi}_j = \frac{\sum_{i=1}^{n} \hat{u}_{ij}}{n}, \qquad j = 1, \ldots, m.$$

Maximization with respect to $\vartheta_1, \ldots, \vartheta_m$ depends on the chosen family of densities $(g_\vartheta)_{\vartheta\in\Theta}$. If we choose for example the univariate Gaussian distribution, where $\vartheta_j = (\mu_j, \sigma_j)$, we obtain a closed form expression for the parameter estimates:

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \hat{u}_{ij} x_i}{\sum_{i=1}^{n} \hat{u}_{ij}}, \qquad \hat{\sigma}_j = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_{ij}(x_i - \hat{\mu}_j)^2}{\sum_{i=1}^{n} \hat{u}_{ij}}}, \qquad j = 1, \ldots, m.$$

The EM-algorithm for hidden Markov models

Due to the dependency structure of the Markov chain, the EM-algorithm for hidden Markov models is much more involved compared to mixture models. Since during the implementation numerical problems occur quickly, we describe the EM-algorithm for hidden Markov models more detailed.

Let $x = (x_1, \ldots, x_T)$ denote a sample drawn from a hidden Markov model as described in Section 1.2. For the unobserved sequence $s = (s_1, \ldots, s_T)$ let us introduce indicator variables $u_{kt} = \mathbb{1}_{\{s_t=k\}}$ $(k = 1, \ldots, K,\ t = 1, \ldots, T)$ and $v_{klt} = \mathbb{1}_{\{s_{t-1}=k, s_t=l\}}$, $k, l = 1, \ldots, K$, $t = 2, \ldots, T$.
Before assigning the EM-procedure to the setting of hidden Markov models, we introduce forward- and backward-probabilities, which will be useful when performing the E-step.

For $t = 1, \ldots, T$, $k = 1, \ldots, K$, the forward-probability a_{tk} is defined as the joint probability of the observation sequence $x_1, \ldots, x_t$ and $s_t = k$. Therefore,

$$a_{tk} = \sum_{k_1=1}^{K} \ldots \sum_{k_{t-1}=1}^{K} \alpha_{k_1} f_{k_1}(x_1) \Big(\prod_{s=2}^{t-1} \gamma_{k_{s-1},k_s} f_{k_s}(x_s)\Big) \gamma_{k_{t-1},k} f_k(x_t).$$

Let $\mathbf{a}_t = (a_{t1}, \ldots, a_{tK})$, $t = 1, \ldots, T$. The forward-probabilities can be calculated using the recursion

$$a_{1k} = \alpha_k f_k(x_1), \qquad a_{tk} = \Big(\sum_{l=1}^{K} a_{(t-1)l} \gamma_{l,k}\Big) f_k(x_t), \tag{1.6}$$

$t = 2, \ldots, T$, $k = 1, \ldots, K$.

For $t = 1, \ldots, T$, $k = 1, \ldots, K$, the backward probability b_{tk} is defined as the conditional probability of the observation sequence $x_{t+1}, \ldots, x_T$ given $s_t = k$. Thus,

$$b_{tk} = \sum_{k_{t+1}=1}^{K} \ldots \sum_{k_T=1}^{K} \gamma_{k,k_{t+1}} f_{k_{t+1}}(x_{t+1}) \prod_{s=t+2}^{T} \gamma_{k_{s-1},k_s} f_{k_s}(x_s).$$

Let $\mathbf{b}_t = (b_{t1}, \ldots, b_{tK})$, $t = 1, \ldots, T$. The backward-probabilities can be calculated using the recursion

$$b_{Tk} = 1, \qquad b_{tk} = \sum_{l=1}^{K} \gamma_{k,l} f_l(x_{t+1}) b_{(t+1)l}, \tag{1.7}$$

$t = T-1, \ldots, 1$, $k = 1, \ldots, K$.
Note that for every $t = 1, \ldots, T$, $k = 1, \ldots, K$, $a_{tk} b_{tk} = P(x, S_t = k)$ and thus $\mathbf{a}_t \mathbf{b}_t^{\mathsf{T}} = L_T(\lambda)$. Consequently, for $k, l = 1, \ldots, K$ we obtain

$$P(S_t = k|x) = \frac{a_{tk} b_{tk}}{L_T(\lambda)}, \qquad t = 1, \ldots, T$$

and

$$P(S_{t-1} = k, S_t = l|x) = \frac{a_{(t-1)k} \gamma_{k,l} f_l(x_t) b_{tl}}{L_T(\lambda)}, \qquad t = 2, \ldots, T. \tag{1.8}$$

When implementing the recursions (1.6) and (1.7), numerical problems occur since for growing t, the multiplication of forward- or backward-probabilities (each smaller than

1) with transition probabilities (also smaller than 1) tends to zero. To overcome this challenge it is recommended to use scaled versions of $\mathbf{a}_t$ and $\mathbf{b}_t$ (see e.g. Rabiner et al., 1993).

For the forward-probabilities we introduce the scaled version $\hat{\mathbf{a}}_t$ by the recursion

$$\begin{aligned} \tilde{\mathbf{a}}_1 &= \mathbf{a}_1, \qquad c_1 = \frac{1}{\sum_{k=1}^K \tilde{a}_{1k}}, \qquad \hat{\mathbf{a}}_1 = c_1 \tilde{\mathbf{a}}_1 \\ \tilde{a}_{tk} &= \sum_{l=1}^K \hat{a}_{(t-1)l} \gamma_{l,k} f_k(x_t), \qquad c_t = \frac{1}{\sum_{k=1}^K \tilde{a}_{tk}}, \qquad \hat{\mathbf{a}}_t = c_t \tilde{\mathbf{a}}_t, \end{aligned} \tag{1.9}$$

$k = 1, \ldots, K$, $t = 2, \ldots, T$.

Using the scaling factors c_t ($t = 1, \ldots, T$), introduced in (1.9), the scaled backward-probabilities $\hat{\mathbf{b}}_t$ can be calculated for $t = T-1, \ldots, 1$ by the recursion

$$\begin{aligned} \tilde{\mathbf{b}}_T &= \mathbb{1}^\mathsf{T}, \qquad \hat{\mathbf{b}}_T = c_T \tilde{\mathbf{b}}_T \\ \tilde{b}_{tk} &= \sum_{l=1}^K \gamma_{k,l} f_l(x_{t+1}) b_{(t+1)l}, \qquad \hat{\mathbf{b}}_t = c_t \tilde{\mathbf{b}}_t, \end{aligned}$$

$k = 1, \ldots, K$. For $t = 1, \ldots, T$ let $C_t := \prod_{s=1}^t c_s$ and $D_t := \prod_{s=t}^T c_s$. Note that $C_T = C_t D_{t+1}$. By induction we can prove that for $k = 1, \ldots, K$,

$$\hat{a}_{(t-1)k} = C_{t-1} a_{(t-1)k} \qquad \text{and} \qquad \hat{b}_{(t+1)k} = D_{t+1} b_{(t+1)k}. \tag{1.10}$$

Thus, $C_T \sum_{k=1}^K a_{Tk} = \sum_{k=1}^K \hat{a}_{Tk} = c_T \sum_{k=1}^K \tilde{a}_{Tk} = 1$ and since $\mathbf{a}_T \mathbf{b}_T^\mathsf{T} = L_T(\lambda)$,

$$\frac{1}{C_T} = L_T(\lambda). \tag{1.11}$$

We are now ready to describe the EM-procedure for hidden Markov models. The complete-data log-likelihood function has the form

$$\begin{aligned} \ell_T^c(\lambda) &= \log(\alpha_{s_1}) + \sum_{t=2}^T \log(\gamma_{s_{t-1},s_t}) + \sum_{t=1}^T \log(f_{s_t}(x_t)) \\ &= \sum_{k=1}^K u_{k1} \log(\alpha_k) + \sum_{k=1}^K \sum_{l=1}^K (\sum_{t=2}^T v_{klt}) \log(\gamma_{k,l}) + \sum_{k=1}^K \sum_{t=1}^T u_{kt} \log(f_k(x_t)). \end{aligned} \tag{1.12}$$

In the E-step we calculate the conditional expectations $\hat{u}_{kt}$ and $\hat{v}_{klt}$, respectively given x. In order to use the scaled versions of forward- and backward-probabilities, applying

(1.8), (1.10) and (1.11) yields

$$\begin{aligned}
\hat{u}_{kt} &= P(S_t = k|x) = \frac{a_{tk}b_{tk}}{L_T(\lambda)} = \frac{\frac{\hat{a}_{tk}}{C_t}\frac{\hat{b}_{tk}}{D_t}}{\frac{1}{C_T}} = \frac{\hat{a}_{tk}\hat{b}_{tk}}{c_t}, \qquad t = 1, \dots, T, \\
\hat{v}_{klt} &= P(S_{t-1} = k, S_t = l|x) = \frac{a_{(t-1)k}\gamma_{k,l}f_l(x_t)b_{tl}}{L_T(\lambda)} = \frac{\frac{\hat{a}_{(t-1)k}}{C_{t-1}}\gamma_{k,l}f_l(x_t)\frac{\hat{b}_{tl}}{D_t}}{\frac{1}{C_T}} \\
&= \hat{a}_{(t-1)k}\gamma_{k,l}f_l(x_t)\hat{b}_{tl}, \qquad t = 2, \dots, T,
\end{aligned} \tag{1.13}$$

$k, l = 1, \dots, K$.

Plugging in these quantities for u_{kt} and v_{klt} in (1.12) respectively, the M-step involves maximization of each of the three terms of the sum individually. For the initial distribution we obtain

$$\hat{\alpha}_k = \hat{u}_{k1}, \qquad k = 1, \dots, K, \tag{1.14}$$

and for the transition probabilities, maximization yields

$$\hat{\gamma}_{k,l} = \frac{\sum_{t=2}^{T}\hat{v}_{klt}}{\sum_{m=1}^{K}\sum_{t=2}^{T}\hat{v}_{kmt}}, \qquad k, l = 1, \dots, K. \tag{1.15}$$

The solution of the last term depends on the choice of $f_1, \dots, f_K$. In case of the multivariate Gaussian distribution ($\vartheta_k = (\mu_k, \Sigma_k)$, $k = 1, \dots, K$), there exists a closed form for the parameter estimates:

$$\hat{\mu}_k = \frac{\sum_{t=1}^{T}\hat{u}_{kt}x_t}{\sum_{t=1}^{T}\hat{u}_{kt}}, \qquad \hat{\Sigma}_k = \frac{\sum_{t=1}^{T}\hat{u}_{kt}x_t x_t^{\mathsf{T}}}{\sum_{t=1}^{T}\hat{u}_{kt}} - \hat{\mu}_k\hat{\mu}_k^{\mathsf{T}}, \qquad k = 1, \dots, K. \tag{1.16}$$

The EM-algorithm can be easily adjusted to modified or non-standard hidden Markov models. We explain the necessary adjustments for our model settings in the respective chapters of this thesis.

Effective initialization of the EM-algorithm is an important issue in order to obtain reasonable parameter estimates, even though the discussion of this aspect in the literature is not very comprehensive. For univariate data, Zucchini and MacDonald (2009) suggested initialization of the state-dependent parameters based on quantiles of the observations, Rabiner et al. (1993) proposed several ways of segmentation of the observations. Especially for multivariate data, performing clustering e.g. using the k-means algorithm might lead to reasonable initial values.

2. Nonparametric maximum likelihood estimation for hidden Markov models

In statistical research, the issue of nonparametric density estimation has been of interest for quite a while now. There are various methods for estimating a density without specifying a parametric structure. Izenman (1991) gave an overview on relevant methods, including kernel density estimation, smoothing methods and restricted maximum likelihood methods.

In recent years, nonparametric estimation of the state-dependent distributions of hidden Markov models has aroused interest in some applications (see for example Jin and Mokhtarian, 2006, Lambert et al., 2003, Lefèvre, 2003), because inappropriate parametric assumptions may lead to biased estimators and misspecification. On the theoretical side, the question of identifiability of nonparametric hidden Markov models had been unexplained, before Gassiat and Rousseau (2016), Gassiat et al. (2016) and Alexandrovich et al. (2016) gave an answer on that issue, see Dannemann et al. (2014) for an overview. Based on these results, some ideas for the estimation of the state-dependent densities have been proposed. For example, Gassiat and Rousseau (2016) considered finite translation hidden Markov models and established a nonparametric estimator based on marginal densities, using the model selection approach by Massart (2007). Vernet (2015) proposed some Bayesian estimation procedures and gave results on posterior consistency. De Castro et al. (2015) investigated a penalized least squares estimation method and gave an oracle inequality for the nonparametric estimator of the state-dependent distributions. Robin et al. (2014) presented an orthogonal-series density estimator for latent-structure models and gave its formulation for hidden Markov models.

So far, there are no theoretical results known for a nonparametric maximum likelihood estimator based on the full likelihood function of a hidden Markov model. We investigate the problem for nonparametric hidden Markov models in two different settings for the state-dependent distributions. This problem is of particular interest, since for maximum likelihood estimation we have a convenient existing computational method given by the EM-algorithm.

In Section 2.1 we summarize results on identifiability of nonparametric hidden Markov models from the literature, which were discussed in Gassiat et al. (2016), Gassiat and

Rousseau (2016) and Alexandrovich et al. (2016). They agreed on the assumption of a full-rank transition probability matrix and differed in assumptions on the state-dependent distributions. The most general statement was given by Alexandrovich et al. (2016) who assumed the state-dependent distributions to be all distinct. In addition, they provided a result on the asymptotic contrast function for maximum likelihood estimation in nonparametric hidden Markov models, which uniquely identifies the underlying parameter of the model. Based on these general results, in Section 2.2 we focus on theoretical properties of hidden Markov models with state dependent mixtures, which are quite popular in applications of speech recognition (see e.g Ajmera and Wooters, 2003, Chatzis and Varvarigou, 2007). In contrast to parametric models, where the state-dependent distributions are chosen from standard parametric families, hidden Markov models with state-dependent mixtures admit more flexible modelling and can approximate unknown densities much better. Especially multimodality, skewness and tail behaviour of the state-dependent densities can be captured by choosing a proper mixture model. We consider the nonparametric maximum likelihood estimator in this context and obtain its existence and shape, next to consistency of the state-dependent mixture densities, applying an approach from Leroux (1992a).
In Section 2.3 we impose a different shape constraint on the state-dependent distributions of the model. We consider the class of log-concave densities, which contains many popular (parametric) examples like Gaussian densities and is thus very flexible. An overview on inference and modelling within this class of densities was given by Walther (2009). We investigate the nonparametric maximum likelihood estimator of hidden Markov models with state-dependent log-concave distributions and give results on its existence and shape, next to some computational examples.

The results of Sections 2.1 and 2.2 are published in Alexandrovich et al. (2016) and the numerical examples for Section 2.3 are part of Dannemann et al. (2014).

2.1. Identification of nonparametric hidden Markov models

In this section we introduce the technical essentials to construct a maximum likelihood estimator for nonparametric hidden Markov models and to develop its theoretical properties. In the literature, the problem of identifiability in the context of nonparametric hidden Markov models is addressed, for example, in Gassiat et al. (2016), Gassiat and Rousseau (2016) and Alexandrovich et al. (2016). The results coincide in the assumption on the Markov chain, which requires the transition probability matrix to be ergodic and to have full rank. To illustrate this assumption, we give the following example:

Example 2.1. For each $K \geq 1$ we construct a $(K+1)$-state transition probability matrix of rank K and two sets of $K+1$ distributions such that the observations in a resulting hidden Markov model with $K+1$ states have the same distribution.
Let $\Gamma = (\gamma_{j,k})_{j,k=1,\dots,K}$ be a K-state ergodic transition probability matrix of full rank and $a, b \in (0,1)$, satisfying $a \neq b$ and set $c = b/(1+b-a)$, which leads to $c \in (0,1)$. Consider the following $(K+1)$-state transition probability matrix

$$\Gamma_1 = \begin{pmatrix} \gamma_{1,1} & \cdots & \gamma_{1,K-1} & c\gamma_{1,K} & (1-c)\gamma_{1,K} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \gamma_{K-1,1} & \cdots & \gamma_{K-1,K-1} & c\gamma_{K-1,K} & (1-c)\gamma_{K-1,K} \\ \gamma_{K,1} & \cdots & \gamma_{K,K-1} & c\gamma_{K,K} & (1-c)\gamma_{K,K} \\ \gamma_{K,1} & \cdots & \gamma_{K,K-1} & c\gamma_{K,K} & (1-c)\gamma_{K,K} \end{pmatrix}$$

of rank K. Let $F_1, \dots, F_{K+1}$ be linearly independent distribution functions (for example Gaussian with distinct parameters) and define a second set of distribution functions $\tilde{F}_1, \dots, \tilde{F}_{K+1}$ according to

$$\begin{aligned} \tilde{F}_1 = F_1, \dots, \tilde{F}_{K-1} = F_{K-1}, \qquad \tilde{F}_K = aF_K + (1-a)F_{K+1}, \\ \tilde{F}_{K+1} = bF_K + (1-b)F_{K+1}. \end{aligned}$$

Then $c\tilde{F}_K + (1-c)\tilde{F}_{K+1} = cF_K + (1-c)F_{K+1}$ and from Holzmann and Schwaiger (2015) the distribution of a hidden Markov model with $(K+1)$ states and transition probability matrix Γ_1, stationary starting distribution and either set of the state-dependent distributions is equal to the distribution of a stationary K-state hidden Markov model with transition probability matrix Γ and state-dependent distributions $F_1, \dots, F_{K-1}, cF_K + (1-c)F_{K+1}$.

The idea of this example is based on Holzmann and Schwaiger (2015), who indicated that hidden Markov models nest independent finite mixtures and thus identifiability of the hidden Markov model could fail if it were possible to reduce some states of the Markov chain to mixture components, since identifiability of mixtures requires strong assumptions.
Thus, we state the assumption on the Markov chain of the model as follows

A1. *The transition probability matrix* Γ *of* $(S_t)_{t\in\mathbb{N}}$ *is ergodic and has full rank.*

The assumptions on the state-dependent distributions which are required to guarantee identifiability of the model are diversely discussed in the literature. Gassiat et al. (2016) assumed the state-dependent distributions to be linearly independent. This assumption is not always easy to verify and might fail for important classes of distributions. When

thinking for example about the class of log-concave densities, containing the Gaussian distribution, convex combinations of two distinct densities might be contained in the class of densities and thus three distinct distributions being linear dependent can easily be constructed.
Alexandrovich et al. (2016) stated a more general assumption:

A2. *The state-dependent distributions $F_1, \ldots, F_K$ are all distinct.*

Based on the two formulated assumptions, they gave the following theorem.

Theorem 2.1. *For a given number of states K, let $\alpha, \Gamma, F_1, \ldots, F_K$ and $\tilde{\alpha}, \tilde{\Gamma}, \tilde{F}_1, \ldots, \tilde{F}_K$ be two sets of parameters for a hidden Markov model with K states. If the joint distribution of $(X_1, \ldots, X_T)$ with $T = (2K+1)(K^2 - 2K + 2) + 1$ is equal under both sets of parameters and Γ and $F_1, \ldots, F_K$ satisfy Assumptions A1 and A2, then both sets of parameters coincide up to label switching.*

For the proof see Alexandrovich et al. (2016).

Remark 2.2. Originally, Alexandrovich et al. (2016) stated the result of Theorem 2.1 for a stationary hidden Markov model. The corresponding result requires the following additional assumption and that the joint distribution of $T = 2K + 1$ observations is equal for both parameter sets.

A3. *The Markov chain $(S_t)_{t\in\mathbb{N}}$ is stationary. Thus, it has the stationary starting distribution δ.*

Denote by $\ell_T(\lambda)$ the log-likelihood function of $x_1, \ldots, x_T$ and write $\lambda_0 = (\alpha_0, \Gamma_0, f_{1,0}, \ldots, f_{K,0})$ for the underlying parametervector of the nonparametric hidden Markov model. Based on their result on identifiability, Alexandrovich et al. (2016) proved that analogously to the parametric setting stated in Leroux (1992a), the generalized Kullback–Leibler divergence for hidden Markov models, which is defined as a limit of normalized log-likelihood functions, uniquely identifies the parameters of the model. To obtain this result two more assumptions are required.

A4. $E_{\lambda_0}|\log(f_{k,0}(x_1))| < \infty$, $k = 1, \ldots, K$.

A5. $E_{\lambda_0}(\log(f(x_1)))^+ < \infty$ *for* $f \in \mathcal{F}$.

Theorem 2.3. *Suppose that $(X_t, S_t)_{t\in\mathbb{N}}$ is a K-state hidden Markov model with transition probability matrix Γ_0 satisfying Assumptions A1 and A3 and that the state-dependent distributions $F_{1,0}, \ldots, F_{K,0}$ satisfy Assumption A2 and have densities $f_{1,0}, \ldots, f_{K,0} \in \mathcal{F}$,*

satisfying Assumption A4. Let α, α_0 be K-state probability vectors with strictly positive entries. Under Assumption A5, given $f_1, \ldots, f_K \in \mathcal{F}$ we have almost surely that

$$\begin{aligned} T^{-1}(\ell_T(\alpha, \Gamma, f_1, \ldots, f_K) - \ell_T(\alpha_0, \Gamma_0, f_{1,0}, \ldots, f_{K,0})) \\ \to -\mathcal{K}((\Gamma_0, f_{1,0}, \ldots, f_{K,0}), (\Gamma, f_1, \ldots, f_K)) \in (-\infty, 0], \end{aligned} \tag{2.1}$$

as $T \to \infty$ and $\mathcal{K}((\Gamma_0, f_{1,0}, \ldots, f_{K,0}), (\Gamma, f_1, \ldots, f_K)) = 0$ if and only if the two sets of parameters are equal up to label switching.

The proof is given in Alexandrovich et al. (2016).

Remark 2.4. The limit in (2.1) defines the generalized Kullback–Leibler divergence of the hidden Markov model.

Based on the statement of this theorem, in the following sections we utilize the generalized Kullback–Leibler divergence as contrast function when investigating the maximum likelihood estimator for nonparametric hidden Markov models.

2.2. Nonparametric maximum likelihood estimation for state-dependent mixtures

In this section we consider the maximum likelihood estimator of a hidden Markov model if the state-dependent distributions are general mixtures of a parametric family, as presented in Section 1.1. In Section 2.2.1 we introduce the model and state assumptions to prove existence of the maximum likelihood estimator in Section 2.2.2, as well as consistency of the state-dependent mixture densities in Section 2.2.3. In Section 2.2.4 we discuss adjustments of the EM-algorithm to our setting and investigate the performance of the maximum likelihood estimator in a simulation study. Section 2.2.5 contains the proofs for the theoretical results given in this section.

2.2.1. Hidden Markov models with state-dependent mixtures

For the hidden Markov model introduced in Section 1.2, we choose the state-dependent densities $f_1, \ldots, f_K$ to be mixture densities in the sense of (1.1). Thus, $(g_\vartheta)_{\vartheta \in \Theta}$ is a parametric family of densities on $\mathcal{X}$ with respect to ν and $\Theta \subset \mathbb{R}^d$ denotes the corresponding parameter space. Assume that the map $(x, \vartheta) \mapsto g_\vartheta(x)$ is continuous on $\mathcal{X} \times \Theta$. Let $\mathcal{G}$ denote a compact set of K-state transition probability matrices.

A6. *The parameter space of the mixture model Θ is compact and $\Gamma \in \mathcal{G}$.*

Now let $\tilde{\Theta}$ denote the set of Borel probability measures on Θ. Since we assume Θ to be compact, $\tilde{\Theta}$ is compact in the weak topology (see e.g. Taylor, 2006, Corollary 13.9). Given a mixing distribution $\pi \in \tilde{\Theta}$, the corresponding mixture density is given by $f_\pi(x) = \int_\Theta g_\vartheta(x) d\pi(\vartheta)$. Thus, in our model the state-dependent densities are from the class of all mixture densities $\mathcal{F} = \{f_\pi : \pi \in \tilde{\Theta}\}$ and the parameter vector of the hidden Markov model is given by $\lambda = (\alpha, \Gamma, \pi_1, \ldots, \pi_K) \in \mathcal{P}_K \times \mathcal{G} \times \tilde{\Theta} \times \ldots \times \tilde{\Theta}$, where $\mathcal{P}_K$ contains all strictly positive probability vectors of length K.
A sample of the described hidden Markov model is denoted by $x = (x_1, \ldots, x_T)$.
Using the introduced notation, we impose an additional assumption, which will be required in the following sections.

A7. *For every $\pi \in \tilde{\Theta}$ and a small enough neighbourhood O_π of π we have*

$$E_{\lambda_0}(\sup_{\tilde{\pi} \in O_\pi} (\log(f_{\tilde{\pi}}(x_1)))^+) < \infty.$$

In particular, this assumption implies Assumption A5.

2.2.2. Existence of the maximum likelihood estimator

In this section we show that for the hidden Markov model with state-dependent mixtures, the nonparametric maximum likelihood estimator exists and that the resulting estimator of the state-dependent mixing distributions has finite support. To prove this result, we use methods from convex analysis. For mixture models, literature on nonparametric estimation of the mixing distributions or mixture densities exists. Lindsay (1983) proved that for a sample of size n, there exists a nonparametric maximum likelihood estimator for the mixing distribution with at most n support points.

Theorem 2.5. *Under Assumption A6, for the parameters of the hidden Markov model described in Section 2.2.1 and any $T \geq 1$, there exists a maximum likelihood estimate $\hat{\lambda}_T = (\hat{\alpha}_T, \hat{\Gamma}_T, \hat{\pi}_{1,T}, \ldots, \hat{\pi}_{K,T})$ for which the state-dependent mixing distributions have the form*

$$\hat{\pi}_{k,T} = \sum_{j=1}^{m} a_j \delta_{\vartheta_{j,k}}, \qquad k = 1, \ldots, K,$$

where $m \in \{1, \ldots, KT+1\}$, $a_j > 0$, $\sum_{j=1}^m a_j = 1$, $\vartheta_{j,k} \in \Theta$ $(j = 1, \ldots, m)$ and where δ_ϑ denotes the point-mass at ϑ.

The proof, which is stated in Section 2.2.5, uses arguments from convex analysis similar to the corresponding proof for existence of the maximum likelihood estimator for mixture models in Lindsay (1983). The result that instead of n components in the mixture model

$KT + 1$ components are required for the hidden Markov model is due to the fact that the model has K distinct states and that the likelihood function is not convex. The result of the theorem justifies applying hidden Markov models with state-dependent finite mixtures, as for example used in Holzmann and Schwaiger (2015) and Volant et al. (2013).

2.2.3. Consistency of the state-dependent densities

We now turn to consistency of the maximum likelihood estimator. We do not focus on estimating the mixing distributions π_k $(k = 1, \dots, K)$ but rather the mixture densities f_{π_k}, since proving their consistency does not require identification of the π_k from the mixture density f_{π_k}. This allows a more flexible choice of the mixture model and includes for example general mixtures of Gaussian densities in mean and variance, even though in this case identification of the mixing distribution fails (see Teicher, 1960).
To prove our result, we review parts of the consistency proof for the parametric hidden Markov model, given by Leroux (1992a).
Let (P) denote convergence in probability.

Theorem 2.6. *For the nonparametric hidden Markov model described in Section 2.2.1 suppose A1–A4 and A6–A7 and let* $\hat{\lambda}_T = (\hat{\alpha}_T, \hat{\Gamma}_T, \hat{\pi}_{1,T}, \dots, \hat{\pi}_{K,T})$ *denote a maximum likelihood estimator.*
Then after relabeling, $\hat{\Gamma}_T \to \Gamma_0$ (P) *and for any* $x \in \mathcal{X}$ *and* $k \in \{1, \dots, K\}$ *we have*

$$f_{\hat{\pi}_{k,T}}(x) \longrightarrow f_{k,0}(x) \ (P).$$

If the mixing distributions π_k *are identified from the mixture densities* f_{π_k} $(k = 1, \dots, K)$*, then* $d_w(\hat{\pi}_{k,T}, \pi_{k,0}) \to 0$ (P)*, where* d_w *denotes a distance that metrizes weak convergence in* $\tilde{\Theta}$*.*

To prove the general result without imposing the additional assumption of identifiability for the mixture models, the statement of the following lemma is required. We provide a proof in Section 2.2.5.

Lemma 2.7. *Let* (Θ, d) *be a metric space. Every bounded and uniformly continuous function* $g : \Theta \to [0, \infty)$ *can be uniformly approximated by Lipschitz-continuous functions.*

Remark 2.8. Lemma 2.7 can also be formulated for bounded and uniformly continuous functions $\tilde{g} : \Theta \to \mathbb{R}$, since its positive and negative parts are bounded and uniformly continuous, too.

The proof of Theorem 2.6, given in Section 2.2.5, follows the arguments of Leroux (1992a) for the parametric case, which help to obtain the consistency of $\hat{\Gamma}_T$ and the convergence $d_w(\hat{\pi}_{k,T}, \tilde{\Theta}_{k,0}) \to 0$ in probability, where $\tilde{\Theta}_{k,0} = \{\pi \in \tilde{\Theta} : \ f_\pi = f_{\pi_{k,0}}\}$. It gets clear from the second part of the theorem that if the mixing distributions are identified from the mixture densities and thus for $k = 1, \ldots, K$, each $\tilde{\Theta}_{k,0}$ contains a single mixing distribution only, consistency follows directly. This part of the proof is based on the fact that the generalized Kullback–Leibler divergence serves as an asymptotic contrast for maximum likelihood estimation in hidden Markov models, as seen in Theorem 2.3. The second part of the proof consists of concluding that $f_{\hat{\pi}_{k,T}}(x) \longrightarrow f_{k,0}(x)$ (P) if $\tilde{\Theta}_{k,0}$ $(k \in \{1, \ldots, K\})$ contains more than a single mixing distribution. For this purpose we use Lemma 2.7 and approximate the function $\vartheta \mapsto g_\vartheta(x)$ by Lipschitz-continuous functions for fixed $x \in \mathcal{X}$. This approach allows to estimate $P(|f_{\hat{\pi}_{k,T}}(x) - f_{k,0}(x)| > \varepsilon)$ by the bounded Lipschitz metric, which metrizes weak convergence on $\tilde{\Theta}$. Thus, using the result of the first part of the proof, consistency of the mixture densities follows.

2.2.4. Simulation

In this section we investigate the numerical properties of the maximum likelihood estimates in hidden Markov models with state-dependent mixtures. Due to the nonparametric setting, the EM-algorithm described in Section 1.3 needs to be modified. We first describe the resulting algorithm before we give some numerical examples in selected simulation scenarios for the nonparametric hidden Markov model.

When formulating an EM-algorithm for our model, which is described in Section 2.2.1, the problem of computing a nonparametric maximum likelihood estimator of the state-dependent mixing distributions and the resulting mixtures occurs. Resulting from Theorem 2.5, the maximum likelihood estimate of the state-dependent mixing distributions is a finite mixture. Thus, we can follow the suggestions by Laird (1978): we start the estimation procedure for the hidden Markov model by estimating state-dependent 1-component mixtures of the chosen parametric family of densities and then increase the number of components in each state as long as the resulting value of the likelihood function increases or the maximal number of components according to Theorem 2.5 is reached. Volant et al. (2013) described an EM-algorithm for hidden Markov models with state-dependent finite mixtures for fixed numbers of components m_k, $k = 1, \ldots, K$. We combine their algorithm and the described estimation procedure for the state-dependent mixtures to obtain a maximum likelihood estimator for our model.

The EM-algorithm for hidden Markov models with state-dependent mixtures

To take account of the mixture model structure of the state-dependent distributions, next to the latent variable S_t, we introduce an additional latent variable Z_t referring to component z of the mixture in state k of the Markov chain, $z = 1, \ldots, m_k$, $k = 1, \ldots, K$. Analogously to Section 1.3 we introduce indicator variables $u_{kt} = \mathbb{1}_{\{s_t=k\}}$ ($k = 1, \ldots, K$, $t = 1, \ldots, T$) and $v_{klt} = \mathbb{1}_{\{s_{t-1}=k, s_t=l\}}$, $k, l = 1, \ldots, K$, $t = 2, \ldots, T$ for the unobserved sequence s. Additionally we define $w_{klt} = \mathbb{1}_{\{z_t=l|s_t=k\}}$ for $l = 1, \ldots, m_k$, $k = 1, \ldots, K$, $t = 1, \ldots, T$. The parameter corresponding to component l in state k is denoted by ϑ_l^k, $l = 1, \ldots, m_k$, $k = 1, \ldots, K$.
As described above, we start the EM-procedure with $m_k = 1$, $k = 1, \ldots, K$.
We maximize the conditional expectation of the complete-data log-likelihood function, which has the form

$$\begin{aligned}\ell_T^c(\lambda) &= \log(\alpha_{s_1}) + \sum_{t=2}^{T} \log(\gamma_{s_{t-1},s_t}) + \sum_{t=1}^{T} (\log(\pi_{z_t}^{s_t}) + \log(g_{\vartheta_{z_t}^{s_t}}(x_t))) \\ &= \sum_{k=1}^{K} u_{k1} \log(\alpha_k) + \sum_{t=2}^{T} \sum_{k=1}^{K} \sum_{l=1}^{K} v_{klt} \log(\gamma_{k,l}) + \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{l=1}^{m_k} u_{kt} w_{klt} \log(\pi_l^k) \\ &\quad + \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{l=1}^{m_k} u_{kt} w_{klt} \log(g_{\vartheta_l^k}(x_t)),\end{aligned}$$

where the π_l^k denote the component weights of the state-dependent mixtures, $l = 1, \ldots, m_k$, $k = 1, \ldots, K$.

The E-step comprises the calculation of conditional expectations $\hat{u}_{kt}$, $\hat{v}_{klt}$ by (1.13) as well as

$$\hat{w}_{klt} = \frac{\pi_l^k g_{\vartheta_l^k}(x_t)}{\sum_{j=1}^{m_k} \pi_j^k g_{\vartheta_j^k}(x_t)}, \qquad l = 1, \ldots, m_k,\ k = 1, \ldots, K,\ t = 1, \ldots, T,$$

which is analogously to (1.5).

In the M-step, we maximize the conditional expectation of the complete-data log-likelihood function using the quantities from the E-step. For the initial distribution and the transition probabilities, we can use (1.14) and (1.15) respectively. For the component weights of the state-dependent mixtures we have

$$\hat{\pi}_l^k = \frac{\sum_{t=1}^{T} \hat{u}_{kt} \hat{w}_{klt}}{\sum_{t=1}^{T} \hat{u}_{kt}}, \qquad l = 1, \ldots, m_k,\ k = 1, \ldots, K.$$

Estimation of ϑ_l^k $(l = 1, \ldots, m_k,\ k = 1, \ldots, K)$ depends on the choice of the parametric family $(g_\vartheta)_{\vartheta \in \Theta}$. When choosing univariate Gaussian distributions, we obtain

$$\hat{\mu}_l^k = \frac{\sum_{t=1}^T \hat{u}_{kt}\hat{w}_{klt}x_t}{\sum_{t=1}^T \hat{u}_{kt}\hat{w}_{klt}}, \qquad \hat{\sigma}_l^k = \sqrt{\frac{\sum_{t=1}^T \hat{u}_{kt}\hat{w}_{klt}(x_t - \hat{\mu}_l^k)^2}{\sum_{t=1}^T \hat{u}_{kt}\hat{w}_{klt}}},\ l = 1, \ldots, m_k,\ k = 1, \ldots, K.$$

Afterwards, in each state $k = 1, \ldots, K$ we increase the number of mixture components m_k and perform a grid search over a range of possible parameters $\vartheta \in \Theta$ for additional mixture components, which are added if the additional component yields an increase of the likelihood function and the maximal number of components is not exceeded.

As usual, the described EM-procedure is iterated until a convergence criterion is satisfied.

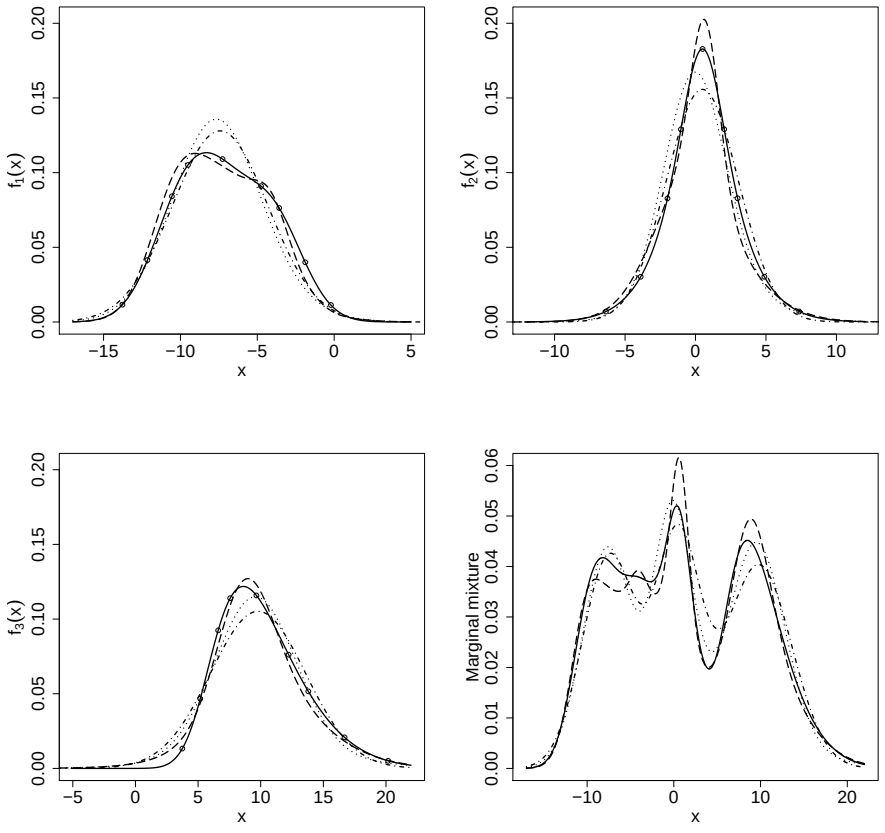


FIGURE 2.1.: State-dependent densities and marginal density of the hidden Markov model Scenario 1, together with estimators for a typical sample. Solid line: true densities, dashed line: nonparametric maximum likelihood estimators, dotted line: two-component mixture maximum likelihood estimators, dot-dashed line: Gaussian maximum likelihood estimators

Numerical results

We consider two different scenarios of three-state hidden Markov models in which the state-dependent densities are mixtures of Gaussian densities $g_\vartheta(x)$, where $\vartheta = (\mu, \sigma)$. Let $h_{\beta(a,b)}(x) = \frac{\mathbf{\Gamma}(a+b)}{\mathbf{\Gamma}(a)\mathbf{\Gamma}(b)} x^{a-1}(1-x)^{b-1}\mathbb{1}_{(0,1)}(x)$ denote the density of the Beta distribution and $h_{\beta(a,b)}(x; l, s) = h_{\beta(a,b)}(\frac{x-l}{s})/s$ the density of the Beta distribution translated by l and scaled by s. Here, $\mathbf{\Gamma}$ denotes the Gamma function.
In both scenarios, we choose for the transition probability matrix

$$\Gamma_0 = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.4 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{pmatrix}$$

and use series of length $T = 1000$ from the models specified below. In the supplementary material for Alexandrovich et al. (2016), simulation results for several choices of T are presented and illustrate consistency of the nonparametric maximum likelihood estimator.

Scenario 1
In the first scenario, the state-dependent densities are chosen as follows

$f_{1,0}(x) = 0.33 g_{(-10,2)}(x) + 0.33 g_{(-7.5,2)}(x) + 0.34 g_{(-4,2)}(x)$,

$f_{2,0}$ is a general mixture of univariate Gaussian densities, where μ follows the Beta distribution $h_{\beta(2,2)}(\mu)$ and σ is uniformly distributed on the interval $(1, 4)$,

$f_{3,0}$ is a general mixture of univariate Gaussian densities, where μ follows the Beta distribution $h_{\beta(2,11)}(\mu; 5, 33)$ and σ is uniformly distributed on $(1.4, 1.6)$.

We apply the EM-algorithm described above to compute the nonparametric maximum likelihood estimator, which is denoted by $f_{\hat{\pi}_{k,T}}$. In addition, we consider two misspecified parametric hidden Markov models. The first one with simple Gaussian distributions, where the estimators are denoted by $f_{\tilde{\pi}_{k,T}}$ and the second with two-component mixtures of Gaussian distributions, where the estimator is denoted by $f_{\bar{\pi}_{k,T}}$, $k = 1, 2, 3$.

Figure 2.1 shows the state-dependent Gaussian mixture densities $f_{k,0}$ and the fits $f_{\hat{\pi}_{k,T}}$, $f_{\tilde{\pi}_{k,T}}$ and $f_{\bar{\pi}_{k,T}}$ $(k = 1, 2, 3)$ for a typical sample. We observe that the nonparametric estimator captures the overall shape of the underlying density, in particular its skewness in states 1 and 3, much better than both parametric estimators, which deviate substantially from it.
To illustrate the consistency for $f_{\hat{\pi}_{k,T}}$ as stated in Theorem 2.6, we evaluate the relative errors over 10000 simulations of selected points indicated in Figure 2.1. The results together with those for the misspecified parametric estimators are presented in Table

x	-15.45	-13.77	-11.22	-9.05	-7.26	-5.3	-2.86	-0.21	1.56
nonpar	109.79	28.00	6.92	12.94	23.93	5.09	43.82	46.40	43.87
2-comp	117.75	28.61	6.26	12.46	25.18	4.94	45.04	48.01	49.49
Gauss	136.66	31.14	5.84	10.68	24.37	4.68	43.15	52.93	37.43

x	-9.36	-6.36	-2.71	-0.68	0.5	1.67	3.71	7.36	10.36
nonpar	65.27	22.20	64.95	9.77	13.44	19.36	25.00	59.64	67.53
2-comp	69.44	22.63	68.76	10.60	13.88	19.48	25.12	61.55	67.06
Gauss	79.61	16.69	74.73	9.60	15.02	19.97	25.32	81.74	98.08

x	2.27	3.74	6	7.99	9.66	11.61	14.93	20.17	22
nonpar	1090.32	166.99	9.90	20.26	13.87	6.38	7.04	33.61	48.31
2-comp	1103.22	175.93	8.29	22.56	15.08	5.95	6.81	37.69	50.26
Gauss	1236.47	202.98	4.79	24.17	18.80	6.69	3.24	34.78	52.51

TABLE 2.1.: Relative errors ($\times 100$) of the three estimators compared to the true densities at selected values for x averaged over 10000 replications. 'Gauss' stands for Gaussian state-dependent distributions, '2-comp' for two-component Gaussian mixtures and 'nonpar' for nonparametric Gaussian mixtures.

2.1.

We observe that the relative errors for $f_{\tilde{\pi}_{k,T}}$ and $f_{\bar{\pi}_{k,T}}$ are higher at most points than those for $f_{\hat{\pi}_{k,T}}$, in particular for states 1 and 3, which reflects the bias of those estimators due to misspecification. The estimators for the transition probability matrices perform rather similarly for the three methods. The averaged absolute errors over 10000 simulations are reported in Table 2.2.

	$\frac{1}{K}\sum_{k=1}^{K}\lvert\hat{\gamma}_{j,k}-\gamma_{j,k,0}\rvert$	$\frac{1}{K}\sum_{k=1}^{K}\lvert\bar{\gamma}_{j,k}-\gamma_{j,k,0}\rvert$	$\frac{1}{K}\sum_{k=1}^{K}\lvert\tilde{\gamma}_{j,k}-\gamma_{j,k,0}\rvert$
State $j=1$	3.89	3.97	3.74
State $j=2$	3.44	3.44	3.12
State $j=3$	2.93	3.08	2.79

TABLE 2.2.: Absolute errors ($\times 100$) of estimated transition probabilities averaged over 10000 simulations. Nonparametric estimator ($\hat{\gamma}_{j,k}$), parametric 2-component mixture model ($\bar{\gamma}_{j,k}$) and parametric Gaussian model ($\tilde{\gamma}_{j,k}$), $j,k=1,\ldots,K$.

Scenario 2

In the second simulation scenario we consider a hidden Markov model in which the state-dependent mixture densities are linearly dependent and differ not in location, as seen in scenario 1, but rather in scale. The state-dependent densities are chosen as follows

> $f_{1,0}$ is a general mixture of univariate Gaussian densities, where μ follows the Beta distribution $h_{\beta(2,11)}(\mu;-3,20)$, while the scale parameter σ is uniformly distributed on the interval $(0.9,1.5)$,

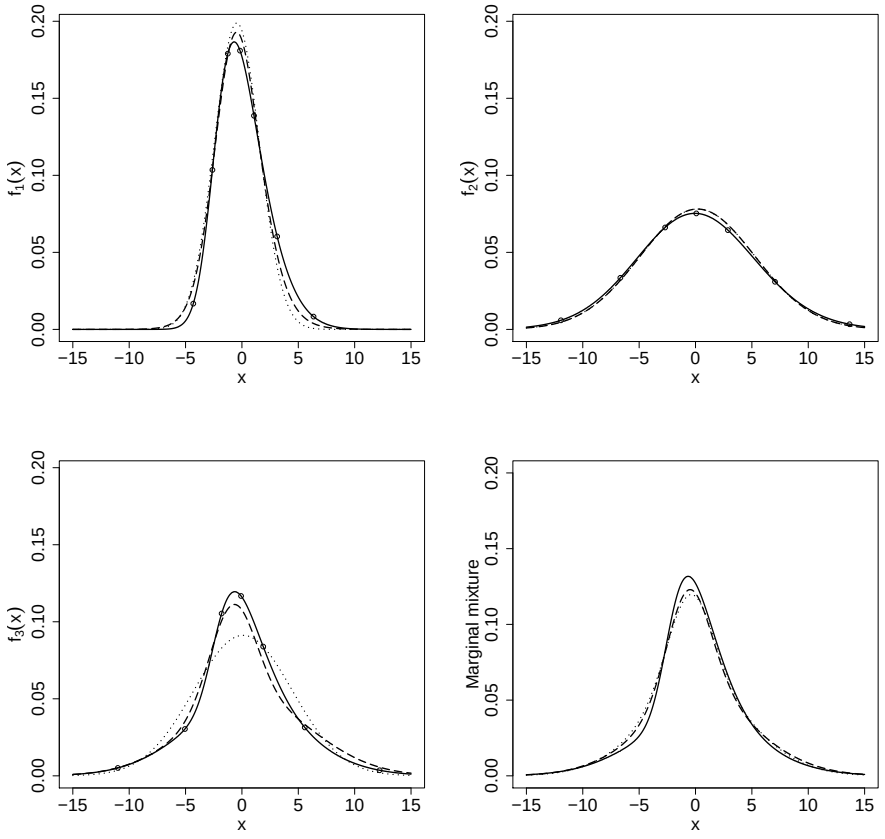


FIGURE 2.2.: State-dependent densities and marginal density of the hidden Markov model Scenario 2, together with estimators for a typical sample. Solid line: true densities, dashed line: nonparametric maximum likelihood estimators, dotted line: Gaussian maximum likelihood estimators

$f_{2,0}$ is a general mixture of univariate Gaussian densities, where μ follows the Beta distribution $h_{\beta(2,11)}(\mu; -3, 20)$ and the scale parameter σ is uniformly distributed on the interval $(4, 6)$,

$f_{3,0}(x) = 0.4f_{1,0}(x) + 0.6f_{2,0}(x)$, thus $f_{3,0}$ is a linear combination of the state-dependent densities of the first and second state.

In this scenario, we only compare the nonparametric and a parametric Gaussian maximum likelihood estimator. The estimated densities and the marginal distribution of the hidden Markov model are plotted in Figure 2.2.

Since the density in the first state is slightly skew, we observe that the nonparametric maximum likelihood estimator performs better than the parametric estimator. Due to the large scale parameters, the density in the second state is nearly symmetric, thus both estimators yield similar results. In the third state, obviously the nonparametric estimator yields a better fit, especially in tracing the left tail and the peak of the density.

x	-4.31	-2.62	-1.25	-0.17	1.07	3.12	6.35
nonparametric	22.87	12.85	7.88	18.34	15.61	27.22	72.45
parametric	27.84	10.48	6.69	19.28	20.16	32.96	94.34

x	-11.94	-6.67	-2.69	0.07	2.87	7.05	13.66
nonparametric	20.61	9.25	4.77	8.12	6.59	7.00	40.76
parametric	21.43	4.92	2.78	4.78	5.38	3.98	33.89

x	-11.01	-5.06	-1.8	-0.08	1.89	5.57	12.21
nonparametric	22.97	37.08	15.74	15.93	5.40	22.23	41.73
parametric	31.77	49.92	20.36	21.09	2.20	30.80	49.30

TABLE 2.3.: Relative errors ($\times 100$) of the two estimators compared to the true densities at selected values for x averaged over 10000 replications.

Table 2.3 shows the relative errors of the estimators evaluated for the points plotted in Figure 2.2, averaged over 10000 replications. We observe that in the first state, except for two points, the nonparametric estimator yields better results than the parametric estimator. When estimating the nearly symmetric density of the second state, the parametric estimator yields somewhat better results, whereas for the density of the third state, the advantage of the nonparametric estimator is obvious.
Again, the estimates of the transition probabilities are very similar for both estimators. In Table 2.4 we report the absolute errors of the estimated probabilities averaged over 10000 simulations.

	$\frac{1}{K}\sum_{k=1}^{K}\lvert\hat{\gamma}_{j,k}-\gamma_{j,k,0}\rvert$	$\frac{1}{K}\sum_{k=1}^{K}\lvert\tilde{\gamma}_{j,k}-\gamma_{j,k,0}\rvert$
State $j=1$	11.93	11.71
State $j=2$	9.65	9.93
State $j=3$	4.52	5.34

TABLE 2.4.: Absolute errors ($\times 100$) of estimated transition probabilities averaged over 10000 simulations. Nonparametric estimator ($\hat{\gamma}_{j,k}$) and parametric Gaussian model ($\tilde{\gamma}_{j,k}$), $j,k=1,\ldots,K$.

2.2.5. Proofs

We first give the proof for Theorem 2.5, stating the existence of a nonparametric maximum likelihood estimator for hidden Markov models with state-dependent mixtures.

Proof of Theorem 2.5. By assumption, the parameter space Θ is compact and $\vartheta \mapsto g_\vartheta(x)$ is continuous for $x \in \mathcal{X}$. Thus, by the Portmanteau Theorem, if $\pi_T \to \pi$ weakly, we

obtain $\int_\Theta g_\vartheta(x)d\pi_T(\vartheta) \to \int_\Theta g_\vartheta(x)d\pi(\vartheta)$, $T \to \infty$. In addition, the affine map

$$\begin{aligned}\Psi : \tilde{\Theta} \times \ldots \times \tilde{\Theta} &\longrightarrow \mathbb{R}^T \times \ldots \times \mathbb{R}^T \\ (\pi_1, \ldots, \pi_K) &\longmapsto ((f_{\pi_1}(x_t))_{t=1,\ldots,T}, \ldots, (f_{\pi_K}(x_t))_{t=1,\ldots,T})\end{aligned}$$

is continuous.
Since $\tilde{\Theta}$ is compact, the image $\Psi(\tilde{\Theta} \times \ldots \times \tilde{\Theta}) =: D \subseteq \mathbb{R}^{KT}$ is compact and convex.

First, for fixed Γ we express the likelihood function by

$$\begin{aligned}\tilde{\mathcal{L}}_T : D &\longrightarrow \mathbb{R} \\ (t_1, \ldots, t_K)^\mathsf{T} &\longmapsto \sum_{k_1=1}^{K} \ldots \sum_{k_T=1}^{K} \alpha_{k_1} t_{k_1,1} \prod_{s=2}^{T} \gamma_{k_{s-1},k_s} t_{k_s,s},\end{aligned}$$

where $t_k = (t_{k,1}, \ldots, t_{k,T})$, $k = 1, \ldots, K$.
Since $\tilde{\mathcal{L}}_T$ is continuous and D is compact, when maximizing $\tilde{\mathcal{L}}_T$ over D, there exists a $t^* = (t_1^*, \ldots, t_K^*)^\mathsf{T} \in D$, $t_k^* \in \mathbb{R}^T$, where $\tilde{\mathcal{L}}_T$ is maximal. By the convexity of D, we can apply Carathéodory's theorem and t^* can be expressed by a convex combination of at most $KT + 1$ extreme points $s_j^* \in D$, so that

$$t^* = \sum_{j=1}^{KT+1} a_j s_j^*, \qquad \sum_{j=1}^{KT+1} a_j = 1,\ a_j \geq 0. \tag{2.2}$$

The s_j^* are images of extreme points in $\tilde{\Theta} \times \ldots \times \tilde{\Theta}$ under the affine map Ψ (see e.g. Simon, 2011). In addition, points in the Cartesian product $\tilde{\Theta} \times \ldots \times \tilde{\Theta}$ are extreme if and only if all coordinates are extreme in $\tilde{\Theta}$. Since the extreme points in $\tilde{\Theta}$ are point masses δ_ϑ, $\vartheta \in \Theta$, there exist $\vartheta_{j,k} \in \Theta$ ($j = 1, \ldots, KT + 1$, $k = 1, \ldots, K$), such that $s_j^* = \Psi(\delta_{\vartheta_{j,1}}, \ldots, \delta_{\vartheta_{j,K}})$.
Let $m \in \{1, \ldots, KT + 1\}$ denote the number of extreme points needed in the convex combination (2.2) to express t^*. Then, after relabeling

$$t^* = \sum_{j=1}^{m} a_j s_j^* = \sum_{j=1}^{m} a_j \Psi(\delta_{\vartheta_{j,1}}, \ldots, \delta_{\vartheta_{j,K}}) = \Psi(\sum_{j=1}^{m} a_j \delta_{\vartheta_{j,1}}, \ldots, \sum_{j=1}^{m} a_j \delta_{\vartheta_{j,K}}),$$

where $a_j > 0$, $\sum_{j=1}^m a_j = 1$.
Since $\sup_{(\Gamma,\pi_1,\ldots,\pi_K)} \mathcal{L}_T(\lambda) = \sup_\Gamma \sup_{\pi_1,\ldots,\pi_K} \mathcal{L}_T(\lambda)$, the claim follows. □

We now provide a proof for Lemma 2.7 using an argument of Garrido and Jaramillo (2008).

Proof of Lemma 2.7. Choose $M > 0$ such that for all $\vartheta \in \Theta$ we have $|g(\vartheta)| < M$. Given $\varepsilon > 0$, let $N \in \mathbb{N}$ so that $(N+1)\varepsilon \geq M$.
For $n = 0, \ldots, N$ we define sets

$$C_n := \{\vartheta \in \Theta : \ (n-1)\varepsilon < g(\vartheta) < (n+1)\varepsilon\},$$

which cover Θ.
By construction, for $|n-m| > 1$ we have $C_n \bigcap C_m = \emptyset$. Since g is uniformly continuous, we can choose a $\delta > 0$, so that for $\eta, \vartheta \in \Theta$ with $d(\eta, \vartheta) < \delta$ we have $|g(\eta) - g(\vartheta)| < \frac{\varepsilon}{2}$.

Let us prove that for every $\vartheta \in \Theta$, there is a $m \in \{0, \ldots, N\}$ satisfying

$$B_\delta(\vartheta) = \{\eta \in \Theta : \ d(\vartheta, \eta) < \delta\} \subseteq C_m. \tag{2.3}$$

Observe that if ϑ is contained in a single set C_m only, we must have $g(\vartheta) = m\varepsilon$. Then $B_\delta(\vartheta) \subseteq C_m$ is obvious by the choice of δ and by definition of C_m.
If $\vartheta \in C_n \bigcap C_{n+1}$ for some $n \in \{0, \ldots, N-1\}$ and $n\varepsilon < g(\vartheta) < (n+1)\varepsilon$, consider two cases:

- If $n\varepsilon < g(\vartheta) \leq (n+\frac{1}{2})\varepsilon$, we take $m = n$, which leads to $(n-\frac{1}{2})\varepsilon < g(\eta) < (n+1)\varepsilon$ for $\eta \in B_\delta(\vartheta)$ and thus $B_\delta(\vartheta) \subseteq C_n$.
- Otherwise take $m = n+1$ with analogue argumentation,

and (2.3) follows.

Now define functions

$$\begin{aligned} g_n : \Theta &\longrightarrow [0,1] \\ \vartheta &\longmapsto \inf\{1, d(\vartheta, \Theta \setminus C_n)\}, \end{aligned}$$

where $d(\vartheta, \emptyset) = \infty$.
The g_n are Lipschitz-continuous with constant 1, since for $\vartheta_1 \neq \vartheta_2$

$$\frac{|g_n(\vartheta_1) - g_n(\vartheta_2)|}{d(\vartheta_1, \vartheta_2)} \leq \frac{|d(\vartheta_1, \Theta \setminus C_n) - d(\vartheta_2, \Theta \setminus C_n)|}{d(\vartheta_1, \vartheta_2)} \leq \frac{d(\vartheta_1, \vartheta_2)}{d(\vartheta_1, \vartheta_2)} = 1.$$

Define $h(\vartheta) := \sum_{n=0}^{N} g_n(\vartheta)$. From (2.3) we have $h(\vartheta) \geq \delta$ for $\vartheta \in \Theta$. Since by construction each $\vartheta \in \Theta$ can be contained in at most two sets C_n, C_{n+1}, observe $h(\vartheta) \leq 2$.
For $\vartheta_1, \vartheta_2 \in \Theta$ we have

$$|h(\vartheta_1) - h(\vartheta_2)| \leq \sum_{n=0}^{N} |g_n(\vartheta_1) - g_n(\vartheta_2)| \leq (N+1) d(\vartheta_1, \vartheta_2),$$

which proves that h is a Lipschitz-continuous function with constant $(N+1)$.
Now set $\tilde{h}(\vartheta) := \frac{1}{h(\vartheta)} \sum_{n=0}^{N} n g_n(\vartheta)$. We show that $\tilde{h}$ is Lipschitz-continuous and that

$$\sup_{\vartheta \in \Theta} |g(\vartheta) - \varepsilon \tilde{h}(\vartheta)| \leq 2\varepsilon. \tag{2.4}$$

Using the properties of h and g_n, we obtain for $\vartheta_1, \vartheta_2 \in \Theta$

$$\begin{aligned}
|\tilde{h}(\vartheta_1) - \tilde{h}(\vartheta_2)| &\leq \sum_{n=0}^{N} |\frac{1}{h(\vartheta_1)} n g_n(\vartheta_1) - \frac{1}{h(\vartheta_2)} n g_n(\vartheta_2)| \\
&= \sum_{n=0}^{N} |\frac{n h(\vartheta_2) g_n(\vartheta_1) - n h(\vartheta_1) g_n(\vartheta_2) + n g_n(\vartheta_1) h(\vartheta_1) - n g_n(\vartheta_1) h(\vartheta_1)}{h(\vartheta_1) h(\vartheta_2)}| \\
&\leq \sum_{n=0}^{N} \frac{n g_n(\vartheta_1) |h(\vartheta_2) - h(\vartheta_1)|}{h(\vartheta_1) h(\vartheta_2)} + \sum_{n=0}^{N} \frac{n |g_n(\vartheta_1) - g_n(\vartheta_2)|}{h(\vartheta_2)} \\
&\leq (\frac{(N+1)^3}{\delta^2} + \frac{(N+1)^2}{\delta}) d(\vartheta_1, \vartheta_2).
\end{aligned}$$

To prove (2.4), suppose that $\vartheta \in C_m$. Then

$$\begin{aligned}
|\varepsilon \tilde{h}(\vartheta) - g(\vartheta)| &= |\varepsilon \tilde{h}(\vartheta) - \varepsilon m + \varepsilon m - g(\vartheta)| \\
&\leq \varepsilon |\frac{(m-1) g_{m-1}(\vartheta) + m g_m(\vartheta) + (m+1) g_{m+1}(\vartheta)}{g_{m+1}(\vartheta) + g_m(\vartheta) + g_{m-1}(\vartheta)} - m| + |\varepsilon m - g(\vartheta)| \\
&\leq \varepsilon |\frac{g_{m-1}(\vartheta) - g_{m+1}(\vartheta)}{g_{m+1}(\vartheta) + g_m(\vartheta) + g_{m-1}(\vartheta)}| + \varepsilon \leq 2\varepsilon,
\end{aligned}$$

which completes the proof. □

We are now ready to give a proof for the consistency result provided in Theorem 2.6.

Proof of Theorem 2.6. Set $\Lambda = (\mathcal{P} \times \mathcal{G} \times \tilde{\Theta} \times \ldots \times \tilde{\Theta})$ and $\Lambda_0 = (\{\alpha_0\} \times \{\Gamma_0\} \times \tilde{\Theta}_{1,0} \times \cdots \times \tilde{\Theta}_{K,0})$, where for $k = 1, \ldots, K$

$$\tilde{\Theta}_{k,0} = \{\pi \in \tilde{\Theta} : \ f_\pi = f_{\pi_{k,0}}\}.$$

Weak convergence on $\tilde{\Theta}$ can be metrized using the bounded Lipschitz metric (see e.g. Van der Vaart and Wellner, 1996),

$$d_{BL}(\pi_1, \pi_2) = \sup\{|\int f d\pi_1 - \int f d\pi_2|; \ f : \Theta \to [0,1], \ |f(\vartheta_1) - f(\vartheta_2)| \leq d(\vartheta_1, \vartheta_2)\}.$$

On $\mathcal{G}$ we take any metric that is equivalent to the Euclidean metric and on Λ we take a product metric denoted by d.
Define $H(\lambda_0, \lambda) = \lim_T \frac{1}{T} E_{\lambda_0}(\ell_T(\lambda))$.

To prove the theorem, we first show that $d(\hat{\lambda}_T, \Lambda_0) \to 0$ (P), $T \to \infty$, which implies that $\hat{\Gamma}_0 \to \Gamma_0$ (P).
In a second step we show that from $d_{BL}(\hat{\pi}_{k,T}, \tilde{\Theta}_{k,0}) \to 0$ (P) it follows that for any $x \in \mathcal{X}$, $f_{\hat{\pi}_{k,T}}(x) \to f_{k,0}(x)$ (P), $k = 1, \ldots, K$.

For the first part of the proof, we follow the argument of Leroux (1992a). The idea is to provide theory so that the generalized Kullback–Leibler divergence serves as asymptotic contrast function for maximum likelihood estimation in hidden Markov models. Therefore, we construct a subadditive process which allows to apply Kingman's subadditive ergodic theorem (Kingman, 1976) and has the same asymptotic behaviour as the log-likelihood function. This subadditive process is thus used to prove existence of the limit $H(\lambda_0, \lambda)$. Based on these steps, we develop asymptotic behaviour in a small open neighbourhood of a parameter λ, to prove consistency using an approach by Wald (1949).

For a parameter vector $\lambda = (\alpha, \Gamma, \pi_1, \ldots, \pi_K)$ and $s, t \in \mathbb{N}_0$ with $s < t$ set

$$\mathcal{L}_{s,t}(\lambda|k) := f_{\pi_k}(x_{s+1}) \sum_{k_2=1}^{K} \cdots \sum_{k_{t-s}=1}^{K} \gamma_{k,k_2} f_{\pi_{k_2}}(x_{s+2}) \prod_{u=3}^{t-s} \gamma_{k_{u-1},k_u} f_{\pi_{k_u}}(x_{s+u})$$

and

$$\mathcal{M}_{s,t}(\lambda) = \max_{1 \le k \le K} \mathcal{L}_{s,t}(\lambda|k).$$

Since $\mathcal{L}_T(\lambda) = \sum_{k=1}^{K} \alpha_k \mathcal{L}_{0,T}(\lambda|k) \ge \mathcal{M}_{0,T}(\lambda) \cdot \min_{1 \le k \le K} \alpha_k$ and $\mathcal{L}_T(\lambda) \le \mathcal{M}_{0,T}(\lambda)$, we have

$$\min_{1 \le k \le K} \alpha_k \le \frac{\mathcal{L}_T(\lambda)}{\mathcal{M}_{0,T}(\lambda)} \le 1 \qquad \text{and} \qquad \log(\min_{1 \le k \le K} \alpha_k) \le \log(\frac{\mathcal{L}_T(\lambda)}{\mathcal{M}_{0,T}(\lambda)}) \le 0.$$

Therefore, $\frac{1}{T}\ell_T(\lambda)$ and $\frac{1}{T}\log(\mathcal{M}_{0,T}(\lambda))$ have the same asymptotic behaviour, just like $\frac{1}{T}E_{\lambda_0}(\ell_T(\lambda))$ and $\frac{1}{T}E_{\lambda_0}(\log(\mathcal{M}_{0,T}(\lambda)))$.

For $s < u < t$, from Leroux (1992a, Lemma 3) we obtain $\mathcal{M}_{s,t}(\lambda) \le \mathcal{M}_{s,u}(\lambda)\mathcal{M}_{u,t}(\lambda)$ so that the process $\log(\mathcal{M}_{s,t}(\lambda))$ is subadditive. By Assumptions A1 and A3 and Lemma 1 in Leroux (1992a), the sequence $(X_t)_{t \in \mathbb{N}}$ is ergodic and thus $(\log(\mathcal{M}_{s,t}(\lambda)))$ is stationary regarding $\log(\mathcal{M}_{s,t}(\lambda)) \to \log(\mathcal{M}_{(s+1),(t+1)}(\lambda))$.
In addition, by Assumption A7,

$$E_{\lambda_0}(\log(\mathcal{M}_{0,1}(\lambda))^+) = E_{\lambda_0}(\log(\max_{1 \le k \le K} \mathcal{L}_{0,1}(\lambda|k))^+) = E_{\lambda_0}(\log(\max_{1 \le k \le K} f_{\pi_k}(x_1))^+) < \infty.$$

Thus, from Kingman's subadditive ergodic theorem (see Kingman, 1976), with probability one, $\lim_T \frac{1}{T}\log(\mathcal{M}_{0,T}(\lambda)) = M < \infty$ exists and $E(M) = \lim_T \frac{1}{T}E(\log(\mathcal{M}_{0,T}))$. In

particular, since $\log(\mathcal{M}_{0,T})$ and ℓ_T have the same asymptotics, $M = H(\lambda_0, \lambda)$ a.s.

Alexandrovich et al. (2016) showed that for the generalized Kullback–Leibler divergence for nonparametric hidden Markov models we have $\mathcal{K}(\lambda_0, \lambda) = H(\lambda_0, \lambda_0) - H(\lambda_0, \lambda) \geq 0$ and $\mathcal{K}(\lambda_0, \lambda) = 0$ if and only if $\lambda \in \Lambda_0$. Thus, $\lim_T \frac{1}{T} E_{\lambda_0}(\log(\mathcal{M}_{0,T})) \overset{a.s.}{=} H(\lambda_0, \lambda) < H(\lambda_0, \lambda_0)$ for $\lambda \notin \Lambda_0$.
For $\lambda \notin \Lambda_0$, there is an $\varepsilon > 0$ and $T_\varepsilon \in \mathbb{N}$ so that

$$\frac{1}{T_\varepsilon} E_{\lambda_0}(\log(\mathcal{M}_{0,T_\varepsilon}(\lambda))) < H(\lambda_0, \lambda_0) - \varepsilon. \tag{2.5}$$

$\mathcal{M}_{0,T_\varepsilon}$ is continuous in λ, since $\pi \mapsto f_\pi(x)$ is continuous. By Assumption A7 we obtain $E_{\lambda_0}(\log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T_\varepsilon}(\lambda'))^+) < \infty$ for O_λ a small neighbourhood of λ. Thus, there is a small open neighbourhood where $\log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T_\varepsilon}(\lambda')) \to \log(\mathcal{M}_{0,T_\varepsilon}(\lambda))$. In addition, since $|\log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T_\varepsilon}(\lambda'))| = \log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T_\varepsilon}(\lambda'))^+ + \log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T_\varepsilon}(\lambda'))^-$, by dominated convergence

$$\lim_{O_\lambda \to \lambda} \frac{1}{T_\varepsilon} E_{\lambda_0}(\log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T_\varepsilon}(\lambda'))) = \frac{1}{T_\varepsilon} E_{\lambda_0}(\log(\mathcal{M}_{0,T_\varepsilon}(\lambda))).$$

We can choose a neighbourhood O_λ and use (2.5) to obtain

$$\frac{1}{T_\varepsilon} E_{\lambda_0}(\log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T_\varepsilon}(\lambda'))) < \frac{1}{T_\varepsilon} E_{\lambda_0}(\log(\mathcal{M}_{0,T_\varepsilon}(\lambda))) + \frac{1}{2}\varepsilon < H(\lambda_0, \lambda_0) - \frac{1}{2}\varepsilon. \tag{2.6}$$

Using the same argument again for $\log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,t-s}(\lambda'))$, Kingman's subadditive ergodic theorem yields that with probability one

$$H(\lambda_0, \lambda; O_\lambda) := \lim_T \frac{1}{T} E_{\lambda_0}(\log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T}(\lambda')))$$

exists and $\lim_T \frac{1}{T} \log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T}(\lambda')) = H(\lambda_0, \lambda; O_\lambda)$.
Using Theorem 1 from Kingman (1976),

$$H(\lambda_0, \lambda; O_\lambda) = \inf_T \frac{1}{T} E_{\lambda_0}(\log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T}(\lambda')))$$

and together with (2.6) we obtain

$$H(\lambda_0, \lambda; O_\lambda) \leq \frac{1}{T_\varepsilon} E_{\lambda_0}(\log(\sup_{\lambda' \in O_\lambda} \mathcal{M}_{0,T_\varepsilon}(\lambda'))) < H(\lambda_0, \lambda_0) - \frac{1}{2}\varepsilon.$$

Since $\mathcal{L}_T$ and $\mathcal{M}_{0,T}$ have the same asymptotics, we have

$$\lim_T \frac{1}{T} \log(\sup_{\lambda' \in O_\lambda} \mathcal{L}_T(\lambda')) = H(\lambda_0, \lambda; O_\lambda) < H(\lambda_0, \lambda_0) - \frac{1}{2}\varepsilon$$

with probability one.

Given $\delta > 0$ let $\Lambda_\delta := \{\lambda \in \Lambda,\ d(\lambda, \Lambda_0) \geq \delta\}$. Since Λ is compact and the distance is continuous, Λ_δ is compact. Therefore, we can find finitely many O_{λ_j}, $j = 1, \ldots, q$ which cover Λ_δ.
We obtain

$$\frac{1}{T} \sup_{\lambda \in \Lambda_\delta} \ell_T(\lambda) \leq \frac{1}{T} \max_{j=1,\ldots,q} \sup_{\lambda \in O_{\lambda_j}} \ell_T(\lambda) \longrightarrow \max_{j=1,\ldots,q} H(\lambda_0, \lambda_j; O_{\lambda_j}) \leq H(\lambda_0, \lambda_0) - \varepsilon.$$

Since $\frac{1}{T}\ell_T(\hat{\lambda}_T) \geq \frac{1}{T}\ell_T(\lambda_0) \to H(\lambda_0, \lambda_0)$ and $\{\hat{\lambda}_T \in \Lambda_\delta\} \subset \{\frac{1}{T}\sup_{\lambda \in \Lambda_\delta} \ell_T(\lambda) \geq \frac{1}{T}\ell_T(\lambda_0)\}$, we obtain $P(\hat{\lambda}_T \in \Lambda_\delta) \longrightarrow 0$.

For the second part of the proof, note that the function $\vartheta \mapsto g_\vartheta(x) =: g(\vartheta)$ is continuous and by Assumption A6 the parameter space Θ is compact. Thus, the function $g(\vartheta)$ is bounded and given an $\varepsilon > 0$, from Lemma 2.7 there is a Lipschitz-continuous function h such that $|g(\vartheta) - h(\vartheta)| < \varepsilon$.
Let $K_1(\varepsilon) := \sup_{\vartheta \in \Theta} |h(\vartheta)|$ and let $K_2(\varepsilon)$ denote the Lipschitz-constant of h.
Set $K(\varepsilon) := \max\{K_1(\varepsilon), K_2(\varepsilon)\}$. Given any $\pi \in \tilde{\Theta}$, there is a $\rho \in \tilde{\Theta}_{k,0}$ for which $d_{BL}(\pi, \rho) \leq d_{BL}(\pi, \tilde{\Theta}_{k,0}) + \frac{\varepsilon}{K(\varepsilon)}$. From the definition of $\tilde{\Theta}_{k,0}$, $f_{k,0}(x) = \int_\Theta g(\vartheta) d\rho(\vartheta)$.
We estimate

$$\begin{aligned}
|f_\pi(x) - f_{k,0}(x)| &= |\int_\Theta g(\vartheta) d\pi(\vartheta) - \int_\Theta g(\vartheta) d\rho(\vartheta)| \\
&\leq \int_\Theta |g(\vartheta) - h(\vartheta)| d\pi(\vartheta) + |\int_\Theta h(\vartheta) d\pi(\vartheta) - \int_\Theta h(\vartheta) d\rho(\vartheta)| + \int_\Theta |h(\vartheta) - g(\vartheta)| d\rho(\vartheta) \\
&\leq 2\varepsilon + K(\varepsilon) d_{BL}(\pi, \rho) \\
&\leq 3\varepsilon + K(\varepsilon) d_{BL}(\pi, \tilde{\Theta}_{k,0}).
\end{aligned}$$

For $\delta := \frac{\varepsilon}{K(\varepsilon)}$ we obtain

$$P(|f_{\hat{\pi}_{k,T}}(x) - f_{\pi_{k,0}}(x)| > 4\varepsilon) \leq P(d_{BL}(\hat{\pi}_{k,T}, \tilde{\Theta}_{k,0}) > \delta) \longrightarrow 0.$$

□

2.3. Nonparametric maximum likelihood estimation for state-dependent log-concave densities

In this section we consider nonparametric maximum likelihood estimation under a shape constraint on the state-dependent densities of the hidden Markov model. Shape-constrained estimation is a quite popular approach in nonparametric density estimation, since most fully nonparametric methods require smoothing parameters which can be difficult to choose.

We consider the class of log-concave densities, which is a very flexible choice containing many commonly used parametric densities like Gaussian densities, but also skewed ones like gamma densities with shape parameter equal or greater than 1.
A function f on $\mathbb{R}^p$, which can be written as

$$f(x) = \exp(\phi(x)) \tag{2.7}$$

for some concave function $\phi : \mathbb{R}^p \to [-\infty, \infty)$, is log-concave. As an example we could think of the Gaussian density, where $\phi(x)$ is a quadratic function in x.

For independent identically distributed observations, Walther (2009) reviewed results on the theory of nonparametric maximum likelihood estimation of univariate and multivariate log-concave densities. In the univariate case, results on existence and shape of the maximum likelihood estimator can be found for example in Walther (2002), Rufibach (2006) or Pal et al. (2007). In addition, there are some results regarding consistency of the estimator in Pal et al. (2007), using the Hellinger distance or in Dümbgen and Rufibach (2009), giving results on uniform consistency on compact subspaces of the interior of the support.
Cule (2010), Cule et al. (2010) considered maximum likelihood estimation of multivariate log-concave densities and proved that the structure of the estimator is analogous to those of the univariate estimator. In Cule and Samworth (2010) they gave further results on theoretical properties of the estimator. Recently, Kim and Samworth (2014) presented results on rates of convergence in log-concave density estimation.

The advantage of the estimation problem in case of independent identically distributed observations $x_1, \ldots, x_n$ is the fact that due to (2.7), the log-likelihood function has the form $\sum_{i=1}^n \log(f(x_i)) = \sum_{i=1}^n \phi(x_i)$, which makes the problem quite accessible. Since the likelihood function for hidden Markov models (see (1.3)) does not allow this simplification, maximum likelihood estimation in this context requires some additional considerations. In Section 2.3.1 we introduce the model and state some assumptions, which allow to state results on the existence and shape of the univariate and multivariate maximum likelihood estimator in Section 2.3.2. These results justify the computation of the estimator, which we approach in Section 2.3.3. The proofs are relegated to Section 2.3.4.

2.3.1. Hidden Markov models with state-dependent log-concave densities

We consider the hidden Markov model introduced in Section 1.2 and analogously to (2.7) assume the state-dependent densities to be log-concave, i.e. for $k = 1, \ldots, K$, $f_k(x) = \exp(\phi_k(x))$ for concave functions $\phi_k : \mathbb{R}^p \to [-\infty, \infty)$.
We impose the stationarity assumption A3 and further state

A8. *The Markov chain $(S_t)_{t\in\mathbb{N}}$ is irreducible and aperiodic.*

By assuming irreducibility and aperiodicity, we have $\lim_{t\to\infty}\Gamma^t = \mathbb{1}\delta$ (see for example Seneta, 2006, Theorem 4.2).

A9. *For some $c > 0$, the transition probabilities satisfy $\min_{i,j\in\{1,\dots,K\}}\gamma_{i,j} \geq c$.*

For $t \geq 2$ and $\Gamma^t = (\gamma_{j,k}^{(t)})_{j,k=1,\dots,K}$, this assumption yields $\gamma_{j,k}^{(t)} \geq c$, $j,k = 1,\dots,K$. To see this, consider the case $t = 2$. We have $\gamma_{j,k}^{(2)} = \sum_{i_1=1}^K \gamma_{j,i_1}\gamma_{i_1,k} \geq c\sum_{i_1=1}^K \gamma_{j,i_1} = c$, because the rows of Γ sum to 1. Since the product of two transition probability matrices is a transition probability matrix again, this argument proves the general case $t \geq 2$. Together with Assumption A8 and the associated remark we also observe that $\delta_k \geq c$ for $k = 1,\dots,K$.
These considerations together with the following assumption are revisited in Section 2.3.4, when we bound the log-likelihood function of the model.

A10. *The concave functions ϕ_k $(k = 1,\dots,K)$ are bounded from above, i.e. there is a $M < \infty$ such that $-\infty \leq \phi_k(x) \leq M$ for $x \in \mathbb{R}^p$.*

Write $\phi = (\phi_1,\dots,\phi_K)$ and let $\mathcal{G}^b$ denote a compact set of K-state transition probability matrices satisfying Assumption A9. We set

$$\mathcal{F}_{lc}^b = \{\phi = (\phi_1,\dots,\phi_K):\ \phi_k : \mathbb{R}^p \to [-\infty,\infty) \text{ concave}, \int \exp(\phi_k(x))dx = 1,$$
$$\phi_k(x) \leq M,\ k = 1,\dots,K\}.$$

Thus, the parameter vector of the stationary hidden Markov model with state-dependent log-concave densities is $\lambda = (\Gamma,\phi) \in \mathcal{G}^b \times \mathcal{F}_{lc}^b$.

2.3.2. Existence and shape of the maximum likelihood estimator

We state that for the hidden Markov model with state-dependent log-concave densities, the nonparametric maximum likelihood estimator exists and specify its shape. The univariate and multivariate problem are considered separately. In each case we start with the introduction of some ideas from the estimation problem for independent identically distributed observations.

Univariate case ($p = 1$)

Existence of the univariate nonparametric maximum likelihood estimator for a log-concave density $f(x) = \exp(\phi(x))$ based on independent identically distributed observations $x_1,\dots,x_n$ is proven in Walther (2002), Rufibach (2006) and Pal et al. (2007). They

stated that the nonparametric maximum likelihood estimator $\hat{f}_n(x) = \exp(\hat{\phi}_n(x))$ exists and that the concave function $\hat{\phi}_n$ is piecewise linear with changes of slope at $x_1, \ldots, x_n$. Thus, the problem of maximizing the log-likelihood function $\sum_{i=1}^{n} \phi(x_i)$ over all concave functions ϕ such that $\exp(\phi(x))$ is a density function can be reformulated to a finite dimensional optimization problem. Details are given in Section 2.3.4.

As mentioned in the introduction of Section 2.3, due to the dependency structure and the resulting form of the likelihood function, some more considerations and assumptions have to be made in order to give a similar result in the context of the nonparametric hidden Markov model. Let $x = (x_1, \ldots, x_T)$ denote a sample from the hidden Markov model introduced in Section 2.3.1 for $p = 1$ and assume $x_1 \leq \cdots \leq x_T$.

Theorem 2.9. *Under Assumptions A3, A8, A9 and A10, for the parameters of the univariate hidden Markov model with state-dependent log-concave densities described in Section 2.3.1, a nonparametric maximum likelihood estimator $\hat{\lambda}_T = (\hat{\Gamma}, \hat{\phi})$ exists. For $k = 1, \ldots, K$, $\hat{\phi}_k$ are piecewise linear functions with knots at $x_1, \ldots, x_T$ and $supp\hat{f}_k \subset [x_1, x_T]$ for $\hat{f}_k = \exp(\hat{\phi}_k)$.*

The proof is given in Section 2.3.4. Before we can apply the ideas from Pal et al. (2007), we show that under the given assumptions, the log-likelihood function can be bounded. The remaining issue is to adjust the proof of the independent identically distributed case to the setting of hidden Markov models.

Multivariate case ($p > 1$)

The multivariate nonparametric estimation problem of finding a maximum likelihood estimator of a log-concave density based on independent identically distributed observations $x_1, \ldots, x_n$ ($n \geq p + 1$), was considered by Cule (2010), Cule and Samworth (2010) and Cule et al. (2010). Similar to the univariate setting they proved that the problem of maximizing the log-likelihood function over the class of all log-concave density functions can be reduced to maximization over a finite-dimensional subclass of functions. They showed that the maximum likelihood estimator $\hat{f}_n(x) = \exp(\hat{\phi}_n(x))$ exists, its support is the convex hull of the $x_1, \ldots, x_n$ and that $\hat{\phi}_n$ is a 'tent function'.
The term 'tent function' means a function $\varphi_y : \mathbb{R}^p \to \mathbb{R}$ for a fixed vector $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, such that φ_y is the smallest concave function satisfying $\varphi_y(x_i) \geq y_i$ for $i = 1, \ldots, n$. The $y_1, \ldots, y_n$ can be associated with the 'tent pole heights'.
We apply this idea when dealing with the multivariate hidden Markov model with state-dependent log-concave densities.

Consider a sample $x = (x_1, \ldots, x_T)$ ($T \geq p + 1$) from the hidden Markov model introduced in Section 2.3.1 and assume the observations to be all distinct. Denote by

$\mathcal{C}_T = conv(x_1, \ldots, x_T)$ the convex hull of the observations and write

$$\mathcal{F}_{tent}(x) = \{\varphi_{y_k} : \mathbb{R}^p \to \mathbb{R};\ y_k \in \mathbb{R}^T, \varphi_{y_k} \text{ least concave function so that } \varphi_{y_k}(x_t) \geq y_{k,t},\ t = 1, \ldots, T\}$$

for the set of 'tent functions' corresponding to the sample x.

Theorem 2.10. *Under Assumptions A3, A8, A9 and A10, for the parameters of the multivariate hidden Markov model with state-dependent log-concave densities described in Section 2.3.1, a nonparametric maximum likelihood estimator $\hat{\lambda}_T = (\hat{\Gamma}, \hat{\phi})$ exists. For $k = 1, \ldots, K$, $\hat{\phi}_k \in \mathcal{F}_{tent}(x)$ and $\hat{f}_k(\bar{x}) = \exp(\hat{\phi}_k(\bar{x})) = 0$ for $\bar{x} \notin \mathcal{C}_T$.*

The proof, which is similar to those of Theorem 2.9, is given in Section 2.3.4. It requires usage of some concepts from convex analysis (see for example Rockafellar, 1970), which were also used by Cule (2010), Cule and Samworth (2010).

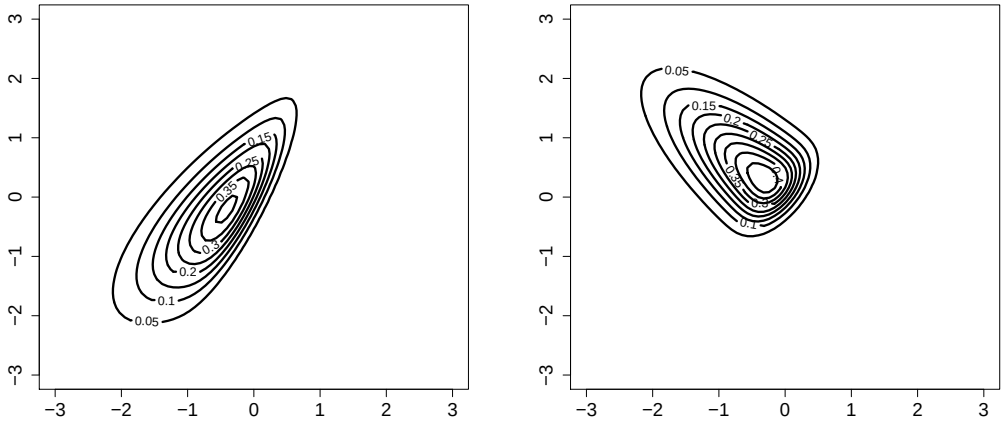
2.3.3. Numerical examples

Existence of a nonparametric maximum likelihood estimator, which was stated in the preceding section, justifies its computation when given observations $x_1, \ldots, x_T$ from a hidden Markov model. Again, the estimator can be computed using an adjusted EM-algorithm, see Dannemann (2012) or Dannemann et al. (2014) for a version for semi- or nonparametric hidden Markov models. Using notation of Section 1.3, when maximizing the complete-data log-likelihood function (1.12) it suffices to adjust the last part of maximizing $\sum_{k=1}^{K} \sum_{t=1}^{T} u_{kt} \log(f_k(x_t))$. This leads to a weighted nonparametric maximum likelihood problem.

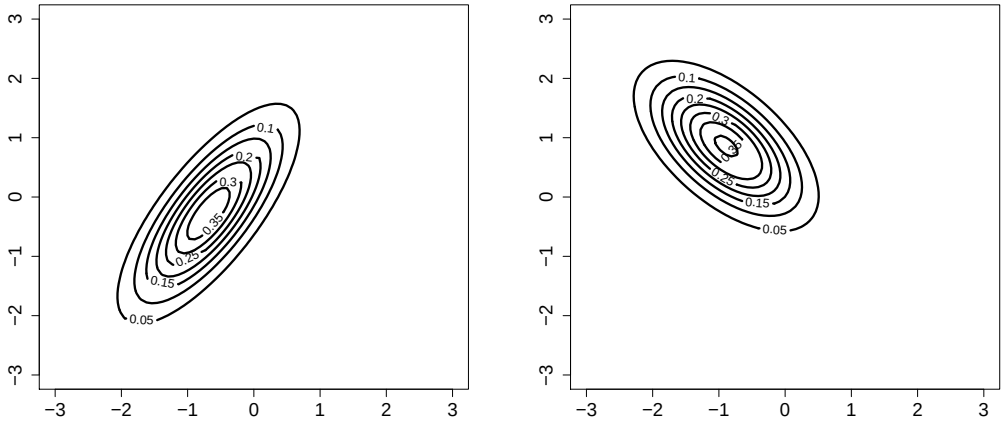
Based on the explicit form of the nonparametric maximum likelihood estimator of log-concave densities explained in Section 2.3.2, precise algorithms have been deduced. For univariate data, Dümbgen and Rufibach (2011) gave an algorithm and the corresponding implementation in the R package *logcondens*. Analogously, for multivariate data the algorithm was given by Cule et al. (2009) and implemented in the R package *LogConcDEAD*. Both implementations allow to put weights on the observations, thus we can easily use these packages for our computations in the context of hidden Markov models. The following numerical examples are published in Dannemann et al. (2014).

Simulation

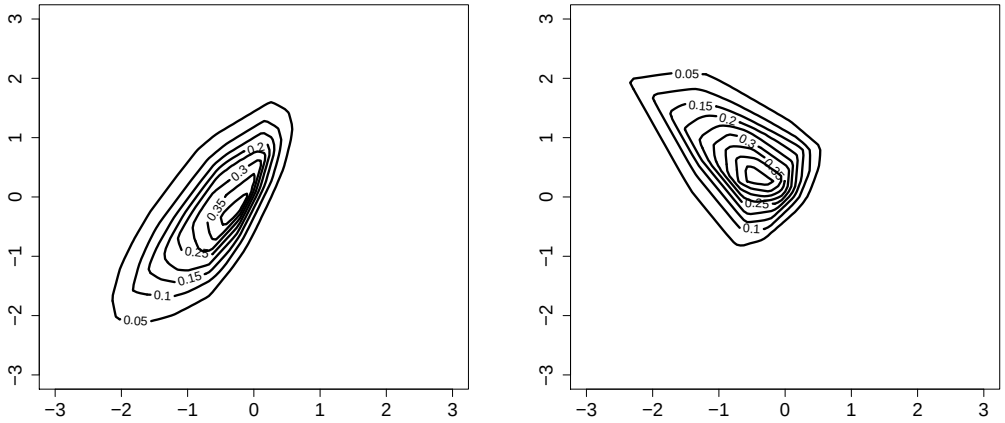
In this section, we use simulated bivariate observations from a two-state hidden Markov model to illustrate the flexibility of the nonparametric estimation procedure.



(A) True skew normal densities.



(B) Parametric estimation (Gaussian densities).



(C) Nonparametric estimation (Log-concave densities).

FIGURE 2.3.: Contour plots of the state-dependent densities in the simulation scenario. Left: state 1, right: state 2.

We choose the state-dependent densities to be skew normal (see Azzalini and Valle, 1996, for the multivariate definition) and set the transition probability matrix to

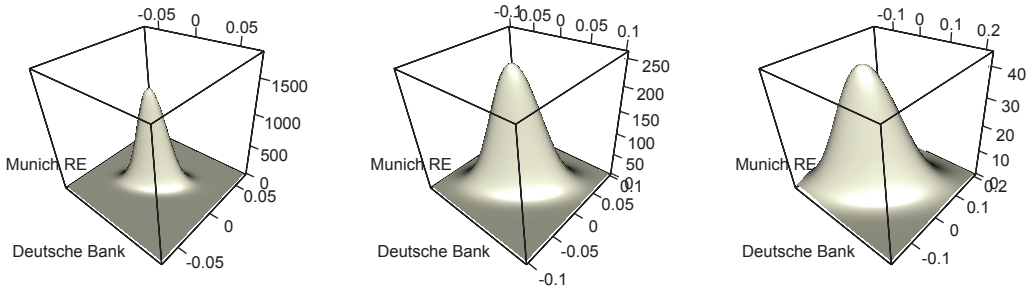
$$\Gamma = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

We generate 5000 bivariate observations from the model and estimate a parametric (Gaussian) model and a nonparametric (log-concave) model using the EM-procedure.

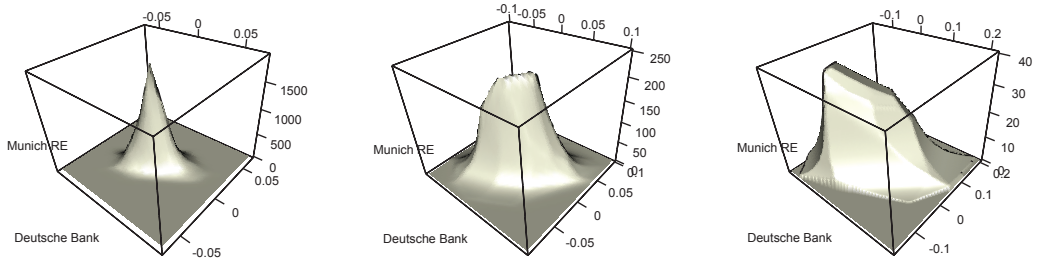
The results show the advantage of the nonparametric estimator. The estimates for the transition probability matrix of the Gaussian model ($\tilde{\Gamma}$) and the log-concave model ($\hat{\Gamma}$) are

$$\tilde{\Gamma} = \begin{pmatrix} 0.75 & 0.25 \\ 0.49 & 0.51 \end{pmatrix} \qquad \text{and} \qquad \hat{\Gamma} = \begin{pmatrix} 0.73 & 0.27 \\ 0.37 & 0.63 \end{pmatrix}.$$

The contour plots in Figure 2.3 illustrate the assumed skewness of the state-dependent densities. Of course, the parametric Gaussian estimate cannot capture the skewness, while the nonparametric log-concave estimate yields a good fit.



(A) Parametric estimation (Gauss)



(B) Nonparametric estimation (log-concave)

FIGURE 2.4.: Estimated state-dependent densities. Left: state 1 (low volatility), mid: state 2 (intermediate volatility), right: state 3 (high volatility).

Real data example

Now, we study a bivariate time series of financial log-returns. We use a database of 2926 log-returns from the stocks of Deutsche Bank and Munich RE on a daily basis from January 21, 2000 to April 23, 2013 [1] and compare two choices for the state-dependent distributions of the model.

First, we fit a three-state parametric hidden Markov model with state-dependent Gaussian densities. The estimated transition probability matrix and the corresponding stationary distribution are

$$\tilde{\Gamma} = \begin{pmatrix} 0.807 & 0.193 & 8 \times 10^{-5} \\ 0.392 & 0.575 & 0.033 \\ 0.056 & 0.477 & 0.467 \end{pmatrix} \qquad \text{and} \qquad \tilde{\delta} = \begin{pmatrix} 0.659 & 0.321 & 0.020 \end{pmatrix}.$$

For the parameters μ_k and Σ_k $(k = 1, 2, 3)$ of the state-dependent distributions we estimate

$$\tilde{\mu}_1 = \begin{pmatrix} -1.80 \times 10^{-4} \\ 1.49 \times 10^{-4} \end{pmatrix}, \qquad \tilde{\Sigma}_1 = \begin{pmatrix} 1.8 \times 10^{-4} & 0.8 \times 10^{-4} \\ 0.8 \times 10^{-4} & 1.0 \times 10^{-4} \end{pmatrix},$$
$$\tilde{\mu}_2 = \begin{pmatrix} -2.4 \times 10^{-4} \\ -2.7 \times 10^{-4} \end{pmatrix}, \qquad \tilde{\Sigma}_2 = \begin{pmatrix} 9.4 \times 10^{-4} & 5.0 \times 10^{-4} \\ 5.0 \times 10^{-4} & 6.7 \times 10^{-4} \end{pmatrix},$$
$$\tilde{\mu}_3 = \begin{pmatrix} 64.4 \times 10^{-4} \\ 52.9 \times 10^{-4} \end{pmatrix}, \qquad \tilde{\Sigma}_3 = \begin{pmatrix} 54.8 \times 10^{-4} & 29.0 \times 10^{-4} \\ 29.0 \times 10^{-4} & 39.6 \times 10^{-4} \end{pmatrix}.$$

The corresponding densities are shown in Figure 2.4a. We observe that state 1 represents a state with low volatility, whereas state 3 corresponds to a state with high volatility.

In the second model, we assume the three state-dependent densities to be log-concave and compute the maximum likelihood estimator applying the R-package *LogConcDEAD* (Cule et al., 2009) in the M-step of the EM-algorithm. The estimates of the parametric

[1] Data access from `http://de.finance.yahoo.com` (23rd April 2013)

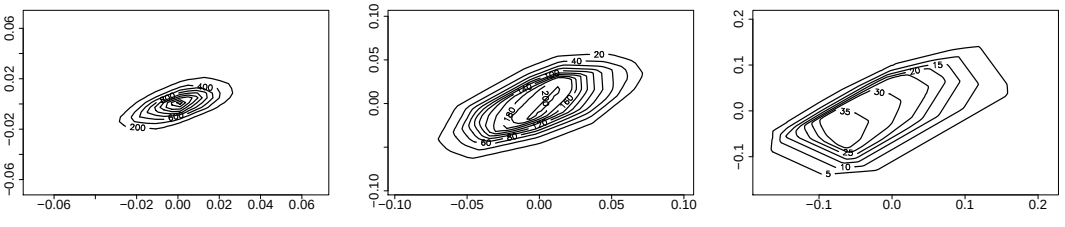


FIGURE 2.5.: Log-concave estimation - contour plots. Left: state 1 (low volatility), mid: state 2 (intermediate volatility), right: state 3 (high volatility).

fit serve as initial values for the estimation procedure.
The estimates for the transition probability matrix and the corresponding stationary distribution are

$$\hat{\Gamma} = \begin{pmatrix} 0.860 & 0.140 & 5 \times 10^{-7} \\ 0.325 & 0.644 & 0.031 \\ 0.046 & 0.482 & 0.472 \end{pmatrix} \qquad \text{and} \qquad \hat{\delta} = \begin{pmatrix} 0.689 & 0.294 & 0.017 \end{pmatrix}.$$

The estimated densities are plotted in Figure 2.4b and in Figure 2.5 we give the corresponding contour plots. The nonparametric fit, especially in state 3 is somewhat skewed, apart from that the estimates are quite similar to those of the parametric model.
Lastly, for both estimated models we perform a state-decoding using the Viterbi-algorithm, which is explained in Section 1.2. A plot of the time series and the resulting global decoding is given in Figure 2.6. We observe that the decoding based on the

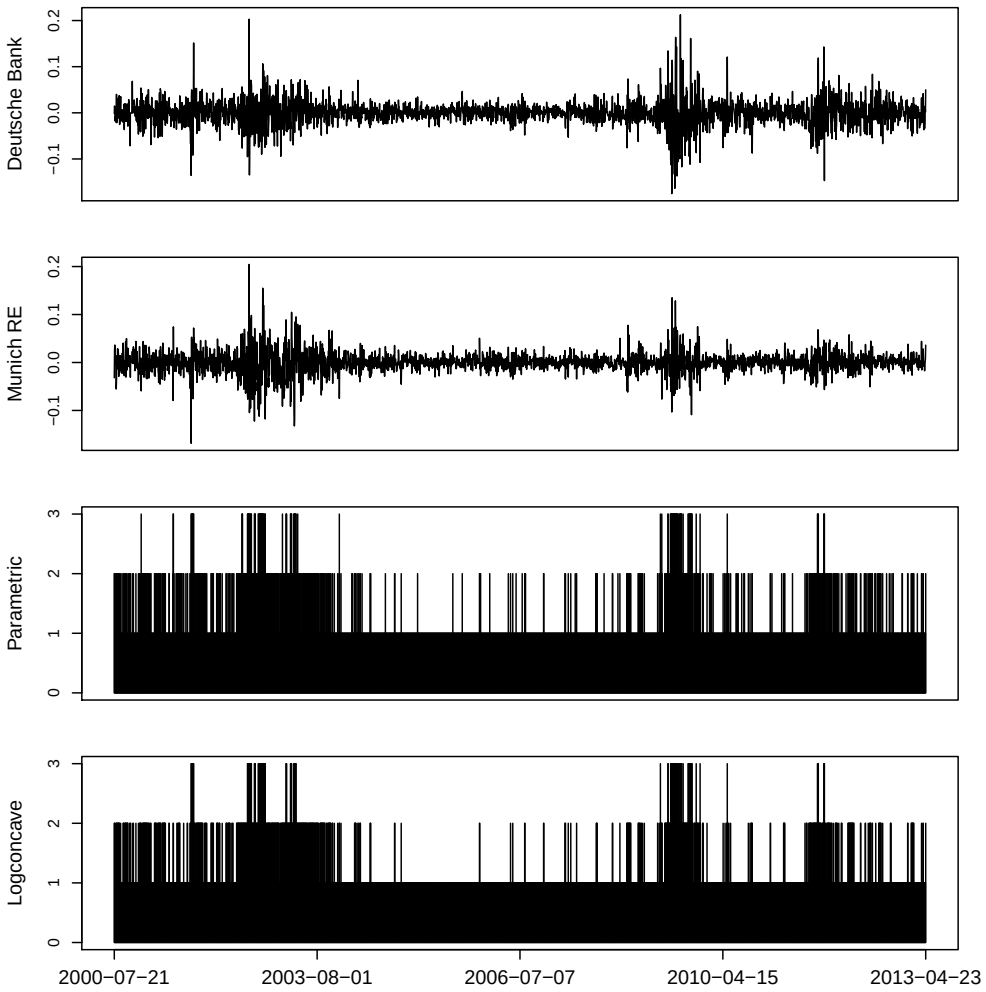


FIGURE 2.6.: From top to bottom: Series of log-returns of the two stocks. Global decoding using the Viterbi-algorithm based on parameter estimates from the parametric fit and the nonparametric log-concave fit.

nonparametric model has slightly less transitions than those based on the parametric fit.

2.3.4. Proofs

Before we give the proofs of Theorem 2.9 and Theorem 2.10, we justify the assumptions that need to be imposed in order to have the log-likelihood function of the hidden Markov model bounded. In addition, we draw a comparison to the nonparametric estimation problem with independent identically distributed observations.

As explained in Section 2.3.2, given an independent identically distributed sample $x_1, \ldots, x_n$ from a univariate log-concave density $f(x) = \exp(\phi(x))$, the nonparameteric estimation problem of maximizing the log-likelihood function $\sum_{i=1}^n \phi(x_i)$ over all concave functions ϕ such that $\exp(\phi(x))$ is a density function can be transferred to a finite dimensional maximization problem. More precisely, writing $y_i = \phi(x_i)$, we maximize $\sum_{i=1}^n y_i$ over the set of $(y_1, \ldots, y_n) \in \mathbb{R}^n$ for which

$$\frac{y_i - y_{i-1}}{x_i - x_{i-1}} \geq \frac{y_{i+1} - y_i}{x_{i+1} - x_i}, \tag{2.8}$$

subject to

$$\sum_{i=1}^{n-1} \frac{x_{i+1} - x_i}{y_{i+1} - y_i}(\exp(y_{i+1}) - \exp(y_i)) = 1, \tag{2.9}$$

details will get more clear in the following proofs. Here, (2.8) guarantees the concavity and (2.9) makes sure that the resulting estimate of the density integrates to 1. From (2.9) it also gets clear, why in this setting it is not necessary to bound the concave function ϕ from above: if for a $i \in \{1, \ldots, n\}$ we had $\phi(x_i) \to \infty$, by a similar argument as used in Rufibach (2006), due to (2.9) in case that $y_i - y_{i-1} > 0$ we had

$$\begin{aligned} 1 &\geq \frac{x_i - x_{i-1}}{y_i - y_{i-1}}(\exp(y_i) - \exp(y_{i-1})) = (x_i - x_{i-1})\exp(y_i)\frac{1 - \exp(y_{i-1} - y_i)}{y_i - y_{i-1}} \\ &\geq \frac{(x_i - x_{i-1})\exp(y_i)}{1 + y_i - y_{i-1}}, \end{aligned}$$

since for $x > 0$, $\frac{1-\exp(-x)}{x} \geq \frac{1}{1+x}$. Thus, $y_{i-1} \leq 1 + y_i - (x_i - x_{i-1})\exp(y_i) \to -\infty$. Analogously, if $y_i - y_{i-1} < 0$, using (2.8) we had

$$1 \geq \frac{x_{i+1} - x_i}{y_{i+1} - y_i}(\exp(y_{i+1}) - \exp(y_i)) \geq \frac{(x_{i+1} - x_i)\exp(y_i)}{1 + y_i - y_{i+1}},$$

which leads to $y_{i+1} \leq 1 + y_i - (x_{i+1} - x_i)\exp(y_i) \to -\infty$. In other words, if there were a x_i with $\phi(x_i) \to \infty$, in order to satisfy (2.9) would be required that $\phi(x_{i-1}) \to -\infty$ or $\phi(x_{i+1}) \to -\infty$ faster than $\phi(x_i) \to \infty$, which gives an important point in proving

existence of a maximum likelihood estimator.

Now, turning to the hidden Markov model, we explain why in our setting this argument does not apply and thus we need to assume A10 in order to bound the log-likelihood function of the model. Using Assumption A9 and the fact that $\gamma_{j,k} \leq 1$ for $j,k = 1,\dots,K$, we can bound the log-likelihood function of the hidden Markov model as follows:

$$\begin{aligned}
&\sum_{t=1}^{T} \log(\sum_{k=1}^{K} \exp(\phi_k(x_t))) = \log(\prod_{t=1}^{T}(\sum_{k=1}^{K} \exp(\phi_k(x_t))) \\
=& \log(\sum_{k_1=1}^{K} \dots \sum_{k_T=1}^{K} \prod_{t=1}^{T} \exp(\phi_{k_t}(x_t))) \\
\geq& \ell_T(\lambda) = \log(\sum_{k_1=1}^{K} \dots \sum_{k_T=1}^{K} \delta_{k_1} \prod_{t=2}^{T} \gamma_{k_{t-1},k_t} \prod_{s=1}^{T} \exp(\phi_{k_s}(x_s))) \\
\geq& \log(c^T \sum_{k_1=1}^{K} \dots \sum_{k_T=1}^{K} \prod_{t=1}^{T} \exp(\phi_{k_t}(x_t))) = \log(c^T) + \sum_{t=1}^{T} \log(\sum_{k=1}^{K} \exp(\phi_k(x_t))).
\end{aligned} \tag{2.10}$$

Assuming restriction (2.9) for each of the K state-dependent densities, which we will prove below, does not help in proving existence as it did in the independent identically distributed setting: If $\phi_k(x_t) \to \infty$ for some $k \in \{1,\dots,K\}$ and $t \in \{1,\dots,T\}$, the result from above, giving that $\phi_k(x_{t-1}) \to -\infty$ or $\phi_k(x_{t+1}) \to -\infty$ faster than $\phi_k(x_t) \to \infty$, does not help in bounding $\sum_{t=1}^{T} \log(\sum_{k=1}^{K} \exp(\phi_k(x_t)))$, since possibly for a $l \in \{1,\dots,K\} \setminus \{k\}$, $\phi_l(x_{t-1}) \to \infty$ (or $\phi_l(x_{t-1}) \to \infty$) faster than $\phi_k(x_{t-1}) \to -\infty$ (or $\phi_k(x_{t+1}) \to -\infty$). Thus, by A10 we assume $-\infty \leq \phi_k(x) \leq M$ $(k = 1,\dots,K)$, where the value $-\infty$ is permitted since we consider $\exp(\phi_k(x))$, which tends to 0 for $\phi_k(x) \to -\infty$. Keeping these remarks in mind, we now prove existence of the nonparametric univariate maximum likelihood estimator, using ideas from Pal et al. (2007).

Proof of Theorem 2.9. Since $\sup_\lambda \mathcal{L}_T(\lambda) = \sup_\Gamma \sup_\phi \mathcal{L}_T((\Gamma,\phi))$, let us fix Γ and focus on $\sup_\phi \mathcal{L}_T((\Gamma,\phi))$ for the moment.
We optimize ϕ over $\mathcal{F}_{lc}^b$, thus $\phi_k \mathbb{1}_{[x_1,x_T]}$ is concave for $k = 1,\dots,K$. If for a $k \in \{1,\dots,K\}$ and some $x \notin [x_1,x_T]$ we have $\exp(\phi_k(x)) > 0$, then, since $\exp(\phi_k)$ is a continuous function,

$$1 = \int_{-\infty}^{\infty} \exp(\phi_k(x))dx > \int_{x_1}^{x_T} \exp(\phi_k(x))dx =: d_k.$$

For $k = 1, \ldots, K$, let $\varphi_k(x) = \begin{cases} \phi_k(x) - \log(d_k), & x \in [x_1, x_T] \\ -\infty, & x \notin [x_1, x_T] \end{cases}$, which is a concave function satisfying $\int_{x_1}^{x_T} \exp(\varphi_k(x))dx = 1$. Then,

$$\begin{aligned} \ell_T((\Gamma, \phi)) &= \log(\sum_{k_1=1}^{K} \cdots \sum_{k_T=1}^{K} \delta_{k_1} \prod_{t=2}^{T} \gamma_{k_{t-1},k_t} \prod_{s=1}^{T} \exp(\phi_{k_s}(x_s))) \\ &= \log(\sum_{k_1=1}^{K} \cdots \sum_{k_T=1}^{K} \delta_{k_1} \prod_{t=2}^{T} \gamma_{k_{t-1},k_t} \prod_{s=1}^{T} d_{k_s} \exp(\varphi_{k_s}(x_s))) \\ &< \log(\sum_{k_1=1}^{K} \cdots \sum_{k_T=1}^{K} \delta_{k_1} \prod_{t=2}^{T} \gamma_{k_{t-1},k_t} \prod_{s=1}^{T} \exp(\varphi_{k_s}(x_s))) = \ell_T((\Gamma, \varphi)). \end{aligned} \tag{2.11}$$

Hence, for the maximizer $\hat{\phi}$ of ℓ_T, $supp(\exp(\hat{\phi}_k)) \subseteq [x_1, x_T]$, $k = 1, \ldots, K$.

By a similar argument, for each component ϕ_k of $\phi \in \mathcal{F}_{lc}^b$, there is a piecewise linear concave function φ_k with knots in $x_1, \ldots, x_T$, satisfying $\phi_k(x_t) = \varphi_k(x_t)$, $k = 1, \ldots, K$, $t = 1, \ldots, T$. Since ϕ_k is concave, $\phi_k(x) \geq \varphi_k(x)$ for $x \in [x_1, x_T]$. Thus,

$$1 = \int_{-\infty}^{\infty} \exp(\phi_k(x))dx \geq \int_{x_1}^{x_T} \exp(\phi_k(x))dx \geq \int_{x_1}^{x_T} \exp(\varphi_k(x))dx =: d_k$$

and there is a $\tilde{\varphi} \in \mathcal{F}_{lc}^b$, which is given by $\tilde{\varphi}_k(x) = \begin{cases} \varphi_k(x) - \log(d_k), & x \in [x_1, x_T] \\ -\infty, & x \notin [x_1, x_T] \end{cases}$ $(k = 1, \ldots, K)$, such that by the same argument as used above, $\ell_T(\phi) \leq \ell_T(\tilde{\varphi})$. This proves that the maximizer $\hat{\phi}$ of ℓ_T consists of K piecewise linear concave functions.

Now write $y_{k,t} := \phi_k(x_t)$ for piecewise linear concave functions ϕ_k, $t = 1, \ldots, T$, $k = 1, \ldots, K$. Since for $x \in [x_t, x_{t+1}]$, $\phi'(x) = \frac{y_{k,t+1} - y_{k,t}}{x_{t+1} - x_t}$, we have

$$\begin{aligned} \int_{x_t}^{x_{t+1}} \exp(\phi_k(x))dx &= \frac{x_{t+1} - x_t}{y_{k,t+1} - y_{k,t}} \int_{x_t}^{x_{t+1}} \exp(\phi_k(x)) \frac{y_{k,t+1} - y_{k,t}}{x_{t+1} - x_t} dx \\ &= \frac{x_{t+1} - x_t}{y_{k,t+1} - y_{k,t}} \int_{y_{k,t}}^{y_{k,t+1}} \exp(x)dx \\ &= \frac{x_{t+1} - x_t}{y_{k,t+1} - y_{k,t}} (\exp(y_{k,t+1}) - \exp(y_{k,t})). \end{aligned}$$

Thus, the condition $\int \exp(\phi_k(x))dx = 1$ can be rewritten as

$$\sum_{t=1}^{T-1} \frac{x_{t+1} - x_t}{y_{k,t+1} - y_{k,t}} (\exp(y_{k,t+1}) - \exp(y_{k,t})) = 1. \tag{2.12}$$

Instead of maximizing ℓ_T over $\phi \in \mathcal{F}_{lc}^b$ we can maximize ℓ_T over $y_{k,t} \in [-\infty, M]$ with respect to (2.12) and

$$\frac{y_{k,t} - y_{k,t-1}}{x_t - x_{t-1}} \geq \frac{y_{k,t+1} - y_{k,t}}{x_{t+1} - x_t},$$

$k = 1, \ldots, K$. Since the function $\sum_{k_1=1}^K \cdots \sum_{k_T=1}^K \delta_{k_1} \prod_{t=2}^T \gamma_{k_{t-1},k_t} \prod_{s=1}^T \exp(y_{k_s,s})$ is continuous on the compact set $[-\infty, M]^T \times \cdots \times [-\infty, M]^T$, the maximum likelihood estimator exists. □

The proof of Theorem 2.10 follows very similar arguments. In order to take account of the dimension $p > 1$ some concepts from convex analysis are required. These can be found in Rockafellar (1970) and were also used in Cule (2010), Cule et al. (2010) when proving existence of the nonparametric maximum likelihood estimator in the independent identically distributed setting.
As introduced above, $\mathcal{C}_T$ denotes the convex hull of the $x_1, \ldots, x_T$. Write $ri(\mathcal{C}_T)$ for the relative interior of $\mathcal{C}_T$ and $rb(\mathcal{C}_T)$ for the relative boundary of $\mathcal{C}_T$. The closure of a function f is denoted by $cl(f)$.
Recall that for any set A in $\mathbb{R}^p$ the closure $cl(A)$ is given by $cl(A) = \bigcap\{A + \varepsilon B \mid \varepsilon > 0\}$, where $B = \{x \mid \|x\| \leq 1\}$ is the Euclidean unit ball in $\mathbb{R}^p$. Then, the relative interior of a convex set A in $\mathbb{R}^p$ is defined as $ri(A) = \{x \in aff(A) \mid \exists \varepsilon > 0, (x + \varepsilon B) \cap (aff(A)) \subset A\}$, where $aff(A)$ denotes the affine hull of A. The relative boundary of A is then given by $(cl(A)) \setminus (ri(A))$ (see Rockafellar, 1970, pp. 43).
The closure $cl(f)$ of a concave function f is the pointwise infimum of all affine functions g such that $g \geq f$ (see Rockafellar, 1970, p. 307).

Proof of Theorem 2.10. First, fix Γ and maximize $\ell_T((\Gamma, \phi))$ over $\phi \in \mathcal{F}_{lc}^b$.
Since the functions ϕ_k are concave, $\phi_k \mathbb{1}_{\mathcal{C}_T}$ are concave functions for $k = 1, \ldots, K$. If for a $k \in \{1, \ldots, K\}$ and some $x \notin \mathcal{C}_T$ we have $\exp(\phi_k(x)) > 0$, due to continuity

$$1 = \int_{\mathbb{R}^p} \exp(\phi_k(x))dx > \int_{C_T} \exp(\phi_k(x))dx =: d_k.$$

Thus, setting $\tilde{\phi}_k = \begin{cases} \phi_k - \log(d_k), & x \in \mathcal{C}_T \\ -\infty, & x \notin \mathcal{C}_T \end{cases}$ $(k = 1, \ldots, K)$ yields $\ell_T((\Gamma, \phi)) < \ell_T((\Gamma, \tilde{\phi}))$, see (2.11). Therefore, the maximizer $\hat{\phi}$ of ℓ_T satisfies $supp(\exp(\hat{\phi}_k)) \subseteq \mathcal{C}_T$, $k = 1, \ldots, K$.

By Rockafellar (1970, p. 37), for any $y_k \in \mathbb{R}^T$ there exists a function $\varphi_{y_k} : \mathbb{R}^p \to \mathbb{R}$, which is the least concave function satisfying $\varphi_{y_k}(x_t) \geq y_{k,t}$, $t = 1, \ldots, T$.
To prove that for $k = 1, \ldots, K$, $\hat{\phi}_k \in \mathcal{F}_{tent}(x)$, assume that $\hat{\phi}_k(x_t) = y_{k,t}$ $(t = 1, \ldots, T)$ and $\hat{\phi}_k \neq \varphi_{y_k}$.

Since $\hat{\phi}_k$ is concave, $\hat{\phi}_k(\bar{x}) \geq \varphi_{y_k}(\bar{x})$ for $\bar{x} \in \mathbb{R}^p$ and hence there exists a $x_0 \in \mathcal{C}_T$ for which $\hat{\phi}_k(x_0) > \varphi_{y_k}(x_0)$.
Assume $x_0 \in ri(\mathcal{C}_T)$: Since $\hat{\phi}_k$ and φ_{y_k} are continuous in x_0,

$$1 = \int_{\mathbb{R}^d} \exp(\hat{\phi}_k(x))dx \geq \int_{\mathcal{C}_T} \exp(\hat{\phi}_k(x))dx > \int_{\mathcal{C}_T} \exp(\varphi_{y_k}(x))dx =: d_k$$

Setting $\tilde{\varphi}_{y_k} = \begin{cases} \varphi_{y_k} - \log(d_k), & x \in \mathcal{C}_T \\ -\infty, & x \notin \mathcal{C}_T \end{cases}$, which is a concave function and where the function $\exp(\tilde{\varphi}_{y_k}(x))$ is a density, by the same argument as above, $\ell_T(\hat{\phi}) \leq \ell_T(\tilde{\varphi}_y)$, which is a contradiction since $\hat{\phi}$ maximizes the log-likelihood function. Therefore, $x_0 \notin ri(\mathcal{C}_T)$ and the only remaining possibility is $x_0 \in rb(\mathcal{C}_T)$. Since φ_{y_k} is upper semi-continuous (see Rockafellar, 1970, Corollary 17.2.1), by Rockafellar (1970, Corollary 7.3.4) we have $\varphi_{y_k} = cl(\varphi_{y_k}) = cl(\hat{\phi}_k) \geq \hat{\phi}_k$, which together with the fact $\hat{\phi}_k(x) \geq \varphi_{y_k}(x)$ for $x \in \mathcal{C}_T$ yields $\hat{\phi}_k = \varphi_{y_k}$.

These considerations prove that maximizing $\ell_T(\lambda)$ over $\mathcal{G}^b \times \mathcal{F}_{lc}^b$ can be reformulated to maximization of $\ell_T(\lambda)$ over $y_{k,t} \in [-\infty, M]$, such that the resulting $\varphi_{y_k} \in \mathcal{F}_{tent}(x)$, $k = 1, \ldots, K$, $t = 1, \ldots, T$.
Due to the continuity of the function $\sum_{k_1=1}^K \cdots \sum_{k_T=1}^K \delta_{k_1} \prod_{t=2}^T \gamma_{k_{t-1},k_t} \prod_{s=1}^T \exp(y_{k_s,s})$ on the compact set $[-\infty, M]^T \times \ldots \times [-\infty, M]^T$, the maximum likelihood estimator exists. □

3. Penalized estimation for hidden Markov models

In recent years, due to the growing availability of high-dimensional data in areas like bioinformatics, climatology, finance or information technology, the introduction of restrictions and sparsity assumptions to statistical models have become quite popular, see for example Bühlmann and van de Geer (2011) for a comprehensive introduction. Tibshirani (1996) introduced the least absolute shrinkage and selection operator (Lasso) in the context of least squares estimation for linear models, which became a common method for variable selection and prediction. Fan and Li (2001) considered alternative penalty functions and its properties in the context of likelihood-based regression models. Penalized estimation methods are also used in the context of sparse covariance matrices and their inverse (called precision matrices), see for example Cai et al. (2011) and Huang et al. (2006). These matrices are relevant for applications in areas like risk management, fMRI, portfolio management and algorithms for web searches. Yuan and Lin (2007) applied penalized likelihood methods for estimating precision matrices in Gaussian graphical models with focus on zero entries indicating conditional independence. They applied two penalties based on the Lasso approach and develop asymptotic properties of their estimators using asymptotic theory for penalized regression estimators from Knight and Fu (2000). Ruan et al. (2011) adopted this idea and applied a Lasso-type penalty to precision matrices of Gaussian mixture models. They gave an EM-algorithm for parameter estimation and numerical results including real data examples for model-based clustering and mixture discriminant analysis. Lotsi and Wit (2013) proved that imposing a Lasso-type penalty on the likelihood function of a Gaussian mixture model does not alter consistency results in the context of mixture models.

In this chapter we introduce the penalty functions used in Fan and Li (2001) to parametric hidden Markov models with state-dependent multivariate Gaussian distributions, in order to estimate sparse state-dependent precision matrices. In Section 3.2 we show that the asymptotic results from Fan and Li (2001) hold true for maximum likelihood estimation in hidden Markov models, while in Section 3.3 we compare three penalty functions for the estimation of sparse precision matrices concerning several criteria and apply our methods to multivariate financial time series. Using the estimated precision

matrices we try to detect conditional independence of financial positions in different volatility states of the market, which are represented by the states of the hidden Markov model.

3.1. Penalized maximum likelihood estimation

We consider parametric hidden Markov models as introduced in Section 1.2. The underlying Markov chain of the model is assumed to be stationary with stationary distribution $\delta = (\delta_1, \ldots, \delta_K)$. We assume the observed process to take values in $\mathbb{R}^p$ and focus on Gaussian hidden Markov models, i.e. $F_k = \mathcal{N}(\mu_k, \Sigma_k)$, where $\mu \in \mathbb{R}^p$ and $\Sigma_k \in \mathbb{R}^{p\times p}$ positive semidefinite, $k = 1, \ldots, K$. Thus, the parameter vector of the model $\lambda = (\Gamma, \mu_1, \Sigma_1, \ldots, \mu_K, \Sigma_K)$ is of dimension $G := K^2 - K + 1.5Kp + 0.5Kp^2$, since there are $K(K-1)$ transition probabilities and $(p + \frac{p(p+1)}{2})K$ parameters of the state-dependent Gaussian distributions. We write f_k for the density of the Gaussian distribution in state k with respect to a σ-finite measure ν, $k = 1, \ldots, K$. The true underlying parameter of our model is denoted by $\lambda_0 = (\Gamma_0, \mu_{1,0}, \Sigma_{1,0}, \ldots, \mu_{K,0}, \Sigma_{K,0})$ and $f_{k,0}$ indicates that the parameters for f_k are $\mu_{k,0}$ and $\Sigma_{k,0}$ (or $\Omega_{k,0}$), $k = 1, \ldots, K$. We denote the state-dependent precision matrices by $\Omega_k = \Sigma_k^{-1}$, $\Omega_k = (\omega_{ij}^{(k)})_{i,j=1,\ldots,p}$, $k = 1, \ldots, K$. Since our focus is on sparse precision matrices, we introduce more notation. For $k = 1, \ldots, K$ let

$$\Omega_{k,0}^{(1)} = \{\omega_{ij,0}^{(k)} \neq 0,\ i,j = 1,\ldots,p\} \text{ and } \Omega_{k,0}^{(2)} = \{\omega_{ij,0}^{(k)} = 0,\ i,j = 1,\ldots,p\}$$

denote the sets of nonzero and zero entries of the precision matrices respectively and write

$$I_{k,0}^{(1)} = \{(i,j) :\ i,j = 1,\ldots,p,\ i \neq j,\ \omega_{ij,0}^{(k)} \neq 0\}$$

and

$$I_{k,0}^{(2)} = \{(i,j) :\ i,j = 1,\ldots,p,\ i \neq j,\ \omega_{ij,0}^{(k)} = 0\}$$

for the corresponding index sets. Set $\lambda_0 = (\lambda_{10}, \lambda_{20})$, where

$$\lambda_{10} := (\lambda_{1,0}, \ldots, \lambda_{H,0}) = (\Gamma_0, \mu_{1,0}, \ldots, \mu_{K,0}, \Omega_{1,0}^{(1)}, \ldots, \Omega_{K,0}^{(1)})$$

contains the transition probabilities, the state-dependent expected values and the non-zero entries of the precision matrices and

$$\lambda_{20} := (\lambda_{H+1,0}, \ldots, \lambda_{G,0}) = (\Omega_{1,0}^{(2)}, \ldots, \Omega_{K,0}^{(2)})$$

contains the zero entries of the precision matrices.
The objective of this chapter is the estimation of λ using the maximum likelihood method. In order to take into account the sparse structure of the precision matrices, we impose a penalty to the likelihood function (1.3) of the model:

$$\mathcal{L}_T(\lambda) - T\rho(\lambda; c_T),$$

where $\rho : \Lambda \to \mathbb{R}$ is a penalty function depending on a tuning parameter $c_T > 0$. Since in our model it is not necessary to penalize the entire parameter vector λ, we restrict penalization to the state-dependent precision matrices and write

$$\mathcal{Q}_T(\lambda) = \mathcal{L}_T(\lambda) - T\sum_{k=1}^{K}\sum_{i\neq j}\rho(|\omega_{ij}^{(k)}|; c_T)$$

for the penalized likelihood function.

Popular choices for the penalty functions on a parameter θ are the l_1-penalty

$$\rho(\theta; c_T) = c_T\|\theta\|_{l_1}$$

or hard thresholding

$$\rho(|\theta|; c_T) = c_T^2 - (|\theta| - c_T)^2\mathbb{1}_{\{|\theta|<c_T\}}, \tag{3.1}$$

where for a matrix $A = (a_{ij})_{i,j=1,\ldots,p}$, $\|A\|_{l_1} = \sum_{i\neq j}|a_{ij}|$ and for a vector $a \in \mathbb{R}^p$, $\|a\|_{l_1} = \sum_{j=1}^{p}|a_{ij}|$. Fan and Li (2001) introduced the so called smoothly clipped absolute

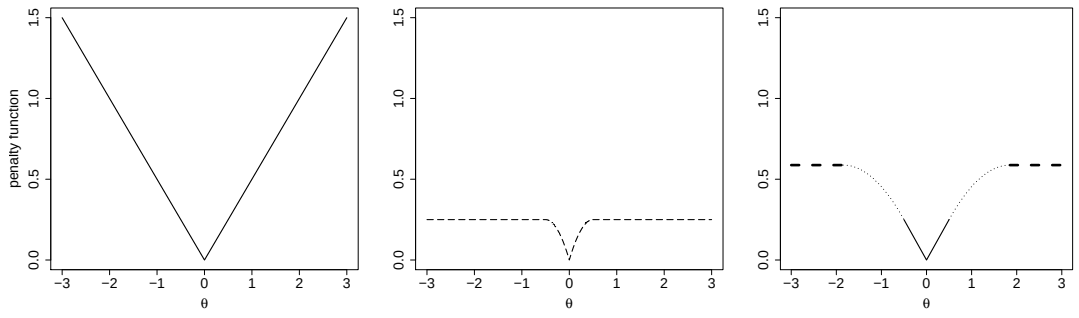


FIGURE 3.1.: Penalty functions for $c_T = 0.5$ and $c^* = 3.7$: Left: l_1, mid: hard thresholding, right: SCAD. Solid line: l_1-penalty, dashed line: hard thresholding, dotted line: quadratic spline.

deviation penalty (SCAD),

$$\rho(\theta; c_T) = \begin{cases} c_T|\theta| & |\theta| \leq c_T \\ -(\frac{|\theta|^2 - 2c^* c_T|\theta| + c_T^2}{2(c^*-1)}) & c_T < |\theta| \leq c^* c_T \\ \frac{(c^*+1)c_T^2}{2} & |\theta| > c^* c_T \end{cases},$$

for an additional tuning parameter $c^* > 2$, which is a combination of the l_1-penalty and hard thresholding. In a neighbourhood $|\theta| < c_T$, l_1-penalization is performed, while for $|\theta| > c^* c_T$ the SCAD-penalty uses hard thresholding. For values in between a quadratic spline with knots c_T and $c^* c_T$ determines the SCAD-penalty function. For $\theta > 0$ the first derivative is

$$\rho'(\theta; c_T) = c_T(\mathbb{1}_{\{\theta \leq c_T\}} + \frac{\max\{c^* c_T - \theta, 0\}}{(c^*-1)c_T}\mathbb{1}_{\{\theta > c_T\}}). \tag{3.2}$$

Figure 3.1 illustrates the relationship of the three penalty functions.

3.2. Asymptotic theory for the penalized estimator

Fan and Li (2001) developed asymptotic theory for penalized maximum likelihood estimators in generalized linear models. They established convergence rates of the penalized maximum likelihood estimator, discussed consistency depending on the chosen penalty function and proved asymptotic normality and an oracle property for consistent estimators. We transfer their theory to our model using the consistency result from Leroux (1992a) and the results on asymptotic normality from Bickel et al. (1998) for unpenalized maximum likelihood estimation in parametric hidden Markov models. For this purpose, we impose the following assumptions.

A11. *The Markov chain $(S_t)_{t\in\mathbb{N}}$ is irreducible and aperiodic.*

A12. *The map $\lambda \mapsto \gamma_{k,l}$ is continuous on Λ and in some neighbourhood $\|\lambda - \lambda_0\| < \varepsilon$, the maps $\lambda \mapsto \gamma_{k,l}$ and $\lambda \mapsto \delta_k$ have two continuous derivatives for all $k, l = 1, \ldots, K$.*

A13. *For all $k = 1, \ldots, K$ and $x \in \mathbb{R}^p$ the map $\lambda \mapsto f_k(x)$ is continuous on the parameter space Λ and has two continuous derivatives in the same neighbourhood as used in Assumption A12.*

A14. *We have $E_{\lambda_0}(|\log(f_{k,0}(x_1))|) < \infty$ $(k = 1, \ldots, K)$ and there exists an $\varepsilon > 0$ such that for $\lambda \in \mathbb{R}^G$ we have*

1. *$E_{\lambda_0}(\sup_{\|\lambda-\lambda_0\|<\varepsilon} |\frac{\partial}{\partial\lambda_i}\log(f_k(x_1))|^2) < \infty$ for $i = 1, \ldots, G$ and $k = 1, \ldots, K$,*
2. *$E_{\lambda_0}(\sup_{\|\lambda-\lambda_0\|<\varepsilon} |\frac{\partial^2}{\partial\lambda_i\partial\lambda_j}\log(f_k(x_1))|) < \infty$ for $i, j = 1, \ldots, G$, $k = 1, \ldots, K$ and*

3. for $j = 1, 2$, $i_l = 1, \dots, G$ *and* $l = 1, \dots, j$, $\int \sup_{\|\lambda - \lambda_0\| < \varepsilon} |\frac{\partial^j}{\partial \lambda_{i_1} \cdots \partial \lambda_{i_j}} f_k(y)| \nu(dy) < \infty$ *for* $k = 1, \dots, K$.

A15. *For some* $\varepsilon > 0$ *and all* $\lambda \in \Lambda$ *we have* $E_{\lambda_0}(\sup_{\|\lambda - \lambda_0\| < \varepsilon}(\log(f_k(x_1))^+)) < \infty$, $k = 1, \dots, K$.

A16. *There is an* $\varepsilon > 0$ *such that for* $k = 1, \dots, K$,

$$P_{k,0}(\sup_{\|\lambda - \lambda_0\| < \varepsilon} \max_{i,j=1,\dots,K} \frac{f_{i,0}(x_1)}{f_j(x_1)} = \infty) < 1.$$

A17. *The parameter space* Λ *is compact and* λ_0 *is an interior point of* Λ.

A18. *The family of mixtures of at most* K *elements of* $\{f(x; \vartheta) : \vartheta \in \Theta\}$ *is identifiable.*

The consistency proof of Leroux (1992a), which is based on the method of Wald (1949) uses Assumptions A11, A15, A18, as well as the first parts of Assumptions A12 and A13 concerning continuity, the compactness assumption from A17 and the first part of A14: $E_{\lambda_0}(|\log(f_{k,0}(x_1))|) < \infty$ for $k = 1, \dots, K$. Note that consistency of the unpenalized maximum likelihood estimator in Gaussian hidden Markov models requires the parameter space to be compactified and in particular the variances need to be bounded away from zero to prevent unboundedness of the likelihood function, see Alexandrovich (2014) for details.
The proof of asymptotic normality in Bickel et al. (1998) is based on a Taylor expansion of the first derivative of $\mathcal{L}_T(\lambda)$ at λ_0 and requires the consistency of the maximum likelihood estimator next to Assumptions A11, A12, A13, A14 1.–3., A16 and the interior point assumption of A17.

We state our first theorem, which follows Fan and Li (2001, Theorem 1).

Theorem 3.1. *Assume that in the hidden Markov model described above Assumptions A11–A18 hold and that the second derivative of the penalty function satisfies*

$$\max\{|\rho''(|\lambda_j|; c_T)| : \lambda_j \in \lambda_{10}\} \to 0.$$

Then there exists a local maximizer $\hat{\lambda}_T$ *of the penalized likelihood function* $\mathcal{Q}_T(\lambda)$ *such that*

$$\|\hat{\lambda}_T - \lambda_0\| = O_P(T^{-\frac{1}{2}} + a_T),$$

where $a_T = \max\{\rho'(|\lambda_j|; c_T) : \lambda_j \in \lambda_{10}\}$.

When choosing SCAD-penalty function or hard thresholding, from (3.1) and (3.2) we observe that if $c_T \to 0$, there exists a root-T-consistent penalized maximum likelihood estimator, since $a_T \to 0$ in Theorem 3.1. For the l_1- penalty function $a_T = c_T$, thus the

penalized maximum likelihood estimator is root-T-consistent, if $c_T = O_P(T^{-\frac{1}{2}})$.

For the next theorem, let

$$A := \text{diag}(\rho''(|\lambda_{1,0}|; c_T), \ldots, \rho''(|\lambda_{H,0}|; c_T)),$$
$$b := (\rho'(|\lambda_{1,0}|; c_T)sgn(\lambda_{1,0}), \ldots, \rho'(|\lambda_{H,0}|; c_T)sgn(\lambda_{H,0}))^\mathsf{T}$$

and denote by $I(\lambda_0)$ the Fisher information matrix and by $I^*(\lambda_{10}) = I^*((\lambda_{10}, 0))$ the Fisher information matrix when $\lambda_{20} = 0$ is known. The following theorem, which follows Fan and Li (2001, Lemma 1, Theorem 2), states that under certain conditions on the penalty function, the root-T-consistent estimator has oracle properties and is asymptotic normal.

Theorem 3.2. *Assume that in the hidden Markov model described above Assumptions A11–A18 hold and that* $\liminf_{T\to\infty} \liminf_{\theta\to 0+} \frac{\rho'(\theta;c_T)}{c_T} > 0$. *If* $c_T \to 0$ *and* $\sqrt{T}c_T \to \infty$ *as* $T \to \infty$, *the* $\sqrt{T}$*-consistent local maximizer* $\hat{\lambda}_T = (\hat{\lambda}_{T,1}, \hat{\lambda}_{T,2})^\mathsf{T}$ *in Theorem 3.1 has the following properties*

1. $\hat{\lambda}_{T,2} = 0$ *(sparsity)*
2. $\sqrt{T}(I^*(\lambda_{10}) + A)(\hat{\lambda}_{T,1} - \lambda_{10} + (I^*(\lambda_{10}) + A)^{-1}b) \to \mathcal{N}(0, I^*(\lambda_{10}))$ *in distribution (asymptotic normality)*

with probability tending to 1.

In combination with the results from Theorem 3.1 we observe that given $\sqrt{T}c_T \to \infty$, the penalized maximum likelihood estimator using hard thresholding or the SCAD-penalty works as well as the maximum likelihood estimates when estimating λ_{10} and $\lambda_{20} = 0$ is known. In addition, the estimator is asymptotically normal. Since the penalized maximum likelihood estimator using l_1-penalization is $\sqrt{T}$-consistent if $c_T = O_P(T^{-\frac{1}{2}})$, there is a contradiction to the assumption $\sqrt{T}c_T \to \infty$, thus in this case the results from Theorem 3.2 do not apply.

3.3. Numerical illustrations

In this section we investigate finite sample performance of the penalized maximum likelihood estimator described in Section 3.1. In the context of l_1-penalization, penalized estimation of Gaussian hidden Markov models was considered by Städler and Mukherjee (2013). Their approach combined the estimation of sparse precision matrices and the model selection problem of choosing the number of states K. They proposed an EM-algorithm for the model and provided a simulation study next to a real data example

from genome biology.

We investigate penalized maximum likelihood estimation in the described model, comparing the performance of the three penalty functions introduced above. For estimation, we implement the EM-algorithm described in Section 1.3 with modified M-step. Instead of the closed-form estimation of the covariance matrices according to (1.16), we perform numerical optimization of the corresponding part of the conditional expectation of the penalized complete-data log-likelihood function.
An additional problem is the selection of the tuning parameters. We stick to the literature and use a penalized version of the Bayesian Information Criterion (BIC), (see Fan and Li, 2001, Ruan et al., 2011, Städler and Mukherjee, 2013, Yuan and Lin, 2007). Bühlmann and van de Geer (2011) proposed the BIC criterion for selection of the tuning parameters as alternative to cross-validation, but they mentioned that it "has no theoretical justification for variable selection with the Lasso"[1].
Since we gain better results allowing for state-dependent tuning parameters, we choose the vector $c_T = (c_{T,1}, \ldots, c_{T,K})$ giving the minimal penalized BIC

$$\begin{aligned} BIC(c_T) = &-2\mathcal{L}_T(\hat{\lambda}_T) + T\rho(\hat{\lambda}_T; c_T) + \log(T) \sum_{k=1}^{K} (p + \sum_{i<j} \mathbb{1}_{\{|\hat{\omega}_{ij}^{(k)}|/(\max|\hat{\omega}_{ij}^{(k)}|)>0.0001\}}) \\ &+ \log(T)K(K-1). \end{aligned}$$

As proposed by Fan and Li (2001) we choose $c^* = 3.7$ for the second tuning parameter of the SCAD-penalty.

3.3.1. Simulations

We start our numerical investigation using simulated data of hidden Markov models from several dimensions as described in Section 3.1, in order to compare the performance of the proposed penalty functions. For all models we fix $K = 2$ and the transition probability matrix is set to $\Gamma = \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix}$. The state-dependent expected values are chosen as $\mu_{1,0} = (0, \ldots, 0)^\intercal \in \mathbb{R}^p$ and $\mu_{2,0} = (0.5, \ldots, 0.5)^\intercal \in \mathbb{R}^p$ for the respective dimension p and we simulate data sets consisting of $T = 3371$ observations.
For each p we run the simulation 500 times and report the averaged spectral norm, Frobenius norm and Kullback–Leibler divergence of the precision matrices over all states according to

$$\frac{1}{K}\sum_{k=1}^{K} \|\hat{\Omega}_k - \Omega_{k,0}\|_S, \qquad \frac{1}{K}\sum_{k=1}^{K} \|\hat{\Omega}_k - \Omega_{k,0}\|_F$$

[1] Bühlmann and van de Geer (2011, p.17)

and

$$\frac{1}{K}\sum_{k=1}^{K}\operatorname{tr}(\Sigma_{k,0}\hat{\Omega}_k) - \log(\det(\Sigma_{k,0}\hat{\Omega}_k)) - p$$

respectively. In addition we transfer the concept of the l_0-'norm' for vectors ($\|x\|_0 = \sum_{i=1}^{p} \mathbb{1}_{\{x_i \neq 0\}}$ for $x \in \mathbb{R}^p$) to matrices,

$$\frac{1}{K}\sum_{k=1}^{K}\|\hat{\Omega}_k\|_0,$$

where we denote by $\|A\|_0 = \sum_{i,j} \mathbb{1}_{\{a_{ij}\neq 0\}}$ for $A = (a_{ij})$. Furthermore, we report the average number of correct and incorrect estimated zero entries.

1. Dimension $p = 2$. The precision matrices of the model are fixed as

$$\Omega_{1,0} = \begin{pmatrix} 600 & 0 \\ 0 & 750 \end{pmatrix}, \qquad \Omega_{2,0} = \begin{pmatrix} 1000 & 0 \\ 0 & 850 \end{pmatrix}.$$

The state-dependent tuning parameters of the penalty functions vary in the ranges reported in Table 3.1. The estimation results averaged over 500 simulations together with the standard errors are reported in Table 3.2. Regarding most of the criteria l_1-penalization yields the best results, only when considering the spectral norm, SCAD-penalization is slightly better. Compared to the first two penalty functions, hard thresholding yields poor results. Together with the observed standard errors we observe that the differences between the estimates are small. To gain further insights we increase the considered dimension.

	l_1	SCAD	hard
$c_{T,1}$	0 – 0.00002	0 – 0.35	0 – 0.9
$c_{T,2}$	0 – 0.0001	0 – 0.2	0 – 0.9

TABLE 3.1.: Ranges for tuning parameters of the penalty functions for $p = 2$

	l_1	SCAD	hard
Frobenius norm	**38.74** (15.91)	42.28 (17.41)	45.69 (16.48)
Spectral norm	212.68 (22.26)	**210.45** (24.36)	215.96 (23.85)
Kullback–Leibler divergence	**0.00148** (0.001)	0.00174 (0.001)	0.002 (0.001)
l_0 (target: 2)	**2.09** (0.29)	2.23 (0.51)	2.69 (0.59)
Zeroes correct (target: 4)	**3.81** (0.58)	3.54 (1.01)	2.62 (1.18)
Zeroes incorrect	0 (0)	0 (0)	0 (0)

TABLE 3.2.: Averaged estimation results and standard errors of 500 simulations, $p = 2$.

2. Dimension $p = 4$. The precision matrices of the model are

$$\Omega_{1,0} = \begin{pmatrix} 650 & 0 & -10 & 0 \\ 0 & 850 & 100 & 0 \\ -10 & 100 & 700 & 120 \\ 0 & 0 & 120 & 800 \end{pmatrix}, \qquad \Omega_{2,0} = \begin{pmatrix} 950 & 20 & 0 & 0 \\ 20 & 750 & 0 & 50 \\ 0 & 0 & 1000 & -20 \\ 0 & 50 & -20 & 800 \end{pmatrix},$$

thus $\|\Omega_{1,0}\|_0 = \|\Omega_{2,0}\|_0 = 10$ and there is a total number of 12 zeroes to be estimated. The selected ranges of tuning parameters are reported in Table 3.3.

Again, the averaged results reported in Table 3.4 show that, regarding most

	l_1	SCAD	hard
$c_{T,1}$	0 – 0.000015	0 – 0.29	0 – 0.8
$c_{T,2}$	0 – 0.00006	0 – 0.2	0 – 0.7

TABLE 3.3.: Ranges for tuning parameters of the penalty functions for $p = 4$

criteria, l_1-penalization seems to be the best choice. Taking a closer look and considering the standard errors we observe that when comparing l_1- and SCAD-penalization, the results are very close. Concerning the l_0-'norm', the result from SCAD-penalization gets closer to the correct number of 10 than l_1-penalization or hard thresholding. Observing the number of correct and incorrect estimated zeroes yields that l_1-penalization gets very close to the true number of 12 correct zeros, but on the contrary estimates a lot of incorrect zero entries as well. Hard thresholding yields the least number of incorrect estimated zeroes, but this is caused by the fact that it overall does not detect many zeroes. SCAD-penalization seems to be a compromise, on the one hand it does not estimate all correct zero entries but on the other hand, not many entries are mistakenly shrunken to zero.

	l_1	SCAD	hard
Frobenius norm	**88.05** (13.72)	89.59 (23.31)	89.98 (21.41)
Spectral norm	**211.72** (19.55)	225.23 (24.53)	225.51 (22.65)
Kullback–Leibler divergence	**0.0065** (0.002)	0.0066 (0.003)	0.0065 (0.003)
l_0 (target: 10)	6.74 (0.85)	**12.58** (1.18)	13.608 (1.05)
Zeroes correct (target: 12)	**11.22** (1.19)	4.96 (2.37)	3.35 (2.01)
Zeroes incorrect	7.3 (1.09)	1.88 (1.51)	**1.43** (1.36)

TABLE 3.4.: Averaged estimation results and standard errors 500 simulations, $p = 4$.

3. Dimension $p = 8$. The precision matrices of the model are set to

$$\Omega_{1,0} = \begin{pmatrix} 650 & 0 & -10 & 0 & 15 & 0 & 0 & 75 \\ 0 & 800 & 100 & 0 & 0 & 50 & -15 & 0 \\ -10 & 100 & 750 & 120 & 0 & 0 & 0 & 15 \\ 0 & 0 & 120 & 820 & -5 & -20 & 5 & -10 \\ 15 & 0 & 0 & -5 & 700 & 0 & -5 & 0 \\ 0 & 50 & 0 & -20 & 0 & 600 & 0 & -25 \\ 0 & -15 & 0 & 5 & -5 & 0 & 650 & 40 \\ 75 & 0 & 15 & -10 & 0 & -25 & 40 & 700 \end{pmatrix},$$

$$\Omega_{2,0} = \begin{pmatrix} 950 & 20 & 0 & 0 & 10 & -25 & 0 & 100 \\ 20 & 750 & 0 & 50 & 0 & 0 & 10 & -50 \\ 0 & 0 & 1000 & -20 & 0 & 10 & 0 & 0 \\ 0 & 50 & -20 & 800 & -5 & 0 & 0 & 0 \\ 10 & 0 & 0 & -5 & 900 & -15 & -150 & 5 \\ -25 & 0 & 10 & 0 & -15 & 850 & 0 & 10 \\ 0 & 10 & 0 & 0 & -150 & 0 & 950 & 0 \\ 100 & -50 & 0 & 0 & 5 & 10 & 0 & 1050 \end{pmatrix},$$

which leads to $\|\Omega_{1,0}\|_0 = 38$ and $\|\Omega_{2,0}\|_0 = 36$, so that a total number of 54 zeroes is to be estimated. The tuning parameters are selected from the values reported in Table 3.5. The averaged results provided in Table 3.6 show that for growing dimension p, SCAD-penalization improves compared to l_1-penalization and hard thresholding. While hard thresholding seems not to be a good choice for detecting zero entries, l_1-penalization mistakenly shrinks many entries to zero. SCAD-penalization performs quite good, the estimate gets close to the number of correct zero entries and does not yield too many incorrect zeroes, while it leads to reasonable values concerning the norm-criteria.

One conclusion from the simulation study is that despite its theoretical properties proven in Section 3.2, the finite sample performance of the maximum likelihood estimator penalized by hard thresholding is not competitive to l_1- and SCAD-penalization and does not lead to the desired sparsity. While in small dimensions l_1-penalization yields good results, which are slightly better than using SCAD-penalization, we observe that for growing dimension SCAD-penalization turns out to be more advantageous.

	l_1	SCAD	hard
$c_{T,1}$	0 – 0.00003	0 – 0.5	0 – 0.7
$c_{T,2}$	0 – 0.00002	0 – 0.8	0 – 0.5

TABLE 3.5.: Ranges for tuning parameters of the penalty functions for $p = 8$.

	l_1	SCAD	hard
Frobenius norm	182.97 (23.28)	**83.94** (0.86)	163.34 (35.72)
Spectral norm	249.53 (18.47)	**242.4** (0.61)	277.58 (24.95)
Kullback–Leibler divergence	0.0256 (0.006)	**0.0054** ($9.7e^{-5}$)	0.0207 (0.008)
l_0 (target: 37)	16.15 (2.25)	**36.77** (0.51)	59.27 (1.99)
Zeroes correct (target: 54)	**53.13** (1.42)	47.41 (1.14)	5.94 (3.77)
Zeroes incorrect	42.57 (3.83)	7.04 (1.13)	**3.53** (2.18)

TABLE 3.6.: Averaged estimation results and standard errors of 500 simulations, $p = 8$.

However, one should be aware of the advantages and disadvantages of both approaches: While l_1-penalization has a very strong shrinkage-effect which partly leads to high error rates, SCAD-penalization might miss some zero entries when estimating the parameters.

3.3.2. Real data example

Now we apply the proposed method to financial data[2]. We use share prices of German stocks on a daily basis from 14th January 2001 to 23rd April 2013, which yields a data set of $T = 3371$ log-returns. Since in our simulation study hard thresholding did not prove beneficial compared to the other penalty functions, we only report results for l_1- and SCAD-penalization.

The main objective is the investigation of conditional independence in different volatility states of the market, which are represented by the states of the Markov chain. The advantage of Gaussian hidden Markov models is the information on conditional independence provided by the precision matrices, which is a popular concept in the context of graphical models (see for example Lauritzen (1996) for detailed information). A zero entry $\omega_{ij}^{(k)} = 0$ indicates that in state k the random variables i and j are conditional independent given all remaining variables of the model ($k = 1, \ldots, K$, $i, j = 1, \ldots, p$).

We choose the number of states of the Markov chain using the BIC criterion, which is for both penalty functions minimal for $K = 3$. Thus, we fit a hidden Markov model with three states, where state 1 represents the state with smallest volatility, while state 3 represents the state with highest volatility.

Below we provide results for several portfolios of different dimensions, consisting of shares from different industry sectors. Surprisingly the precision matrices are not very sparse, thus we cannot make a clear point on conditional independence among the shares.

At first, we consider a portfolio consisting of four shares from the financial sector, namely Commerzbank, Deutsche Bank, Baader Bank and Deutsche Balaton. Estimation yields

[2]Available from `http://de.finance.yahoo.com/`, data access 23rd April 2013

only one zero entry in the highest volatility state, corresponding to conditional independence of Baader Bank and Deutsche Balaton given the observations of Commerzbank and Deutsche Bank. This result is given by the unpenalized estimator as well as when using l_1- or SCAD-penalization, thus in this example penalization does not yield additional insights.

Next we construct a four-dimensional portfolio consisting of shares from Merck, Bayer, Curasan and Evotec from the biotechnology sector. Unpenalized estimation ($\hat{\Omega}_k$) yields no zero entries in the precision matrices, while l_1- and SCAD-penalization ($\hat{\Omega}_k^{l_1}$ and $\hat{\Omega}_k^S$ respectively) suggest several zero entries in the highest volatility state. As previously observed in the simulation study, l_1-penalization yields more zero entries than SCAD-penalization:

$$\hat{\Omega}_3 = \begin{pmatrix} 346 & -46 & -20 & -10 \\ -46 & 101 & -4 & -1 \\ -20 & -4 & 21 & -1 \\ -10 & -1 & -1 & 36 \end{pmatrix}, \qquad \hat{\Omega}_3^{l_1} = \begin{pmatrix} 292 & 0 & -12 & 0 \\ 0 & 80 & -6 & 0 \\ -12 & -6 & 17 & 0 \\ 0 & 0 & 0 & 33 \end{pmatrix},$$

$$\hat{\Omega}_3^S = \begin{pmatrix} 368 & -54 & -27 & -4 \\ -54 & 107 & 0 & 0 \\ -27 & 0 & 23 & 0 \\ -4 & 0 & 0 & 39 \end{pmatrix}.$$

TABLE 3.7.: Estimated precision matrices state 3, $p = 4$ biotechnology portfolio: Unpenalized estimator, l_1-penalization, SCAD-penalization

Both penalized estimates suggest conditional independence of Evotec and Bayer and Evotec and Curasan given the respective remaining portfolios. In addition, l_1-penalization yields conditional independence of Merck and Bayer as well as Merck and Evotec given the respective remaining portfolios, while SCAD-penalization yields a zero entry concerning the shares from Bayer and Curasan. Compared to the unpenalized estimator, the SCAD-penalized estimator seems more plausible than the l_1-penalized estimator, since it shrinks those entries to zero, which posses the smallest absolute values in the unpenalized estimate.
The remaining estimates are given in Appendix A. Concerning expected values and transition probabilities, the three procedures, especially the two penalized estimators, yield very similar results.

Now, we extend the portfolio in two different ways. First, we add two shares of a related industry sector to see if the observed results remain. Second, we merge the two portfolios from dimension four to study conditional independence of shares from different industry sectors.

Adding the shares of BASF and K+S to the biotechnology portfolio, unpenalized estimation yields a zero entry which was not detected before and corresponds to the shares from Bayer and Evotec in the highest volatility state. Penalization by l_1 yields one zero entry in the lowest volatility state, which suggests conditional independence of Curasan and BASF given the remaining shares. The other estimated zeroes concern the highest volatility state:

$$\hat{\Omega}_3 = \begin{pmatrix} 455 & -39 & -22 & -15 & -41 & -80 \\ -39 & 140 & -1 & 0 & -44 & 2 \\ -22 & -1 & 29 & -1 & -7 & -2 \\ -15 & 0 & -1 & 49 & -5 & 11 \\ -41 & -44 & -7 & -5 & 120 & -65 \\ -80 & 2 & -2 & 11 & -65 & 313 \end{pmatrix},$$

$$\hat{\Omega}_3^{l_1} = \begin{pmatrix} 416 & -28 & -26 & -16 & -37 & -42 \\ -28 & 130 & -7 & 1 & -43 & 0 \\ -26 & -7 & 25 & -1 & -4 & 0 \\ -16 & 1 & -1 & 45 & -2 & 8 \\ -37 & -43 & -4 & -2 & 112 & -60 \\ -42 & 0 & 0 & 8 & -60 & 288 \end{pmatrix},$$

$$\hat{\Omega}_3^{S} = \begin{pmatrix} 455 & -29 & -36 & -17 & -32 & -82 \\ -29 & 132 & 0 & 0 & -48 & 0 \\ -36 & 0 & 25 & -2 & -6 & -3 \\ -17 & 0 & -2 & 45 & -3 & 11 \\ -32 & -48 & -6 & -3 & 118 & -69 \\ -82 & 0 & -3 & 11 & -69 & 316 \end{pmatrix}.$$

TABLE 3.8.: Estimated precision matrices state 3, $p = 6$ biotechnology portfolio: Unpenalized estimator, l_1-penalization, SCAD-penalization

We observe that when performing l_1-penalization, the zero entries from the four-dimensional portfolio all vanish and the zero entry detected in the unpenalized estimation does not occur. Instead, conditional independence of Bayer and K+S as well as Curasan and K+S given the respective remaining portfolios is estimated. SCAD-penalization reestimates two of the three zero entries from the four-dimensional portfolio, corresponding to Bayer and Curasan and Bayer and Evotec, in addition there is one zero entry concerning Bayer and K+S.

These observations show that SCAD-penalized estimation might yield more reliable results, since the results from dimension four and six mainly coincide and are in line with the unpenalized estimator, while using l_1-penalization yields two very different results. The remaining parameter estimates are given in Appendix A.

Finally, we consider a portfolio of dimension $p = 8$, consisting of the shares from the

two four-dimensional portfolios from above, namely Merck, Bayer, Curasan, Evotec, Commerzbank, Deutsche Bank, Baader Bank and Deutsche Balaton. Performing unpenalized estimation, there is only one zero entry in the highest volatility state, which states conditional independence of Bayer and Curasan given the remaining shares. The l_1-penalized estimator suggests conditional independence in each of the three states but the zero entries are not constantly estimated at the same positions. In the first volatility state we observe conditional independence of Merck and Commerzbank, Merck and Deutsche Balaton as well as Deutsche Bank and Deutsche Balaton given the respective remaining portfolio. In the second volatility state, the only zero entry suggests conditional independence of Bayer and Baader Bank given the remaining shares. The remaining zero entries are estimated in the highest volatility state:

$$\hat{\Omega}_3 = \begin{pmatrix} 542 & -63 & -28 & -12 & -25 & -55 & 13 & 95 \\ -63 & 172 & 0 & -1 & 34 & -68 & -14 & -25 \\ -28 & 0 & 35 & -1 & 2 & -6 & -8 & -5 \\ -12 & -1 & -1 & 55 & -3 & 5 & -5 & -18 \\ -25 & 34 & 2 & -3 & 206 & -195 & -17 & 6 \\ -55 & -68 & -6 & 5 & -195 & 312 & -13 & -125 \\ 13 & -14 & -8 & -5 & -17 & -13 & 88 & -36 \\ 95 & -25 & -5 & -18 & 6 & -125 & -36 & 1537 \end{pmatrix},$$

$$\hat{\Omega}_3^{l_1} = \begin{pmatrix} 443 & -24 & -20 & -4 & -17 & -34 & 0 & 0 \\ -24 & 141 & -2 & 0 & 0 & -33 & -9 & 0 \\ -20 & -2 & 31 & -1 & 0 & -4 & -6 & 0 \\ -4 & 0 & -1 & 50 & -1 & 0 & 0 & 0 \\ -17 & 0 & 0 & -1 & 154 & -131 & -13 & 0 \\ -34 & -33 & -4 & 0 & -131 & 211 & -14 & -49 \\ 0 & -9 & -6 & 0 & -13 & -14 & 75 & 0 \\ 0 & 0 & 0 & 0 & 0 & -49 & 0 & 1424 \end{pmatrix},$$

$$\hat{\Omega}_3^{S} = \begin{pmatrix} 551 & -63 & -38 & -12 & -30 & -59 & 11 & 96 \\ -63 & 169 & 0 & 0 & 35 & -65 & -12 & -33 \\ -38 & 0 & 32 & 0 & 0 & -5 & -9 & -14 \\ -12 & 0 & 0 & 54 & -3 & 3 & -6 & -15 \\ -30 & 35 & 0 & -3 & 200 & -190 & -18 & 21 \\ -59 & -65 & -5 & 3 & -190 & 307 & -15 & -137 \\ 11 & -12 & -9 & -6 & -18 & -15 & 91 & -36 \\ 96 & -33 & -14 & -15 & 21 & -137 & -36 & 1556 \end{pmatrix}.$$

TABLE 3.9.: Estimated precision matrices state 3, $p = 8$: Unpenalized estimator, l_1-penalization, SCAD-penalization

Considering l_1-penalization the zero entry from the unpenalized estimator is not identified. The conditional independence structure from the four-dimensional portfolio of the biotechnology sector is destroyed, only the zero entry concerning Bayer and Evotec remains. The conditional independence of Deutsche Balaton and Baader Bank remains

as suggested by the four-dimensional financial portfolio. In addition, Commerzbank and Deutsche Balaton are conditional independent given the remaining shares. Moreover there are 9 zero entries which correspond to conditional independence when considering two shares of different industry sectors. SCAD-penalization yields different results. In contrast to l_1-penalization there are no zero entries in the first and second volatility state. In the third volatility state, the zero entries we estimated in the four-dimensional biotechnology portfolio remain, while those of the financial sector get lost. Between the two industry sectors, there is conditional independence of Commerzbank and Curasan given the remaining shares. Note that the zero entry of the unpenalized estimate is detected here as well.
The remaining estimates are shown in Appendix A.

In summary, the simulation study and real data example yield very similar results. In almost every example l_1-penalization estimates more zeroes than SCAD-penalization. Furthermore, SCAD-penalization yields more stable results when extending or merging the portfolio, while l_1-penalized estimated zero-positions change much more. Altogether, the real data example shows that in times of high volatility, conditional independence is much more likely than in low volatility states, but all in all the estimated precision matrices are not very sparse, so that the effect one might expect, for example, in the context of portfolio allocation is probably very limited.

3.4. Proofs

We give a proof of Theorem 3.1, which follows the arguments of Fan and Li (2001, Theorem 1).

Proof of Theorem 3.1. Let $a_T^* := T^{-\frac{1}{2}} + a_T$. We prove that for $\varepsilon > 0$ and a constant c

$$P(\sup_{\|u\|=c} \mathcal{Q}_T(\lambda_0 + a_T^* u) < \mathcal{Q}_T(\lambda_0)) \geq 1 - \varepsilon. \tag{3.3}$$

Write $\Delta_T := \mathcal{Q}_T(\lambda_0 + a_T^* u) - \mathcal{Q}_T(\lambda_0)$. Keeping in mind that $\rho(0; c_T) = 0$, by Taylor expansion of $\mathcal{Q}_T(\lambda_0 + a_T^* u)$ at λ_0 we have

$$\begin{aligned}
\Delta_T =& \mathcal{L}_T(\lambda_0 + a_T^* u) - T\sum_{k=1}^{K}\sum_{i\neq j}\rho(|\omega_{ij,0}^{(k)} + a_T^* u_i|; c_T) - \mathcal{L}_T(\lambda_0) + T\sum_{k=1}^{K}\sum_{i\neq j}\rho(|\omega_{ij,0}^{(k)}|; c_T) \\
=& \mathcal{L}_T(\lambda_0 + a_T^* u) - T\sum_{k=1}^{K}\sum_{(i,j)\in I_{k,0}^{(1)}}(\rho(|\omega_{ij,0}^{(k)} + a_T^* u_i|; c_T) - \rho(|\omega_{ij,0}^{(k)}|; c_T)) - \mathcal{L}_T(\lambda_0) \\
& - T\sum_{k=1}^{K}\sum_{(i,j)\in I_{k,0}^{(2)}}(\rho(|\omega_{ij,0}^{(k)} + a_T^* u_i|; c_T) - \rho(|\omega_{ij,0}^{(k)}|; c_T))
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathcal{L}_T(\lambda_0 + a_T^* u) - T\sum_{k=1}^{K} \sum_{(i,j)\in I_{k,0}^{(1)}} (\rho(|\omega_{ij,0}^{(k)} + a_T^* u_i|; c_T) - \rho(|\omega_{ij,0}^{(k)}|; c_T)) - \mathcal{L}_T(\lambda_0) \\
&= \mathcal{L}_T(\lambda_0) + a_T^* \mathcal{L}_T'(\lambda_0)^\mathsf{T} u + \frac{1}{2}(a_T^*)^2 u^\mathsf{T} \mathcal{L}_T''(\lambda_0) u + (a_T^*)^2 u^\mathsf{T} u o_P(1) - \mathcal{L}_T(\lambda_0) \\
&\quad - T\sum_{k=1}^{K} \sum_{(i,j)\in I_{k,0}^{(1)}} (\rho(|\omega_{ij,0}^{(k)}|; c_T) + a_T^* \rho'(|\omega_{ij,0}^{(k)}|; c_T) sgn(\omega_{ij,0}^{(k)}) u_i + (a_T^*)^2 \rho''(|\omega_{ij,0}^{(k)}|; c_T) u_i^2 \\
&\quad + (a_T^*)^2 u_i^2 o_P(1) - \rho(|\omega_{ij,0}^{(k)}|; c_T)) \\
&= a_T^* \mathcal{L}_T'(\lambda_0)^\mathsf{T} u + \frac{1}{2}(a_T^*)^2 u^\mathsf{T} \mathcal{L}_T''(\lambda_0) u + (a_T^*)^2 u^\mathsf{T} u o_P(1) - T\sum_{k=1}^{K} \sum_{(i,j)\in I_{k,0}^{(1)}} ((a_T^*)^2 u_i^2 o_P(1) \\
&\quad + a_T^* \rho'(|\omega_{ij,0}^{(k)}|; c_T) sgn(\omega_{ij,0}^{(k)}) u_i + (a_T^*)^2 \rho''(|\omega_{ij,0}^{(k)}|; c_T) u_i^2).
\end{aligned}
$$

Bickel et al. (1998) proved that $\mathcal{L}_T'(\lambda_0) = O_P(T^{\frac{1}{2}})$ and $T^{-1}\mathcal{L}_T''(\lambda) = O_P(1)$. Thus, for the first summand we have $a_T^* \mathcal{L}_T'(\lambda_0)^\mathsf{T} u = O_P(T^{\frac{1}{2}} a_T)$ and for the second summand $\frac{1}{2}(a_T^*)^2 u^\mathsf{T} \mathcal{L}_T''(\lambda_0) u = O_P((a_T^*)^2 T)$. For the last summand,

$$
\begin{aligned}
&\sum_{k=1}^{K} (T a_T^* \sum_{(i,j)\in I_{k,0}^{(1)}} \rho'(|\omega_{ij,0}^{(k)}|; c_T) sgn(\omega_{ij,0}^{(k)}) u_i + T(a_T^*)^2 \sum_{(i,j)\in I_{k,0}^{(1)}} \rho''(|\omega_{ij,0}^{(k)}|; c_T) u_i^2 \\
&\quad + T(a_T^*)^2 \sum_{(i,j)\in I_{k,0}^{(1)}} u_i^2 o_P(1)) \\
&\leq \sum_{k=1}^{K} (T a_T^* a_T \sum_{(i,j)\in I_{k,0}^{(1)}} sgn(\omega_{ij,0}^{(k)}) u_i + T(a_T^*)^2 \max\{\rho''(|\omega_{ij,0}^{(k)}|; c_T) : \ \omega_{ij,0}^{(k)} \neq 0\} \sum_{(i,j)\in I_{k,0}^{(1)}} u_i^2 \\
&\quad + T(a_T^*)^2 \|u\|^2 o_P(1)) \\
&\leq \sum_{k=1}^{K} (T a_T^* a_T \|u\| \sqrt{H} + T(a_T^*)^2 \max\{\rho''(|\omega_{ij,0}^{(k)}|; c_T) : \omega_{ij,0}^{(k)} \neq 0\} \|u\|^2 + T(a_T^*)^2 \|u\|^2 o_P(1)).
\end{aligned}
$$

By assumption, $\max\{\rho''(|\omega_{ij,0}^{(k)}|; c_T) : \ \omega_{ij,0}^{(k)} \neq 0\} \to 0$, thus $\Delta_T = O_P(a_T^* T)$ and for a sufficient large constant c, (3.3) follows and there exists a local maximizer $\hat{\lambda}_T$ of $\mathcal{Q}_T(\lambda)$ such that $\|\hat{\lambda}_T - \lambda_0\| = O_P(a_T^*)$. □

The following proof of Theorem 3.2 is based on the proofs of Fan and Li (2001, Lemma 1, Theorem 2)

Proof of Theorem 3.2. 1. Sparsity: We prove that for some $c > 0$

$$
\mathcal{Q}_T((\hat{\lambda}_{T,1}, 0)^\mathsf{T}) = \max_{\|\hat{\lambda}_{T,2}\| \leq c T^{-\frac{1}{2}}} \mathcal{Q}_T((\hat{\lambda}_{T,1}, \hat{\lambda}_{T,2})^\mathsf{T}) \tag{3.4}
$$

with probability tending to 1.

Let $\varepsilon_T := cT^{-\frac{1}{2}}$. We prove that for $j = H+1, \ldots, G$,

$$\frac{\partial \mathcal{Q}_T(\lambda)}{\partial \lambda_j} \begin{cases} < 0, & 0 < \lambda_j < \varepsilon_T \\ > 0, & -\varepsilon_T < \lambda_j < 0 \end{cases},$$

which implies that the penalized likelihood function is maximized at $\hat{\lambda}_{T,2} = 0$. Taylor expansion for the first derivative of the likelihood function at λ_0 yields

$$\begin{aligned}
\frac{\partial \mathcal{Q}_T(\lambda)}{\partial \lambda_j} =& \frac{\partial \mathcal{L}_T(\lambda)}{\partial \lambda_j} - T\rho'(|\lambda_j|; c_T) sgn(\lambda_j) \\
=& \frac{\partial \mathcal{L}_T(\lambda_0)}{\partial \lambda_j} + \left(\frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_j \partial \lambda_1}, \ldots, \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_j \partial \lambda_G} \right) \begin{pmatrix} \lambda_1 - \lambda_{1,0} \\ \vdots \\ \lambda_G - \lambda_{G,0} \end{pmatrix} \\
& + o_P(1) \mathbb{1}^\intercal \begin{pmatrix} \lambda_1 - \lambda_{1,0} \\ \vdots \\ \lambda_G - \lambda_{G,0} \end{pmatrix} - T\rho'(|\lambda_j|; c_T) sgn(\lambda_j) \\
\leq & Tc_T \Big(\frac{1}{c_T} \frac{1}{T} \frac{\partial \mathcal{L}_T(\lambda_0)}{\partial \lambda_j} + \frac{1}{c_T} \frac{1}{T} \sum_{l=1}^{G} \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_j \partial \lambda_l} (\lambda_l - \lambda_{l,0}) \\
& - \frac{1}{c_T} \rho'(|\lambda_j|; c_T) sgn(\lambda_j) + \frac{1}{c_T} \frac{1}{T} o_P(1) \|\lambda - \lambda_0\| \sqrt{G} \Big).
\end{aligned}$$

Following Bickel et al. (1998), $\frac{1}{T} \frac{\partial \mathcal{L}_T(\lambda_0)}{\partial \lambda_j} = O_P(T^{-\frac{1}{2}})$ and $\frac{1}{T} \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_j \partial \lambda_l} = O_P(1)$. By assumption, $Tc_T \to \infty$, $\|\lambda - \lambda_0\| = O_P(T^{-\frac{1}{2}})$ and $c_T^{-1} \rho'(\theta; c_T) > 0$, thus

$$\frac{\partial \mathcal{Q}_T(\lambda)}{\partial \lambda_j} = Tc_T (O_P(c_T^{-1} T^{-\frac{1}{2}}) - c_T^{-1} \rho'(|\lambda_j|; c_T) sgn(\lambda_j)),$$

which shows that the sign of the derivative is determined by the sign of λ_j, leading to (3.4).

2. Asymptotic normality: Denote by $\hat{\lambda}_T$ a $\sqrt{T}$-consistent estimator from Theorem 3.1, which is a local maximizer of $\mathcal{Q}_T((\lambda_1, 0)^\intercal)$ satisfying $\frac{\partial \mathcal{Q}_T(\lambda)}{\partial \lambda_j} \mid_{\lambda = (\hat{\lambda}_{T,1}, 0)^\intercal} = 0$, $j = 1, \ldots, H$. Taylor expansion of the first derivative at λ_0 yields

$$\begin{aligned}
0 =& \frac{\partial \mathcal{Q}_T(\lambda)}{\partial \lambda_j} \mid_{\lambda = (\hat{\lambda}_{T,1}, 0)^\intercal} = \frac{\mathcal{L}_T(\lambda)}{\partial \lambda_j} \mid_{\lambda = (\hat{\lambda}_{T,1}, 0)^\intercal} - T\rho'(|\hat{\lambda}_{T,j}|; c_T) sgn(\hat{\lambda}_{T,j}) \\
=& \frac{\mathcal{L}_T(\lambda_0)}{\partial \lambda_j} + \sum_{l=1}^{H} \Big(\frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_j \partial \lambda_l} + o_P(1) \Big) (\hat{\lambda}_{T,l} - \lambda_{l,0}) - T\rho'(|\lambda_{j,0}|; c_T) sgn(\lambda_{j,0}) \\
& - T(\rho''(|\lambda_{j,0}|; c_T) + o_P(1))(\hat{\lambda}_{T,j} - \lambda_{j,0}),
\end{aligned}$$

in matrix notation

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\partial \mathcal{L}_T(\lambda_0)}{\partial \lambda_1} \\ \vdots \\ \frac{\partial \mathcal{L}_T(\lambda_0)}{\partial \lambda_H} \end{pmatrix} + \begin{pmatrix} \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_1 \partial \lambda_1} & \cdots & \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_1 \partial \lambda_H} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_H \partial \lambda_1} & \cdots & \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_H \partial \lambda_H} \end{pmatrix} (\hat{\lambda}_{T,1} - \lambda_{10}) + o_P(1)\mathbb{1}^\mathsf{T}(\hat{\lambda}_{T,1} - \lambda_{10}) - Tb - TA(\hat{\lambda}_{T,1} - \lambda_{10}),$$

thus,

$$b + (-\frac{1}{T} \begin{pmatrix} \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_1 \partial \lambda_1} & \cdots & \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_1 \partial \lambda_H} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_H \partial \lambda_1} & \cdots & \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_H \partial \lambda_H} \end{pmatrix} + A)(\hat{\lambda}_{T,1} - \lambda_{10}) + o_P(1)\mathbb{1}^\mathsf{T}(\hat{\lambda}_{T,1} - \lambda_{10}) = \frac{1}{T} \begin{pmatrix} \frac{\partial \mathcal{L}_T(\lambda_0)}{\partial \lambda_1} \\ \vdots \\ \frac{\partial \mathcal{L}_T(\lambda_0)}{\partial \lambda_H} \end{pmatrix}.$$

Since by Bickel et al. (1998) $T^{-\frac{1}{2}}\mathcal{L}'_T(\lambda_0) \to \mathcal{N}(0, I^*(\lambda_{10}))$ weakly, we have

$$\sqrt{T}(b - \frac{1}{T}(\begin{pmatrix} \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_1 \partial \lambda_1} & \cdots & \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_1 \partial \lambda_H} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_H \partial \lambda_1} & \cdots & \frac{\partial^2 \mathcal{L}_T(\lambda_0)}{\partial \lambda_H \partial \lambda_H} \end{pmatrix} + A)(\hat{\lambda}_{T,1} - \lambda_{10})) \to \mathcal{N}(0, I^*(\lambda_{10}))$$

weakly, which completes the proof.

□

4. A hidden Markov model for panel data: Modelling income distributions and classification

In this chapter we consider an application of hidden Markov models to GDP data for modelling monetary welfare distributions in the context of economic growth. This investigation is part of a DFG-project "Structure, trends and determinants of monetary and non-monetary welfare distributions" with the objective to provide statistical methods and models to analyze welfare distributions. We generalize the hidden Markov model introduced in Chapter 1 in order to allow for a longitudinal structure and the inclusion of covariables.

The panel data under consideration describe the GDP of countries of the world over a time horizon from 1970 to 2010. We analyze the data in four models in order to investigate different dependency structures and the influence of covariables on the welfare distribution. Objectives of the study are the construction of models to explain the income distribution at a fixed year, classification of the countries to income groups, and the investigation of determinants of a welfare distribution and their influence on the development of countries.

After introducing the data and the general setting of the models in Section 4.1, we start with a cross sectional model without covariables. The idea is to fit a finite mixture model for each year independently, to properly model subpopulations in the income distribution. This approach has been used for example by Paapaa and van Dijk (1998), who used a two-component mixture of a truncated normal distribution and a Weilbull distribution to model the income distribution or by Pittau et al. (2010), Vollmer et al. (2013), modelling the log-income using three-component Gaussian mixture models. In order to allow for dependence over time while marginally retaining the structure of a mixture model, we extend the analysis by introducing a nonhomogeneous hidden Markov model for panel data. Both models allow convenient methods to perform a-posteriori assignment of the countries to income groups and classification results show advancement and decline of some countries concerning the income group. An attempt to explain this behaviour is given in Sections 4.5 and 4.6, where we expand our models by the use of covariables. For mixture models, this approach is given by switching regression models,

see for example Quandt and Ramsey (1978), DeSarbo and Cron (1988) or Bloom et al. (2003) for an example in a similar context of economic growth. While usually covariables are introduced to explain the component-dependent parameters of the distributions, we find it reasonable to use covariables for the mixing probabilities. For more details see Section 4.4, where we illustrate the selection of covariables in our model. In a final step we give an idea how to transfer this approach to the transition probabilities in the nonhomogeneous hidden Markov model. It turns out that the inclusion of covariables to the model with dependence structure over time is difficult for the available data.

4.1. Data

In our models we consider random variables $X_{t,i}$ being the log-GDP (base 10) of country $i \in \{1, \dots, I\}$ at time $t \in \{1, \dots, T\}$. Thus, we deal with panel data, giving the income of all countries at time t in row t, while column i provides the time series of log-GDP for country i. For a fixed $t \in \{1, \dots, T\}$ we assume the income of all countries to be independent to each other, having marginal distribution $f_t(x)$.
The GDP data are given in Penn World Table 7.1 (Heston et al., 2012). We choose the PPP converted GDP per capita (chain series) at 2005 constant prices (variable rgdpch in the data set mentioned) and consider $T = 41$ years from 1970 to 2010. For our analysis, similarly as Vollmer et al. (2013), we exclude small oil-exporting countries (Bahrain, Brunei, Equatorial Guinea, Gabon, Kuwait, Qatar, Suriname, Trinidad and Tobago), since their GDP is heavily affected by the oil-price. In addition, we remove those countries for which between 1970 and 2010 GDP data are missing. The resulting data set contains GDP data for $I = 152$ countries over time, see Appendix B for a list

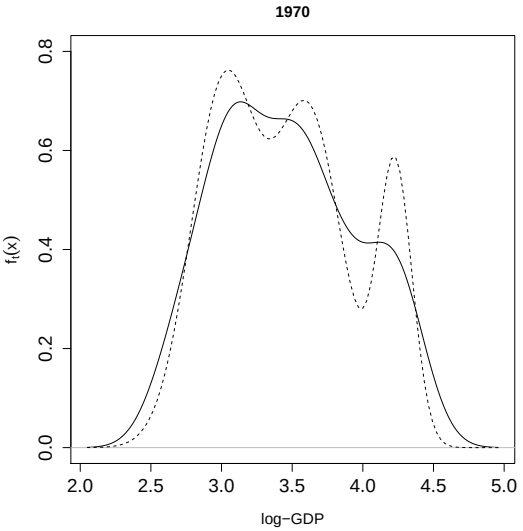


FIGURE 4.1.: GDP data for 1970: Kernel density estimator (solid line) and estimated mixture density (dashed line).

of the countries and the corresponding codes based on the ISO standard.
When considering models with covariables, additional data sets are required. The choice of covariables is discussed in more detail in Section 4.4. The corresponding data sources are Penn World Table 7.1 (Heston et al., 2012) for the variable *investment share of GDP*, the Barro–Lee data set (Barro and Lee, 2013) for the variable *average years of schooling*, a data base for world development indicators provided by the World Bank[1] for the variables *life expectancy* and *fertility rate*, as well as a data set on geography data[2] for the variable *latitude*. When combining the data sets we obtain a collection of 107 countries for which all variables are available.

4.2. Cross sectional analysis using mixture models

In this section we formulate our first model for the welfare distribution. We perform cross-sectional analysis, assuming the log-income of each country to be independent over time. Thus, in order to properly fit K subpopulations in the income distribution, we model finite Gaussian mixture models for each year t independently. The marginal distribution is then given by

$$f_t(x) = \sum_{k=1}^{K} \pi_k^{(t)} g(x; \vartheta_k^{(t)}),$$

[1] `http://data.worldbank.org/data-catalog/world-development-indicators` (04.06.2015)
[2] `http://www.pdx.edu/econ/country-geography-data` (04.06.2015)

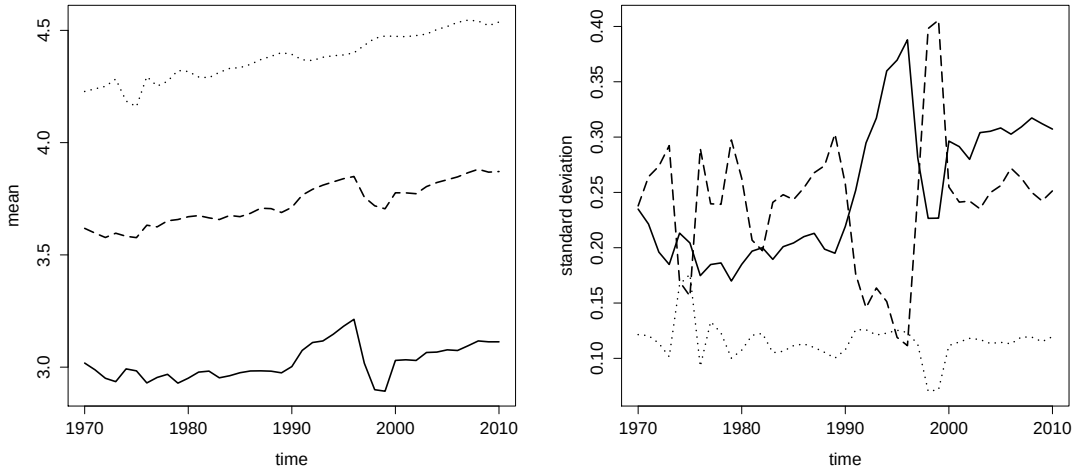


FIGURE 4.2.: Parameter estimates mixture model - left: means, right: standard deviations. Solid line: income group 1, dashed line: income group 2, dotted line: income group 3.

where the mixing probabilities $\pi_k^{(t)} \geq 0$ sum up to one for each t and $\vartheta_k^{(t)} = (\mu_k^{(t)}, \sigma_k^{(t)})$ $(k = 1, \ldots, K,\ t = 1, \ldots, T)$ are the component-dependent parameters of the Gaussian distributions, which are denoted by $g(\cdot; \vartheta_k^{(t)})$.

Our first issue is the estimation of the parameters $(\pi_1^{(t)}, \ldots, \pi_K^{(t)}, \vartheta_1^{(t)}, \ldots, \vartheta_K^{(t)})$ $(t = 1, \ldots, T)$, thus $2KT + (K-1)T$ parameters are considered. We apply the maximum likelihood approach, where the likelihood function of the model is

$$\mathcal{L}_T^{(mix)} = \prod_{i=1}^{I} \prod_{t=1}^{T} (\sum_{k=1}^{K} \pi_k^{(t)} g(x_{t,i}; \vartheta_k^{(t)})).$$

We perform maximization using the R-package mclust (Fraley and Raftery, 2002, Fraley et al., 2012). In each year, the estimation procedure selects a three-component mixture model, where the components correspond to three income groups of countries with low/mid/high income. Showing a kernel density estimate for the GDP data in the year 1970 together with the estimated mixture density, Figure 4.1 indicates that this model is a plausible choice. Figure 4.2 shows an overview over the estimated parameters for the Gaussian distributions over the 41 years. We observe that the means of the income groups are relatively stable, except for the 1990s, when especially the means of income group 1 and 2 rise at first and then drop sharply. The estimated standard deviations are quite volatile. Based on the parameter estimates we perform a-posteriori classification, estimating the most likely sequence of income groups for each country in the mixture model by

$$\underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \frac{\hat{\pi}_k^{(t)} g(x_{t,i}; \hat{\vartheta}_k^{(t)})}{\sum_{l=1}^{K} \hat{\pi}_l^{(t)} g(x_{t,i}; \hat{\vartheta}_l^{(t)})}, \qquad t = 1, \ldots, T.$$

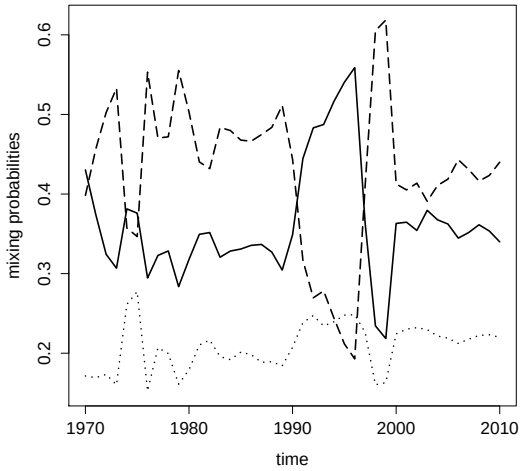


FIGURE 4.3.: Mixture model: shares of income groups. Solid line: income group 1, dashed line: income group 2, dotted line: income group 3.

Classification results are presented in Section 4.9, where we report the income group for each country together with potential switches of income group and the associated year. Figure 4.3 reports the shares of income groups. We observe that income group 2 is the largest group over time except for the years 1990–1998, when income group 1 dominates. From Figure 4.2 we observe that this fact might be due to the rising mean of group 1 during this period together with its rising standard deviation.

We observe that due to the assumed independence over time, many switches of the income group occur when performing a-posteriori classification: 76 of the 152 countries perform at least one switch of income groups, while the remaining 76 countries stay in the same income group over the T years. The consequences from our observations in Figures 4.2 and 4.3 are visible in Table 4.2 (Section 4.9) too: many countries (e.g. Albania (ALB), Bolivia (BOL), Djibouti (DJI), Namibia (NAM), Paraguay (PRY)) switch from income group 2 to group 1 in the late 1980s/ early 1990s and return to group 2 around 1998. The overall impression is that the assumption of independence over time leads to many redundant switches of income groups, see for example Angola (AGO) where in the 1980s switches occur almost every year. This aspect is further discussed when using hidden Markov models in the following section.

4.3. Serial dependence in nonhomogeneous hidden Markov models

Now we drop the independence assumption and fit a nonhomogeneous parametric hidden Markov model to allow for serial dependence of a countries' income over time, while keeping the structure of a mixture model for fixed time t. Since we allow the parameters

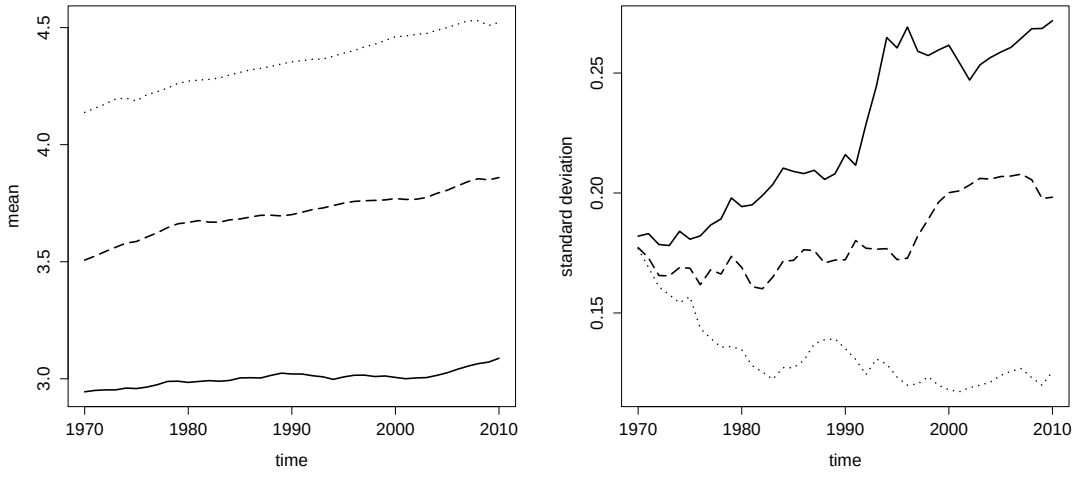


FIGURE 4.4.: Parameter estimates HMM - left: means, right: standard deviations. Solid line: income group 1, dashed line: income group 2, dotted line: income group 3.

of the model to change over time, we adjust the notation introduced in Section 1.2. Let $\Gamma^{(t)} = (\gamma_{k,l}^{(t)})_{k,l=1,\ldots,K}$ denote the transition probability matrix of the Markov chain at time $t \geq 2$ and $\alpha = (\alpha_1, \ldots, \alpha_K)$ its initial distribution. We assume the state-dependent distributions to be Gaussian with parameters $\vartheta_k^{(t)} = (\mu_k^{(t)}, \sigma_k^{(t)})$, $k = 1, \ldots, K$, $t = 1, \ldots, T$. Thus, there are $K - 1 + (T-1)K(K-1) + 2TK$ parameters, which we estimate maximizing the penalized log-likelihood function

$$\ell_T^{(HMM)} = \sum_{i=1}^{I} \log\Big(\sum_{k_1=1}^{K} \cdots \sum_{k_T=1}^{K} \alpha_k \prod_{t=2}^{T} \gamma_{k_{t-1},k_t}^{(t)} \prod_{t=1}^{T} g(x_{t,i}; \vartheta_{k_t}^{(t)})\Big) + \frac{c}{K} \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \log(\gamma_{j,k}^{(t)}).$$

We introduce the penalty term with tuning parameter $c \geq 0$, since in the hidden Markov model switches of the income groups are rarely observed but should still be enabled in the estimated model. It penalizes small values for the transition probabilities to prevent the estimation of diagonal transition probability matrices.

We use the EM-algorithm for parameter estimation. The algorithm described in Section 1.3 needs to be modified to be suitable for panel data, the nonhomogeneous structure of the model and the penalty term. Maruotti (2011) gave an EM-algorithm for longitudinal hidden Markov models. For our model, further modifications are required due to the nonhomogeneity and the introduced penalty term on the transition probabilities. Details are given in Section 4.8.1. The resulting parameter estimates are shown in Figure 4.4. We observe that the income means are of the same order as in the mixture model but growing more steady. The estimated standard deviations are less volatile compared to the results from the mixture model. The standard deviation of income group 1 rises over the 41 years, with very high slope in the early 1990s, while for income group 2 the estimates are more or less at a constant level from 1970 to the late 1990s and then begin

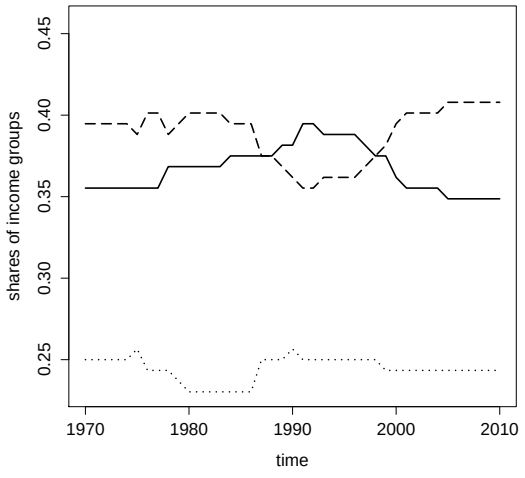


FIGURE 4.5.: Hidden Markov model: Shares of income groups. Solid line: income group 1, dashed line: income group 2, dotted line: income group 3.

to increase. The estimated standard deviations for income group 3 show a downward tendency.

Classification is performed for each country independently using the Viterbi algorithm described in Section 1.2. Based on these estimates we report the shares of income groups of each year in Figure 4.5. Again, we observe that income group 2 is the largest group, except for a time horizon between the late 1980s until the late 1990s, when income group 1 dominates, as already observed in the mixture model. From Table 4.2 we see that in the hidden Markov model only 24 countries switch income group over the 41 years, while the remaining 128 countries are constantly assigned to the same income group. In particular, the assignments are more stable since Angola (AGO) and Iraq (IRQ) are the only countries that switch income group more than once. Around the year 1990 there are four countries switching from group 2 to group 1 (Djibouti (DJI), Iraq (IRQ), Mongolia (MNG) and Nicaragua (NIC)), as mentioned Iraq is the only country that switches back to group 2 in 1997. Further, China (CHN), Sri Lanka (LKA) and the Maldives (MDV) ascend from group 1 to group 2 around 1998.

From the estimated transition probabilities we observe that apart from a peak at the beginning of the time horizon, the probability to ascend from income group 1 to group 2 rises starting in 1985 until 2000, while the probability to ascend from group 2 to group 3 reaches its peak in the late 1980s and then declines and stays close to zero from 1995 on. On the other hand we observe that the probability of a decline from group 2 to group 1 drops in the early 1990s, while the probability to switch from group 3 to group 2 is quite volatile over the 41 years.

Switching model: Hidden Markov model versus fixed state model

Since estimation results from the hidden Markov model show that only a minority of countries switch income groups, we modify the model in order to take account of the countries remaining in one state: We allow each country to either be in a hidden Markov model or to stay in one fixed income group. For this purpose, we introduce an additional variable $\pi_i \in \{0,1\}$, which switches between the hidden Markov model and the fixed state model, depending on country $i \in \{1,\dots,I\}$. For parameter estimation we maximize the penalized log-likelihood function of the model

$$\begin{aligned}\ell_T^{(SHMM)} = \sum_{i=1}^{I} \log(\pi_i(\sum_{k_1=1}^{K}\dots\sum_{k_T=1}^{K}\alpha_{k_1}\prod_{t=2}^{T}\gamma^{(t)}_{k_{t-1},k_t}\prod_{t=1}^{T}g(x_{t,i};\vartheta^{(t)}_{k_t}))\\ + (1-\pi_i)\max_{k=1,\dots,K}\prod_{t=1}^{T}g(x_{t,i};\vartheta^{(t)}_{k})) + \frac{c}{KI}\sum_{i=1}^{I}\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\pi_i\log(\gamma^{(t)}_{j,k}),\end{aligned}$$

using an EM-algorithm where the π_i are treated as additional latent variable, which is assumed to be independent of $S_{t,i}$. Details on the algorithm are given in Section 4.8.1.

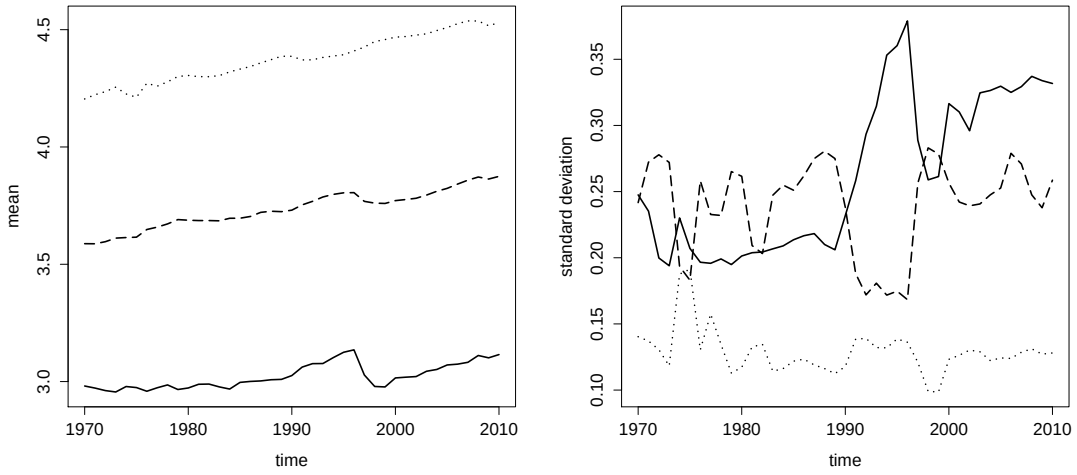


FIGURE 4.6.: Parameter estimates switching HMM - left: means, right: standard deviations.

The estimated parameters are shown in Figure 4.6. We observe that in contrast to the previously estimated hidden Markov model, the means are more volatile. Especially the mean of income group 1 shows similar behaviour as the estimate in the mixture model, where in the late 1990s the mean of income group 1 drops and at the same time the estimated standard deviation of group 1 reaches its peak. A-posteriori analysis shows that only seven countries (namely Hong Kong (HKG), Iran (IRN), Republic of Korea (KOR), Lebanon (LBN), Macao (MAC), Maldives (MDV) and Taiwan (TWN)) are assigned to the hidden Markov model part of the model, while the remaining 145 countries stay in a fixed income group over time. Due to this fact, there is hardly any

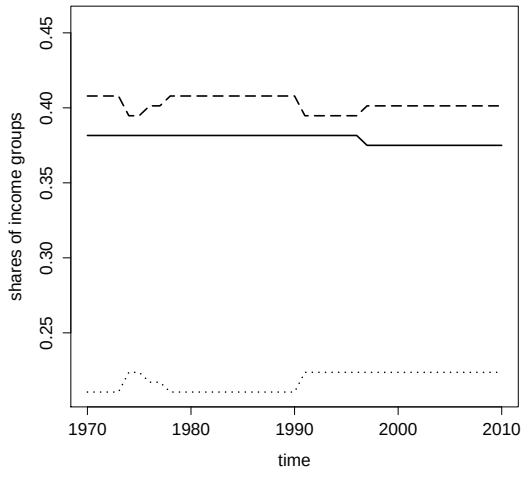


FIGURE 4.7.: Switching hidden Markov model: Shares of income groups. Solid line: income group 1, dashed line: income group 2, dotted line: income group 3.

dynamic and the shares of income groups plotted in Figure 4.7 rarely change over the 41 years. In particular, in contrast to the mixture model and the hidden Markov model, income group 2 is the largest group over the complete time horizon.
Since this model does not seem to capture the dynamics of the data, we omit to report the results in Table 4.2.

Split model: Separated hidden Markov models for advancement and decline

Obviously, the model described above is not able to capture all relevant switches of income groups we observed in the mixture model and the hidden Markov model. Thus, we split the model into three parts: One hidden Markov model part for countries which experience an advancement, where the transition probability matrices are assumed to be upper triangular matrices, one hidden Markov model part for declining countries, where the transition probability matrices are assumed to be lower triangular matrices and one part for countries remaining in one income group over the time horizon. We denote by $\Gamma^{(t,m)} = (\gamma_{k,l}^{(t,m)})_{k,l=1,\dots,K}$, $m = 1, 2$ the transition probability matrices at time $t \geq 2$ for the first and second part of the model, respectively. Similar to the model described above, we introduce an additional variable $\pi_{im} \in \{0, 1\}$ ($m \in \{1, 2, 3\}$), where $\sum_{m=1}^{3} \pi_{im} = 1$, which selects the model for each country $i = 1, \dots, I$. The penalized log-likelihood function of the model is

$$\begin{aligned}
\ell_T^{(HMMud)} = &\sum_{i=1}^{I} \log(\pi_{i1}(\sum_{k_1=1}^{K} \dots \sum_{k_T=1}^{K} \alpha_{k_1} \prod_{t=2}^{T} \gamma_{k_{t-1},k_t}^{(t,1)} \prod_{t=1}^{T} g(x_{t,i}; \vartheta_{k_t}^{(t)})) \\
&+ \pi_{i2}(\sum_{k_1=1}^{K} \dots \sum_{k_T=1}^{K} \alpha_{k_1} \prod_{t=2}^{T} \gamma_{k_{t-1},k_t}^{(t,2)} \prod_{t=1}^{T} g(x_{t,i}; \vartheta_{k_t}^{(t)})) \\
&+ \pi_{i3} \max_{k=1,\dots,K} \prod_{t=1}^{T} g(x_{t,i}; \vartheta_k^{(t)})) \\
&+ \frac{c}{\sqrt{K^2 + K - 2I}} \sum_{t=2}^{T} \sum_{j=1}^{K} (\sum_{k=j}^{K} \pi_{i1} \log(\gamma_{j,k}^{(t,1)}) + \sum_{k=1}^{j} \pi_{i2} \log(\gamma_{j,k}^{(t,2)})),
\end{aligned}$$

where for $k > l$ $\gamma_{k,l}^{(t,1)} = 0$ and for $k < l$ $\gamma_{k,l}^{(t,2)} = 0$.
The modifications for the EM-algorithm are described in Section 4.8.1. The estimated parameters shown in Figure 4.8 are very similar to those from the previous model. A-posteriori analysis yields the following classification:

- 129 countries are assigned to a fix income group,
- 12 countries are assigned to the advancement hidden Markov model, these are China (CHN), Cyprus (CYP), Egypt (EGY), Hong Kong (HKG), Indonesia (IDN),

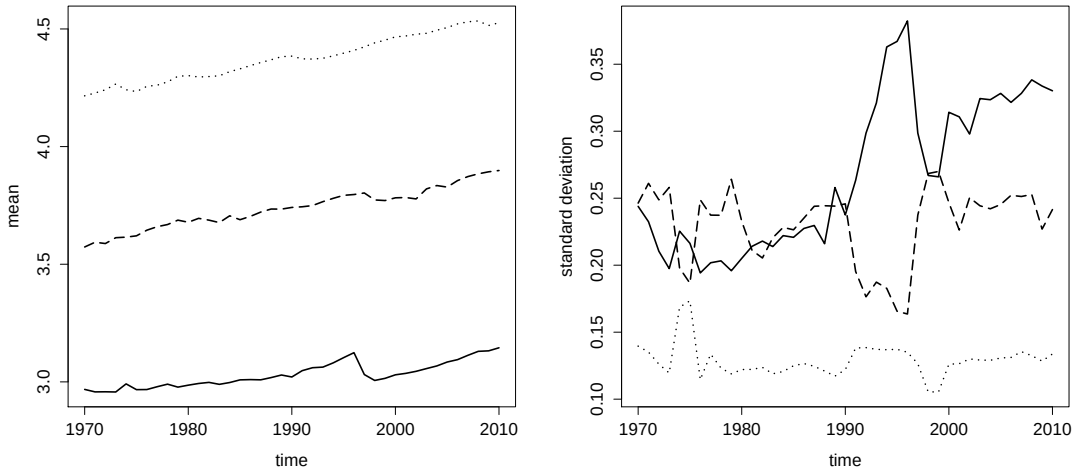


FIGURE 4.8.: Parameter estimates in the split hidden Markov model - left: means, right: standard deviations.

Republic of Korea (KOR), Macao (MAC), Maldives (MDV), Malta (MLT), Oman (OMN), Portugal (PRT), Taiwan (TWN),

- 11 countries are assigned to the declining hidden Markov model, namely Angola (AGO), Djibouti (DJI), Iran (IRN), Iraq (IRQ), Lebanon (LBN), Mongolia (MNG), Nicaragua (NIC), Philippines (PHL), Palau (PLW), Papua New Guinea (PNG) and Venezuela (VEN),

see Section 4.9 for details.

The resulting shares of income groups shown in Figure 4.9 confirm the observations from the mixture model and the hidden Markov model: Group 2 is the largest income group except for the years 1989–1998, since during this period there are a couple of countries switching from income group 2 to income group 1 (DJI, IRQ, MNG, NIC) and at the same time some countries ascend from group 2 to group 3 (CYP, KOR, MLT, PRT, TWN), thus during this period income group 1 is the dominating income group. This effect is compensated in the late 1990s, when some countries ascend from group 1 to group 2 (CHN, EGY, IDN, MDV).

The estimated transition probabilities show that the probability $\gamma_{1,2}$ starts to rise in the early 1980s and reaches its peak in the late 1990s, while the probability $\gamma_{2,3}$ stays at a relatively low level over time with a higher period in the early 1990s. The transition probability $\gamma_{2,1}$ is close to zero over the 41 years with a higher phase around 1990, while the probability $\gamma_{3,2}$ sharply rises after 1995.

Taken as a whole, we observe that nonhomogeneous hidden Markov models are suitable to analyze the GDP of countries over a time horizon of several years, to perform classification to income groups, and to examine switches of countries between these income

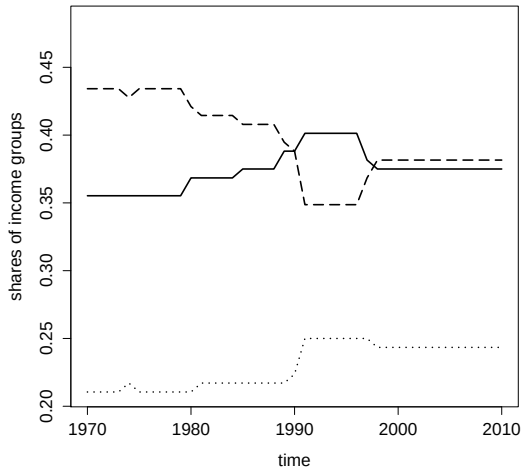


FIGURE 4.9.: Split model: Shares of income groups. Solid line: income group 1, dashed line: income group 2, dotted line: income group 3.

groups. Due to the assumed serial dependence, hidden Markov models yield much more stable estimates and classification results compared with the mixture model from Section 4.2.

The first extension of the model, which allows each country to either run in a hidden Markov model or in a fix income group seems not to be able to capture all the aspects we observed in the mixture model and in the hidden Markov model. Splitting the hidden Markov part of the extended model into an advancement and a declining part yields the desired results. A drawback of this model is that countries can now either ascend or decline, thus multiple switches of one country like Angola or Iraq perform in the general hidden Markov model are impossible. In addition, one should be aware of the fact that the parameter estimates of the hidden Markov parts of the split model are based on a few observations only, since most of the countries are assigned to the fix-state part of the model.

4.4. Selection of covariables

In the following sections we extend our analysis by the use of covariables, which might be helpful in explaining the group membership of countries and their switching behaviour. In this section we try to get an idea of a reasonable choice of covariables. To gain intuition of their possible impacts in our models, we investigate the three income groups separated from each other, before we combine the results to comprehensive models in the following sections.

The data sources of the chosen covariables are given in Section 4.1. Including covariables to the models reduces the number of countries in our analysis to $I = 107$, due to

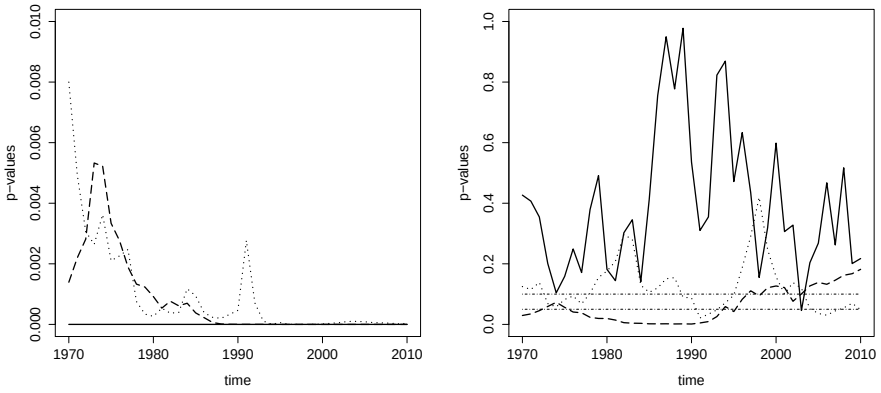


FIGURE 4.10.: Mean regression GDP: p-values. Left: Intercept (solid line), years of schooling (dashed line), life expectancy (dotted line). Right: Investment share of GDP (solid line), latitude (dashed line), fertility rate (dotted line). Dash-dotted: 5% and 10% level.

availability issues. In Table 4.2 those countries which are not considered in the following models are indicated by '-'.

To improve comparability of the results we standardize all covariables to mean 0 and standard deviation 1.

Mean regression in the income groups

The first step when including covariables to the models is the choice of explanatory variables. To get an idea of a reasonable model, we study the influence of the variables *investment share of GDP*, *average years of schooling*, *life expectancy*, *latitude* and *fertility rate* on the response variable *GDP* of all countries. In addition, we add an intercept to our model.

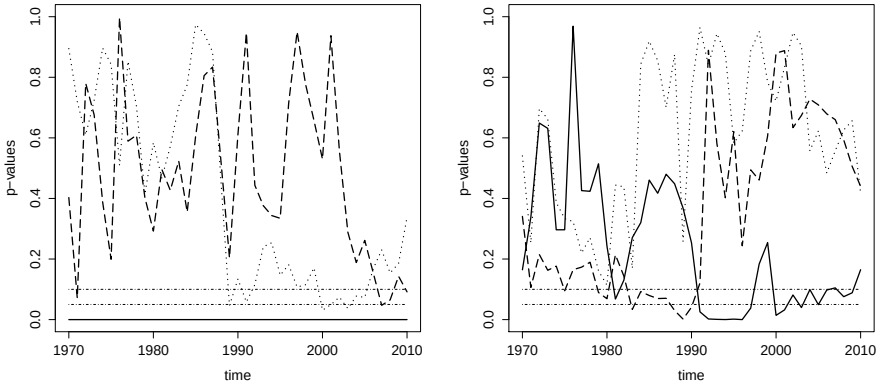


FIGURE 4.11.: Mean regression component 1: p-values. Left: Intercept (solid line), investment share of GDP (dashed line), years of schooling (dotted line). Right: Life expectancy (solid line), latitude (dashed line), fertility (dotted line). Dashed-dotted lines: 5% and 10% level.

The p-values of the estimated linear model are shown in Figure 4.10. They indicate that the variables *years of schooling*, *life expectancy* and *latitude* might affect the GDP of a country. To gain more insight, we use the classification of the mixture model to divide the countries into three income groups and perform linear regression in each group, using the covariables from above.
Once the countries are divided into income groups, none of the variables seems to be significant for explaining the GDP in mean. As an example, Figure 4.11 shows the p-values of the model in income group 1, the other income groups yield similar results. Thus, it is probably more reasonable to perform regression on the mixing probabilities.

Regression for the mixing probabilities

Now we use the a-posteriori mixing probabilities for each country $i = 1, \dots, I$, from the estimated mixture model in Section 4.2,

$$\hat{\pi}_{k,i}^{(t)} = \frac{\hat{\pi}_k^{(t)} g(x_{t,i}^{(t)}; \hat{\vartheta}_k^{(t)})}{\sum_{k=1}^{K} \hat{\pi}_k^{(t)} g(x_{t,i}^{(t)}; \hat{\vartheta}_k^{(t)})}$$

and perform a linear regression for each component $k = 1, \dots, K$. The response variable is the (probit-)transformed a-posteriori probability $\Phi^{-1}(\hat{\pi}_{k,i}^{(t)})$, where Φ denotes the distribution function of a Gaussian distribution and the covariables are chosen as in the model above. The corresponding p-values for component 2 are shown in Figure 4.12. Since income groups 1 and 3 yield similar results, we observe that next to the intercept, the variables *years of schooling*, *latitude* and *life expectancy* might influence the probability of a country to be in a certain income group. Thus, we reduce the model and again perform linear regressions of the transformed a-posteriori probabilities on the

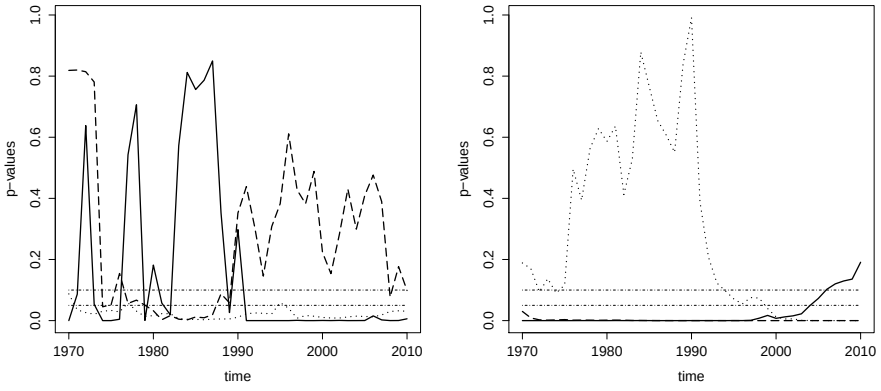


FIGURE 4.12.: Regression for mixing probabilities: p-values component 2. Left: Intercept (solid line), investmentshare of GDP (dashed line), years of schooling (dotted line). Right: Life expectancy (solid line), latitude (dashed line), fertility rate (dotted line). Dashed-dotted lines: 5% and 10% level.

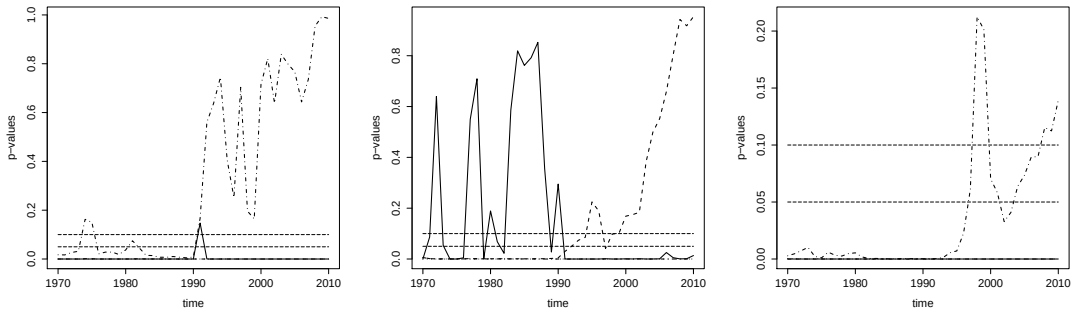


FIGURE 4.13.: Reduced model: p-values of regression for mixing probabilities. Left: income group 1, mid: income group 2, right: income group 3. Covariables intercept (solid line), years of schooling (dashed line), life expectancy (dotted line), latitude (dash-dotted line). Longdash: 5% and 10% level.

variables intercept, *years of schooling*, *latitude* and *life expectancy*.

The p-values of the reduced models are shown in Figure 4.13. We observe that the variable *life expectancy* is highly significant in the three income groups with p-value close to zero. In the first income group beginning in 1990 the p-value of the variable *latitude* rises over the 10% significance level, while the remaining covariables are significant at the 5% level over time (except of the intercept in a period between 1990 and 1992). In the second income group the p-value of the intercept is quite high from 1970–1992 and while close to zero in the beginning of the observed time horizon, from 1990 the p-value of the variable *years of schooling* rises and stays over the 10% level after 1999. In the third income group all variables have p-values close to zero over the 41 years, except of the variable *latitude* with p-value close to zero from 1970 until 1993, rising above the 10% significance level between 1997 and 2000 and after 2007.

The estimated coefficients of the reduced models show that the chosen covariables *years of schooling*, *life expectancy* and *latitude* seem to have positive effects on the GDP of a country. We observe that in income group 1 the signs of the estimated coefficients are almost always negative. Thus, increasing years of schooling, life expectancy and latitude lowers the probability of a country to be in income group 1. In income group 2 the variables *years of schooling* and *latitude* still have negative signs, while the variable *life expectancy* has positive influence on the probability of a country to be part of income group 2. In income group 3 the effects of all three covariables have positive signs.

4.5. Switching Regression: Cross sectional analysis with covariables

The mixture model from Section 4.2 is now combined with the covariables selected in Section 4.4 in order to formulate a switching regression model for the welfare distribution of the countries. Let L denote the number of considered covariables (including the

intercept) and write $\chi_i^{(t)} = (\chi_{1,i}^{(t)}, \ldots, \chi_{L,i}^{(t)})^\mathsf{T}$ for the data of country $i \in \{1, \ldots, I\}$ at time $t \in \{1, \ldots, T\}$.

For each year we formulate a K-component Gaussian mixture model

$$f(x_{t,i}) = \sum_{k=1}^{K} \pi_{k,i}^{(t)} g(x_{t,i}; \vartheta_k^{(t)}),$$

with mixing probabilities modelled by categorial logit regression: Let r denote the reference income group of the model and $M = \{1, \ldots, K\} \setminus \{r\}$. In this model, we choose $r = 2$, thus the results are to be interpreted relative to income group 2. Then, for $k \in M$,

$$\pi_{k,i}^{(t)} = \frac{\exp(\chi_i^{(t)\mathsf{T}} \beta_k^{(t)})}{1 + \sum_{l \in M} \exp(\chi_l^{(t)} \beta_l^{(t)})}, \qquad \pi_{r,i}^{(t)} = 1 - \sum_{l \in M} \pi_{l,i}^{(t)}, \qquad (i = 1, \ldots, I,\ t = 1, \ldots, T), \tag{4.1}$$

thus $\beta_k^{(t)} = (\beta_{k,1}^{(t)}, \ldots, \beta_{k,L}^{(t)})$ $(k \in M)$ denote the parameter vectors of the regression part of the model.

For estimation of the $(K-1)TL + 2KT$ parameters $\beta_k^{(t)}$ $(k \in M)$ and $\vartheta_k^{(t)}$ $(k = 1, \ldots, K,$ $t = 1, \ldots, T)$, we maximize the likelihood function of the model, given by

$$\begin{aligned} \mathcal{L}_T^{SR} = \prod_{t=1}^{T} \prod_{i=1}^{I} [& \sum_{k \in M} \frac{\exp(\chi_i^{(t)\mathsf{T}} \beta_k^{(t)})}{1 + \sum_{l \in M} \exp(\chi_i^{(t)\mathsf{T}} \beta_l^{(t)})} g(x_{t,i}; \vartheta_k^{(t)}) \\ & + (1 - \sum_{l \in M} \frac{\exp(\chi_i^{(t)\mathsf{T}} \beta_l^{(t)})}{1 + \sum_{m \in M} \exp(\chi_i^{(t)\mathsf{T}} \beta_m^{(t)})}) g(x_{t,i}; \vartheta_r^{(t)})]. \end{aligned}$$

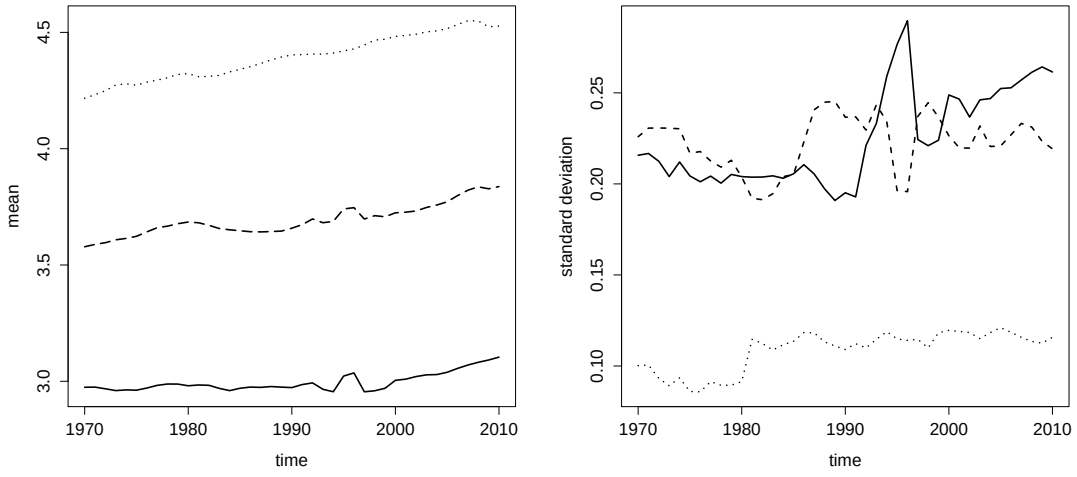
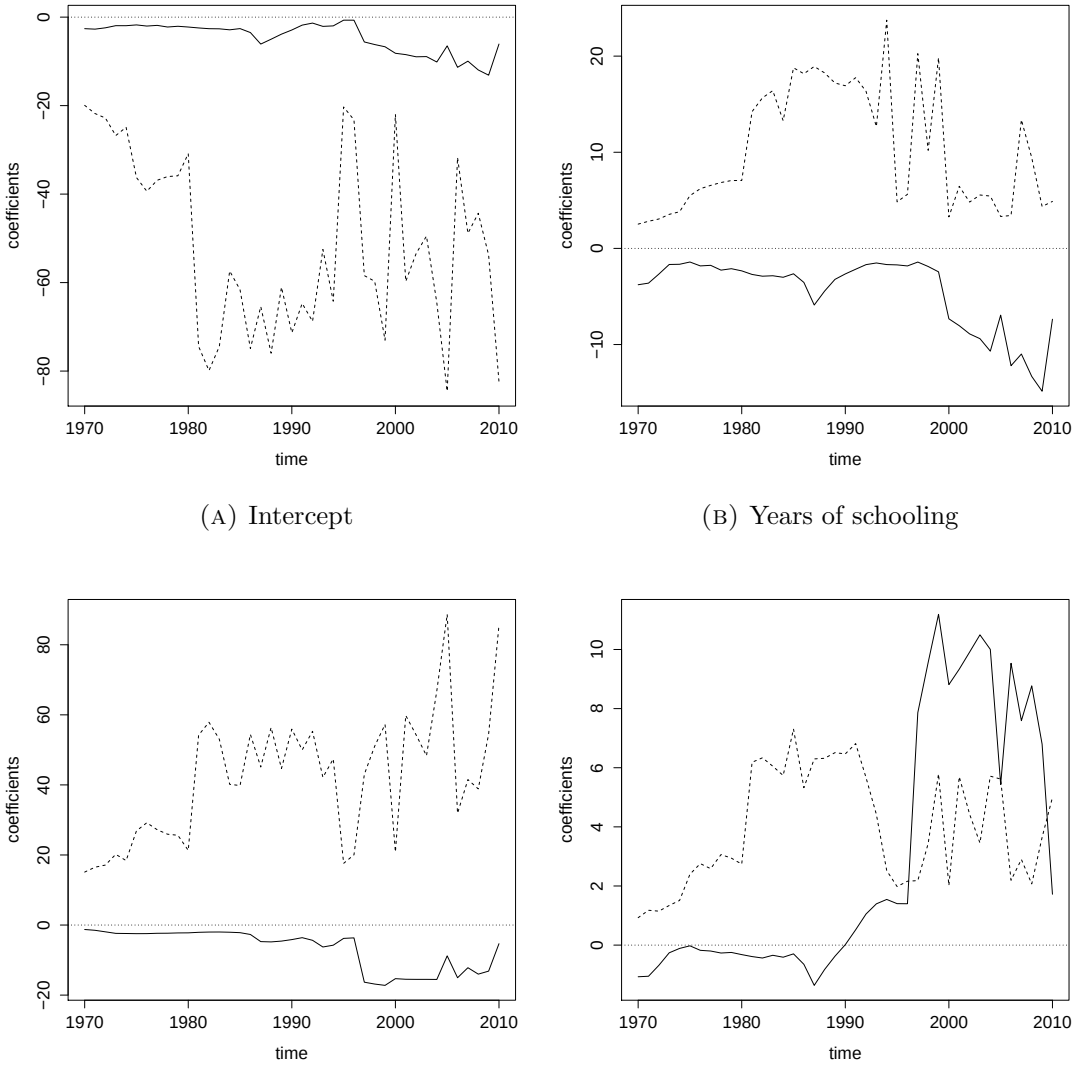


FIGURE 4.14.: Parameter estimates switching regression - left: means, right: standard deviations. Income group 1 (solid line), income group 2 (dashed line), income group 3 (dotted line).

Maximization is performed using an EM-algorithm where the group membership is the latent variable and due to the independence assumption, each year can be modelled separately. Details on the computation are given in Section 4.8.2.
The estimated parameters of the component-dependent distributions are shown in Figure 4.14. We observe that the estimated means are less volatile than in the mixture model, but of the same magnitude. In addition, the rising mean in the 1990s followed by a decline after 1996 combined with a high standard deviation reminds of the results in the mixture model and the modified hidden Markov models presented in the previous sections.



(A) Intercept

(B) Years of schooling

(C) Life expectancy

(D) Latitude

FIGURE 4.15.: Switching Regression: Estimated parameters in multinomial regression. Income group 1 (solid line), income group 3 (dashed line).

The estimated parameters of the multinomial regression are shown in Figure 4.15. Since income group 2 is the reference category of our model, the estimates can be interpreted relative to this income group. The estimates for variable *years of schooling* in income group 1 are negative, thus the odds for income group 1 relative to income group 2 decrease with increasing years of schooling, while the odds for income group 3 relative to income group 2 increase, due to the positive sign of the estimated parameters. The same effect holds for the variable *life expectancy*. For the variable *latitude* we observe that for income group 1 the sign of the parameter estimate changes from negative to positive in 1990. Thus, from 1970 to 1990, increasing latitude decreases the odds for income group 1 relative to income group 2, while after 1990, the odds increase. The parameter for income group 3 is positive from 1970–2010.

Based on the estimation results we perform maximum-aposteriori analysis and report the results in Section 4.9. We observe that compared to the results from the mixture model, the number of switches of income group decreased dramatically. One reason is the number of countries for which we observe data, which decreased from 152 in models without covariables to $I = 107$ when using covariables. In the mixture model we observed switches of income groups for 76 countries, 24 of these can not be modelled in the switching regression model due to missing data. 35 countries which switched income group at least once in the mixture model are assigned to a fix income group over the 41 years in the switching regression model. This effect might be a result of the parameter estimates which are much more stable, as mentioned above. In addition to the decreasing number of countries which switch income group, the number of back-and-forth-switches as observed very often in the mixture model decreased, but the effect still occurs (see for example China (CHN), Iraq (IRQ), Morocco (MAR), Nicaragua

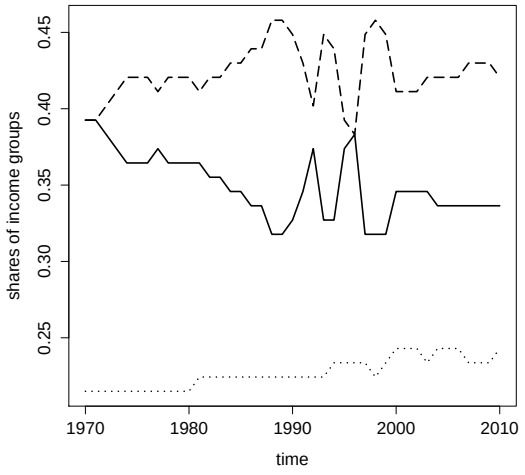


FIGURE 4.16.: Switching Regression: shares of income groups. Income group 1 (solid line), income group 2 (dashed line), income group 3 (dotted line).

(NIC), Portugal (PRT) and Vietnam (VNM)). Even though switching income group during the 41 years, 8 countries end up in the same group in 2010 as they started in 1970 and thus do not experience an advancement in the end. These are Republic of Congo (COG), Iraq (IRQ), Sri Lanka (LKA), Morocco (MAR), Mongolia (MNG), Nicaragua (NIC), Papua new guinea (PNG) and Swaziland (SWZ). It bears mentioning that the 9 countries which experience a switch of income group in the switching regression model all ascent. Namely these are Botswana (BWA), China (CHN), Cyprus (CYP), Egypt (EGY), Indonesia (IDN), Republic of Korea (KOR), Portugal (PRT), Thailand (THA) and Vietnam (VNM). Five of these countries were also modelled as one of the 24 ascending countries in the hidden Markov model (CHN, CYP, EGY, IDN and KOR). The hidden Markov model which modelled ascending and declining countries separately also covered the advancement of PRT in addition to these five countries.
The shares of income group are shown in Figure 4.16. We observe that income group 2 is the largest income group all over the time, except for 1998, when income group 1 and income group 2 have the same share. The share of income group 1 decreases from 1970 until 1989 and then fluctuates around 1/3. Over the 41 years, the share of income group 3 slightly rises from 21.5% to 25.3%.

4.6. Nonhomogeneous hidden Markov models with covariables

While in the previous section we used covariables to model the mixing probabilities and thus focused on the membership of countries to certain income groups, we now aim at investigating the influence of the chosen covariables on the switching behaviour of countries between income groups. Thus, we include covariables to the hidden Markov model from Section 4.3 for modelling the transition probabilities. Since splitting the hidden Markov model did not lead to significant advantages in Section 4.3, we now stick to the nonhomogeneous hidden Markov model for all countries.
Zucchini and MacDonald (2009) proposed an approach to include covariables in the transition probabilities of a two-state hidden Markov model using a logit model. Since we fit a three-state model, we need to extend this idea. In our context, transitions skipping one income group (i.e. $1 \to 3$ and $3 \to 1$) do not occur, thus we simplify our transition probability matrices to

$$\Gamma^{(t)} = \begin{pmatrix} \gamma_{1,1}^{(t)} & \gamma_{1,2}^{(t)} & 0 \\ \gamma_{2,1}^{(t)} & \gamma_{2,2}^{(t)} & \gamma_{2,3}^{(t)} \\ 0 & \gamma_{3,2}^{(t)} & \gamma_{3,3}^{(t)} \end{pmatrix}, \qquad t = 2, \ldots, T.$$

The parameters of the first and the third row can then be modelled using binary regression. Analogously to Section 4.5, let L denote the number of considered covariables and write $\chi_i^{(t)} = (\chi_{1,i}^{(t)}, \ldots, \chi_{L,i}^{(t)})^\intercal$ for the data of country $i \in \{1, \ldots, I\}$ at time $t \in \{1, \ldots, T\}$.

Set

$$\begin{aligned}\gamma_{1,2,i}^{(t)} &= \frac{\exp(\chi_i^{(t)\mathsf{T}}\beta_{(12)}^{(t)})}{1+\exp(\chi_i^{(t)\mathsf{T}}\beta_{(12)}^{(t)})}, \qquad \gamma_{1,1,i}^{(t)} = 1-\gamma_{1,2,i}^{(t)},\\ \gamma_{3,2,i}^{(t)} &= \frac{\exp(\chi_i^{(t)\mathsf{T}}\beta_{(32)}^{(t)})}{1+\exp(\chi_i^{(t)\mathsf{T}}\beta_{(32)}^{(t)})}, \qquad \gamma_{3,3,i}^{(t)} = 1-\gamma_{3,2,i}^{(t)},\end{aligned} \tag{4.2}$$

$t = 2, \dots, T$, $i = 1, \dots, I$. For the parameters of the second row, we use a multinomial regression, where the reference transition is to remain in state 2 (i.e. $\gamma_{2,2}^{(t)}$). Thus,

$$\begin{aligned}\gamma_{2,1,i}^{(t)} &= \frac{\exp(\chi_i^{(t)\mathsf{T}}\beta_{(21)}^{(t)})}{1+\exp(\chi_i^{(t)\mathsf{T}}\beta_{(21)}^{(t)})+\exp(\chi_i^{(t)\mathsf{T}}\beta_{(23)}^{(t)})},\\ \gamma_{2,3,i}^{(t)} &= \frac{\exp(\chi_i^{(t)\mathsf{T}}\beta_{(23)}^{(t)})}{1+\exp(\chi_i^{(t)\mathsf{T}}\beta_{(21)}^{(t)})+\exp(\chi_i^{(t)\mathsf{T}}\beta_{(23)}^{(t)})}\end{aligned} \tag{4.3}$$

and $\gamma_{2,2,i}^{(t)} = 1-\gamma_{2,1,i}^{(t)}-\gamma_{2,3,i}^{(t)}$ for $i = 1, \dots, I$ and $t = 2, \dots, T$.
In addition, we perform multinomial regression for the initial distribution, where income group 2 is the reference group:

$$\alpha_1 = \frac{\exp(\chi_i^{(1)\mathsf{T}}\beta_1^{(1)})}{1+\exp(\chi_i^{(1)\mathsf{T}}\beta_1^{(1)})+\exp(\chi_i^{(1)\mathsf{T}}\beta_3^{(1)})}, \quad \alpha_3 = \frac{\exp(\chi_i^{(1)\mathsf{T}}\beta_3^{(1)})}{1+\exp(\chi_i^{(1)\mathsf{T}}\beta_1^{(1)})+\exp(\chi_i^{(1)\mathsf{T}}\beta_3^{(1)})}, \tag{4.4}$$

$\alpha_2 = 1-\alpha_1-\alpha_3$.
In Section 4.3 we observed that transitions do not occur in each year from 1970 to 2010. Therefore, to reduce the complexity of the optimization problem, we define a set of years $\mathcal{T} \subset \{2, \dots, T\}$, where we estimate the transition probability matrices according to (4.2) and (4.3), whereas for the remaining years, we do not perform regression but rather estimate a transition probability matrix for all countries: $\gamma_{j,k,i}^{(t)} := \gamma_{j,k}^{(t)}$, $j,k = 1, \dots, K$, $i = 1, \dots, I$, $t \notin \mathcal{T}$. Thus, parameter estimation comprises estimation of $\beta_1, \beta_3, \beta_{(12)}^{(t)}, \beta_{(32)}^{(t)}, \beta_{(21)}^{(t)}, \beta_{(23)}^{(t)}$ ($t \in \mathcal{T}$), $\Gamma^{(t)}$ ($t \notin \mathcal{T}$) and $\vartheta_k^{(t)}$ ($k = 1, \dots, K$, $t = 1, \dots, T$), which yields $2L + 4L(\#\mathcal{T}) + (T-1-(\#\mathcal{T}))K(K-1) + 2KT$ parameters. We maximize the penalized log-likelihood function

$$\begin{aligned}\ell_T^{HMMreg} =& \sum_{i=1}^{I} \log\Big(\sum_{k_1=1}^{K} \dots \sum_{k_T=1}^{K} \alpha_{k_1} \prod_{t=2}^{T} \gamma_{k_{t-1},k_t,i}^{(t)} \prod_{t=1}^{T} g(x_{t,i}; \vartheta_{k_t}^{(t)})\Big)\\ &+ \frac{c}{KI} \sum_{i=1}^{I} \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \log(\gamma_{j,k,i}^{(t)}).\end{aligned}$$

Again, we modify the EM-algorithm to estimate the parameters of the model. Plugging

in the quantities (4.2), (4.3) and (4.4) into the complete-data log-likelihood function of the hidden Markov model, the regression parameters are estimated using numerical optimization of the respective part of the function. The parameters of the state-dependent Gaussian distributions as well as the transition probabilities for years not in $\mathcal{T}$ are estimated using closed formulas, see (4.5) in Section 4.8.

In this model, due to the large number of parameters, estimation is a complex problem. The number of parameters and the flexibility to fit a countryspecific transition probability matrix using covariables leads to significant differences compared to the hidden Markov model without covariables. The estimated parameters are unstable and do not yield very plausible a-posteriori results. Obviously, our dataset does not posses enough data to yield reasonable results for this complex model.

The estimated state-dependent parameters for the Gaussian distributions are given in Figure 4.17. The estimated means are similar those of the previous models. In 1999, the mean of income group 1 rises sharply and returns to the previous level in 2000. This observation together with the high standard deviation of income group 1 during this period leads to many switches of income group around the year 2000. Many countries switch from income group 2 to income group 1 in the late 1990s and switch back to income group 2 shortly afterwards (see for example Albania (ALB), Bulgaria (BGR), Belize (BLZ), Costa Rica (CRI), Cuba (CUB), Iran (IRN), Jamaica (JAM), Jordan (JOR), Mexico (MEX), Malaysia (MYS), Panama (PAN), Sri Lanka (LKA) in Table 4.2). In contrast to previous models, in Figure 4.18 we observe that from 1970 to 2001 income group 1 is the largest income group. The phenomenon described above is also visible, since in the late 1990s the share of income group 1 rises sharply, while the share

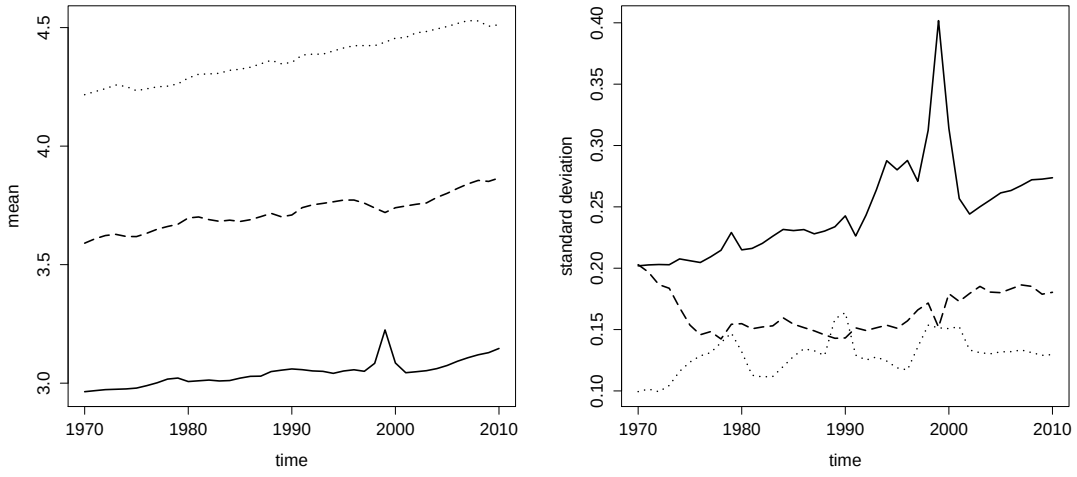


FIGURE 4.17.: Parameter estimates hidden Markov model with covariables - left: means, right: standard deviations. Income group 1 (solid line), income group 2 (dashed line), income group 3 (dotted line).

of income group 2 drops and in 2001 both groups equalize. In 1999, there are even more countries assigned to income group 3 than to income group 2.
In total, 26 countries switch income group at least once during the time from 1970 to 2010, with 16 countries switching income group during that time, but returning to the same income group as they started in 1970. 10 countries of those 16 switch up and down due to the estimated mean of income group 1 in 1999.
Taken as a whole, much more countries switch income group more than once, compared to the hidden Markov model from Section 4.3. This might be due to the fact that the regression approach allows for individual transition probabilities for all countries.
In total, 8 countries experience advancement to a higher income group, namely China (CHN), Egypt (EGY), Hungary (HUN), Korea (KOR), Sri Lanka (LKA), Portugal (PRT), Swaziland (SWZ) and Thailand (THA). Nicaragua (NIC) is the only country which ends up in a lower income group than it started in 1970 (decline from group 2 in 1987 to group 1 in 1988). These effects were also observed in several previous models.
The estimated coefficients for the multinomial regression of the initial distribution show that the variable *years of schooling* increases the odds of group 3 relative to group 2, while the odds of group 1 relative to group 2 decrease. The same effect holds for the variable *life expectancy*. For the variable *latitude* both coefficients are positive, while the odds of income group 3 relative to group 2 grow faster than the odds for group 1 relative to group 2.
For the estimated coefficients of the regression models concerning the transition probabilities for selected years, we observe that not every regression yields reasonable results for our model. Since some coefficients vary strongly over time, a meaningful interpretation is hardly possible.

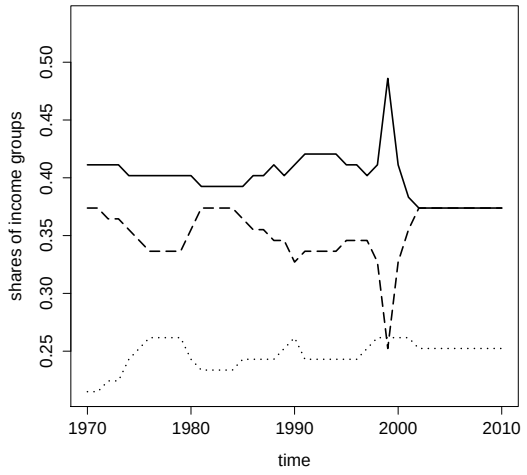


FIGURE 4.18.: Hidden Markov model with covariables: shares of income groups. Income group 1 (solid line), income group 2 (dashed line), income group 3 (dotted line).

Concerning transition probability $\gamma_{1,2}^{(t)}$ estimated coefficients for the variable *life expectancy* has a positive sign for all considered years, thus the odds for an advancement from group 1 to group 2 relative to remaining in group 1 tend to increase when life expectancy improves. For almost every year, regression coefficients of the variables *years of schooling* and *latitude* are positive too, which leads to the same interpretation. But somehow there are selected years when the regression coefficients are negative, which is possibly due to numerical problems during the optimization process.
The coefficients of the binary logit regression concerning transition probability $\gamma_{3,2}^{(t)}$ are barely interpretable in a reasonable way, since signs are switching permanently. Especially for the variable *years of schooling*, there is a switch of sign almost every year. Anyhow, switches from income group 3 to income group 2 are very rarely observed, only HUN, IRN, POL and VEN are affected, and for all of these countries it is only a temporal effect since they switch back to group 3 later on.
Similar problems occur when considering the estimated coefficients for the multinomial regression concerning transition starting from income group 2, when transition $2 \to 2$ serves as reference. Regarding variable *years of schooling*, we observe that the odds for a transition from group 2 to group 3 relative to remaining in group 2 increase with growing number of years of schooling, since the estimated parameter possesses positive sign. The coefficients for the transition from group 2 to group 1 are not so clearly interpretable, since starting with a positive sign in the 1970s, in the 1980s, 1990s and 2000s the estimated coefficients have negative sign, while switching to positive again in between. The same holds for variable *life expectancy* and transition $2 \to 1$: the coefficients are hardly informative due to the permanently changing sign, while the odds of transition $2 \to 3$ relative to $2 \to 2$ decrease with increasing life expectancy (except for year 2000, where the estimated coefficient is positive, possibly due to numerical issues). While the coefficients for variable *latitude* concerning the odds of transition $2 \to 3$ relative to $2 \to 2$ are positive in the beginning and ending of the observed period of time (1975–1984 and 1998–2005), they switch sign in between and thus do not give a clear interpretation. The odds of transition $2 \to 1$ relative to $2 \to 2$ increase with increasing latitude.

In this section, we provided an approach to include covariables into the nonhomogeneous hidden Markov model and experienced that parameter estimation in the resulting model is rather complex and not always easy to interpret. For the concrete example of modelling income distributions one might need to include different or more covariables, possibly depending on the income group from which transitions start, to gain more insight. In addition, to handle the complex optimization problem, a larger database is required.

4.7. Conclusion

We introduced four different approaches to model monetary welfare distributions in the context of economic growth. When considering a mixture model, we experienced that the underlying independence structure lead to quite volatile parameter estimates and many redundant switches of countries between the three income groups. The hidden Markov models in Section 4.3 revealed more stable parameter estimates over time and reduced the number of switches between income groups. It turned out that the dependence over time is crucial to gain reasonable a-posteriori results regarding the assignment of the countries to income groups. Ideas to extend the model for a better handling of the given data did not achieve significant effort.
In order to explain the results we observed from classification in mixture models and hidden Markov models, we included covariables to both approaches. In the literature Bloom et al. (2003) focused on geographical, cultural and climatic factors, while Owen et al. (2009) investigated the influence of institutional variables. We considered several variables and observed that information on education, life expectancy and geography of the countries seems to be more influential on the income than the investment share of GDP or the fertility rate. The resulting switching regression model improved the results from the mixture model: the parameter estimates were much more stable and the number of redundant switches between income groups reduced. In addition, the estimated regression coefficients yielded reasonable results when explaining the countries' membership to the income groups. The inclusion of covariables to the hidden Markov model appeared to be much more complex. Combining the extensions of the model which consider dependence over time and covariables is a difficult approach, since the result is a model with a huge number of parameters. Our data set did not contain enough observations to handle this complexity, which made reasonable interpretation of the results nearly impossible.

Altogether, when considering Tables 4.1 and 4.2 classification results of all models were very similar. Advantages of the models with underlying Markov dependence structure were the stable classification results, avoiding many group changes, while the independence structure of mixture models and switching regression facilitated many transitions between income groups.

All models coincided in the result that in all three income groups the mean income grew from 1970 to 2010. However, some countries benefited more or less than the remaining countries of the respective groups and thus switched to a higher or lower income group. For some countries all considered models agreed on certain switches of income groups. An advancement from group 1 to group 2 was the result for China (late 1990s or around the year 2000), Egypt (during the 1990s), Indonesia (during the 1980/90s, except for the HMM with covariables) and Bhutan (around 2000, due to missing data

only in models without covariables). The opposite movement from group 2 to group 1 was experienced by Nicaragua (during the 1980/90s except for the Switching Regression model) and Djibouti (end 1980s, due to missing data only in models without covariables). In addition, a couple of countries advanced from group 2 to group 3. These were Cyprus (between 1985 and 1991), Republic of Korea (during the 1990s) and the following countries, which were only considered in models without covariables: Malta (end of 1980s/during the 1990s), Oman (before 1990) and Taiwan (during the 1980/90s). Moreover, in the models without covariables, Palau switched from income group 3 to group 2 in the 1990s.

4.8. Modifications of the EM-algorithm

In this section we provide details on the modifications of the general EM-algorithm described in Chapter 1, to take into account the structures of the models from Sections 4.3–4.6.

4.8.1. Nonhomogeneous hidden Markov model

The EM-algorithm for the nonhomogeneous hidden Markov model introduced in Section 4.3 needs to take into account the longitudinal data structure, the nonhomogeneity of the model and the introduced penalization of the transition probabilities.

At first, analogously to (1.6) and (1.7) we calculate forward and backward probabilities $a_{tk,i}$, $b_{tk,i}$ $(t = 1, \ldots, T,\ k = 1, \ldots, K)$ for each country $i = 1, \ldots, I$, taking into account the time-dependent parameters of the model. Based on these quantities, the E-step can be performed for each country according to (1.13), yielding conditional expectations $\hat{u}_{kt,i}$ $(t = 1, \ldots, T)$, $\hat{v}_{klt,i}$ $(t = 2, \ldots, T)$, $i = 1, \ldots, I$, $k, l = 1, \ldots, K$. In the M-step, partial differentiating of the conditional expectation of the complete-data log-likelihood function yields

$$\begin{aligned} \hat{\alpha}_k &= \frac{1}{I}\sum_{i=1}^{I} \hat{u}_{k1,i}, \qquad \hat{\gamma}_{k,l}^{(t)} = \frac{\sum_{i=1}^{I} \hat{v}_{jkt,i} + \frac{c}{K}}{\sum_{i=1}^{I}(\sum_{j=1}^{K} \hat{v}_{kjt,i} + \frac{c}{K})}, \qquad t = 2, \ldots, T \\ \hat{\mu}_k^{(t)} &= \frac{\sum_{i=1}^{I} \hat{u}_{kt,i} x_{t,i}}{\sum_{i=1}^{I} \hat{u}_{kt,i}}, \qquad \hat{\sigma}_k^{(t)} = \sqrt{\frac{\sum_{i=1}^{I} \hat{u}_{kt,i} x_{t,i}^2}{\sum_{i=1}^{I} \hat{u}_{kt,i}} - \hat{\mu}_k^2}, \qquad t = 1, \ldots, T, \end{aligned} \tag{4.5}$$

$k, l = 1, \ldots, K$.

Switching model: Hidden Markov model versus fixed state model

We introduce modifications of the algorithm, when each country can either run in a hidden Markov model or stay in one fixed state. The introduced variables π_i are treated as

additional latent variable, which is assumed to be independent of $S_{t,i}$. The introduction of indicator variables $w_i = \mathbb{1}_{\{\pi_i=1\}}$, $i = 1, \ldots, I$, yields a complete-data log-likelihood function of the form

$$\begin{aligned}\ell_T^{(SHMM)c} = &\sum_{i=1}^{I}\sum_{k=1}^{K} w_i u_{k1,i} \log(\alpha_k) + \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{t=2}^{T} w_i v_{klt,i} \log(\gamma_{k,l}^{(t)}) \\ &+ \frac{c}{KI}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{t=2}^{T} w_i \log(\gamma_{k,l}^{(t)}) + \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} w_i u_{kt,i} \log(g(x_{t,i}; \vartheta_k^{(t)})) \\ &+ \sum_{i=1}^{I}(1 - w_i) \log(\max_{k=1,\ldots,K} \prod_{t=1}^{T} g(x_{t,i}; \vartheta_k^{(t)})).\end{aligned}$$

In the E-step, next to the calculation of the conditional expectations $\hat{u}_{kt,i}$ and $\hat{v}_{klt,i}$ as before, we calculate $\hat{w}_i = P(\pi_i = 1 \mid x)$. Plugging in these quantities, the M-step can be split into three maximization problems. Partial differentiating yields for $k, l = 1, \ldots, K$

$$\hat{\alpha}_k = \frac{\sum_{i=1}^{I} \hat{w}_i \hat{u}_{k1,i}}{\sum_{i=1}^{I} \hat{w}_i}, \qquad \hat{\gamma}_{k,l}^{(t)} = \frac{\sum_{i=1}^{I} \hat{w}_i (\hat{v}_{klt,i} + \frac{c}{KI})}{\sum_{j=1}^{K}\sum_{i=1}^{I} \hat{w}_i (\hat{v}_{kjt,i} + \frac{c}{KI})}, \qquad t = 2, \ldots, T.$$

Since the maximum function is not differentiable, we perform numerical maximization of the remaining part of the conditional expectation of the complete-data log-likelihood function

$$\sum_{i=1}^{I}(\hat{w}_i \sum_{t=1}^{T}\sum_{k=1}^{K} \hat{u}_{kt,i} \log(g(x_{t,i}; \vartheta_k^{(t)}))) + (1 - \hat{w}_i) \log(\max_{k=1,\ldots,K} \prod_{t=1}^{T} g(x_{t,i}; \vartheta_k^{(t)}))),$$

in order to obtain $\hat{\mu}_k^{(t)}$, $\hat{\sigma}_k^{(t)}$, $k = 1, \ldots, K$, $t = 1, \ldots, T$.

Split model: Separated hidden Markov models for advancement and decline

For the model consisting of three parts as described in Section 4.3, we perform the following modification of the EM-algorithm.
For the model-selection variable π_{im}, we introduce indicator variables

$$\begin{aligned} u_{kt,i}^{(1)} &= \mathbb{1}_{\{s_{t,i}=k|\pi_{i1}=1\}}, & u_{kt,i}^{(2)} &= \mathbb{1}_{\{s_{t,i}=k|\pi_{i2}=1\}} \qquad (t = 1, \ldots, T), \\ v_{klt,i}^{(1)} &= \mathbb{1}_{\{s_{t-1,i}=k, s_{t,i}=l|\pi_{i1}=1\}}, & v_{klt,i}^{(2)} &= \mathbb{1}_{\{s_{t-1,i}=k, s_{t,i}=l|\pi_{i2}=1\}} \qquad (t = 2, \ldots, T) \end{aligned}$$

$(k, l = 1, \ldots, K)$, for the latent variable regarding the income group of country i and $w_{im} = \mathbb{1}_{\{\pi_{im}=1\}}$ for the latent variable regarding the part of the model, $m = 1, 2, 3$,

$i = 1, \ldots, I$. This leads to the complete-data log-likelihood function

$$\begin{aligned}\ell_T^{(HMMud)c} =& \sum_{i=1}^{I}\sum_{k=1}^{K}(w_{i1}u_{k1,i}^{(1)} + w_{i2}u_{k1,i}^{(2)})\log(\alpha_k) + \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=k}^{K}\sum_{t=2}^{T} w_{i1}v_{klt,i}^{(1)}\log(\gamma_{k,l}^{(t,1)}) \\ &+ \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{k}\sum_{t=2}^{T} w_{i2}v_{klt,i}^{(2)}\log(\gamma_{k,l}^{(t,2)}) \\ &+ \frac{c}{\sqrt{K^2+K-2I}}\sum_{i=1}^{I}\sum_{t=2}^{T}\sum_{k=1}^{K}(\sum_{l=k}^{K} w_{i1}\log(\gamma_{k,l}^{(t,1)}) + \sum_{l=1}^{k} w_{i2}\log(\gamma_{k,l}^{(t,2)})) \\ &+ \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{t=1}^{T}((w_{i1}u_{kt,i}^{(1)} + w_{i2}u_{kt,i}^{(2)})\log(g(x_{t,i};\vartheta_k^{(t)}))) \\ &+ \sum_{i=1}^{I} w_{i3}\log(\max_{k=1,\ldots,K}\prod_{t=1}^{T} g(x_{t,i};\vartheta_k^{(t)}))).\end{aligned}$$

The E-step works analogously to the algorithms introduced before and in the M-step, for the initial distribution and the transition probabilities we have

$$\begin{aligned}\hat{\alpha}_k &= \frac{\sum_{i=1}^{I}(\hat{w}_{i1}\hat{u}_{k1,i}^{(1)} + \hat{w}_{i2}\hat{u}_{k1,i}^{(2)})}{\sum_{i=1}^{I}\sum_{l=1}^{K}(\hat{w}_{i1}\hat{u}_{l1,i}^{(1)} + \hat{w}_{i2}\hat{u}_{l1,i}^{(2)})}, \qquad k = 1, \ldots, K, \\ \hat{\gamma}_{j,k}^{(t,1)} &= \frac{\sum_{i=1}^{I}\hat{w}_{i1}(\hat{v}_{jkt,i}^{(1)} + \frac{c}{\sqrt{K^2+K-2I}})}{\sum_{l=j}^{K}\sum_{i=1}^{I}\hat{w}_{i1}(\hat{v}_{jlt,i}^{(1)} + \frac{c}{\sqrt{K^2+K-2I}})}, \qquad k = j, \ldots, K, \\ \hat{\gamma}_{j,k}^{(t,2)} &= \frac{\sum_{i=1}^{I}\hat{w}_{i2}(\hat{v}_{jkt,i}^{(2)} + \frac{c}{\sqrt{K^2+K-2I}})}{\sum_{l=1}^{j}\sum_{i=1}^{I}\hat{w}_{i2}(\hat{v}_{jlt,i}^{(2)} + \frac{c}{\sqrt{K^2+K-2I}})}, \qquad k = 1, \ldots, j,\end{aligned}$$

while for $k = 1, \ldots, K$, $t = 1, \ldots, T$ the parameters $\hat{\vartheta}_k^{(t)}$ are calculated using numerical optimization of

$$\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{t=1}^{T}((\hat{w}_{i1}\hat{u}_{kt,i}^{(1)} + \hat{w}_{i2}\hat{u}_{kt,i}^{(2)})\log(g(x_{t,i};\vartheta_k^{(t)}))) + \sum_{i=1}^{I}(\hat{w}_{i3}\log(\max_{k=1,\ldots,K}\prod_{t=1}^{T} g(x_{t,i};\vartheta_k^{(t)}))).$$

4.8.2. Switching Regression

For the switching regression model formulated in Section 4.5 we introduce indicator variables $u_{k,i}^{(t)} = \mathbb{1}_{\{\text{country } i \text{ belongs to income group } k \text{ at time } t\}}$, which yields the complete-data log-likelihood function

$$\ell_T^{SR,c} = \sum_{i=1}^{I}\sum_{k=1}^{K} u_{k,i}^{(t)}\log(\pi_{k,i}^{(t)}) + \sum_{i=1}^{I}\sum_{k=1}^{K} u_{k,i}^{(t)}\log(g(x_{t,i};\vartheta_k^{(t)})), \qquad t = 1, \ldots, T.$$

In the E-step we use the current parameter estimates to calculate

$$\hat{u}_{k,i}^{(t)} = \frac{\pi_{k,i}^{(t)} g(x_{t,i}; \vartheta_k^{(t)})}{\sum_{l=1}^{K} \pi_{l,i}^{(t)} g(x_{t,i}; \vartheta_l^{(t)})}, \quad k = 1, \ldots, K, i = 1, \ldots, I \text{ and } t = 1, \ldots, T.$$

In the M-step, the optimal parameters for the income group-dependent parameters of the Gaussian distributions are

$$\hat{\mu}_k^{(t)} = \frac{\sum_{i=1}^{I} \hat{u}_{k,i}^{(t)} x_{t,i}}{\sum_{i=1}^{I} \hat{u}_{k,i}^{(t)}} \qquad \text{and} \qquad \hat{\sigma}_k^{(t)2} = \frac{\sum_{i=1}^{I} \hat{u}_{k,i}^{(t)} (x_{t,i} - \hat{\mu}_k^{(t)})^2}{\sum_{i=1}^{I} \hat{u}_{k,i}^{(t)}},$$

while the regression parameters are estimated using numerical optimization of the first part of conditional expectation of the complete-data log-likelihood function with (4.1) plugged in.

4.9. Classification results

The following tables illustrate the classification results and potential switches of income groups together with the corresponding years of the considered models. At first, in Table 4.1 we report countries, which do not switch income group in all considered models together with the corresponding income groups. Table 4.2 provides the results on switches of income groups for the remaining countries. The '-' indicates those countries for which some data for the models with covariables were missing.

Income group 1	AFG, BDI, BEN, BFA, BGD, CAF, COM, ETH, GIN, GMB, GNB, KEN, KHM, LAO, LBR, LSO, MDG, MLI, MOZ, MWI, NER, NPL, RWA, SEN, SLE, SOM, TCD, TGO, TZA, UGA, ZAR, ZWE
Income group 2	BRA, CHL, COL, DMA, DOM, GRD, GTM, KNA, LCA, MHL, MUS, ROM, TON, TUR, URY, VUT, ZAF
Income group 3	AUS, AUT, BEL, BHS, BMU, BRB, CAN, CHE, DEU, DNK ESP, FIN, FRA, GBR, ISL, ITA, JPN, NLD, NZL, SWE, USA

TABLE 4.1.: Classification results of the different models for panel data: Countries which do not switch income group.

	AGO	ALB	ARG	ATG	BGR
Mixture	80-81: 2→1 82-83: 1→2 85-86: 2→1 87-88: 1→2 89-90: 2→1 97-98: 1→2 99-00: 2→1 04-05: 1→2	90-91: 2→1 96-97: 1→2	73-74: 2→3 75-76: 3→2	95-96: 2→3 96-97: 3→2	2
HMM	77-78: 2→1 04-05: 1→2	2	2	2	2
Split HMM	79-80: 2→1	2	2	2	2
S. Regression	-	2	2	-	2
HMM Regr.	-	97-98: 2→1 98-99: 1→2 99-00: 2→1 00-01: 1→2	2	-	97-98: 2→1 99-00: 1→2

	BLZ	BOL	BTN	BWA	CHN
Mixture	2	90-91: 2→1 96-97: 1→2	97-98: 1→2 99-00: 2→1 00-01: 1→2	71-72: 1→2	97-98: 1→2
HMM	2	2	00-01: 1→2	2	99-00: 1→2
Split HMM	2	2	1	2	97-98: 1→2
S. Regression	2	2	-	72-73: 1→2	85-86: 1→2 90-91: 2→1 92-93: 1→2 94-95: 2→1 96-97: 1→2
HMM Regr.	98-99: 2→1 99-00: 1→2	2	-	2	99-00: 1→2

	CIV	CMR	COG	CPV	CRI
Mixture	75-76: 1→2 76-77: 2→1 78-79: 1→2 79-80: 2→1 97-98: 1→2 99-00: 2→1	72-73: 1→2 73-74: 2→1 75-76: 1→2 89-90: 2→1 97-98: 1→2 99-00: 2→1	72-73: 1→2 73-74: 2→1 75-76: 1→2 76-77: 2→1 78-79: 1→2 90-91: 2→1 97-98: 1→2 99-00: 2→1	97-98: 1→2 99-00: 2→1 01-02: 1→2 02-03: 2→1 05-06: 1→2	2
HMM	1	1	1	1	2
Split HMM	1	1	1	1	2

S. Regression	1	1	81-82: 1→2 86-87: 2→1	-	2
HMM Regr.	1	1	1	-	99-00: 2→1 00-01: 1→2

	CUB	CYP	DJI	DZA	ECU
Mixture	2	90-91: 2→3 96-97: 3→2 00-01: 2→3 02-03: 3→2	89-90: 2→1 97-98: 1→2 99-00: 2→1	94-95: 2→1 96-97: 1→2	95-96: 2→1 96-97: 1→2
HMM	2	86-87: 2→3	88-89: 2→1	2	2
Split HMM	2	90-91: 2→3	88-89: 2→1	2	2
S. Regression	2	80-81: 2→3	-	2	2
HMM Regr.	98-99: 2→1 99-00: 1→2	84-85: 2→3	-	2	2

	EGY	FJI	FSM	GHA	GRC
Mixture	82-83: 1→2 90-91: 2→1 96-97: 1→2	91-92: 2→1 96-97: 1→2	90-91: 2→1 96-97: 1→2	72-73: 1→2 73-74: 2→1 98-99: 1→2 99-00: 2→1	97-98: 3→2 99-00: 2→3
HMM	88-89: 1→2	2	2	1	3
Split HMM	96-97: 1→2	2	2	1	3
S. Regression	83-84: 1→2	-	-	1	3
HMM Regr.	94-95: 1→2	-	-	1	3

	GUY	HKG	HND	HTI	HUN
Mixture	89-90: 2→1 96-97: 1→2	73-74: 2→3 75-76: 3→2 76-77: 2→3 78-79: 3→2 79-80: 2→3	90-91: 2→1 96-97: 1→2	78-79: 1→2 80-81: 2→1	73-74: 2→3 75-76: 3→2 81-82: 2→3 82-83: 3→2
HMM	2	3	2	1	90-91: 3→2
Split HMM	2	76-77: 2→3	2	1	2
S. Regression	2	-	2	1	2
HMM Regr.	85-86: 2→1 01-02: 1→2	-	2	1	74-75: 2→3 90-91: 3→2 96-97: 2→3

	IDN	IND	IRL	IRN	IRQ
Mixture	87-88: 1→2 90-91: 2→1 96-97: 1→2	97-98: 1→2 99-00: 2→1 08-09: 1→2	71-72: 2→3 72-73: 3→2 73-74: 2→3 75-76: 3→2 76-77: 2→3	73-74: 2→3 75-76: 3→2 76-77: 2→3 77-78: 3→2	90-91: 2→1 96-97: 1→2 02-03: 2→1 03-04: 1→2

HMM	92-93: 1→2	1	3	78-79: 3→2	90-91: 2→1
Split HMM	96-97: 1→2	1	3	76-77: 3→2	90-91: 2→1
S. Regression	87-88: 1→2	1	3	2	90-91: 2→1 92-93: 1→2 94-95: 2→1 96-97: 1→2
HMM Regr.	1	1	3	71-72: 2→3 79-80: 3→2 98-99: 2→1 00-01: 1→2	89-90: 2→1 96-97: 1→2

	ISR	JAM	JOR	KIR	KOR
Mixture	97-98: 3→2 99-00: 2→3	73-74: 2→3 75-76: 3→2	90-91: 2→1 96-97: 1→2	73-74: 2→3 75-76: 3→2 91-92: 2→1 96-97: 1→2	91-92: 2→3 97-98: 3→2 99-00: 2→3
HMM	3	75-76: 3→2	2	2	89-90: 2→3
Split HMM	3	2	2	2	90-91: 2→3
S. Regression	3	2	2	-	93-94: 2→3 97-98: 3→2 98-99: 2→3
HMM Regr.	3	98-99: 2→1 99-00: 1→2	98-99: 2→1 99-00: 1→2	-	89-90: 2→3

	LBN	LKA	LUX	MAC	MAR
Mixture	72-73: 3→2 73-74: 2→3 75-76: 3→2	96-97: 1→2	97-98: 3→2 99-00: 2→3	73-74: 2→3 75-76: 3→2 76-77: 2→3 78-79: 3→2 79-80: 2→3 98-99: 3→2 99-00: 2→3	75-76: 1→2 80-81: 2→1 81-82: 1→2 90-91: 2→1 97-98: 1→2 99-00: 2→1 05-06: 1→2
HMM	75-76: 3→2	99-00: 1→2	3	3	1
Split HMM	74-75: 3→2	1	3	73-74: 2→3	1
S. Regression	-	76-77: 2→1 77-78: 1→2	3	-	87-88: 1→2 91-92: 2→1 92-93: 1→2 94-95: 2→1 96-97: 1→2 99-00: 2→1
HMM Regr.	-	97-98: 1→2 98-99: 2→1 99-00: 1→2	3	-	1

	MDV	MEX	MLT	MNG	MRT
Mixture	88-89: 1→2 89-90: 2→1 96-97: 1→2	73-74: 2→3 75-76: 3→2	90-91: 2→3 97-98: 3→2 99-00: 2→3 05-06: 3→2 06-07: 2→3	71-72: 1→2 73-74: 2→1 75-76: 1→2 90-91: 2→1 97-98: 1→2 99-00: 2→1 00-01: 1→2	75-76: 1→2 76-77: 2→1 78-79: 1→2 79-80: 2→1 97-98: 1→2 99-00: 2→1
HMM	97-98: 1→2	2	86-87: 2→3	90-91: 2→1	1
Split HMM	96-97: 1→2	2	90-91: 2→3	90-91: 2→1	1
S. Regression	-	2	-	91-92: 2→1 03-04: 1→2	1
HMM Regr.	-	98-99: 2→1 99-00: 1→2	-	88-89: 1→2 90-91: 2→1	1

	MYS	NAM	NGA	NIC	NOR
Mixture	2	91-92: 2→1 96-97: 1→2	71-72: 1→2 73-74: 2→1 75-76: 1→2 77-78: 2→1 78-79: 1→2 79-80: 2→1	89-90: 2→1 97-98: 1→2 99-00: 2→1	97-98: 3→2 98-99: 2→3
HMM	2	2	1	88-89: 2→1	3
Split HMM	2	2	1	88-89: 2→1	3
S. Regression	2	2	-	91-92: 2→1 92-93: 1→2 94-95: 2→1 96-97: 1→2	3
HMM Regr.	98-99: 2→1 99-00: 1→2	2	-	87-88: 2→1	3

	OMN	PAK	PAN	PER	PHL
Mixture	74-75: 2→3 75-76: 3→2 81-82: 2→3 82-83: 3→2 83-84: 2→3 86-87: 3→2 90-91: 2→3 97-98: 3→2 00-01: 2→3 02-03: 3→2 07-08: 2→3	97-98: 1→2 99-00: 2→1	2	94-95: 2→1 96-97: 1→2	71-72: 1→2 73-74: 2→1 74-75: 1→2 90-91: 2→1 96-97: 1→2 99-00: 2→1 01-02: 1→2 02-03: 2→1 05-06: 1→2 07-08: 2→1 09-10: 1→2
HMM	74-75: 2→3	1	2	2	83-84: 2→1

Split HMM	80-81: 2→3	1	2	2	84-85: 2→1
S. Regression	-	1	2	2	2
HMM Regr.	-	1	98-99: 2→1 99-00: 1→2	2	1

	PLW	PNG	POL	PRI	PRT
Mixture	72-73: 3→2 73-74: 2→3 75-76: 3→2 76-77: 2→3 78-79: 3→2 81-82: 2→3 82-83: 3→2 89-90: 2→3 92-93: 3→2 94-95: 2→3 96-97: 3→2	80-81: 2→1 82-83: 1→2 83-84: 2→1 88-89: 1→2 89-90: 2→1 97-98: 1→2 99-00: 2→1	74-75: 2→3 75-76: 3→2	73-74: 2→3 75-76: 3→2 76-77: 2→3 78-79: 3→2 80-81: 2→3 82-83: 3→2 84-85: 2→3	73-74: 2→3 75-76: 3→2 81-82: 2→3 82-83: 3→2 90-91: 2→3 97-98: 3→2 99-00: 2→3 05-06: 3→2
HMM	98-99: 3→2	77-78: 2→1	2	3	3
Split HMM	97-98: 3→2	79-80: 2→1	2	3	89-90: 2→3
S. Regression	-	96-97: 1→2 99-00: 2→1	2	-	99-00: 2→3 02-03: 3→2 03-04: 2→3 06-07: 3→2 09-10: 2→3
HMM Regr.	-	1	75-76: 2→3 80-81: 3→2 88-89: 2→3 90-91: 3→2 97-98: 2→3 01-02: 3→2	-	73-74: 2→3

	PRY	SDN	SGP	SLB	SLV
Mixture	91-92: 2→1 96-97: 1→2	98-99: 1→2 99-00: 2→1	73-74: 2→3 75-76: 3→2 76-77: 2→3 78-79: 3→2 79-80: 2→3	78-79: 1→2 80-81: 2→1 97-98: 1→2 99-00: 2→1	95-96: 2→1 96-97: 1→2
HMM	2	1	3	1	2
Split HMM	2	1	3	1	2
S. Regression	2	1	-	-	2
HMM Regr.	2	1	-	-	2

	STP	SWZ	SYC	SYR	THA
Mixture	75-76: 1→2 80-81: 2→1	71-72: 1→2 91-92: 2→1 96-97: 1→2	73-74: 2→3 75-76: 3→2 76-77: 2→3	71-72: 1→2 90-91: 2→1 96-97: 1→2	71-72: 1→2 73-74: 2→1 75-76: 1→2
HMM	1	2	3	2	2
Split HMM	1	2	3	2	2
S. Regression	-	73-74: 1→2 95-96: 2→1 96-97: 1→2 99-00: 2→1	-	2	71-72: 1→2
HMM Regr.	-	73-74: 1→2	-	2	80-81: 1→2

	TUN	TWN	VCT	VEN	VNM
Mixture	94-95: 2→1 96-97: 1→2	90-91: 2→3 97-98: 3→2 99-00: 2→3	74-75: 2→1 75-76: 1→2 94-95: 2→1 96-97: 1→2	73-74: 2→3 75-76: 3→2 76-77: 2→3 77-78: 3→2	97-98: 1→2 99-00: 2→1
HMM	2	86-87: 2→3	2	79-80: 3→2	1
Split HMM	2	90-91: 2→3	2	2	1
S. Regression	2	-	-	2	86-87: 1→2 89-90: 2→1 92-93: 1→2 94-95: 2→1 96-97: 1→2
HMM Regr.	2	-	-	73-74: 2→3 79-80: 3→2	1

	WSM	ZMB			
Mixture	94-95: 2→1 96-97: 1→2	71-72: 1→2 73-74: 2→1			
HMM	2	1			
Split HMM	2	1			
S. Regression	-	1			
HMM Regr.	-	1			

Table 4.2.: Classification results of the different models for panel data: Countries which switch income group.

Discussion and outlook

In this work, several modifications of standard hidden Markov models were considered and theoretical and computational results for maximum likelihood estimation in the presented settings were given. The suggested models illustrated the flexibility of hidden Markov models and their adaptability to various settings for many classes of data. The theoretical results were mainly based on maximum likelihood theory for parametric models, given by Leroux (1992a), which were extended and adjusted when necessary.
The proposed nonparametric classes for the state-dependent densities gave theoretical justification for the computation of maximum likelihood estimators in more general settings than proposed in the literature so far. Thus, existing nonparametric estimation theory for hidden Markov models using Bayesian methods (Vernet, 2015), least squares estimation (De Castro et al., 2015) or orthogonal-series density estimation (Robin et al., 2014) was extended by theoretical results in a maximum likelihood context. While for state-dependent mixtures, a consistency result was given, for log-concave densities this question remained unanswered. Another open issue is the consideration of rates of convergence for the estimator. De Castro et al. (2015) used an approach by Massart (2007) to develop rates for their penalized least squares estimator in nonparametric hidden Markov models, whereas up to now there seem to be no results for the nonparametric maximum likelihood estimator. A first approach on that issue could be to examine a blockwise log-likelihood function first, which means to build blocks $y_i = (x_{iN+1}, \dots, x_{iN+N})$ $(i = 0, \dots, T-1)$ of length $N \in \mathbb{N}$ with joint distribution $h(y_i) = \sum_{k_0=1}^{K} \cdots \sum_{k_{N-1}=1}^{K} \delta_{k_0} \prod_{j=1}^{N-1} \gamma_{k_{j-1},k_j} \prod_{j=0}^{N-1} f_{k_j}(x_{iN+j})$ and then consider the log-likelihood function $\sum_{i=0}^{T-1} \log(h(y_i))$. When trying to process a maximal inequality using arguments from Massart (2007), one task would be to consider an appropriate Bernstein-type inequality for this setting.

Since for hidden Markov models mostly parametric settings are considered, the results on nonparametric models can further be used when comparing both approaches. One could think of testing the goodness of fit by using likelihood ratio tests. First considerations on that issue were investigated in a simulation scenario in Alexandrovich et al. (2016).

The chapter on penalized estimation proposed one approach how to regard structural assumptions on the parameters of hidden Markov models. Results on l_1-penalized estimation in Gaussian mixture models (Ruan et al., 2011, Yuan and Lin, 2007) and Gaussian hidden Markov models (Städler and Mukherjee, 2013) were extended by the introduction of different penalty functions to Gaussian hidden Markov models and a theoretical and computational comparison of l_1-penalization, hard thresholding and SCAD-penalization as proposed by Fan and Li (2001). While the focus was on sparsity of state-dependent precision matrices, another interesting consideration would be assumptions on zero entries of the transition probability matrix in scenarios where certain transitions are impossible or specific structures should be taken into account. An additional interesting aspect is the theoretical justification for using BIC or AIC as selection criterion of the tuning parameters.

The last chapter of this thesis covered computational aspects when dealing with hidden Markov models. It was shown that several assumptions on the model structure can be implemented by adjusting the EM-algorithm. Many of these adjustments even lead to only small modifications of the closed-form solutions, which made computation very convenient.
In the certain example of GDP panel data, the advantage of hidden Markov models over mixture models (used for example by Paapaa and van Dijk (1998), Pittau et al. (2010) and Vollmer et al. (2013)) concerning stability of the estimates over time was observed. When considering covariables, it was shown that their inclusion to mixture models is quite simple, while for hidden Markov models it entailed some computational problems. In the given context, in contrast to Bloom et al. (2003) who performed regression on the location parameters of the mixture model, explaining the mixing probabilities of the model using covariables yielded some good insights when investigating economic growth. For the explanation of advancement and decline of countries' income groups in the panel hidden Markov model, the considerations were not sufficiently comprehensive. It would be very interesting to investigate this question in a more detailed study.

A. Additional parameter estimates for chapter 3

We provide additional parameter estimates for sparse Gaussian hidden Markov models from section 3.3.

1. Biotechnology sector $p = 4$. Portfolio: Merck, Bayer, Curasan, Evotec. Unpenalized estimation:

$$\hat{\Omega}_1 = \begin{pmatrix} 7454 & -2390 & -160 & -425 \\ -2390 & 6683 & -108 & -1119 \\ -160 & -108 & 1346 & -68 \\ -425 & -1119 & -68 & 3011 \end{pmatrix}, \qquad \hat{\Omega}_2 = \begin{pmatrix} 1224 & -315 & -29 & -50 \\ -315 & 1360 & -79 & -126 \\ -29 & -79 & 423 & -23 \\ -50 & -126 & -23 & 328 \end{pmatrix}$$

$$\hat{\mu}_1 = \begin{pmatrix} 0.00085 & 0.00118 & -0.00061 & -0.00240 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_2 = \begin{pmatrix} -0.00017 & -0.00132 & -0.00284 & -0.00043 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_3 = \begin{pmatrix} -0.00155 & -0.00054 & 0.02188 & 0.01430 \end{pmatrix}^\mathsf{T},$$

$$\hat{\Gamma} = \begin{pmatrix} 0.78 & 0.22 & 0.00 \\ 0.41 & 0.57 & 0.02 \\ 0.29 & 0.53 & 0.18 \end{pmatrix}.$$

l_1-penalized estimation:

$$\hat{\Omega}_1 = \begin{pmatrix} 7243 & -2312 & -147 & -417 \\ -2312 & 6492 & -116 & -1077 \\ -147 & -116 & 1322 & -61 \\ -417 & -1077 & -61 & 2889 \end{pmatrix}, \qquad \hat{\Omega}_2 = \begin{pmatrix} 1157 & -265 & -22 & -45 \\ -265 & 1268 & -69 & -111 \\ -22 & -69 & 403 & -21 \\ -45 & -111 & -21 & 309 \end{pmatrix},$$

$$\hat{\mu}_1 = \begin{pmatrix} 0.00087 & 0.00123 & -0.00056 & -0.00253 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_2 = \begin{pmatrix} -0.00028 & -0.00155 & -0.00299 & 0.00002 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_3 = \begin{pmatrix} -0.00110 & 0.00093 & 0.02796 & 0.01519 \end{pmatrix}^\mathsf{T},$$

$$\hat{\Gamma} = \begin{pmatrix} 0.80 & 0.20 & 0.00 \\ 0.38 & 0.60 & 0.02 \\ 0.49 & 0.36 & 0.15 \end{pmatrix}.$$

SCAD-penalization:

$$\hat{\Omega}_1 = \begin{pmatrix} 7519 & -2408 & -172 & -428 \\ -2408 & 6732 & -109 & -1087 \\ -172 & -109 & 1355 & -73 \\ -428 & -1087 & -73 & 2980 \end{pmatrix}, \qquad \hat{\Omega}_2 = \begin{pmatrix} 1226 & -313 & -26 & -54 \\ -313 & 1366 & -80 & -128 \\ -26 & -80 & 425 & -24 \\ -54 & -128 & -24 & 332 \end{pmatrix},$$

$$\hat{\mu}_1 = \begin{pmatrix} 0.00087 & 0.00125 & -0.00054 & -0.00253 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_2 = \begin{pmatrix} -0.00021 & -0.00144 & -0.00295 & -0.00018 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_3 = \begin{pmatrix} -0.00137 & -0.00046 & 0.02376 & 0.01504 \end{pmatrix}^\mathsf{T},$$

$$\hat{\Gamma} = \begin{pmatrix} 0.80 & 0.20 & 0.00 \\ 0.38 & 0.60 & 0.02 \\ 0.44 & 0.40 & 0.16 \end{pmatrix}.$$

2. Biotechnology sector $p = 6$. Portfolio: Merck, Bayer, Curasan, Evotec, BASF, K+S. Unpenalized estimation:

$$\hat{\Omega}_1 = \begin{pmatrix} 6916 & -1468 & -65 & -297 & -944 & -461 \\ -1468 & 10596 & -140 & -559 & -6359 & -886 \\ -65 & -140 & 1208 & -82 & 50 & -39 \\ -297 & -559 & -82 & 2645 & -490 & -338 \\ -944 & -6359 & 50 & -490 & 12140 & -1932 \\ -461 & -886 & -39 & -338 & -1932 & 5611 \end{pmatrix},$$

$$\hat{\Omega}_2 = \begin{pmatrix} 1273 & -180 & -16 & -37 & -209 & -69 \\ -180 & 2326 & -52 & -83 & -1618 & -84 \\ -16 & -52 & 437 & -18 & -26 & -40 \\ -37 & -83 & -18 & 329 & -90 & -41 \\ -209 & -1618 & -26 & -90 & 3429 & -557 \\ -69 & -84 & -40 & -41 & -557 & 1259 \end{pmatrix},$$

$$\hat{\mu}_1 = \begin{pmatrix} 0.00107 & 0.00116 & -0.00063 & -0.00215 & 0.00140 & 0.00134 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_2 = \begin{pmatrix} -0.00071 & -0.00167 & -0.00334 & -0.00070 & -0.00244 & -0.00048 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_3 = \begin{pmatrix} -0.00033 & 0.00232 & 0.02273 & 0.01071 & 0.02053 & 0.00097 \end{pmatrix}^\mathsf{T},$$

$$\hat{\Gamma} = \begin{pmatrix} 0.81 & 0.19 & 0.00 \\ 0.42 & 0.56 & 0.02 \\ 0.15 & 0.66 & 0.19 \end{pmatrix}.$$

l_1-penalized estimation:

$$\hat{\Omega}_1 = \begin{pmatrix} 6667 & -1374 & -41 & -269 & -869 & -417 \\ -1374 & 10053 & -115 & -521 & -5925 & -852 \\ -41 & -115 & 1204 & -72 & 0 & -26 \\ -269 & -521 & -72 & 2526 & -464 & -314 \\ -869 & -5925 & 0 & -464 & 11567 & -1813 \\ -417 & -852 & -26 & -314 & -1813 & 5442 \end{pmatrix},$$

$$\hat{\Omega}_2 = \begin{pmatrix} 1244 & -172 & -16 & -36 & -208 & -69 \\ -172 & 2250 & -51 & -81 & -1575 & -79 \\ -16 & -51 & 424 & -18 & -22 & -40 \\ -36 & -81 & -18 & 320 & -84 & -40 \\ -208 & -1575 & -22 & -84 & 3310 & -549 \\ -69 & -79 & -40 & -40 & -549 & 1223 \end{pmatrix},$$

$$\hat{\mu}_1 = \begin{pmatrix} 0.00110 & 0.00119 & -0.00058 & -0.00222 & 0.00145 & 0.00139 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_2 = \begin{pmatrix} -0.00086 & -0.00185 & -0.00348 & -0.00054 & -0.00266 & -0.00072 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_3 = \begin{pmatrix} 0.00024 & 0.00331 & 0.02419 & 0.01133 & 0.02244 & 0.00248 \end{pmatrix}^\mathsf{T},$$

$$\hat{\Gamma} = \begin{pmatrix} 0.83 & 0.17 & 0.00 \\ 0.40 & 0.58 & 0.02 \\ 0.16 & 0.65 & 0.19 \end{pmatrix}.$$

SCAD-penalization:

$$\hat{\Omega}_1 = \begin{pmatrix} 6844 & -1452 & -65 & -295 & -940 & -455 \\ -1452 & 10547 & -138 & -552 & -6354 & -885 \\ -65 & -138 & 1204 & -80 & 55 & -45 \\ -295 & -552 & -80 & 2610 & -484 & -329 \\ -940 & -6354 & 55 & -484 & 12120 & -1934 \\ -455 & -885 & -45 & -329 & -1934 & 5561 \end{pmatrix},$$

$$\hat{\Omega}_2 = \begin{pmatrix} 1266 & -176 & -12 & -35 & -211 & -67 \\ -176 & 2294 & -54 & -80 & -1596 & -82 \\ -12 & -54 & 432 & -19 & -26 & -38 \\ -35 & -80 & -19 & 326 & -86 & -42 \\ -211 & -1596 & -26 & -86 & 3370 & -549 \\ -67 & -82 & -38 & -42 & -549 & 1246 \end{pmatrix},$$

$$\hat{\mu}_1 = \begin{pmatrix} 0.00108 & 0.00120 & -0.00055 & -0.00224 & 0.00145 & 0.00140 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_2 = \begin{pmatrix} -0.00078 & -0.00177 & -0.00346 & -0.00051 & -0.00260 & -0.00071 \end{pmatrix}^\mathsf{T},$$

$$\hat{\mu}_3 = \begin{pmatrix} 0.00020 & 0.00308 & 0.02376 & 0.01098 & 0.02234 & 0.00248 \end{pmatrix}^\mathsf{T},$$

$$\hat{\Gamma} = \begin{pmatrix} 0.83 & 0.17 & 0.00 \\ 0.39 & 0.59 & 0.02 \\ 0.16 & 0.65 & 0.19 \end{pmatrix}.$$

3. Merged portfolio $p = 8$. Portfolio: Merck, Bayer, Curasan, Evotec, Commerzbank, Deutsche Bank, Baader Bank, Deutsche Balaton. Unpenalized estimation:

$$\hat{\Omega}_1 = \begin{pmatrix} 6952 & -1518 & -116 & -225 & 22 & -1075 & -282 & 19 \\ -1518 & 7903 & -158 & -545 & -528 & -2365 & -277 & -433 \\ -116 & -158 & 1110 & -22 & -82 & 113 & -1 & -16 \\ -225 & -545 & -22 & 2787 & -188 & -644 & -204 & 52 \\ 22 & -528 & -82 & -188 & 5940 & -4458 & -182 & -78 \\ -1075 & -2365 & 113 & -644 & -4458 & 9852 & -219 & 3 \\ -282 & -277 & -1 & -204 & -182 & -219 & 2131 & -56 \\ 19 & -433 & -16 & 52 & -78 & 3 & -56 & 3555 \end{pmatrix},$$

$$\hat{\Omega}_2 = \begin{pmatrix} 1372 & -176 & -9 & -33 & -79 & -190 & -32 & -18 \\ -176 & 1997 & -40 & -69 & -146 & -734 & -2 & -10 \\ -9 & -40 & 498 & -20 & -18 & -50 & -8 & -46 \\ -33 & -69 & -20 & 378 & -36 & -91 & -68 & -13 \\ -79 & -146 & -18 & -36 & 1507 & -1076 & -40 & -10 \\ -190 & -734 & -50 & -91 & -1076 & 2373 & -88 & -30 \\ -32 & -2 & -8 & -68 & -40 & -88 & 512 & -33 \\ -18 & -10 & -46 & -13 & -10 & -30 & -33 & 811 \end{pmatrix},$$

$$\hat{\mu}_1 = (0.00108 \quad 0.00126 \quad -0.00006 \quad -0.00198 \quad -0.00042 \quad 0.00015 \quad -0.00004 \quad 0.00037)^\mathsf{T},$$

$$\hat{\mu}_2 = (-0.00019 \quad -0.00140 \quad -0.00383 \quad -0.00096 \quad -0.00332 \quad -0.00176 \quad -0.00311 \quad -0.00078)^\mathsf{T},$$

$$\hat{\mu}_3 = (-0.00512 \quad -0.00047 \quad 0.01826 \quad 0.00853 \quad 0.00930 \quad 0.00909 \quad 0.00465 \quad 0.00123)^\mathsf{T},$$

$$\hat{\Gamma} = \begin{pmatrix} 0.79 & 0.21 & 0.00 \\ 0.38 & 0.58 & 0.03 \\ 0.05 & 0.72 & 0.23 \end{pmatrix}.$$

l_1-penalization:

$$\hat{\Omega}_1 = \begin{pmatrix} 6798 & -1482 & -108 & -206 & 0 & -1035 & -271 & 0 \\ -1482 & 7800 & -158 & -528 & -538 & -2329 & -270 & -409 \\ -108 & -158 & 1108 & -21 & -76 & 97 & -1 & -16 \\ -206 & -528 & -21 & 2701 & -179 & -638 & -203 & 40 \\ 0 & -538 & -76 & -179 & 5832 & -4359 & -179 & -63 \\ -1035 & -2329 & 97 & -638 & -4359 & 9661 & -220 & 0 \\ -271 & -270 & -1 & -203 & -179 & -220 & 2091 & -53 \\ 0 & -409 & -16 & 40 & -63 & 0 & -53 & 3484 \end{pmatrix},$$

$$\hat{\Omega}_2 = \begin{pmatrix} 1330 & -160 & -6 & -32 & -74 & -180 & -24 & -7 \\ -160 & 1854 & -33 & -65 & -120 & -671 & 0 & -9 \\ -6 & -33 & 483 & -20 & -18 & -46 & -8 & -42 \\ -32 & -65 & -20 & 363 & -35 & -83 & -69 & -15 \\ -74 & -120 & -18 & -35 & 1407 & -1008 & -42 & -9 \\ -180 & -671 & -46 & -83 & -1008 & 2209 & -77 & -26 \\ -24 & 0 & -8 & -69 & -42 & -77 & 493 & -34 \\ -7 & -9 & -42 & -15 & -9 & -26 & -34 & 798 \end{pmatrix},$$

$$\hat{\mu}_1 = (0.00109 \quad 0.00125 \quad -0.00007 \quad -0.00198 \quad -0.00045 \quad 0.00013 \quad -0.00006 \quad 0.00039)^\mathsf{T},$$

$$\hat{\mu}_2 = (-0.00024 \quad -0.00147 \quad -0.00389 \quad -0.00088 \quad -0.00337 \quad -0.00181 \quad -0.00310 \quad -0.00085)^\mathsf{T},$$

$$\hat{\mu}_3 = (-0.00565 \quad 0.00005 \quad 0.02061 \quad 0.00889 \quad 0.01102 \quad 0.01070 \quad 0.00486 \quad 0.00166)^\mathsf{T},$$

$$\hat{\Gamma} = \begin{pmatrix} 0.79 & 0.21 & 0.00 \\ 0.39 & 0.58 & 0.03 \\ 0.05 & 0.71 & 0.24 \end{pmatrix}.$$

SCAD-penalization:

$$\hat{\Omega}_1 = \begin{pmatrix} 6925 & -1512 & -117 & -221 & 22 & -1074 & -280 & 17 \\ -1512 & 7886 & -155 & -543 & -532 & -2361 & -275 & -430 \\ -117 & -155 & 1110 & -21 & -82 & 111 & -1 & -17 \\ -221 & -543 & -21 & 2771 & -187 & -644 & -204 & 51 \\ 22 & -532 & -82 & -187 & 5914 & -4436 & -181 & -76 \\ -1074 & -2361 & 111 & -644 & -4436 & 9812 & -219 & 3 \\ -280 & -275 & -1 & -204 & -181 & -219 & 2123 & -56 \\ 17 & -430 & -17 & 51 & -76 & 3 & -56 & 3540 \end{pmatrix},$$

$$\hat{\Omega}_2 = \begin{pmatrix} 1363 & -175 & -10 & -33 & -78 & -188 & -31 & -17 \\ -175 & 1967 & -39 & -68 & -138 & -724 & -2 & -10 \\ -10 & -39 & 496 & -21 & -18 & -49 & -6 & -46 \\ -33 & -68 & -21 & 374 & -35 & -89 & -70 & -14 \\ -78 & -138 & -18 & -35 & 1491 & -1072 & -40 & -10 \\ -188 & -724 & -49 & -89 & -1072 & 2348 & -84 & -28 \\ -31 & -2 & -6 & -70 & -40 & -84 & 509 & -34 \\ -17 & -10 & -46 & -14 & -10 & -28 & -34 & 809 \end{pmatrix},$$

$$\hat{\mu}_1 = (0.00108 \quad 0.00126 \quad -0.00006 \quad -0.00198 \quad -0.00042 \quad 0.00015 \quad -0.00004 \quad 0.00037)^\mathsf{T},$$

$$\hat{\mu}_2 = (-0.00018 \quad -0.00140 \quad -0.00384 \quad -0.00091 \quad -0.00332 \quad -0.00178 \quad -0.00309 \quad -0.00077)^\mathsf{T},$$

$$\hat{\mu}_3 = (-0.00532 \quad -0.00040 \quad 0.01861 \quad 0.00810 \quad 0.00943 \quad 0.00938 \quad 0.00440 \quad 0.00110)^\mathsf{T},$$

$$\hat{\Gamma} = \begin{pmatrix} 0.79 & 0.21 & 0.00 \\ 0.38 & 0.59 & 0.03 \\ 0.05 & 0.72 & 0.23 \end{pmatrix}.$$

B. Overview ISO codes

We give a list of the countries which were considered in the models of Chapter 4. By (*) we indicate countries, for which due to missing data only models without covariables were considered.

Code	Country	Code	Country
AFG	Afghanistan	KNA(*)	St. Kitts & Nevis
AGO(*)	Angola	KOR	Korea, Republic of
ALB	Albania	LAO	Laos
ARG	Argentina	LBN(*)	Lebanon
ATG(*)	Antigua and Barbuda	LBR	Liberia
AUS	Australia	LCA(*)	St. Lucia
AUT	Austria	LKA	Sri Lanka
BDI	Burundi	LSO	Lesotho
BEL	Belgium	LUX	Luxembourg
BEN	Benin	MAC(*)	Macao
BFA(*)	Burkina Faso	MAR	Morocco
BGD	Bangladesh	MDG(*)	Madagascar
BGR	Bulgaria	MDV(*)	Maldives
BHS(*)	Bahamas	MEX	Mexico
BLZ	Belize	MHL(*)	Marshall Islands
BMU(*)	Bermuda	MLI	Mali
BOL	Bolivia	MLT(*)	Malta
BRA	Brazil	MNG	Mongolia
BRB(*)	Barbados	MOZ	Mozambique
BTN(*)	Bhutan	MRT	Mauritania
BWA	Botswana	MUS(*)	Mauritius
CAF	Central African Republic	MWI	Malawi
CAN	Canada	MYS	Malaysia
CHE	Switzerland	NAM	Namibia
CHL	Chile	NER	Niger
CHN	China	NGA(*)	Nigeria
CIV	Cote d'Ivoire	NIC	Nicaragua
CMR	Cameroon	NLD	Netherlands

COG	Congo, Republic of	NOR	Norway
COL	Colombia	NPL	Nepal
COM(*)	Comoros	NZL	New Zealand
CPV(*)	Cape Verde	OMN(*)	Oman
CRI	Costa Rica	PAK	Pakistan
CUB	Cuba	PAN	Panama
CYP	Cyprus	PER	Peru
DEU	Germany	PHL	Philippines
DJI(*)	Djibouti	PLW(*)	Palau
DMA(*)	Dominica	PNG	Papua New Guinea
DNK	Denmark	POL	Poland
DOM	Dominican Republic	PRI(*)	Puerto Rico
DZA	Algeria	PRT	Portugal
ECU	Ecuador	PRY	Paraguay
EGY	Egypt	ROM(*)	Republic of Moldova
ESP	Spain	RWA	Rwanda
ETH(*)	Ethiopia	SDN	Sudan
FIN	Finland	SEN	Senegal
FJI(*)	Fiji	SGP(*)	Singapore
FRA	France	SLB(*)	Solomon Islands
FSM(*)	Micronesia, Fed. Sts.	SLE	Sierra Leone
GBR	United Kingdom	SLV	El Salvador
GHA	Ghana	SOM(*)	Somalia
GIN(*)	Guinea	STP(*)	Sao Tome and Principe
GMB	Gambia, The	SWE	Sweden
GNB(*)	Guinea-Bissau	SWZ	Swaziland
GRC	Greece	SYC(*)	Seychelles
GRD(*)	Grenada	SYR	Syria
GTM	Guatemala	TCD(*)	Chad
GUY	Guyana	TGO	Togo
HKG(*)	Hong Kong	THA	Thailand
HND	Honduras	TON(*)	Tonga
HTI	Haiti	TUN	Tunisia
HUN	Hungary	TUR	Turkey
IDN	Indonesia	TWN(*)	Taiwan
IND	India	TZA	Tanzania
IRL	Ireland	UGA	Uganda
IRN	Iran	URY	Uruguay
IRQ	Iraq	USA	United States

ISL	Iceland	VCT(*)	St.Vincent & Grenadines
ISR	Israel	VEN	Venezuela
ITA	Italy	VNM	Vietnam
JAM	Jamaica	VUT(*)	Vanuatu
JOR	Jordan	WSM(*)	Samoa
JPN	Japan	ZAF	South Africa
KEN	Kenya	ZAR(*)	Congo, Dem. Rep.
KHM	Cambodia	ZMB	Zambia
KIR(*)	Kiribati	ZWE	Zimbabwe

Table B.1.: Countries: ISO codes.

Bibliography

J. Ajmera and C. Wooters. A robust speaker clustering algorithm. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 411–416. IEEE, 2003.

G. Alexandrovich. *Identification and estimation of hidden Markov models.* PhD thesis, Philipps-Universität Marburg, 2014. URL `http://archiv.ub.uni-marburg.de/diss/z2014/0393/pdf/dgael.pdf`.

G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 2016. doi: doi:10.1093/biomet/asw001.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37:3099–3132, 2009.

A. Azzalini and A. D. Valle. The multivariate skew-normal distribution. *Biometrika*, 83 (4):715–726, 1996.

R. J. Barro and J. W. Lee. A new data set of educational attainment in the world, 1950 - 2010. *Journal of Development Economics*, 104(0):184 – 198, 2013. URL `http://www.barrolee.com/data/full1.htm(19.09.2014)`.

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37(6):1554–1563, 12 1966.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 02 1970.

R. Bhar and S. Hamori. *Hidden Markov Models: Applications to Financial Economics.* Advanced Studies in Theoretical and Applied Econometrics. Springer US, 2010.

P. J. Bickel, Y. Ritov, and T. Ryden. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):pp. 1614–1635, 1998.

D. E. Bloom, D. Canning, and J. Sevilla. Geography and poverty traps. *Journal of Economic Growth*, 8(4):355–378, 2003.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, 2011.

T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, 2005.

S. Chatzis and T. Varvarigou. A robust to outliers hidden Markov model with application in text-dependent speaker identification. In *Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on*, pages 804–807, Nov 2007.

M. Cule. *Maximum likelihood estimation of a multivariate log-concave density*. PhD thesis, University of Cambridge, 2010.

M. Cule and R. Samworth. Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Statist.*, 4:254–270, 2010.

M. Cule, R. Gramacy, and R. Samworth. LogConcDEAD: An R package for maximum likelihood estimation of a multivariate log-concave density. *Journal of Statistical Software*, 29(2), 2009.

M. Cule, R. Samworth, and M. Stewart. Maximum likelihood estimation of a multidimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607, 2010.

J. Dannemann. Semiparametric hidden Markov models. *Journal of Computational and Graphical Statistics*, 21(3):677–692, 2012.

J. Dannemann, H. Holzmann, and A. Leister. Semiparametric hidden Markov models: identifiability and estimation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):418–425, 2014.

Y. De Castro, E. Gassiat, and C. Lacour. Minimax adaptive estimation of nonparametric hidden Markov models. *Journal of Machine Learning Research*, 2015. to appear.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1): 1–38, 1977.

R. DerSimonian. Algorithm as 221: Maximum likelihood estimation of a mixing distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 35(3): pp. 302–309, 1986.

W. DeSarbo and W. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282, 1988.

L. Dümbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15 (1):40–68, 02 2009.

L. Dümbgen and K. Rufibach. logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software*, 39(6):1–28, 2011.

R. J. Elliott, L. Aggoun, and J. B. Moore. *Hidden Markov Models - Estimation and Control*. Springer Science & Business Media, Berlin Heidelberg, 1st ed. 1995. corr. 3rd printing 2008 edition, 1995.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

M. Fiecas, J. Franke, R. von Sachs, and J. Tadjuidje. Shrinkage estimation for multivariate hidden Markov mixture models. *ISBA Discussion Paper 2012/16*, 2012.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):pp. 611–631, 2002.

C. Fraley, R. A. E., M. B. T., and L. Scrucca. mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation technical report no. 597. Technical report, Department of Statistics, University of Washington, 2012.

J. Franke, J.-P. Stockis, and J. Tadjuidje. Nonparametric estimation for Markov switching AR-ARCH models. submitted, 2012.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models - Modeling and Applications to Random Processes*. Springer Science & Business Media, Berlin Heidelberg, 2006 edition, 2006.

M. I. Garrido and J. A. Jaramillo. Lipschitz-type functions on metric spaces. *Journal of Mathematical Analysis and Applications*, 340(1):282 – 290, 2008.

E. Gassiat and J. Rousseau. About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, 20(4):2039–2075, 11 2014.

E. Gassiat and J. Rousseau. Nonparametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1):193–212, 02 2016.

E. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden Markov models and applications. *Statistics and Computing*, 26(1-2):61–71, 2016.

J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

A. Heston, R. Summers, and B. Aten. Penn world table version 7.1 center for international comparisons of production, income and prices at the university of pennsylvania. Website (16.12.2014), 2012. URL `http://www.rug.nl/research/ggdc/data/pwt/pwt-7.1`.

H. Holzmann and F. Schwaiger. Testing for the number of states in hidden Markov models. *Computational Statistics & Data Analysis*, (0):–, 2014.

H. Holzmann and F. Schwaiger. Hidden Markov models with state-dependent mixtures: minimal representation, model testing and applications to clustering. *Statistics and Computing*, 25(6):1185–1200, 2015.

J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

A. J. Izenman. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991.

N. Jin and F. Mokhtarian. A non-parametric HMM learning method for shape dynamics with application to human motion recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 29–32, 2006.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906, 12 1956.

A. K. H. Kim and R. J. Samworth. Global rates of convergence in log-concave density estimation, 2014. URL `http://arxiv.org/abs/1404.2298`.

J. Kingman. Subadditive processes. In P.-L. Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour V-1975*, volume 539 of *Lecture Notes in Mathematics*, pages 167–223. Springer Berlin Heidelberg, 1976.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000.

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):pp. 805–811, 1978.

M. F. Lambert, J. P. Whiting, and A. V. Metcalfe. A non-parametric hidden Markov model for climate state identification. *Hydrology and Earth System Sciences*, 7(5): 652–667, 2003.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

F. Lefèvre. Non-parametric probability estimation for HMM-based automatic speech recognition. *Computer Speech & Language*, 17(23):113 – 136, 2003.

B. G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40(1):127 – 143, 1992a.

B. G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 09 1992b.

B. Lindsay. *Mixture Models - Theory, Geometry, and Applications*. IMS, 1995.

B. G. Lindsay. The geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, 11(1):86–94, 03 1983.

A. Lotsi and E. Wit. High dimensional sparse Gaussian graphical mixture model. *ArXiv e-prints*, Aug. 2013.

A. Maruotti. Mixed hidden Markov models for longitudinal data: An overview. *International Statistical Review*, 79(3):427–454, 2011.

P. Massart. *Concentration Inequalities and Model Selection - Ecole D'Et de Probabilits de Saint-Flour XXXIII - 2003*. Springer, Berlin, Heidelberg, 2007. aufl. edition, 2007.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, New York, 2007.

G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2004.

A. L. Owen, J. Videras, and L. Davis. Do all countries follow the same growth process? *Journal of Economic Growth*, 14(4):265–286, 2009.

R. Paapaa and H. K. van Dijk. Distribution and mobility of wealth of nations. *European Economic Review*, 42(7):1269 – 1293, 1998.

J. K. Pal, M. Woodroofe, and M. Meyer. *Estimating a Polya frequency function*$_2$, volume Volume 54 of *Lecture Notes–Monograph Series*, pages 239–249. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007.

J. Pfanzagl. Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *Journal of Statistical Planning and Inference*, 19(2): 137 – 158, 1988.

M. G. Pittau, R. Zelli, and P. A. Johnson. Mixture models, convergence clubs, and polarization. *Review of Income and Wealth*, 56(1):102–122, 2010.

R. E. Quandt and J. B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):pp. 730–738, 1978.

L. Rabiner, B. H. Juang, and B.-H. Juang. *Fundamentals of Speech Recognition*. PTR Prentice Hall, Englewod Cliffs, New Jersey, new. edition, 1993.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

G. Ritter. *Robust Cluster Analysis and Variable Selection*. CRC Press, Boca Raton, Fla, 2014.

J.-M. Robin, S. Bonhomme, and K. Jochmans. Estimating Multivariate Latent-Structure Models. working paper or preprint, Dec. 2014.

R. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, reprint edition, 1970.

L. Ruan, M. Yuan, and H. Zou. Regularized parameter estimation in high-dimensional Gaussian mixture models. *Neural Comput.*, 23(6):1605–1622, June 2011.

K. Rufibach. *Log-concave Density Estimation and Bump Hunting for i.i.d. Observations*. Phd thesis, University of Bern, Switzerland and Georg-August University of Göttingen, Germany, 2006.

T. Rydén, T. Tersvirta, and S. Åsbrink. Stylized facts of daily return series and the hidden Markov model. Working Paper Series in Economics and Finance 117, Stockholm School of Economics, June 1996.

E. Seneta. *Non-negative Matrices and Markov Chains*. Springer Science & Business Media, Berlin Heidelberg, 2006.

B. Simon. *Convexity: An Analytic Viewpoint*. Cambridge Tracts in Mathematics. Cambridge University Press, 2011.

N. Städler and S. Mukherjee. Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. *Ann. Appl. Stat.*, 7(4):2157–2179, 12 2013.

M. E. Taylor. *Measure Theory and Integration*. American Mathematical Society, Heidelberg, 2006.

H. Teicher. On the mixture of distributions. *The Annals of Mathematical Statistics*, 31 (1):55–73, 03 1960.

H. Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1): 244–248, 03 1961.

H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34 (4):1265–1269, 12 1963.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996.

A. W. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes - With Applications to Statistics*. Springer, Berlin, Heidelberg, 1996.

E. Vernet. Posterior consistency for nonparametric hidden Markov models with finite state space. *Electron. J. Statist.*, 9(1):717–752, 2015.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, April 1967.

S. Volant, C. Bérard, M.-L. Martin-Magniette, and S. Robin. Hidden Markov models with mixtures as emission distributions. *Statistics and Computing*, 24(4):493–504, 2013.

S. Vollmer, H. Holzmann, and F. Schwaiger. Peaks vs components. *Review of Development Economics*, 17(2):352–364, 2013.

A. Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 12 1949.

G. Walther. Detecting the presence of mixing with multiscale maximum likelihood. *Journal of the American Statistical Association*, 97:508–513, 2002.

G. Walther. Inference and modeling with log-concave distributions. *Statist. Sci.*, 24(3): 319–327, 08 2009.

S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 02 1968.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):pp. 19–35, 2007.

W. Zucchini and I. MacDonald. *Hidden Markov Models for Time Series: An Introduction Using R, 2. Auflage*. Boca Raton, Chapman & Hall/CRC, 2009.

W. Zucchini, D. Raubenheimer, and I. L. MacDonald. Modeling time series of animal behavior by means of a latent-state model with feedback. *Biometrics*, 64(3):807–815, 2008.

Zusammenfassung

In dieser Arbeit befassen wir uns mit Hidden Markov Modellen, einer beliebten Klasse stochastischer Modelle, die sich gut für die Behandlung von Daten aus Zeitreihen eignen. Wir betrachten einen bivariaten Prozess $(X_t, S_t)_{t\in\mathcal{T}}$, wobei die erste Komponente $(X_t)_{t\in\mathcal{T}}$ die Beobachtungen des Prozesses darstellen und $(S_t)_{t\in\mathcal{T}}$ unbeobachtet ist. Die Abhängigkeitsstruktur des Prozesses wird durch die Annahme modelliert, dass der unbeobachtete Prozess eine Markovkette ist. Die Verteilung der Zufallsvariablen X_t wird durch den Zustand, den die Markovkette zu dem Zeitpunkt annimmt, bestimmt. Detaillierte Einführungen der Modellklasse sind beispielsweise in Zucchini and MacDonald (2009), Cappé et al. (2005) oder Elliott et al. (1995) nachzulesen. In dieser Arbeit beschränken wir uns auf die Betrachtung von Markovketten mit endlichem Zustandsraum in diskreter Zeit und konzentrieren uns auf die Schätzung der Parameter in diesen Modellen.
Anwendungen der Hidden Markov Modelle finden sich gehäuft im Kontext der Spracherkennung (Rabiner et al., 1993), in der biologischen Verhaltensforschung (Zucchini et al., 2008), der Signalverarbeitung (Cappé et al., 2005) und in der Ökonomie und Finanzwissenschaft (Bhar and Hamori, 2010, Rydén et al., 1996).
Das bereits gut erforschte Standardmodell umfasst die Betrachtung einer homogenen Markovkette und parametrischer zustandsbedingter Verteilungen (vorwiegend Gaußverteilungen). In diesem Kontext ist die Theorie über Maximum Likelihood Schätzung schon weit erforscht. Leroux (1992a) lieferte ein Konsistenzresultat, während Bickel et al. (1998) asymptotische Normalität des Schätzers bewiesen. Ein Vorteil der Maximum Likelihood Schätzung ist, dass durch den EM-Algorithmus ein sehr flexibles Verfahren für die Berechnung des Schätzers zur Verfügung steht. Für viele parametrische Verteilungsklassen liefert dieses sogar geschlossene Formeln für die Parameterschätzung.

In dieser Arbeit wollen wir von einigen dieser Standardannahmen abrücken und in den formulierten Modellen Eigenschaften des Maximum Likelihood Schätzers untersuchen. Diese Betrachtungen verdeutlichen die Flexibilität der Modellklasse und öffnen diese für eine breitere Menge an Datensätzen.
In Kapitel 2 behandeln wir nichtparametrische Maximum Likelihood Schätzung in Hidden Markov Modellen, die in der Vergangenheit bereits in einigen Anwendungen (z.B. Jin and Mokhtarian, 2006, Lambert et al., 2003, Lefèvre, 2003) diskutiert wurde, bislang

aber theoretisch wenig untersucht ist. Grund dafür ist, dass die Frage der Identifizierbarkeit in nichtparametrischen Hidden Markov Modellen erst kürzlich geklärt wurde (siehe Alexandrovich et al., 2016, Gassiat and Rousseau, 2016, Gassiat et al., 2016). Basierend auf diesen Resultaten betrachten wir zwei nichtparametrische Dichteklassen für die zustandsbedingten Verteilungen des Hidden Markov Modells und deren Maximum Likelihood Schätzung. Zunächst untersuchen wir die Klasse der allgemeinen Mischungsmodelle, die sehr flexible Dichten, insbesondere in Bezug auf Multimodalität oder Schiefe, zulassen. In diesem Kontext beweisen wir die Existenz und Konsistenz eines Maximum Likelihood Schätzers, liefern nötige Anpassungen des EM-Algorithmus für dessen Berechnung und führen diese in einigen Simulationsszenarien durch. Insbesondere vergleichen wir dabei den nichtparametrischen Ansatz mit dem parametrischen Standardmodell. Danach untersuchen wir die Klasse der logkonkaven Dichten und beweisen die Existenz eines Maximum Likelihood Schätzers, sowie dessen konkrete Gestalt. Zusätzlich liefern wir numerische Beispiele in Simulationen und anhand eines realen Datensatzes.

In Kapitel 3 behandeln wir penalisierte Schätzungen, insbesondere in Gaußschen Hidden Markov Modellen. Dabei fokussieren wir uns auf Penalisierung der zustandsbedingten Precision Matrizen (Inversen der Kovarianzmatrizen), um bedingte Unabhängigkeit der Zufallsvariablen zu untersuchen. Wir vergleichen die populäre Lassopenalisierung mit dem Hardthresholding Verfahren und der von Fan and Li (2001) eingeführten SCAD-Penalisierung. Zunächst übertragen wir die theoretischen Resultate für den Maximum Likelihood Schätzer aus Fan and Li (2001) auf Hidden Markov Modelle und untersuchen die Eigenschaften des penalisierten Schätzers dann in einer Simulation, sowie anhand eines Datensatzes multivariater Finanzzeitreihen.

Im letzten Kapitel nutzen wir unterschiedliche Modelle, um ein Anwendungsbeispiel aus der Wohlfahrtsökonomik zu behandeln. Für einen Paneldatensatz, der die BIP Daten vieler Länder der Welt von 1970–2010 beinhaltet, konstruieren wir Modelle mit unterschiedlichen Abhängigkeitsstrukturen und der Möglichkeit zur Berücksichtigung von Kovariablen, um die Einkommen der Länder und deren Entwicklung zu untersuchen. Wir beginnen mit der Anpassung von endlichen Mischungsmodellen für jedes Jahr, um die Subpopulationen geeignet modellieren zu können. Um die zeitliche Abhängigkeit besser erfassen zu können, gehen wir dann zu inhomogenen Hidden Markov Modellen über, die wir im Verlauf versuchen besser an die Daten anzupassen. Für beide Modellklassen können handliche Verfahren für eine a-posteriori Klassifikation der Länder in drei Einkommensgruppen angewendet werden. Die Resultate zeigen, dass einige Länder über den Zeitverlauf die Einkommensgruppe durch Auf- oder Abstieg wechseln. Dieses Verhalten versuchen wir durch die Aufnahme von Kovariablen in die Modelle zu erklären. Im Kontext der Mischungsmodelle nutzen wir Switching Regressionsmodelle, um die

Mischungsgewichte durch Kovariablen zu modellieren. Danach integrieren wir die Kovariablen in die Übergangswahrscheinlichkeiten der Hidden Markov Modelle, um Wechsel zwischen den Einkommensgruppen zu erklären. Dabei entstehen allerdings sehr komplexe Modelle mit vielen Parametern, die die Schätzung stark verkomplizieren und für den gegebenen Datensatz keine zufriedenstellenden Ergebnisse liefern.
Für alle behandelten Modelle stellen wir die nötigen Anpassungen des EM-Algorithmus dar.