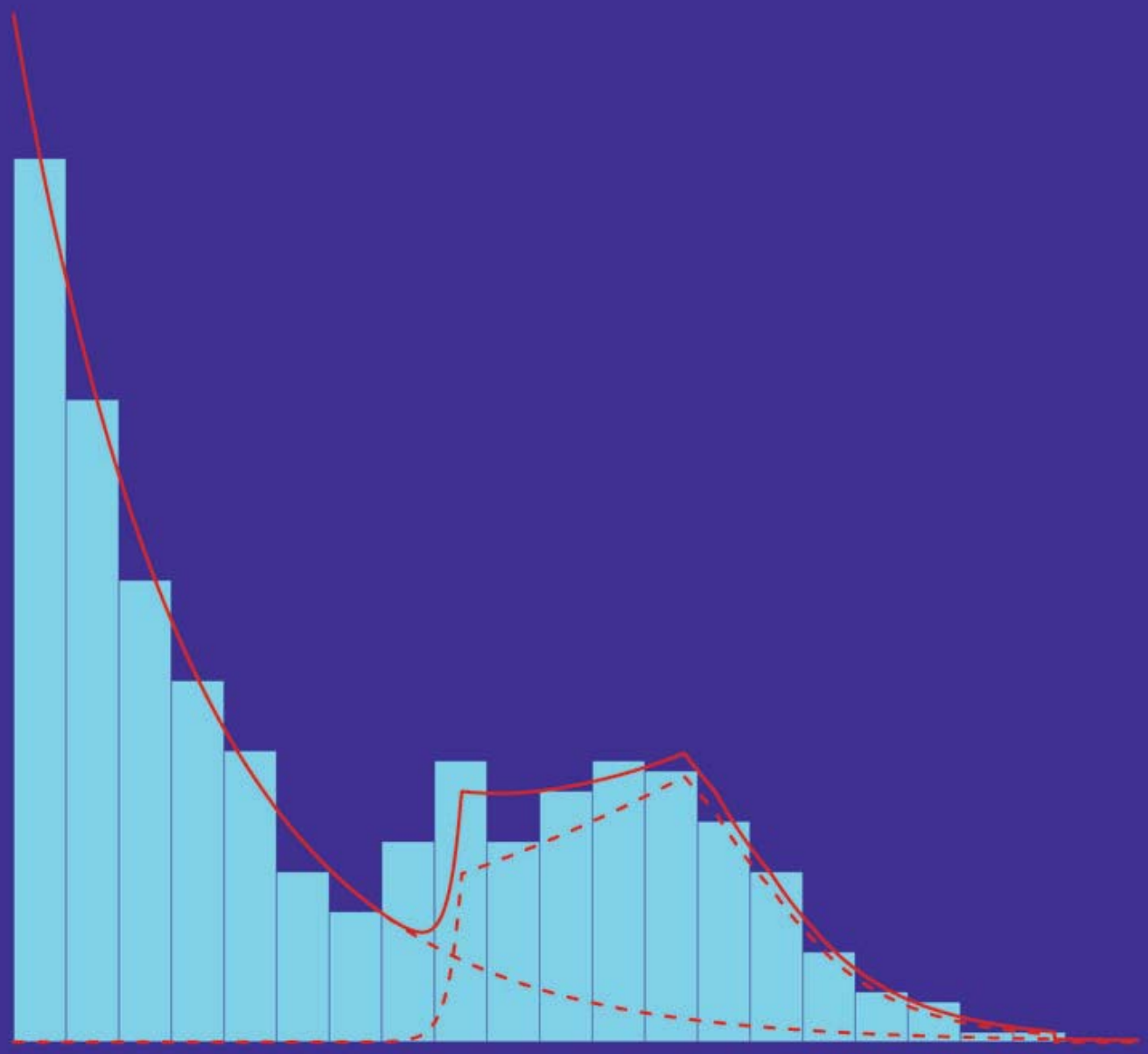
Jörn Dannemann

Inference for Hidden Markov Models and related Models



Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Inference for Hidden Markov Models and related Models

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

vorgelegt von

Jörn Dannemann
aus Eutin

Göttingen 2009

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Aufl. - Göttingen: Cuvillier, 2010
Zugl.: Göttingen, Univ., Diss., 2009

978-3-86955-247-7

D7

Referent:	Prof. Dr. Axel Munk
Koreferent:	Prof. Dr. Hajo Holzmann
Tag der mündlichen Prüfung:	14. 07. 2009

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2010
Gedruckt auf säurefreiem Papier

978-3-86955-247-7

Contents

Introduction

Hidden Markov models (HMMs) and other latent variable models form complex, flexible frameworks for univariate and multivariate data structures. A major advantage of latent structures is the principle simplicity and the accessibility to practitioners and their application-driven interpretations rather than black box systems. The statistical analysis is not apparent as such models do not belong to standard parametric classes of independent identical distributed (i.i.d.) random variables. Therefore, many statistical issues for HMMs and related models are not implied by results of the standard literature, but have been developed step by step by several authors in the last decades, among others Baum and Petrie (1966), Leroux (1992b), Rydén (1994, 1995), Bickel et al. (1998, 2002), Douc and Matias (2001), Gassiat and Keribin (2000), MacKay (2002), Douc et al. (2004), Poskitt and Zhang (2005), Altman (2007). We describe the models and illustrate their importance in various applications in Chapter 1. For a state of the art overview see Cappé et al. (2005).

This thesis is mainly concerned with three topics.

1. Testing for HMMs under nonstandard conditions, namely when the true parameter lies on the boundary.

2. Testing for the number of states in HMMs and switching regression models, in particular

   - testing for two states in an HMM, and
   - testing for two components in switching regression models with independent or Markov-dependent regime.

3. Modeling HMMs with flexible state-dependent distributions.

In practical applications of HMMs, non-standard testing problems are frequently encountered, e.g. testing for the probability of staying in a certain unobserved state being zero.

Such testing problems involving the boundary of the parameter space have achieved much attention for i.i.d. data (e.g. Self and Liang, 1987) . In the first part of the thesis we consider testing problems in HMMs

$$H : \vartheta \in \Theta_0 \quad \text{against} \quad K : \vartheta \in \Theta \setminus \Theta_0,$$

when the true parameter ϑ_0 lies on the boundary of Θ_0 and possibly also of Θ. We derive the relevant asymptotic distribution theory for the likelihood ratio test (LRT) in HMMs under these conditions. Considering the prominent example of testing for a transition probability being zero we derive for the LRT under the hypothesis

$$\text{LRT} \xrightarrow{\mathcal{L}} \frac{1}{2}\,\chi_0^2 + \frac{1}{2}\,\chi_1^2$$

with χ_k^2 denoting the χ^2-distribution with k degrees of freedom and χ_0^2 denoting the point mass at zero accordingly. Apart from this specific situation a number of examples with particular relevance in the HMM framework are examined.

The specification of the number of states is very important in all models with discrete latent variables and performing statistical testing of such hypotheses is one way to deal with this problem. In the second part of the thesis we give an introduction to these specific testing problems and discuss their non-standard nature. We present the work by Chen et al. (2001) and Zhu and Zhang (2004), who develop methods for testing for homogeneity in finite mixture models and switching regression models based on the asymptotic analysis of a modified likelihood ratio test (MLRT). Chen et al. (2004) investigate testing for two components in finite mixtures of one-parametric families. Following this approach we propose a test for two states in HMMs, i.e.

$$H : m = 2 \quad \text{against} \quad K : m \geq 3,$$

where m denotes the number of states. Testing for two states is of primary interest in particular for HMMs, since a two-state HMM represents the smallest non-trivial member of this model class. We show that the asymptotic distribution for the MLRT with independence assumption under the hypothesis is given by

$$\text{MLRT} \xrightarrow{\mathcal{L}} \big(\frac{1}{2} - p\big)\,\chi_0^2 + \frac{1}{2}\chi_1^2 + p\,\chi_2^2,$$

depending on the mixing weight p, which may be replaced by an estimate for applications of the MLRT. We make the surprising and novel observation that the dependency structure of the HMM does not influence the behavior of the MLRT. In addition, we extend

the results by Chen et al. (2004) to switching regression models and propose a test for two components in switching regression models with independent or Markov-dependent regime, also called Markov-switching models. We verify the asymptotic distribution of the MLRT under the hypothesis and discuss an application to a dental health trail data set, for which the classical model selection criteria support different models.

The first two topics are concerned with testing problems within the classical parametric framework, i.e. the distribution of the observations conditioned on the latent variables belong to some parametric family. In the third part of the thesis we relax the parametric assumptions, aiming for more flexible models, which reduce systematic errors caused by model misspecification and give rise to model validation techniques. We propose a parametric as well as a semiparametric approach to this problem. In particular, the latter one introduces a new flavor to hidden Markov modeling by linking recently developed semiparametric mixture models, introduced by Hall and Zhou (2003) and Bordes et al. (2006a), to the HMM framework. Firstly, we show that their identifiability results for two-component semiparametric mixtures transfer to two-state HMMs. Secondly, we propose an estimation procedure to semiparametric HMMs based on the expectation-maximization algorithm that enables extensions of estimation techniques in semiparametric mixtures to HMMs, for example the methods suggested by Chang and Walther (2007) and Cule et al. (2008) for mixtures with log-concave components.

This thesis is organized as follows. In Chapter 1 we introduce the latent variable models that are mainly treated in this thesis, namely finite mixture models, HMMs and switching regression models with independent and Markov-dependent regime. The following chapters correspond each with one of the topics displayed above. For all three topics the analysis includes simulation studies of the finite sample performance of the proposed techniques. As partially mentioned above some illustrative applications from various fields, including epileptic seizures, financial time series and a dental health trail, are presented.

Some results of this thesis have been presented at the GOCPS (Aachen, 2008), WCPS (Singapore, 2008) and the Biometrisches Kolloquium (Hannover, 2009) and are published in Dannemann and Holzmann (2008b, Scand. J. Statist.), Dannemann and Holzmann (2008c, Canad. J. Statist.) and Dannemann and Holzmann (2010, Comput. Statist. Data Anal.).

Acknowledgements

I am grateful to my advisors Axel Munk and Hajo Holzmann for giving me the opportunity to carry out this research, for constant encouragement and for being open for discussion and questions. After Hajo Holzmann left Göttingen, several invitations to the Universities of Karlsruhe and Marburg made a vivid collaboration possible. Beyond professional and non-material support I would like to express my special gratitude to Axel Munk for the opportunity to present some results of this thesis at the World Congress '09 in Singapore. To Sabine Güttler I am indebted for fruitful discussions on statistical modeling of railway data. I would like to thank my office-mates Matthias Mielke, Andreas Dahmen, Sophie Bruns, Achim Wübker and Carsten Gottschlich for helpful discussions and frequent cheering up; and all my colleagues at the Institute for Mathematical Stochastics for the inspiring and warm atmosphere.
During my time as a Ph.D. student I was a member of the Ph.D. Program "Applied Statistics and Empirical Methods" in the Centre of Statistics and I gratefully acknowledge financial support from the "Georg-Lichtenberg-Programm" and the possibility of interesting scientific discourse. I would also like to thank the Verein zur Förderung der Mathematischen Statistik und Versicherungsmathematik e.V. for supporting the publication of the thesis.
Last and important thanks go to my family for various forms of support, above all from my wife Jenny as well as from my parents Waltraut and Werner and parents-in-law Marlene and Wolfgang.

Chapter 1

Hidden Markov models and related models

Hidden Markov models (German: verborgene Markow Modelle, French: modèle de Markov caché) belong to the big class of latent variable models. Models with latent variables have entered almost all fields of statistical applications. It is common for these models that unobserved variables are introduced to model a complex data structure given by the observables. In many applications these unobserved variables have a concrete theoretical justification or are motivated by some desirable interpretation. In other cases, hidden variables are used as a technical tool to build complex models.
In this chapter we present the type of models which are mainly treated in this thesis as well as models closely related. These models have in common that the hidden variables form a discrete time stochastic process on some finite set, e.g. $\{1, \ldots, m\}$. A very simple model of this kind is the finite mixture model, which somehow serves as a nutshell model for a whole bunch of models including hidden Markov models, switching regression models and many others.

1.1 Finite mixture models

Finite mixture models are used extensively for describing populations with unobserved heterogeneity. A number of monographs are available discussing all kinds of properties which appeared in the literature, for example Titterington et al. (1985), Böhning (1999), McLachlan and Peel (2000), Frühwirth-Schnatter (2006). More recent surveys of the topic are given by Böhning et al. (2007) and Young (2008).
One of the oldest examples of a finite mixture model was introduced by Pearson (1894). He

observed in the analysis of some data set that the normal distribution did not give a good fit to the data. He deduced the presence of heterogeneity w.r.t. the normal distribution and claimed that each observation belongs to one of two populations, where the distribution within each population follows a normal distribution.
Denoting the observations by Y_i and the membership to one of the populations by U_i this formalizes to $P(Y_i \leq y|U_i = 0) = F_{\mu_1,\sigma_1^2}(y)$, $P(Y_i \leq y|U_i = 1) = F_{\mu_2,\sigma_2^2}(y)$ and $P(U_i = 0) = 1 - P(U_i = 1) = \pi_1$, where F_{μ,σ^2} denotes a normal distribution with mean μ and variance σ^2. Assuming that $(U_i)_i$ and $(Y_i)_i$ are two independent sequences (but not independent of each other) leads to a univariate two component mixture model of two normal distributions with distribution function:

$$\begin{aligned} G(y) = P(Y_i \leq y) &= P(U_i = 0)P(Y_i \leq y|U_i = 0) + P(U_i = 1)P(Y_i \leq y|U_i = 1) \\ &= \pi_1 F_{\mu_1,\sigma_1^2}(y) + (1 - \pi_1)F_{\mu_2,\sigma_2^2}(y) \end{aligned}$$

with parameters $(\pi_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$. In general, an m-component finite mixture distribution is given by

$$G(y) = \pi_1 F_1(y) + \ldots + \pi_m F_m(y), \tag{1.1}$$

where $\pi_k \geq 0$, $\sum_{k=1}^m \pi_k = 1$ and F_k specifies the distribution of the kth component. As in the example above the latent variable here represents the unobservable membership to one of the components standing for each of the populations and (1.1) arises from

$$G(y) = P(Y_i \leq y) = P(U_i = 1)P(Y_i \leq y|U_i = 1) + \ldots + P(U_i = m)P(Y_i \leq y|U_i = m),$$

where $U_i \sim Mult(\pi)$ are i.i.d. multinomial distributed r.v. on $\{1, \ldots, m\}$. Usually one assumes that the distributions $P(Y_i \leq y|U_i = k) = F_k(y) = F_{\theta_k}(y)$ belong to some parametric family, such that the parameter of interest of an m-component finite mixture model is $\vartheta = (\pi_1, \ldots, \pi_m, \theta_1, \ldots, \theta_m)$.

Example 1.1. Let $\varepsilon_i \sim \mathcal{N}(0, 1)$ and $U_i \sim Mult(\pi)$ i.i.d., and $(\mu_1, \sigma_1^2), \ldots, (\mu_m, \sigma_m^2) \in \mathbb{R} \times \mathbb{R}^+$ pairwise distinct, then

$$Y_i = \mu_{U_i} + \sigma_{U_i}\varepsilon_i$$

follows an m-component Gaussian mixture with parameter $(\pi_1, \ldots, \pi_m, \mu_1, \sigma_1^2, \ldots, \mu_m, \sigma_m^2)$.

For the statistical analysis of finite mixture models identifiability as well as estimation procedures and their computational feasibility are crucial.

Identifiability of finite mixture models

Identifiability means that in principle the true parameter is uniquely determined by the probability law of the observations, i.e.

$$P_{\vartheta}^{(Y_i)} = P_{\vartheta_0}^{(Y_i)} \quad \Longrightarrow \quad \vartheta = \vartheta_0.$$

As ϑ_0 is unknown one should require this property for all $\vartheta \in \Theta$. In many statistical models this is just a question of choosing an appropriate parametrization of the considered model. In the context of finite mixture models identifiability is by far not an obvious issue. One reason for this is that an m-component finite mixture can be represented as a $(m+1)$-component mixture in various ways (either by putting one of the weights to zero or choosing two distributions of the components as equal). Hence, to achieve the identifiability of a particular parametrization for finite mixture models it is a necessary condition that the number of components m is known, i.e. for a m-component mixture $\pi_{k_1} > 0$ and $F_{k_1} \neq F_{k_2}$ for $1 \leq k_1, k_2 \leq m$.

For classical finite mixture models, where F_ks belong to some parametric family, identifiability of finite mixtures is established for most common families. The following finite mixtures are identifiable: finite mixtures of Poisson distributions (Feller, 1943), finite mixtures of univariate normal and gamma distributions (Teicher, 1963) and finite mixtures of multivariate normal distributions (Yakowitz and Spragins, 1968). A general helpful characterization of identifiability of finite mixtures is also given by Yakowitz and Spragins (1968), who proved that the class of finite mixtures of distributions is identifiable if and only if the underlying parametric family is linearly independent over the field of real numbers.

However, the fact, that identifiability is not trivial, is also illustrated by a number of examples of non-identifiable mixtures. These examples usually make use of a linear relationship between y and $F(y)$ to construct linearly dependent distributions and are often based on uniform distributions (see Everitt and Hand, 1981) or triangular distributions (see Holzmann, Munk and Gneiting, 2006a, Ex. 6). Also the $\mathbf{n}$-binomial distribution finite mixtures are not identifiable, if $\mathbf{n} < 2m - 1$ (Teicher, 1963).

Another phenomenon concerning identifiability is the so called *label-switching*, i.e. the states of the latent process can be permuted without changing the law of the observed model. Standard approaches to overcome the problem are changing the nomenclature to equivalence classes w.r.t. label-switching (Leroux, 1992b) or imposing ordering constraints of the parameters of the distributions of the components, e.g. $\theta_1 > \ldots > \theta_m$. While the consideration of equivalence classes is not very comfortable from a practical view point, ordering constraints seem to be an easy way out. However, as pointed out by Frühwirth-

Schnatter (2006) such constraints may not be desirable in some applications and may influence statistical inference especially if the distributions of two components are close.

Estimation in finite mixture models

Based on an i.i.d. sample $(Y_i)_i$ from a finite mixture the parameter of interest

$$\vartheta = (\pi_1, \ldots, \pi_m, \theta_1, \ldots, \theta_m)$$

can be estimated by different methods. Classical approaches are method of moments, maximum likelihood estimation as well as Bayes estimation (Frühwirth-Schnatter, 2006, for a summary cf.). All of them have different advantages and disadvantages. For the maximum likelihood estimator (MLE) we consider the log-likelihood function

$$L_n^{(m)}(\vartheta) = L^{(m)}(\vartheta; Y_1, \ldots, Y_n) = \sum_{i=1}^{n} \log \sum_{k=1}^{m} \pi_k f_{\theta_k}(Y_i), \tag{1.2}$$

where $f_{\theta_k}(\cdot)$ denotes the density corresponding to the conditional distribution $F_{\theta_k}(\cdot)$. The MLE is then defined as the maximizer of the log-likelihood, $\hat{\vartheta} := \arg\max_\vartheta L_n^{(m)}(\vartheta)$. In contrast to other statistical models, namely exponential families, typically the MLE cannot be calculated explicitly. There are different techniques proposed in the literature to evaluate the MLE. Besides direct maximization techniques based on Newton's method, the expectation- maximization (EM) algorithm introduced by Dempster et al. (1977) is very popular among statisticians. It is designed for models with incomplete information and extensions are often straight forward to implement. However, in general the speed of the EM algorithm is comparably low. The methodology of the EM algorithm will be discussed in some detail in Section 4.2.2. Since finite mixture models are i.i.d. models, the usual asymptotic theory applies, i.e. that under the usual regularity conditions the MLE is consistent and centered asymptotic normal with covariance matrix given by the inverse of the Fisher-information-matrix (cf. Frühwirth-Schnatter, 2006, p.52).

1.2 Hidden Markov models

Hidden Markov models (HMMs) generalize the idea of finite mixture models. They model the hidden variable U_i, which is just an i.i.d. multinomial r.v. for finite mixture models, as a Markov chain. This tiny extension yields a surprisingly rich class of non linear processes (Frühwirth-Schnatter, 2006). The main advantage is that these models can incorporate the dependency structure of the data, which occur in various fields, for example for financial

or other time series or biological sequence analysis. HMMs provide a flexible and widely used class of models for dependent data, in particular in the presence of overdispersion (for series of count data) or unobserved heterogeneity.
In the first place HMMs were applied in speech recognition and the famous paper by Rabiner (1989) is with 3463 citations (source: ISI Web of Knowledge^{SM}, 29.04.09) by far still the most cited paper in the context of HMMs. Other important applications can be found in computational and molecular biology, especially in biological sequence alignment (Durbin et al., 1999) as well as ion channel applications (De Gunst, Künsch and Schouten, 2001). Moreover, applications can be found in econometrics (Rydén et al., 1998), medical statistics (Altman, 2007), biology (Holzmann et al., 2006b) and computer science (Dainotti et al., 2008). For a comprehensive summary see MacDonald and Zucchini (1997) and Cappé et al. (2005).
The ISI Web of Knowledge^{SM}database demonstrates that HMMs are still a field of lively research, as the Figures 1.1 and 1.2 highlight. They display the distribution of 4886 publications whose topic contains the phrase "hidden Markov" over the years and their citations (in total 82.970).

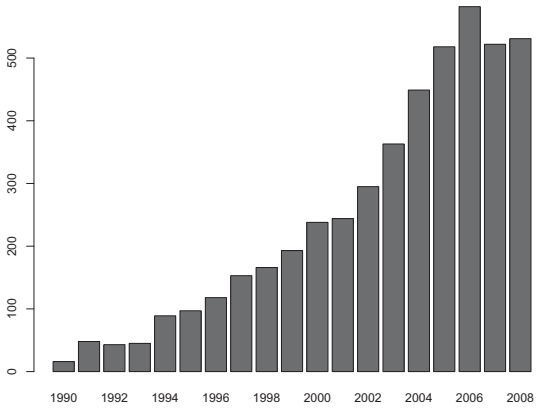


Figure 1.1: Published items in each year. (Source: ISI Web of Knowledge^{SM}, 29.04.09)

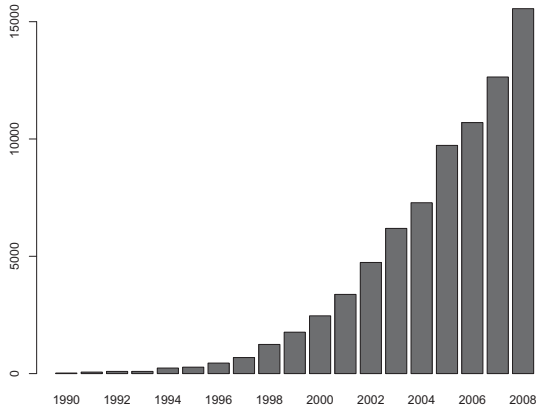


Figure 1.2: Citations in each year. (Source: ISI Web of Knowledge^{SM}, 29.04.09)

As already indicated an HMM can be seen as a finite mixture, where the unobserved switch is not independent but a Markov chain with finite-state space. Therefore, HMMs may be interpreted as an extension of finite mixture, but they can also be seen as *Markov chains observed in noise* (Cappé et al., 2005). As *probabilistic functions of Markov chains* HMMs

were studied for the first time by Baum and Petrie (1966), who considered HMMs with finite state space and finite sample space.
Formally, HMMs consist of two ingredients, an unobservable finite-state Markov chain (U_i) and an observable stochastic process (Y_i), such that

1. the (Y_i) are conditionally independent, given the (U_i) and
2. given the (U_i), the distribution of Y_j depends on U_j only.

This dependency structure can be represented by an undirected graph, where the missing edges between two nodes represent the conditional independence of two r.v.s conditioned on the rest. From the decomposition

$$\begin{aligned} & P(U_1 = u_1, Y_1 = y_1, U_2 = u_2, Y_2 = y_2, \ldots) \\ = \;& P(U_1 = u_1)P(Y_1 = y_1|U_1 = u_1)P(U_2 = u_2|U_1 = u_1)P(Y_2 = y_2|U_2 = u_2)\cdots \end{aligned}$$

we may also represent the dependency structure by an directed graph (see Figure 1.3).

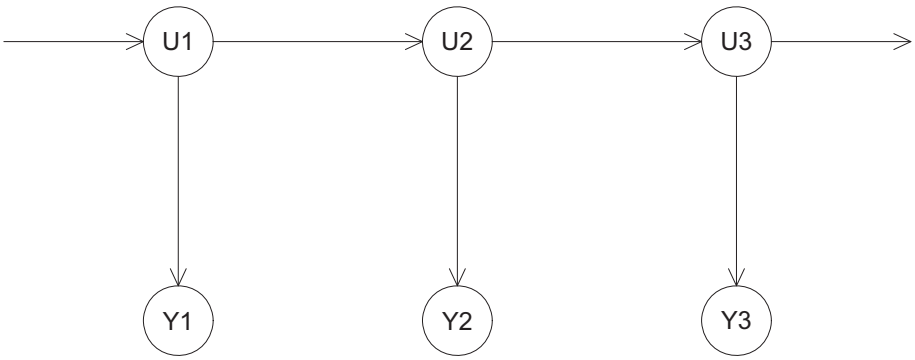


Figure 1.3: Dependency structure of an HMM.

The unobserved process (U_i) is sometimes called *regime*. As realizations of finite Markov chains are called *states*, the conditional distribution functions $P(Y_i \leq y|U_i = k)$ for $k = 1, \ldots, m$ are called *state-dependent distribution functions (sdfs)*. Usually the sdfs come from a parametric family $(F_\theta)_{\theta\in\Theta}$ of distributions, e.g. the normal or the Poisson distribution. In this case the parameter of interest of the model consists of the transition matrix of the Markov chain P and the parameters of the sdfs.
Note that every finite mixture model can be expressed as an HMM just by choosing the transition matrix of the Markov chain to have identical columns, i.e. transition probabilities

do not depend on the state one starts with. Hence, independence can be expressed by a number of equalities on the transition probabilities and can be easily tested via the likelihood ratio procedure, if one assumes the number of states m to be fixed.

Identifiability and estimation in HMMs

As for finite mixtures identifiability is an important issue in the HMM framework. Petrie (1969) gives an identifiability result for HMMs with finite sample space, e.g. $Y_i \in \{1, \dots, N\}$ for some $N \in \mathbb{N}$. For standard HMMs, i.e. with sdfs from the same parametric family, Leroux (1992b) shows how an argument by Teicher (1967) can be used to establish identifiability if it is assumed to hold for the corresponding finite mixture. Teicher (1967) shows that finite mixtures of product distributions $F^*(y) = F_1(y_1) \cdots F_r(y_r)$ from some family $\mathfrak{F}$ are identifiable if and only if finite mixtures of $\mathfrak{F}$ are identifiable. It turns out that this result with $r = 2$ implies identifiability in the HMM framework, as long as it holds for finite mixtures. Hence, HMMs with Gaussian, gamma, Poisson or other classical distributions as sdfs are identifiable. Note, that for HMMs with $\mathbf{n}$-binomial sdfs with $\mathbf{n} < 2m - 1$ Leroux's argument cannot be applied, since the corresponding mixture is not identifiable. However, Petrie (1969) shows identifiability for HMMs with observations in $\{0, \dots, N\}$ for the parameter space $\Theta \setminus \Theta_{Petrie}$ with

$$\Theta = \{(P, F_1, \dots, F_m) | P \text{ ergodic transition matrix, } F_k \text{ distributions on } \{0, \dots, N\}\}$$

and $\Theta_{Petrie} \subset \Theta$ is a null set w.r.t. the Lebesgue measure on $\Theta \subset \mathbb{R}^{m(N+m-1)}$. Since $\{\vartheta \in \Theta | P_{ij} \text{ does not depend on } i \text{ for all } j\} \subset \Theta_{Petrie}$ this result does not contradict Teicher's results on mixtures of binomials. Note that this model is discussed in Ex. 2.2 motivated by MacDonald and Zucchini (1997, pp. 140–144). Some discussion on identifiability as well as treatment of the illustrating example of Gaussian HMMs can be found in Cappé et al. (2005).

There are several estimation techniques proposed in the literature to infer the parameter of interest, i.e. the transition matrix of the chain and the parameters of the sdfs. As for mixtures methods based on maximum likelihood and Bayesian analysis are well elaborated. However, all techniques as well as theoretical analysis are more involved than for finite mixtures. In this thesis we consider maximum likelihood estimation, for detailed discussions of Bayesian methods usual based on MCMC sampling techniques see Frühwirth-Schnatter (2006, Chp. 11.5) and Cappé et al. (2005, Chp. 13). Evaluation of the MLE is much more complicated for HMMs than for mixture models, as the likelihood function is not simply a product as in the independent case. However, it turns out that it can be represented

as a product of matrices, such that computation speed increases only linearly in sample size, not exponentially as one may think of in the first place. For this computation *forward and backward algorithms* are proposed and commonly used (see for example MacDonald and Zucchini, 1997). These algorithms are the building block of many techniques and are closely connected to filtering problems, e.g. the *Viteribi algorithm* computes the most likely sequence of hidden states given the observations. For a profound discussion see also Cappé et al. (2005, Chp. 10). Further discussion of the MLE of HMMs we defer to Chapter 2, especially to Section 2.3.1 for its numerical evaluation.

1.3 Switching regression models

Switching regression models (SRMs), also called mixture regression model, are another extension of finite mixture models. They arise if in addition to population heterogeneity covariates should be taken into account. For a Gaussian response, i.e. linear switching regression, these were introduced by Quandt and Ramsey (1978) and Kiefer (1978). Further, switching regression models are also extensively used for count data, in particular Poisson switching regression as e.g. in Le et al. (1992), Wang and Puterman (2001, 1999), or for binary responses, i.e. switching logistic regression as e.g. in Wang and Puterman (1998). These models allow to incorporate overdispersion relative to the corresponding generalized linear model and can often be nicely interpreted.

We denote the observations by (Y_i, X_i), where Y_i is the response variable and X_i the (multidimensional) covariate, and the latent variable by U_i. The conditional distributions are given by $P(Y_i \leq y | X_i = x_i, U_i = k) = F_k(y|x)$. Usually $F_k(y|x)$ belongs to some parametric family, i.e. $F_k(y|x) = F_{\beta,\theta_k}(y|x)$, where θ is the switching parameter while β coincides in all components. If the U_is are i.i.d. copies following an multinomial distribution on $\{1, \ldots, m\}$ with weights π the density of the observation (Y_i, X_i) is given by

$$f_{\text{switch}}(y_i, x_i) = \left(\pi_1 f(y_i, x_i; \beta, \theta_1) + \ldots + \pi_m f(y_i, x_i; \beta, \theta_m)\right) h(x_i), \tag{1.3}$$

where $f(y_i, x_i; \beta, \theta_k)$ denotes the density of the conditional distribution $F_{\beta,\theta_k}(y_i|x_i)$ and $h(x_i)$ the density of the covariates.

The form of (1.3) indicates a close connection between switching regression models and mixture models with structural parameters, for example a normal mixture with distinct means, but equal variances in the components (Chen and Chen, 2003).

We may also connect the switching regression framework to HMMs by considering a switching regression model, where the switching regime U_i follows a Markov chain and the dependency conditions 1. and 2. in Section 1.2 hold for $(Y_i, X_i)_i$. Such a model is called

Markov-switching regression model (MSRM) and can either be seen as an extension of the switching regression model, by allowing for a Markovian regime, or as an extension of HMMs by embedding covariates into HMMs. The dependency structure of an SRM and an MSRM is displayed in Figure 1.4.

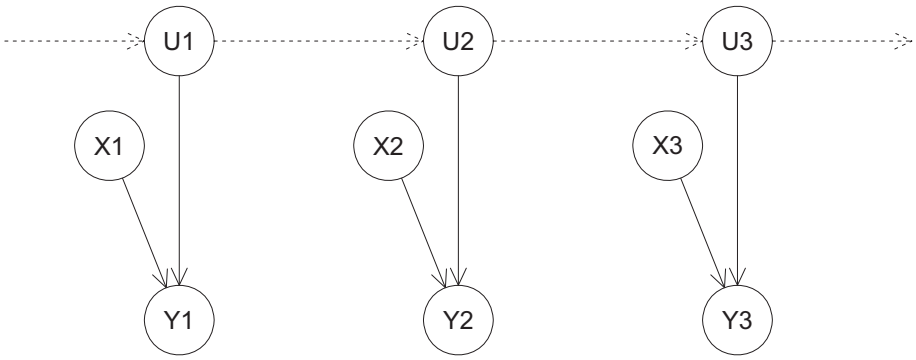


Figure 1.4: Dependency structure of an SRM (without dashed lines) and an MSRM (with dashed lines).

The issue of identifiability essential transfers from mixtures to switching regression models. For example a binomial regression model is only identifiable if $\mathbf{n} \geq 2m - 1$. Additional difficulties occur if the design matrix formed by the X_is does not allow identifiability of the parameters. An interesting example of which is given for the linear regression by Hennig (2000), for a discussion see also Frühwirth-Schnatter (2006, Chp. 8.2.2.).
For estimation techniques based on Bayesian analysis via MCMC algorithms we again refer to Frühwirth-Schnatter (2006, Chp. 8.3.4.). Maximum likelihood estimation is discussed in Section 3.3.

1.4 Other related models

General mixture models and general state-space models

A class closely related to finite mixture models is the class of general mixtures, which will not be discussed further in this thesis but should be mentioned. Assume that H is a probability measure on the parameter space Θ and we associate with each $\theta \in \Theta$ a distribution function F_θ, then the equation (1.1) translates to

$$G(y) = \int_\Theta F_\theta(y) H(d\theta).$$

Inference concerning H based on G is a special case of a statistical inverse problem. Note that, if H is discrete with m support points G represents a finite mixture with m components.
General state-space models extend HMMs in the similar way by replacing the hidden Markov chain with finite state space by a general Markov process on a parameter space Θ with Markov transition kernel Q. As these models are not in the main focus of the thesis we refer to Cappé et al. (2005) for precise definitions and notations.
A prominent class of these models are Gaussian linear state-space models where the conditional distribution functions are Gaussian, e.g.

$$U_i = U_{i-1} + \varepsilon_{1i} \text{ and } Y_i = U_i + \varepsilon_{2i}$$

with $U_1, \varepsilon_{1i}, \varepsilon_{2i}$ are independent zero-mean and homogeneous variance normal r.v.s.
Cappé et al. (2005) treat HMMs with finite state space and general state-space models in their book simultaneously. Their analysis is therefore quite general and applicable to a wide range of models, but of course notionally and technically more involved. Cappé et al. (2005) do not distinguish between HMMs with finite state space and general state-space models conceptually. In contrast to that MacDonald and Zucchini (1997) reserve the term HMM to models with finite state space.

Hidden semi-Markov models

Hidden semi-Markov models extend HMMs in another way. Here the assumption of the Markov property is somehow weakened. The Markov chain is replaced by a semi-Markov chain, which is loosely speaking a Markov chain with arbitrarily specified sojourn time, i.e. the probability of staying in some state over a period of i steps does not necessarily decrease exponentially in i as for Markov chains. Hidden semi-Markov models are therefore enabled to incorporate longer memory into a model. For the application of financial time series models, e.g. stock prices, this is of special interest as "long memory effects" are assumed to be present in this kind of data. The statistical properties of hidden semi-Markov models are investigated by Barbu and Limnios (2006). Applications to Financial Time Series are extensively discussed in Bulla (2006).

Switching autoregression and other models with feedback

So far, all presented models have in common that conditioned on the hidden process the observations are independent. This implies that no feedback between the observations is

possible. Examples for models that incorporate feedback in this context are switching and Markov-switching autoregression models (Douc et al., 2004) and latent state models with feedback (Zucchini et al., 2008). While the latter model has been applied to animal behavior, Markov-switching autoregression models are mainly discussed in econometrics literature (for an overview see Piger, 2009), however the kind of feedback in the two models differs conceptually.

1.5 Notation and standing assumptions

The latent process

Let us start with the latent process $(U_i)_{i\in I}$, usually with $I = \{1, \dots, n\}$ for the sample size $n \in \mathbb{N}$. We assume that (U_i) is a stochastic process with values in $\{1, \dots, m\}$ either simply i.i.d. in case of finite mixture models and switching regression models or for HMMs and Markov switching regression models a Markov chain with transition probabilities $\alpha_{jk}^{(i)} := P(U_{i+1} = k|U_i = j)$ and initial distribution $\pi_k^* = P(U_1 = k)$ for $1 \le j, k \le m$. Throughout the thesis we assume that the Markov chain (U_i) is homogeneous, i.e. the transition probabilities coincide for all i and form a transition matrix $P = (\alpha_{jk})_{jk}$. Moreover, we assume that Markov chain (U_i) is irreducible and aperiodic. This condition ensures that (U_i) is an ergodic process with a unique stationary distribution $\pi = (\pi_1, \dots, \pi_m)$ with $\pi_k > 0$ for all $1 \le k \le m$. We usually assume that the initial distribution π^* coincides with π which yields together with the homogeneity assumption the stationarity of the process (U_i). If (U_i) is i.i.d. π denotes the distribution on $\{1, \dots, m\}$ which coincides with the interpretation of the i.i.d. process as a Markov chain with transitions α_{jk} being equal for all j. In this case ergodicity is simply ensured if $\pi_k > 0$ for all k. We see that ergodicity guarantees that the number of states or components m is a well-defined number.

The observed process

We assume that the observed process (Y_i) takes values in a Borel measurable subset $\mathcal{Y} \subset \mathbb{R}^d$ of a Euclidean space. For all models, namely those discussed in Sections 1.1, 1.2, 1.3, the crucial properties are that the (Y_i) are independent conditioned on the latent process (U_i) and the distribution of Y_j depends on $(U_i)_i$ only through the corresponding U_j

$$P(Y_j \le y|U_1, \dots, U_n, Y_1, \dots, Y_{j-1}, Y_{j+1}, \dots, Y_n) = P(Y_j \le y|U_j) =: F_{U_j}(y).$$

We assume that the F_{U_j}s are also homogeneous, i.e. equal for all j, and for $U_j = k$ have densities f_k for $1 \leq k \leq m$ w.r.t. some σ-finite measure ν on $\mathcal{Y}$. Usually the $f_k = f_{\theta_k}$ belongs to some parametric family of densities $\{f_\theta(y)|\theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^p$. For example $\{f(y;(\mu,\sigma^2)) = 1/(2\pi\sigma)\exp(-(y-\mu)^2/2\sigma^2)|(\mu,\sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}\}$ gives rise to Gaussian mixtures and HMMs. For count data the Poisson family $\{f(y;\lambda) = \lambda^y/y!\exp(-\lambda)|\lambda \in \mathbb{R}_{>0}\}$ is commonly used (e.g. by Le et al., 1992).
In case of parametric HMMs for notational convenience we may view (U_i) as a process on a finite set $\{\theta_1,\ldots,\theta_m\} \subset \Theta \subset \mathbb{R}^d$ with $\theta_j \neq \theta_k$ for $1 \leq j < k \leq m$. Here we may denote the stationary distribution of (U_i) on Θ as $G \in \mathfrak{M}_m$ the set of probability measures on Θ with m support points.

Additional notation for the regression models

In addition to the processes introduced above we assume the presence of covariates $(X_i)_{i\in I}$ with values in a Borel-measurable subset $\mathcal{X} \subset \mathbb{R}^r$. Although the previous notion for the distribution of Y_i conditioned on U_i might be still valid (by integrating over the covariates) we prefer to change the notation slightly by considering the conditional distribution of $Y_i|X_i = x_i, U_i = k$

$$F_k(y_i|x_i) := P(Y_i \leq y_i|X_i = x_i, U_i = k)$$

and the distribution of X_i given by F^X on $\mathcal{X}$ with density h w.r.t. some σ-finite measure on $\mathcal{X}$. We assume that the (Y_i, X_i) are independent for the switching regression model or independent conditioned on $(U_i)_i$ for the Markov switching regression model. Moreover, the distributions of (Y_i, X_i) conditioned on $U_i = k$ possess densities w.r.t. a σ-finite measure on $\mathcal{Y} \times \mathcal{X}$ of the form

$$f_k(y_i, x_i)h(x_i) = f(y_i, x_i; \beta, \theta_k)h(x_i)$$

from some parametric family $\{f(y, x; \beta, \theta)|\beta \in \mathcal{B}, \theta \in \Theta\}$ where we distinguish between the switching parameter $\theta \in \Theta \subset \mathbb{R}^p$ and the structural parameter $\beta \in \mathcal{B} \subset \mathbb{R}^q$ which is equal in all components. We again assume that the model is homogeneous, i.e. the densities are the same for all i. This setting excludes the longitudinal setup based on n_i measurements per observation unit which is of major relevance in the regression context (cf. Zhu and Zhang, 2004). We will indicate how the switching regression model defined above can be adopted to longitudinal data structures in Remark 3.9 in Section 3.3.2.

Chapter 2

Testing in hidden Markov models under nonstandard conditions

In this chapter we introduce maximum likelihood estimation and hypothesis testing based on the likelihood ratio in the context of HMMs. The main focus is to investigate the asymptotic behavior of the maximum likelihood estimator (MLE) and the likelihood ratio test (LRT) under so-called nonstandard conditions. In these cases usually the asymptotic normal or χ^2-distribution does not hold. This occurs for example if the true value lies on the boundary of the parameter space.
Before formally introducing these concepts we may begin with a motivating example representing some relevant testing problem where crucial boundary constraints are present.

A first example

We want to investigate whether a hidden state k is always left immediately, i.e. the (k,k)th entry of the transition matrix is zero:

$$\alpha_{kk} := P(U_{i+1} = k | U_i = k) = 0.$$

Clearly, α_{kk} lies in $[0,1]$, such that this problem is concerned with the boundary of the parameter space.
As HMMs can either be seen as a noisy version of a Markov chain or as a mixture with not i.i.d. but Markovian regime we may watch out for analogous situations in both directions. Let us for a moment assume that $(U_i)_i$ is directly observed, then our testing problem becomes rather trivial. Under the hypothesis $H : \alpha_{kk} = 0$ the event $\{U_i = k, U_{i+1} = k\}$ has probability zero, such that a reasonable testing procedure based on a sample $U_1, \dots, U_n$

would reject H if and only if $T_n = \#\{i|U_i = k, U_{i+1} = k\} > 0$. The distribution of the test statistic T_n under H coincides therefore with the Dirac measure concentrated at zero. Formally this testing procedure can be seen as LRT.
Finding an analogy to our testing problem in context of i.i.d. mixtures is less straight forward, since the notion of transition probabilities is, of course, meaningless in this context. Also, the testing problem for components having zero weights $\pi_k = 0$ does not give a valid analogous setup because in this case crucial regularity conditions are violated, since the number of components is not well-defined (cf. Chapter 3). We may discuss testing

$$H' : \pi_k = \frac{1}{2} \quad \text{against} \quad K' : \pi_k > \frac{1}{2}.$$

By restricting the parameter space $\bar{\Theta} = [1/2, 1]$ this testing problem also appears as a boundary case. The general theory discussing boundary situations for i.i.d. r.v.s (e.g. Self and Liang, 1987) shows that under certain regularity conditions the LRT-statistic behaves under the hypothesis asymptotically as a mixture of a χ^2_0- and χ^2_1-distributed r.v.s with equal weights, where the subindex denotes the number of the degrees of freedom of the χ^2-distribution, the notation χ_0 consistently denotes the Dirac measure at zero.
Summarizing this we note that the i.i.d. analogue suggests that the LRT-statistic for testing $H : \alpha_{kk} = 0$ in an HMM is asymptotically zero with probability 1/2, while the Markov chain analogue yields a distribution degenerated at zero.
In our analysis we actually find both cases represented. On one hand we will show that the results from Self and Liang (1987) and others can be extended to the HMM framework, such that the LRT w.r.t. the likelihood function of an HMM follows asymptotically the "one-half-one-half" mixture under H. On the other hand simulations show that the finite sample behavior of the LRT for many parameter settings exhibits intermediate stages between the two described cases. Especially if the state-dependent distributions are well-separated the weight of χ_0 appears to be close to one even for moderately large sample sizes, such that the theoretical result is a matter of huge sample sizes (cf. Section 2.3.2).

Introductory remarks

As this example indicates, testing problems involving the boundary are frequently encountered in practice of HMMs. Other relevant testing problems might be whether the underlying Markov chain tends to stay in the state k, or whether the state j is on average more frequently visited than the state k. One requires testing for zero-entries of the transition matrix as in the introductory example, testing a one-sided hypothesis on the parameters of the transition matrix and on the parameters of the stationary distribution

of the underlying Markov chain, respectively. All these testing problems require procedures where the boundary situation is taken into account.

For i.i.d. r.v.s testing hypotheses, when the true parameter lies on the boundary or under similar nonstandard conditions, is widely discussed. Classical theoretical contributions are Chernoff (1954), Self and Liang (1987), Shapiro (1985), Shapiro (1988) and others, more recently Drton (2009) introduces algebraic geometric techniques to this field for the analysis of the parameter space and especially its singularities. Boundary situations achieve also a strong interest from the view of applications as demonstrated by many publications, for example in the context of econometrics (Demos and Sentana, 1998), geosciences (Kitchens, 1998, p.812) and clinical trials (Balabdaoui, Mielke and Munk, 2009). More references can be found in the monograph by Silvapulle and Sen (2005).

As the LRT based on the MLE is a major approach for testing hypothesis in the i.i.d. setup for various reasons we may also focus on LRT procedures. In the context of HMMs parameter estimation via likelihood-based methods is well-established. For general HMMs, strong consistency of the MLE was proved by Leroux (1992b). Bickel et al. (1998) established asymptotic normality of the score with limit covariance matrix $\mathcal{J}_0$, as well as a uniform law of large numbers for the Hessian of the log-likelihood with limit matrix $-\mathcal{J}_0$ (for related results see also Douc and Matias, 2001). Once these major results are obtained, the standard likelihood theory such as asymptotic normality of the MLE with limit covariance $\mathcal{J}_0^{-1}$ (Bickel et al., 1998) and the asymptotic χ^2-approximation to the distribution of the LRT under regularity conditions (Giudici et al., 2000) follows as in the i.i.d. setting.

We will show that the likelihood theory under nonstandard conditions with parameters on the boundary, as developed by Chernoff (1954) and Self and Liang (1987), can be extended from the i.i.d. case to HMMs by using the results of Bickel et al. (1998). In particular, we derive the asymptotic distribution theory for the LRT for general, nonlinear hypotheses with parameters on the boundary, and these parameters might also involve the parameters of the state-dependent distributions.

In the following, after introducing to likelihood inference of HMMs, we discuss how the asymptotic distribution theory for the LRT for HMMs under nonstandard conditions. An extensive list of examples is given and simulation results as well as an illustrative application of the tests for a series of epileptic seizure count data, previously analyzed by Le et al. (1992), are presented.

The main results of this chapter are published in Dannemann and Holzmann (2008b).

2.1 Likelihood inference for HMMs

As introduced in Section 1.5 we denote the HMM as bivariate process $(U_i, Y_i)_i$, where $(U_i)_i$ is the unobserved Markov chain and $(Y_i)_i$ the observed data. Throughout the chapter we consider parametric HMMs, i.e. the state-dependent distribution functions (sdfs) are from some parametric family $(f_\theta)_\theta$. The parameter of interest is constituted of the transition matrix $P = (\alpha_{jk})_{1\le j,k\le m}$ and the parameters of sdfs $\theta_k \in \Theta \subset \mathbb{R}^d$ for $k = 1, \dots, m$. We denote the parameter by

$$\vartheta = (\alpha_{11}, \dots, \alpha_{1,m-1}, \alpha_{21}, \dots, \alpha_{m,m-1}, \theta_1, \dots, \theta_m)$$

and assume $\vartheta \in \bar{\Theta} \subset \mathbb{R}^{\bar{d}}$ with $\bar{d} = d + m(m-1)$. In general, ϑ may also denote a parametrization of the HMM that differs from the standard parametrization as defined above, for example if some elements are known and fixed or exhibit a priori equality constraints, e.g. $\alpha_{12}(\vartheta) = \alpha_{32}(\vartheta)$. In this case one may understand in the following the transition probabilities $\alpha_{jk}(\vartheta)$ as well as the parameters of the sdfs $\theta_k(\vartheta)$ as functions of ϑ. The subindex 0 indicates the true value ϑ_0 and the true distribution P_0 of the bivariate process $\big(U_i, Y_i\big)_i$. Note that since the parameters of the transition matrix $\alpha_{jk}(\vartheta)$ depend on ϑ, so do the components of the unique stationary distribution $\pi_k = \pi_k(\vartheta)$.
The joint density of $(U_1, \dots, U_n, Y_1, \dots, Y_n)$ (w.r.t. (counting measure)$^n \times \nu^n$) is given by

$$\begin{aligned} p_n(u_1, \dots, u_n, y_1, \dots, y_n; \vartheta) &= p_n(u_1, \dots, u_n, y_1, \dots, y_n; \alpha_{11}, \dots, \alpha_{m,m-1}, \theta_1, \dots, \theta_m) \\ &= \pi_{u_1} f_{\theta_{u_1}}(y_1) \prod_{i=2}^{n} \alpha_{u_{i-1},u_i} f_{\theta_{u_i}}(y_i) \\ &= \pi_{u_1} \prod_{i=1}^{n-1} \alpha_{u_i,u_{i+1}} \prod_{i=1}^{n} f_{\theta_{u_i}}(y_i), \end{aligned}$$

the joint density of $(Y_1, \dots, Y_n)$ (w.r.t. ν^n) by

$$p_n(y_1, \dots, y_n; \vartheta) = \sum_{u_1=1}^{m} \cdots \sum_{u_n=1}^{m} p_n(u_1, \dots, u_n, y_1, \dots, y_n; \vartheta), \tag{2.1}$$

and the log likelihood is denoted by $L_n(\vartheta) = \log p_n(y_1, \dots, y_n; \vartheta)$. A maximum likelihood estimator (MLE) $\hat{\vartheta}$ is any value of $\vartheta \in \bar{\Theta}$ which maximizes $L_n(\vartheta)$:

$$\hat{\vartheta} := \arg\max_{\vartheta \in \bar{\Theta}} L_n(\vartheta).$$

Computational issues concerning the evaluation of the log likelihood and its maximizer is discussed in Section 2.3.

2.1.1 ML-estimation and LR-testing under regular conditions for HMMs

ML-estimation is well established in the context of HMMs. Baum and Petrie (1966) consider HMMs where the sample space of the observables Y_i is finite. They elaborated the essential techniques for the analysis of MLEs for HMMs. Leroux (1992b) considers, as we do, HMMs with finite state space and general observation space and shows that the MLE is strongly consistent, i.e.

$$\hat{\vartheta} \longrightarrow \vartheta_0 \quad P_0 - \text{a.s.}, \qquad \text{when} \quad n \to \infty$$

under classical Wald-type assumptions (for a detailed discussion of the result see Dannemann, 2006, pp.7-17). Leroux (1992b) also discusses the important issue of identifiability and shows that it holds if (and only if) the corresponding family of m-component mixtures is identifiable.

Asymptotic normality of the MLE

When we speak about asymptotic normality of the MLE we always mean that the sequence $\sqrt{n}(\hat{\vartheta} - \vartheta_0)$ is asymptotically normally distributed with mean zero and finite covariance matrix. Bickel et al. (1998) shows asymptotic normality of the MLE for HMMs. As this result is the corner stone to establish the asymptotic theory for LR-testing under standard and nonstandard conditions we may discuss this result in some detail. We begin with a description of the assumptions under which asymptotic normality is proved by Bickel et al. (1998). Besides ergodicity of the Markov chain they mainly suggest the following regularity conditions:

Assumption 2.1. The maps $\vartheta \mapsto \alpha_{jk}(\vartheta)$ and $\vartheta \mapsto \pi_k(\vartheta)$ for $1 \leq j, k \leq m$ have two continuous derivatives and the maps $\vartheta \mapsto f_{\theta_k(\vartheta)}(y)$ for $1 \leq k \leq m$ and $y \in \mathcal{Y}$ have two continuous derivatives.

Assumption 2.2. Let $\vartheta = (\vartheta_1, \ldots, \vartheta_{\bar{d}})$. There exists a $\delta > 0$ such that
1.) for all $i \in \{1, \ldots, \bar{d}\}$ and for all $1 \leq k \leq m$

$$E_0 \left[\sup_{\vartheta \in B_\delta(\vartheta_0)} \left| \frac{d}{d\vartheta_i} \log f_{\theta_k(\vartheta)}(Y_1) \right|^2 \right] < \infty;$$

2.) for all $i, j \in \{1, \ldots, \bar{d}\}$ and for all $1 \leq k \leq m$

$$E_0 \left[\sup_{\vartheta \in B_\delta(\vartheta_0)} \left| \frac{d^2}{d\vartheta_i d\vartheta_j} \log f_{\theta_k(\vartheta)}(Y_1) \right| \right] < \infty;$$

3.) for $j = 1, 2$, all $i_l \in \{1, \ldots, \bar{d}\}$, $l = 1, \ldots, j$ and for all $1 \leq k \leq m$

$$\int \sup_{\vartheta \in B_\delta(\vartheta_0)} \left| \frac{d^j}{d\vartheta_{i_1} \cdots d\vartheta_{i_j}} f_{\theta_k(\vartheta)}(y) \right| d\nu(y) < \infty.$$

Assumption 2.3. There exists a $\delta > 0$ such that with

$$\rho_0(y) \quad = \quad \sup_{\vartheta \in B_\delta(\vartheta_0)} \max_{1 \leq j,k \leq m} \frac{f_{\theta_j(\vartheta)}(y)}{f_{\theta_k(\vartheta)}(y)},$$

$P_0(\rho_0(Y_1) = \infty | U_1 = k) < 1$ for all $1 \leq k \leq m$.

Following Self and Liang (1987) we formulate in addition conditions on the third derivatives, where the derivatives are meant to be taken from the appropriate side, if ϑ is on the boundary of the parameter space.

Assumption 2.1' The maps $\vartheta \mapsto \alpha_{jk}(\vartheta)$ and $\vartheta \mapsto \pi_k(\vartheta)$ for $1 \leq j, k \leq m$ have three continuous derivatives and the maps $\vartheta \mapsto f_{\theta_k(\vartheta)}(y)$ for $1 \leq k \leq m$ and $y \in \mathcal{Y}$ have three continuous derivatives.

Assumption 2.2' Let $\vartheta = (\vartheta_1, \ldots, \vartheta_{\bar{d}})$. In addition to Assumption 2.2, there exists a $\delta > 0$ such that for all $i, j, l \in \{1, \ldots, \bar{d}\}$ and for all $1 \leq k \leq m$

$$E_0 \left[\sup_{\vartheta \in B_\delta(\vartheta_0)} \left| \frac{d^3}{d\vartheta_i d\vartheta_j d\vartheta_l} \log f_{\theta_k(\vartheta)}(Y_1) \right| \right] < \infty.$$

Note that the Assumptions 2.1, 2.2 and 2.1', 2.2' are so called Cramér-type conditions and appear natural from the classical theory of i.i.d. samples (for discussion cf. also Dannemann, 2006, p.19). Apart from the classical regularity conditions, i.e. mainly existence and boundedness of the derivatives of the log densities, van der Vaart (1998) discusses based on LeCam's work an alternative condition. Based on the notion of differentiability in quadratic mean, i.e. for densities p_ϑ, $p_{\vartheta+h}$ there exists a function g_ϑ with $E[|g_\vartheta|^2] < \infty$ such that

$$E_\vartheta \left[(\sqrt{p_{\vartheta+h}}/\sqrt{p_\vartheta} - 1 - 1/2 h g_\vartheta)^2 \right] = o(|h|^2),$$

van der Vaart shows that the results from Self and Liang (1987) can be derived from this condition for i.i.d experiments (van der Vaart, 1998, see Thm. 7.12 and Thm 16.7). However, extending this concept to dependent data models like HMMs has not been established in the literature so far.

Assumption 2.3 is not very demanding, as pointed out by Bickel and Ritov (1996), it is for example violated if the sdfs of two states have distinct supports. However, Douc

and Matias (2001) and Bickel et al. (2002) give conditions under which results implying asymptotic normality hold that include this case.
Under the Assumptions 2.1-2.3 and assuming that ϑ_0 lies in the interior of $\bar{\Theta}$, the strong consistency of the MLE and the positive definiteness of the Fisher information matrix

$$\mathcal{J}_0 := -\lim_{n\to\infty} n^{-1} D^2_\vartheta L_n(\vartheta_0).$$

Bickel et al. (1998) showed

$$\sqrt{n}(\hat{\vartheta} - \vartheta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{J}_0^{-1}) \quad P_0\text{-weakly.} \tag{2.2}$$

To achieve this Bickel et al. (1998) prove under the presented regularity conditions a central limit theorem (CLT) for the score:

$$\frac{1}{\sqrt{n}} D_\vartheta L_n(\vartheta_0) \xrightarrow{\mathcal{L}} N(0, \mathcal{J}_0) \quad P_0\text{-weakly,} \tag{2.3}$$

and a uniform law of large numbers (ULLN) for the Fisher information, i.e. for any strongly consistent sequence $(\tilde{\vartheta}_n)_n$

$$\frac{1}{n} D^2_\vartheta L_n(\tilde{\vartheta}_n) \to -\mathcal{J}_0 \quad \text{in } P_0\text{-probability.} \tag{2.4}$$

For almost sure convergence results for this law of large numbers see Douc and Matias (2001) and Bickel et al. (2002). After establishing these two lemmas asymptotic normality of the MLE is just a matter of the standard Taylor expansion technique, since

$$0 = D_\vartheta L_n(\hat{\vartheta}) = D_\vartheta L_n(\vartheta_0) + D^2_\vartheta L_n(\bar{\vartheta})(\hat{\vartheta} - \vartheta_0)$$

with $\bar{\vartheta}$ lying on the line segment $[\vartheta_0, \hat{\vartheta}]$. This yields

$$\begin{aligned} \sqrt{n}(\hat{\vartheta} - \vartheta_0) &= (-n^{-1} D^2_\vartheta L_n(\bar{\vartheta}))^{-1} \sqrt{n}^{-1} D_\vartheta L_n(\vartheta_0) \\ &= \mathcal{J}_0^{-1} \sqrt{n}^{-1} D_\vartheta L_n(\vartheta_0) + o_P(1). \end{aligned}$$

by (2.4) and combining this with (2.3) proves (2.2). Note, that if ϑ_0 lies on the boundary of $\bar{\Theta}$ the maximum is not longer necessarily achieved at an inner point of $\bar{\Theta}$ (not even for large n) such that $D_\vartheta L_n(\hat{\vartheta}) = 0$ fails and hence (2.2) may not hold.

LR-testing under standard conditions

We call testing problems as *under standard conditions*, if the parameter space under the hypothesis $\bar{\Theta}_0 \subset \bar{\Theta}$ is given by a smooth manifold with ϑ_0 lying in the interior of $\bar{\Theta}_0$ and $\bar{\Theta}$ (w.r.t. to the relative topologies). For testing the hypothesis

$$H : \vartheta \in \bar{\Theta}_0 \quad \text{against} \quad K : \vartheta \in \bar{\Theta} \setminus \bar{\Theta}_0$$

we denote the LRT-statistic T_n by

$$T_n = 2\left(\sup_{\vartheta\in\bar{\Theta}} L_n(\vartheta) - \sup_{\vartheta\in\bar{\Theta}_0} L_n(\vartheta)\right) \tag{2.5}$$

Suppose that $\bar{\Theta}_0$ is (at least locally around ϑ_0) parametrized by some smooth function $s : \bar{\Theta} \to \mathbb{R}^k$ for $k \leq \bar{d}$ with derivative of rank k s.t. $\bar{\Theta}_0 = \{\vartheta \in \bar{\Theta} : \ s(\vartheta) = 0\}$, where k corresponds to the codimension of $\bar{\Theta}_0$. Then the usual asymptotic χ^2–approximation applies:

$$T_n \xrightarrow{\mathcal{L}} \chi^2_k.$$

Giudici et al. (2000) show this result for HMMs in analogy to the i.i.d. case based on quadratic expansions of the likelihood function and on the fundamental lemmas (2.3,2.4). Dannemann and Holzmann (2008a) extended this methodology to two-sample problems.

2.1.2 ML-estimation and LR-testing under nonstandard conditions for HMMs

As mentioned above, asymptotic normality of $\hat{\vartheta}$ may not hold if the true parameter ϑ_0 lies on the boundary of the parameter space. In order to derive the asymptotic distribution of the MLE and the LRT under such *nonstandard conditions* we introduce the following definition of approximating cones.

Definition 2.1. A set $\bar{\Theta} \subset \mathbb{R}^{\bar{d}}$ is said to be approximated at ϑ_0 by a cone with vertex at ϑ_0, denoted by $C_{\bar{\Theta},\vartheta_0} = C_{\bar{\Theta}}$, if

$$\inf_{z\in C_{\bar{\Theta}}} \|z - y\| = o(\|y - \vartheta_0\|), \qquad \text{for all } y \in \bar{\Theta},$$

and

$$\inf_{y\in\bar{\Theta}} \|z - y\| = o(\|z - \vartheta_0\|), \qquad \text{for all } z \in C_{\bar{\Theta}}.$$

A cone C with vertex at ϑ_0 is given if for $z \in C$, we have $a(z - \vartheta_0) + \vartheta_0 \in C$ for all $a > 0$. Further, if C_Θ is a cone with vertex at ϑ we denote the corresponding centered cone by $C_\Theta - \vartheta$. The concept for approximating cones was introduced to conduct likelihood inference by Chernoff (1954) and is by now the standard approach used in many related results (Self and Liang, 1987; Silvapulle and Sen, 2005).

Self and Liang (1987) show for i.i.d. models that the asymptotic distribution of the MLE, restricted to a subset $\Theta \subset \mathbb{R}^d$, which is approximated at ϑ_0 by a cone C_Θ, is related to the

distribution of the MLE restricted to the cone $C_\Theta - \vartheta_0$ based on a single observation from a normal experiment with known covariance matrix $\mathcal{J}^{-1}$. Let us denote the density of a multivariate normal with mean ϑ and covariance matrix $\mathcal{J}^{-1}$ by

$$f(z;\vartheta) = \frac{|\mathcal{J}^{1/2}|}{(2\pi)^{n/2}} \exp(-\frac{1}{2}(z-\vartheta)^T \mathcal{J}(z-\vartheta))$$

then the restricted MLE for ϑ based on an observation $Z \sim \mathcal{N}\big(0, \mathcal{J}^{-1}\big)$ is given by

$$\begin{aligned} \hat{\vartheta} &= \arg\max_{\vartheta\in\Theta} f(Z;\vartheta) \\ &= \arg\min_{\vartheta\in C_\Theta - \vartheta_0} (Z-\vartheta)^T \mathcal{J}(Z-\vartheta). \end{aligned}$$

Self and Liang (1987) show that the calculation of the asymptotic distribution of the restricted MLE in a general i.i.d model boils down to examine the asymptotic behavior of the corresponding restricted MLE in a normal experiment. They establish this result by proving $\sqrt{n}$-consistency of the MLE, verifying a quadratic expansion of the log-likelihood and using the exchangeability of Θ and C_Θ in the maximization step. This strategy is strongly related to arguments in Chernoff (1954). The corner stones of this analysis are built by a CLT for the score and an ULLN for the Fisher information, which are fulfilled in i.i.d. models under reasonable regularity conditions. Since these two ingredients are also available for HMMs we may formulate the following theorem related to Theorem 2 in Self and Liang (1987).

Theorem 2.1. *Suppose that we have an HMM with ergodic regime fulfilling Assumptions 2.1', 2.2', 2.3 with positive definite Fisher information $\mathcal{J}_0$ and assume that $\bar{\Theta}$ can be approximated at the true value ϑ_0 by a cone $C_{\bar{\Theta}}$ with vertex at ϑ_0. Then, if the MLE $\hat{\vartheta}$ is strongly consistent, we have*

$$\sqrt{n}\big(\hat{\vartheta} - \vartheta_0\big) \xrightarrow{\mathcal{L}} \arg\min_{\vartheta\in C_{\bar{\Theta}} - \vartheta_0} (Z-\vartheta)^T \mathcal{J}_0 (Z-\vartheta), \tag{2.6}$$

where $Z \sim \mathcal{N}\big(0, \mathcal{J}_0^{-1}\big)$.

The proof follows the steps in Self and Liang (1987) by using (2.3,2.4) from Bickel et al. (1998). It is deferred to Section 2.4.

Remark 2.1. Note that the asymptotic distribution of the MLE as displayed in (2.6) depends strongly on the form of the cone as well as on the Fisher Information matrix $\mathcal{J}_0$, which is for HMMs not as straight forward accessible as for i.i.d. models. In order to compute the limit distribution, one has to know the form of the cone and thus whether ϑ lies on the boundary in advance. Self and Liang (1987) give some examples for explicit calculations. An example in the context of HMMs is discussed in Section 2.2.

Similarly as the asymptotic behavior of the MLE we may also investigate the LRT under nonstandard conditions in the context of HMMs by reformulating the related result from the i.i.d. setup. Let us again consider the general testing problem

$$H : \vartheta \in \bar{\Theta}_0 \quad \text{against} \quad K : \vartheta \in \bar{\Theta} \setminus \bar{\Theta}_0$$

but now we do not assume that $\bar{\Theta}_0$ is smooth around ϑ_0, but we allow that ϑ_0 lies on the boundary of $\bar{\Theta}_0$ and possibly also of $\bar{\Theta}$. Then we can derive the asymptotic distribution of the LRT-statistic T_n given in (2.5).

Theorem 2.2. *Suppose that we have an HMM with ergodic regime fulfilling Assumptions 2.1', 2.2', 2.3 with positive definite Fisher information $\mathcal{J}_0$ and assume that $\bar{\Theta}_0$ and $\bar{\Theta}$ can be approximated at the true value ϑ_0 by cones $C_{\bar{\Theta}_0}$ and $C_{\bar{\Theta}}$ with vertex at ϑ_0, respectively. Then, if the restricted MLE as well as unrestricted MLE are strongly consistent, we have*

$$T_n \xrightarrow{\mathcal{L}} \inf_{\vartheta \in C_{\bar{\Theta}_0} - \vartheta_0} \big(Z - \vartheta\big)^T \mathcal{J}_0 \big(Z - \vartheta\big) - \inf_{\vartheta \in C_{\bar{\Theta}} - \vartheta_0} \big(Z - \vartheta\big)^T \mathcal{J}_0 \big(Z - \vartheta\big), \tag{2.7}$$

where $Z \sim \mathcal{N}\big(0, \mathcal{J}_0^{-1}\big)$.

The proof follows the same arguments as in the proof of Theorem 2.1 and is along the proof of Theorem 3 in Self and Liang (1987). Again essential ingredients are (2.3, 2.4) from Bickel et al. (1998). Details are given in Section 2.4.

Remark 2.2. Similar to Theorem 2.1, the asymptotic distribution of T_n depends on the form of the cone at ϑ_0, such that for exact evaluation this needs to be known. Of course, it also depends on the Fisher information matrix for HMMs $\mathcal{J}_0$. In some examples (cf. Example 2.3 in the Section 2.2), it can be evaluated algebraically, otherwise, if required it has to be estimated, for example, by using a version of the forward algorithm for computing the observed information matrix (cf. Lystig and Huges, 2002). Alternatively, if one uses direct numerical maximization for computation of the MLE $\hat{\vartheta}$, most algorithms give an estimate of the Hessian matrix at $\hat{\vartheta}$ in addition. From (2.4) one can then in principle determine an estimate of $\mathcal{J}_0$. Still, estimation of $\mathcal{J}_0$ is a difficult problem, and the approximation (2.7) works best if the right side does not depend on $\mathcal{J}_0$, e.g. if the asymptotic distribution reduces to mixture of χ^2-distributions with different degrees of freedom and known weights (e.g. Example 2.1).
Moreover, the asymptotic distribution is derived in Theorem 2.2 for a single true value $\vartheta_0 \in \bar{\Theta}_0$. Different $\vartheta_0 \in \bar{\Theta}_0$ surely lead to different distributions, such that for testing the hypothesis of the form above it is advocated to compute the critical values for all $\vartheta_0 \in \bar{\Theta}_0$

and reject H if the test statistic exceeds the largest one (cf. Self and Liang, 1987, p.610). However, as for ϑ_0 being an inner point of $\bar{\Theta}_0$ the testing problem is trivial (for large n) just by consistency, the analysis concentrates on the boundary of $\bar{\Theta}_0$.

2.2 Examples

Self and Liang (1987) and others treat several examples of the asymptotic distributions presented in the previous section for the i.i.d. framework. An important role in the analysis of (2.7) play so called $\bar{\chi}^2$-distributions, which are mixtures of χ^2-distributions with different degrees of freedom (cf. Silvapulle and Sen, 2005, Chp. 3). In this section we discuss a extensive list of examples which arise specifically in the HMM context. We will see that the right hand side of (2.7) can often, but not always expressed as a $\bar{\chi}^2$-distribution.

Example 2.1 (*Zero entries of the transition matrix*)**.** In order to reduce the number of parameters in an HMM, and sometimes also from the context of the statistical problem, it is reasonable to restrict attention to transition matrices with certain prespecified zero entries (for an example see Eq. (2.19)). Therefore, testing for zeros in the transition matrix is evidently of some practical interest as also argued in the introductory example. Such boundary cases for LR testing were also studied by Bartolucci (2006, Sec. 4) for latent Markov models, here we investigate them in a general HMM framework. Before considering the testing problem, we briefly discuss the asymptotic behavior of the MLE $\hat{\alpha}_{jk}$ of an entry $\alpha_{jk}(\vartheta_0) = 0$ with $1 \leq j, k \leq m$, where it is assumed that the transition matrix still is ergodic, and the parameters of the sdfs are allowed to vary. In case of two states, if $\alpha_{21} \neq 0$ and $\alpha_{22} \neq 0$ the regime is ergodic for $\alpha_{11} = 0$ but evidently not for $\alpha_{12} = 0$. With Theorem 2.1 we have

$$\begin{aligned} \sqrt{n}\,\hat{\alpha}_{jk} \quad &\xrightarrow{\mathcal{L}} \quad \arg\min_{\alpha_{jk} \geq 0} \; (Z - \alpha_{jk})^2 \\ &= \begin{cases} Z & \text{if } Z \geq 0 \\ 0 & Z < 0 \end{cases} \end{aligned} \tag{2.8}$$

with $Z \sim \mathcal{N}\big(0, \sigma^2\big)$, where σ^2 denotes the asymptotic variance of $\hat{\alpha}_{jk}$. Hence the MLE follows a mixture of a point measure at 0 and a half-normal distribution with equal weights. Now, let us focus on testing the hypothesis that a certain entry α_{jk} of the transition matrix is zero:

$$H_{jk}^0 : \quad \alpha_{jk} = 0 \qquad \text{against} \quad K_{jk}^0 : \quad \alpha_{jk} > 0.$$

Under H_{jk}^0, we have one parameter of interest on the boundary and $\bar{d}-1$ nuisance parameter (not on the boundary). Hence the centered cones $C_{\bar{\Theta}_0} - \vartheta_0$ and $C_{\bar{\Theta}} - \vartheta_0$ are up to affine and orthonormal transformations given by

$$\tilde{C}_{\bar{\Theta}_0} = \{0\} \times \mathbb{R}^{\bar{d}-1} \qquad \text{and} \qquad \tilde{C}_{\bar{\Theta}} = [0, \infty) \times \mathbb{R}^{\bar{d}-1}.$$

As described in Self and Liang (1987, case 5) this ensures together with Theorem 2.2

$$T_n \xrightarrow{\mathcal{L}} \frac{1}{2}\,\chi_0^2 + \frac{1}{2}\,\chi_1^2. \tag{2.9}$$

If one combines several of the H_{jk}^0–type hypotheses, T_n will have a $\bar{\chi}^2$-distribution (Bartolucci, 2006), where the weights in the $\bar{\chi}^2$-distribution can be determined from the entries of the Fisher information matrix. Some results for $\bar{\chi}^2$-distributions as well as simulation procedures accessing the weights are discussed in Silvapulle and Sen (2005, Chp. 3). Thus, a joint test would involve estimation of the Fisher information matrix. A simpler, though less efficient method would be to test several of the H_{jk}^0–type hypotheses via some multiple testing procedure.

Example 2.2 (*Boundary cases for parameters of the state-dependent distributions)***.** Another possibility for model reduction is to test whether certain parameters of the sdfs are on the boundary of their parameter spaces.
As a first example, suppose that the underlying Markov chain has two states, and that the sdfs are binomial $B(\mathbf{n}, p)$

$$f_{\theta_k}(y) = \binom{\mathbf{n}}{y} p_k^y\,(1-p_k)^{\mathbf{n}-y}, \quad p_k \in [0,1] \quad k = 1, 2.$$

Then it might be of interest to test whether in one of the states of the Markov chain (e.g. state 1) the outcome is just deterministic. This can be formulated by testing

$$H : p_1 = 0 \qquad \text{against} \qquad K : p_1 > 0.$$

If the regularity conditions are fulfilled, in particular if the model is identifiable, the corresponding LRT again has the asymptotic distribution (2.9).
Note that an HMM with binomial sdfs is identifiable, if $\mathbf{n} \geq 2m-1 \geq 3$, since identifiability then holds for the corresponding finite mixture of binomials. For an HMM with Bernoulli sdfs, i.e. $\mathbf{n} = 1$, as used for example in MacDonald and Zucchini (1997, pp. 140–144), Petrie (1969) shows identifiability up to a Lebesgue null set Θ_{Petrie}, where Θ_{Petrie} includes all finite mixture models of Bernoulli distributed r.v.s (cf. Section 1.2). From a theoretical

viewpoint such a result is quite satisfying. However, in practice the assumption $\vartheta_0 \notin \Theta_{Petrie}$ might be problematic. For ϑ_0 being close to Θ_{Petrie} the finite sample performance of the MLE and also the LRT might be affected.
Another example arises in the analysis of count data (Leroux and Puterman, 1992). Often, overdispersion (relative to Poisson), i.e. the variance in the sample exceeds the mean, is present in such data sets and is modeled by finite mixtures of Poisson distributions. HMMs then provide a natural generalization to the time-series context. A closer analysis may indicate that overdispersion mainly arises since there are too many zero-observations, which can be modeled by zero-inflation of the Poisson distribution (e.g., van den Broek, 1995). This can be interpreted as a two-component Poisson mixture or two-state Poisson HMM with $\lambda = 0$ for one of the components or states. Thus, in the context of overdispersed count series, testing for zero inflation against general overdispersion structure can be accomplished by testing

$$H : \lambda_1 = 0 \qquad \text{against} \qquad K : \lambda_1 > 0.$$

The LRT exhibits under H again the asymptotic distribution (2.9).

Example 2.3 (*One-sided tests for the transition probabilities*)**.** We consider one-sided hypotheses for the entries of the transition matrix. First consider

$$H : \alpha_{jl} \geq \alpha_{kl} \qquad \text{against} \qquad K : \alpha_{jl} < \alpha_{kl},$$

where $1 \leq j, k, l \leq m$, i.e. it is more probable under the alternative to have reached state l coming from state k than coming from state j. On the boundary of H, i.e. for $\alpha_{jl} = \alpha_{kl}$, T_n has the asymptotic distribution (2.9), and if $\alpha_{jl} > \alpha_{kl}$, by strong consistency of the MLE, $T_n \to 0$ in probability. Similarly, one can test

$$H : \alpha_{jk} \geq \alpha_{jl} \qquad \text{against} \qquad K : \alpha_{jk} < \alpha_{jl},$$

for $1 \leq j, k, l \leq m$ that under the alternative it is more likely to go from j to l than to k. Again T_n has the asymptotic distribution (2.9) on the boundary $\{\alpha_{jk} = \alpha_{jl}\}$ and tends to zero in the interior of the hypothesis $\{\alpha_{jk} > \alpha_{jl}\}$.
Next we examine hypotheses of the form

$$H_{jk}^q : \quad \alpha_{jk} \leq q \qquad \text{against} \qquad K_{jk}^q : \quad \alpha_{jk} > q \tag{2.10}$$

for some $q \in (0, 1)$ and $1 \leq j, k \leq m$. For $\alpha_{jk}(\vartheta_0) = q$, we again have the asymptotic distribution (2.9). For $\alpha_{jk}(\vartheta_0) < q$, we have $T_n \to 0$ in probability as in the previous

examples. The asymptotic distribution (2.9) also holds true if the testing problem (2.10) is slightly changed into $H^q_{jk} : \alpha_{jk} = q$ against $K^q_{jk} : \alpha_{jk} > q$.

As mentioned, the evaluation of the asymptotic distribution in Theorem 2.2 can be complicated and may involve calculations of the Fisher Information matrix for HMMs. We illustrate this in the following by considering several joint tests of H^q_{jk}-type hypotheses. We will see that the resulting asymptotic distributions are $\bar{\chi}^2$–distributions for some, but not all cases.
For the H^q_{jk}-type hypothesis, a relevant special case is when $j = k$ and $q = 1/2$, since in this case under the alternative the HMM tends to stay in state j. We shall call such a state stable. Let us consider joint tests on two states $j, k \in \{1 \dots m\}$. First we examine the testing problem

$$H_{j\wedge k}: \quad \alpha_{jj} = 1/2 \ \wedge \alpha_{kk} = 1/2 \qquad \text{against} \qquad K_{j\wedge k}: \quad \alpha_{jj} > 1/2 \ \wedge \alpha_{kk} > 1/2. \tag{2.11}$$

We shall only derive the limit law in a special situation, namely if there are only two states and only the transition probabilities are allowed to vary. Here, under $H_{j\wedge k}$, we have both parameters lying on the boundary. Hence, the derivations of Self and Liang (1987, case 7) apply. They consider the transformed cones

$$\tilde{C}_{\bar{\Theta}_0} = \{(0,0)\} \qquad \text{and} \qquad \tilde{C}_{\bar{\Theta}} = \mathcal{J}^{1/2}\left([0,\infty) \times [0,\infty)\right),$$

where $\mathcal{J}$ denotes the Fisher information. As this also applies to the testing problem (2.11) w.r.t. the Fisher information for HMMs, we have by Theorem 2.2 that under $H_{j\wedge k}$

$$T_n \xrightarrow{\mathcal{L}} \Big(\frac{1}{2} - p\Big)\,\chi^2_0 + \frac{1}{2}\chi^2_1 + p\,\chi^2_2, \tag{2.12}$$

where the mixing quantity p is determined by the Fisher information matrix $\mathcal{J}_0$ and can be evaluated as

$$p = \big(\cos^{-1}\rho\big)/(2\pi), \tag{2.13}$$

where ρ is the correlation coefficient in the covariance matrix $\mathcal{J}_0$. Algebraic evaluation of $\mathcal{J}_0$ in the special case of an HMM with two states, where only the transition probabilities are allowed to vary, shows that

$$\rho = -\frac{\int f_{\theta_1}(y) f_{\theta_2}(y) / \big(f_{\theta_1}(y) + f_{\theta_2}(y)\big) dy}{\Big(\int f^2_{\theta_1}(y) / \big(f_{\theta_1}(y) + f_{\theta_2}(y)\big)\, dy \int f^2_{\theta_2}(y) / \big(f_{\theta_1}(y) + f_{\theta_2}(y)\big)\, dy\Big)^{1/2}}, \tag{2.14}$$

for $\theta_1 = \theta(j)$, $\theta_2 = \theta(k)$. A proof for this is given in Section 2.4. Note that $-\rho$ is always nonnegative, so that $1/4 \le p \le 1/2$, and in particular the asymptotic distribution in (2.12)

will be stochastically larger than $\frac{1}{4}\,\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$, although if the distributions f_{θ_1} and f_{θ_2} almost have disjoint support, these distributions will be close. An asymptotic upper stochastic bound of T_n following (2.12) is in general simply given by $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$.

The testing problem (2.11) is somewhat artificial, and if one intends to test the hypothesis that neither state is stable against the alternative that both states are stable, the testing problem should be formulated as

$$H_{j\wedge k}: \quad \alpha_{jj} \le 1/2 \ \wedge \alpha_{kk} \le 1/2 \qquad \text{against} \qquad K_{j\wedge k}: \quad \alpha_{jj} > 1/2 \ \wedge \alpha_{kk} > 1/2. \quad (2.15)$$

It turns out that in this testing problem, the asymptotic distribution given in (2.7) is no longer a $\bar{\chi}^2$-distribution. A similar phenomenon was observed by Self and Liang (1987) in case when a nuisance parameter lies on the boundary of the hypothesis. Note that this is not true for the testing problem (2.15). In our case the reason is that the whole parameter space under investigation, $\bar{\Theta} = K_{j\wedge k} \cup H_{j\wedge k}$ (which is its approximating cone at $(1/2, 1/2)$) is not convex. The asymptotic distribution of T_n for the testing problem (2.15) in case of two states with known sdfs for $\alpha_{jj}(\vartheta_0) = \alpha_{kk}(\vartheta_0) = 1/2$ turns out to be

$$T_n \xrightarrow{\mathcal{L}} \frac{1}{2}\,\chi_0^2 + \frac{\pi-\phi}{2\pi}\chi_2^2 + \frac{\pi-\phi}{2\pi}P_1(\phi) + \frac{2\phi-\pi}{2\pi}P_2(\phi), \quad (2.16)$$

where $\phi = \cos^{-1}\rho \in [\pi/2, \pi)$, ρ is given in (2.14), $P_1(\phi)$ has density

$$h_1(t;\phi) = \frac{2}{\pi-\phi}\int_{\phi}^{(\phi+\pi)/2} \frac{1}{2\,a_1(\psi,\theta)} \exp\Big(-\frac{t}{2\,a_1(\psi,\phi)}\Big)\,d\psi,$$

and $P_2(\phi)$ has density

$$h_2(t;\phi) = \frac{2}{2\phi-\pi}\int_{\pi/2}^{\phi} \frac{1}{2\,a_2(\psi)} \exp\Big(-\frac{t}{2\,a_2(\psi)}\Big)\,d\psi,$$

and the functions $a_1(\psi,\phi)$ and $a_2(\psi)$ are given by

$$\begin{aligned} a_1(\psi,\phi) &= \sin^2(\pi-\psi) - \sin^2(\psi-\phi), \\ a_2(\psi) &= \cos^2(\psi-\pi/2). \end{aligned}$$

The proof is given in the Section 2.4. An asymptotic upper stochastic bound of T_n is given by $\frac{1}{2}\,\chi_0^2 + \frac{1}{2}\,\chi_2^2$, this bound is not attained for any possible value of ϕ. For other parameter constellations under $H_{j\wedge k}$, the limit distribution is stochastically smaller.
The asymptotic distributions (2.12) and (2.16) of the closely related testing problems (2.11) and (2.15) differ surprisingly strongly. Since testing problems which lead to these

asymptotics arise in other contexts as well, this example is of more general interest, also for the i.i.d. setting.
Next let us consider the hypothesis that neither state is stable against the alternative that at least one state is stable:

$$H_{j\wedge k}: \quad \alpha_{jj} \le 1/2 \ \wedge \alpha_{kk} \le 1/2 \qquad \text{against} \qquad K_{j\vee k}: \quad \alpha_{jj} > 1/2 \ \vee \alpha_{kk} > 1/2.$$

Then for all parameter values on the boundary of $H_{j\wedge k}$ except for $\alpha_{jj}(\vartheta_0) = 1/2$ and $\alpha_{kk}(\vartheta_0) = 1/2$, the asymptotic distribution in (2.9) holds. However, the relevant asymptotics are those under $\alpha_{jj}(\vartheta_0) = 1/2$ and $\alpha_{kk}(\vartheta_0) = 1/2$, since again by Self and Liang (1987, case 7), T_n follows (2.12), which is asymptotically stochastically larger then the limit law in (2.9).
Finally, we want to test the hypothesis that at most one state is stable against the alternative that both states are stable:

$$H_{j\vee k}: \quad \alpha_{jj} \le 1/2 \ \vee \alpha_{kk} \le 1/2 \qquad \text{against} \qquad K_{j\wedge k}: \quad \alpha_{jj} > 1/2 \ \wedge \alpha_{kk} > 1/2. \quad (2.17)$$

For $\alpha_{jj}(\vartheta_0) = 1/2$ and $\alpha_{kk}(\vartheta_0) = 1/2$ and in case of two states with known sdfs we have under $H_{j\vee k}$

$$T_n \xrightarrow{\mathcal{L}} \frac{2\pi - \phi}{2\pi}\,\chi_0^2 + \frac{\phi}{2\pi} P_3(\phi), \quad (2.18)$$

where again $\phi = \cos^{-1}\rho \in [\pi/2, \pi)$, ρ is given in (2.14) and $P_3(\phi)$ has density

$$h_3(t;\phi) = \frac{2}{\phi}\int_0^{\phi/2} \frac{1}{2\,\sin^2\psi} \exp\Big(-\frac{t}{2\,\sin^2\psi}\Big)\,d\psi.$$

The proof requires an analysis similar to the one leading to (2.16) and is given in Section 2.4. Note that $P_3(\pi/2)$ coincides with the distribution of the minimum of two independent χ_1^2-distributed r.v.s. An asymptotic upper stochastic bound of T_n is given by $\frac{1}{2}\,\chi_0^2 + \frac{1}{2}\,\chi_1^2$. Since for all parameter values on the boundary of $H_{j\vee k}$ except for $\alpha_{jj} = 1/2$ and $\alpha_{kk} = 1/2$, the asymptotics (2.9) hold true, a test decision based on the asymptotic critical value in (2.9) will keep the level all over $\bar{\Theta}_0$.
The asymptotic distribution (2.18) illustrates once more that limiting distributions other than those of $\bar{\chi}^2$-type may occur as for example minima of χ^2-r.v.s. The relevance of distributions of this type was also recently observed by Drton (2009) and Balabdaoui, Mielke and Munk (2009).

Example 2.4 (*Tests on the stationary distribution).* For an ergodic Markov chain, the transition probability matrix uniquely determines the stationary distribution π. Hence,

tests on the entries of π can be reformulated into tests for the entries of the transition probability matrix, and Theorem 2.2 in principle allows to test one-sided hypotheses such as $\pi_j \geq \pi_k$ for $1 \leq j, k \leq m$. However, the formulas for π in terms of the α_{jk}s are highly nonlinear for more than two states, which makes explicit maximization under the hypothesis difficult. We illustrate this issue for two states and for a certain type of transition matrices in case of three states. A general, but different approach to this problem based on a likelihood function under independence assumption will be discussed in Section 3.2.1.

First consider the case of two states, and suppose that we want to test certain restrictions on π_1 (or equivalently on $\pi_2 = 1 - \pi_1$), where the parameters of the sdfs are allowed to vary. Let $\alpha_{12} = \alpha$ and $\alpha_{21} = \beta$, then $\pi_1 = \beta/(\alpha + \beta)$. Consider testing the hypothesis that state 1 is on average at least as often visited as state 2, i.e.

$$H_1: \quad \pi_1 \geq \pi_2 \qquad \text{against} \qquad K_1: \quad \pi_1 < \pi_2.$$

Evidently, H_1 is equivalent to the linear restriction $\beta \geq \alpha$, and on the boundary of H_1, i.e. for $\alpha = \beta = 1/2$, one has the asymptotic distribution given in (2.9). Similarly, general restrictions $H_{1,p} : \pi_1 \leq p$ for some $p \in (0, 1)$ can be formulated into linear restrictions $H_{1,p} : (1 - p)\beta \leq p\alpha$, and on the boundary of the hypothesis, (2.9) applies as well.

Next we consider HMMs with three states, where the transition matrix is supposed to be given by

$$\begin{pmatrix} 1-\alpha & \alpha & 0 \\ \beta & 1-\beta-\gamma & \gamma \\ 0 & \delta & 1-\delta \end{pmatrix}, \tag{2.19}$$

Here state 2 can be interpreted as a transitory state, through which every transition from state 1 to state 3 has to pass. The stationary distribution is given by

$$\pi_1 = \frac{\delta\beta}{\delta\beta + \delta\alpha + \gamma\alpha}, \quad \pi_2 = \frac{\delta\alpha}{\delta\beta + \delta\alpha + \gamma\alpha}, \quad \pi_3 = \frac{\gamma\alpha}{\delta\beta + \delta\alpha + \gamma\alpha}.$$

Evidently, even linear restrictions on the parameters π_1 will lead to nonlinear restrictions for the parameters of the transition matrix. For example, consider the hypothesis $H_{1,3} : \pi_1 \geq \pi_3$, which is equivalent to $H_{1,3} : \delta\beta \geq \gamma\alpha$. Using approximating cones, Theorem 2.2 yields that on the boundary of the hypothesis, one again obtains the asymptotic distribution (2.9) for T_n.

Another hypothesis which might be relevant in this situation is to test whether the transition state is prevalent while the boundary states are equally likely:

$$H_{1,2,3}: \quad \pi_1 = \pi_3 \ \wedge \ \pi_1 \geq \pi_2 \qquad \text{against} \qquad K_{1,2,3}: \quad \pi_1 \neq \pi_3 \ \vee \ \pi_1 < \pi_2.$$

which in terms of the parameters of the transition matrix can be formulated as

$$H_{1,2,3}: \quad \delta\beta = \gamma\alpha \ \wedge \ \alpha \geq \beta \qquad \text{against} \qquad K_{1,2,3}: \quad \delta\beta \neq \gamma\alpha \ \vee \ \alpha < \beta$$

In this situation, we have one parameter of interest on the boundary, one parameter of interest not on the boundary as well as $\bar{d} - 2$ nuisance parameter (not on the boundary). Here the centered cones are (up to coordinate transformation) given by

$$\tilde{C}_{\bar{\Theta}_0} = \{0\} \times [0, \infty) \times \mathbb{R}^{\bar{d}-2} \qquad \text{and} \qquad \tilde{C}_{\bar{\Theta}} = \mathbb{R}^{\bar{d}}.$$

Analogously to Self and Liang (1987, case 6) this implies that (2.7) is represented by a mixture of χ^2-distributions with one and two degrees of freedom, respectively, and equal weights. Hence, with Theorem 2.2, on the boundary of the hypothesis $H_{1,2,3}$ we have

$$T_n \xrightarrow{\mathcal{L}} \frac{1}{2}\,\chi_1^2 + \frac{1}{2}\,\chi_2^2.$$

2.3 Simulations and empirical illustration

We now investigate the finite sample properties of the proposed estimation and testing procedures, namely the MLE and LRT under nonstandard conditions in a Monte Carlo study. As the basis of this analysis is built by numerical evaluation of the likelihood function for HMMs $L_n(\vartheta)$ and its maximizer we briefly comment on that issue at the beginning of this section. This issue is also discussed in the monographs by Cappé et al. (2005); Durbin et al. (1999); MacDonald and Zucchini (1997) with different emphases.

2.3.1 Some notes on numerical evaluation

From the definition of the likelihood function $p_n(y_1, \ldots, y_n; \vartheta)$ it is not clear that its evaluation even for moderate n is feasible, since p_n is given as a sum of m^n summands, each summand representing one possible path $\{U_1, \ldots, U_n\}$.
However, it is very advantageous that p_n can be written as product of matrices

$$p_n(y_1, \ldots, y_n; \vartheta) = \pi^* B_1 \, P \, B_2 \, P \, B_3 \, P \cdots P \, B_{n-1} \, P \, B_n \, \bar{1}$$

where π^* denotes the initial distribution, $B_i = \text{Diag}(f_{\theta_1}(y_i), \ldots, f_{\theta_m}(y_i))$ $m \times m$ matrices with non-zero entries on the diagonal only, for $i = 1, \ldots, n$, P the transition matrix and $\bar{1} = (1, \ldots, 1)^T \in \mathbb{R}^m$.

Evaluating this product starting from the left is known as *forward algorithm*, whereas starting from the right is called *backward algorithm*, since stopping at i gives the forward probabilities

$$a_i(k) = P(Y_1 = y_1, \ldots, Y_i = y_i, U_i = k) = \pi^* B_1 \, P \, B_2 \, P \cdots B_i e_k$$

and backward probabilities

$$b_i(k) = P(Y_{i+1} = y_{i+1}, \ldots, Y_n = y_n | U_i = k) = e_k^T \, P \, B_{i+1} \, P \cdots B_n \, \bar{1}$$

respectively, where $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)$ denotes the k^{th} unit vector. Note that a_i, b_i can be calculated recursively from a_{i-1} or b_{i+1} starting with $a_1 = \pi^* B_1$ and $b_n = \bar{1}$, respectively. These algorithms allow efficient evaluation where the number of operations increases only linearly in n (cf. Durbin et al., 1999, Chp. 3.2).
Although these algorithms allow a fast evaluation of the likelihood function one faces in practice the problem that the probabilities $a_i(k), b_i(k)$ as well as the likelihood itself become extremely small even for moderate n, resulting in numerical underflow. Since we cannot easily pass over to the log-likelihood, turning products into sums, as for i.i.d. models, one needs to implement a scaling procedure, for example by replacing $a_i(k)$ by

$$a_i(k)/c_i \quad \text{with} \quad c_i = \sum_k a_i(k)$$

in each step (cf. Durbin et al., 1999, Chp. 3.6). In this case the log-likelihood is given by $L_n = \log p_n = \sum_i \log c_i$. Note that it might be advantageous from computational viewpoint to apply the scaling not for every observation but only for every 10^{th} or 100^{th} observation, when the underflow problem really becomes an issue.
We shall now discuss maximization of the log-likelihood L_n. Since HMMs can be seen as extensions of finite mixtures models it is clear that we have to face similar problems as in this framework. Clearly, there is no hope to find an explicit form for the MLE, hence numerical maximization techniques must be applied. Moreover, the MLE might not exist if L_n is unbounded. This is the case for HMMs with Gaussian sdfs, when the variance is allowed to be arbitrarily close to zero. Here, one usually assumes a priori that the variance is bounded away from zero. Since L_n is in general not convex, we cannot expect that L_n has a unique local and global maximum. Although the Theorems 2.1 and 2.2 are valid for any global maximizer, local maxima are problematic in practice, since many maximization algorithms get stuck in those. Hence the choice of the starting value of the algorithms becomes very important. As common in the literature we choose the true value as starting

value in the simulation study. For real data applications several starting values have been implemented.

Since L_n can be evaluated efficiently general purpose maximization techniques can be used to provide an estimate $\hat{\vartheta}$. In **R** (R Development Core Team, 2009) the function *nlm* implements a Newton-type algorithm, an alternative function to conduct maximization is *optim*. Both algorithms offer various choices of features and settings. However, as the functions perform unconstrained maximization we usually must reparametrize the parameter ϑ using log- and logit-transformations. In the simulation study below *nlm* is used to compute the MLE- and LRT-values.

An alternative to general purpose techniques provide HMM-specific algorithms, namely the *Baum-Welch-Algorithm*. This algorithm is historically the standard algorithm for parameter estimation (learning) for HMMs. It can be seen as an EM-type algorithm and is based on the missing information data structure of HMMs. As the forward and backward probabilities $a_i(k), b_i(k)$ are its major ingredients the algorithm is also called *forward-backward-algorithm*.

Another class of HMM-specific algorithms is formed by gradient-based methods, which in contrast to its general purpose relatives make use of explicit calculations for the gradient and the Hessian of L_n. Turner (2008) suggested such a procedure and claims its good performance in terms of reliability and speed. However, Cappé et al. (2005) note that EM algorithms are often easier to implement from scratch and deal with parameter constraints in a natural way (e.g. for the transition probabilities).

2.3.2 Quality of asymptotic approximation for the MLE and LRT

Testing for zero-entries (Example 2.1)

Consider testing the hypothesis $H : \alpha_{11} = 0$ in a stationary two-state normal HMM, as described in Example 2.1, where the asymptotic distribution for T_n is given in (2.9). The transition matrix is taken as

$$A = \begin{pmatrix} 0 & 1 \\ 0.3 & 0.7 \end{pmatrix}.$$

As discussed in the introductory example, it appears intuitive that the finite sample behavior of the MLE and LRT is strongly influenced by the degree of separation of the sdfs f_{θ_1} and f_{θ_2}. At first we fix the parameters of the sdfs at their true values, and let only the parameters of the transition matrix vary to choose the setup as simple as possible. Here, for the parameters of the sdfs we choose $\sigma^2 = 1$ and mean values $\mu_1 = 0$, $\mu_2 = 1$ in the first setting, where the sdfs strongly overlap, and $\mu_1 = 0$, $\mu_2 = 2$ in the second setting,

which corresponds to sufficiently well-separated sdfs. Secondly, we also allow the mean values of the sdfs to vary and treat them as nuisance parameters, but we keep σ^2 fixed. In this setup it appears that if the sdfs strongly overlap, estimation is quite unstable for reasonable sample sizes. In the case where we estimate the means of the sdfs, we therefore treat the settings $\mu_1 = 0$, $\mu_2 = 2$ and $\mu_1 = 0$, $\mu_2 = 3$. The Gaussian sdfs as well as the marginal densities of the two-state HMMs are displayed in Figures 2.1 and 2.2.

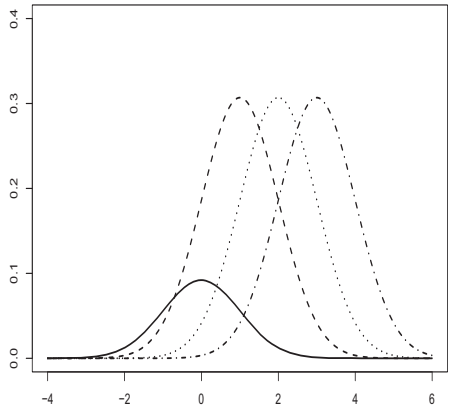


Figure 2.1: Plot of Gaussian sdfs with $\mu_1 = 0$ (solid line), $\mu_2 = 1$ (dashed line), $\mu_2 = 2$ (dotted line), $\mu_2 = 3$ (dash-dotted line). The sdf with $\mu_1 = 0$ is weighted by $\pi_1 = 0.3/1.3 \approx 0.23$, the other ones by $\pi_2 = \approx 0.77$.

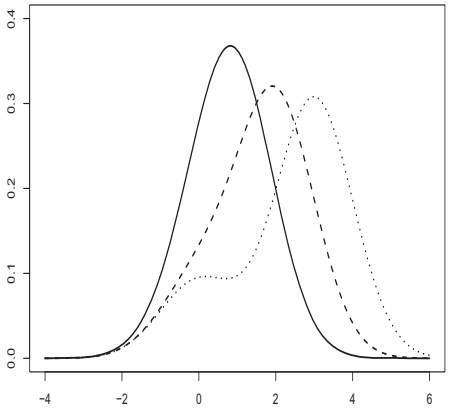


Figure 2.2: Plot of the marginal densities for two-state normal HMMs with $\pi \approx (0.23, 0.77)$ and $\mu_1 = 0$, $\mu_2 = 1$ (solid line), $\mu_1 = 0$, $\mu_2 = 2$ (dashed line), $\mu_1 = 0$, $\mu_2 = 3$ (dotted line).

We generate $N = 10000$ samples of various sizes, and for visualization of the behavior of the MLE $\hat{\alpha}_{11}$ we plot a kernel estimator scaled corresponding to the weight of the non-zero component estimated by $\hat{\pi} = \#\left\{\hat{\alpha}_{11} > 10^{-3}\right\}/N$. We also add a density of the half normal distribution to the plot, which is given by $f(x, \sigma) = 2\varphi(x/\sigma)$ for $x > 0$, where $\sigma > 0$ and $\varphi(\cdot)$ denotes the density of the standard normal distribution. We estimate σ via some robust estimator $\hat{\sigma}$ and scale $f(x, \hat{\sigma})$ by the asymptotic weight $1/2$. To visualize the behavior of the LRT we use PP-plots, which show for each nominal level $1 - \alpha$ the empirical probability that the LRT statistic $T_n \leq q_{1-\alpha}$, where $q_{1-\alpha}$ is the $1 - \alpha$-quantile of the asymptotic distribution.

The results for the MLE $\hat{\alpha}_{11}$ for the model with fixed sdfs are displayed in Figures 2.3 and 2.4. They indicate that in case of well-separated sdfs the estimators are less varying than

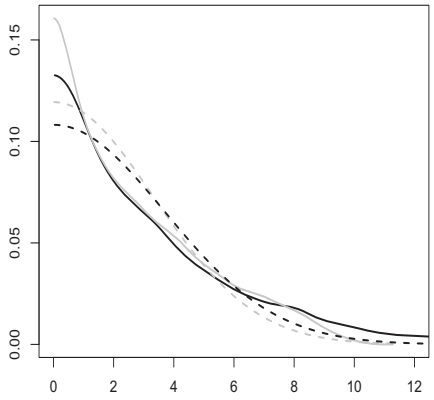


Figure 2.3: Plot of a scaled kernel estimator $\hat{\pi}\hat{f}$ of the density of $\sqrt{n}\hat{\alpha}_{11}$ for $n = 100$ (gray line, $\hat{\pi}$=0.51) and $n = 500$ (black line, $\hat{\pi}$=0.51) for $\alpha_{11}^0 = 0$ in case $\mu_1 = 0$, $\mu_2 = 1$ known. The dotted lines indicate halfnormal distributions with weight 0.5.

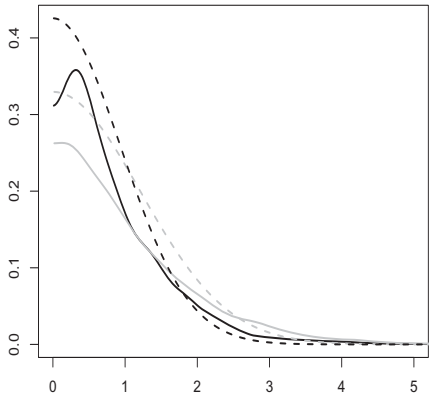


Figure 2.4: Plot of a scaled kernel estimator $\hat{\pi}\hat{f}$ of the density of $\sqrt{n}\hat{\alpha}_{11}$ for $n = 100$ (grey line, $\hat{\pi}$=0.39) and $n = 500$ (black line, $\hat{\pi}$=0.43) for $\alpha_{11}^0 = 0$ in case $\mu_1 = 0$, $\mu_2 = 2$ known. The dotted lines indicate halfnormal distributions with weight 0.5.

if the sdfs strongly overlap, which is intuitively clear, since the labeling of the observations becomes easier, when the sdfs are separated. However, the weight $\hat{\pi}$ appears significantly lower than 1/2 for the case of separated sdfs.

The results for the LRT T_n for the model with fixed sdfs are displayed in Figures 2.5 and 2.6. It turns out that the asymptotic approximation for well separated sdfs is relatively poor, even for large sample sizes such as $n = 500$ and in this simple situation with fixed parameters for the sdfs, while for strongly overlapping sdfs the approximation is quite good already for $n = 100$.

As argued in the introductory example the simulations show that the finite-sample behavior of the MLE and the LRT especially in models with well-separated distributions somehow lies between the case, where the regime can be observed directly resulting in a distribution consisting of a point mass at zero, and the asymptotic case the 1/2-1/2-mixtures (2.8) and (2.9), respectively.

The results for the models where the mean values of the sdfs are treated as nuisance parameter are displayed in Figures 2.7 - 2.10. In general, they confirm the previous findings. As one should expect the presence of nuisance parameters results in stronger variations in the estimates and slower convergence to the asymptotic distributions. Especially, comparing

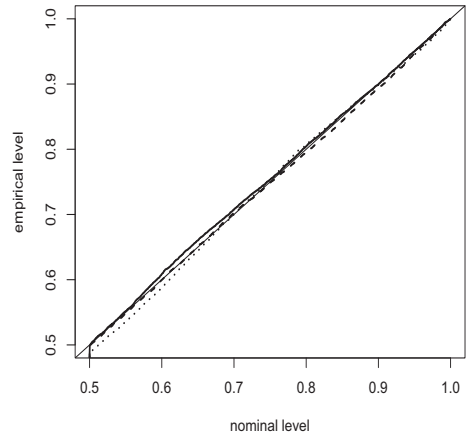


Figure 2.5: PP Plot of distribution of T_n for $n = 100$ (solid line), $n = 200$ (dashed line) and $n = 500$ (dotted line) for hypothesis $H : \alpha_{11} = 0$ in case $\mu_1 = 0$, $\mu_2 = 1$ known.

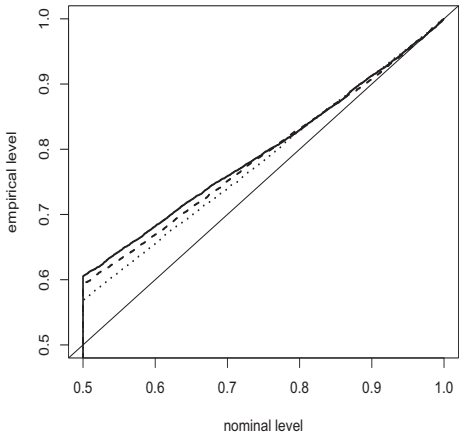


Figure 2.6: PP Plot of distribution of T_n for $n = 100$ (solid line), $n = 200$ (dashed line) and $n = 500$ (dotted line) for hypothesis $H : \alpha_{11} = 0$ in case $\mu_1 = 0$, $\mu_2 = 2$ known.

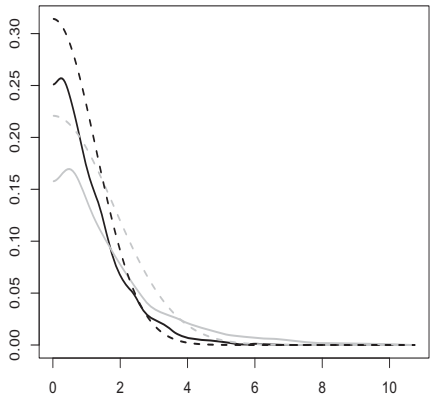


Figure 2.7: Plot of a scaled kernel estimator $\hat{\pi}\hat{f}$ of the density of $\sqrt{n}\hat{\alpha}_{11}$ for $n = 100$ (grey line, $\hat{\pi} = 0.39$) and $n = 500$ (black line, $\hat{\pi} = 0.41$) for $\alpha^0_{11} = 0$ in case $\mu_1 = 0$, $\mu_2 = 2$ estimated. The dotted lines indicate halfnormal distributions with weight 0.5.

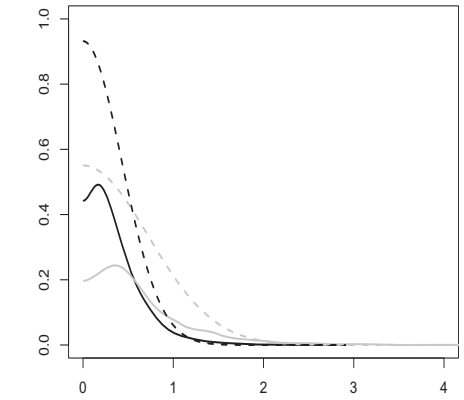


Figure 2.8: Plot of a scaled kernel estimator $\hat{\pi}\hat{f}$ of the density of $\sqrt{n}\hat{\alpha}_{11}$ for $n = 100$ (grey line, $\hat{\pi}$=0.22) and $n = 500$ (black line, $\hat{\pi} = 0.28$) for $\alpha^0_{11} = 0$ in case $\mu_1 = 0$, $\mu_2 = 3$ estimated. The dotted lines indicate halfnormal distributions with weight 0.5.

Figures 2.9 and 2.10 indicates once more that more separated sdfs give rise to much bigger zero components of the finite sample distribution of the LRT.

Different asymptotic distributions (Example 2.3)

Next we consider testing the hypotheses (2.11), (2.15) and (2.17) for a stationary two-state Poisson HMM. The parameters for the sdfs are fixed at their true values, which were taken as $\lambda_1 = 1, \lambda_2 = 2$ and $\lambda_1 = 1, \lambda_2 = 6$, respectively. Further $\alpha_{12} = \alpha_{21} = 1/2$. For the first scenario, the mixing proportion p given in (2.13) equals 0.384, whereas for the second scenario, it is 0.269. Again we generated $N = 10000$ samples of various sizes. In order to illustrate the different forms of the asymptotic distributions, we show the empirical distribution functions of T_n.
Figures 2.11 and 2.12 give the results for the testing problem (2.11), where the limit distribution is as in (2.12). In most cases, the approximation is quite accurate already for a sample size of $n = 100$, for higher sample sizes, it is almost indistinguishable from its limit version.
Figures 2.13 and 2.14 show the simulation results for (2.15). Here, the asymptotic distributions cannot be evaluated easily analytically, but it is simple to sample from these distributions via the formula (2.16). Figures 2.15 and 2.16 show the simulation results for (2.17). Again, the asymptotic distributions is determined via sampling from (2.18).
The figures illustrate that the (not so) different hypotheses (2.11), (2.15) and (2.17) lead to (quite) different limit distributions of T_n (2.12), (2.16) and (2.18), also the value of p determined by the Fisher information matrix (cf. (2.13)) influences the limit distribution apparently.

Tests on the stationary distribution in a three-state HMM (Example 2.4)

Further, we examine testing the hypothesis $H_{1,3} : \pi_1 \geq \pi_3$ as described in Example 2.4 for a stationary three-state Poisson HMM. We use a transition matrix as in (2.19), with $\alpha = 0.4$, $\beta = 0.2$, $\gamma = 0.3$ and $\delta = 0.6$, yielding for the stationary distribution $\pi_1 = \pi_3 = 0.25$, $\pi_2 = 0.5$. The parameters of the sdfs were specified with $\lambda_1 = 2$, $\lambda_2 = 5$ and $\lambda_3 = 11$. The asymptotic distribution of the likelihood ratio statistic T_n on the boundary of the hypothesis $H_{1,3}$ is (2.9).

Firstly, we consider the described model for fixed and known values of the λ's. Secondly, the λ's are considered as unknown parameters one has to estimate. For both cases we generate $N = 5000$ samples of sizes $n = 100, 500$. The results are displayed in Figs. 2.17

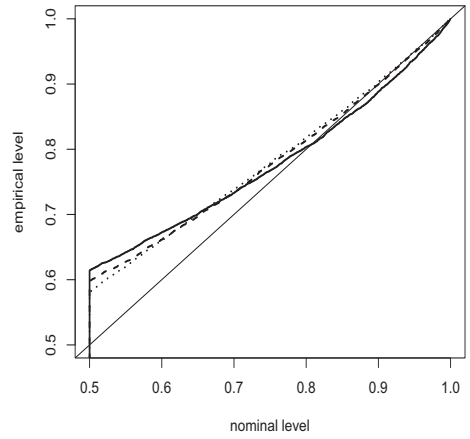


Figure 2.9: PP Plot of distribution of T_n for $n = 100$ (solid line), $n = 200$ (dashed line) and $n = 500$ (dotted line) for hypothesis $H : \alpha_{11} = 0$ in case $\mu_1 = 0$, $\mu_2 = 2$ estimated.

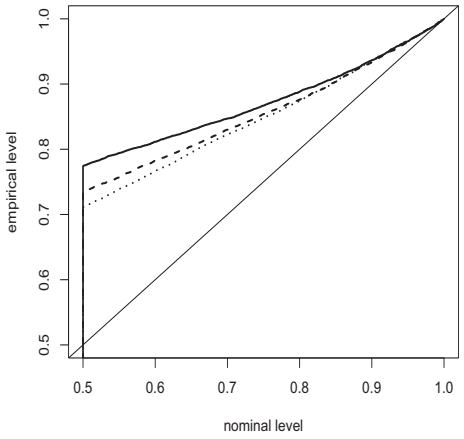


Figure 2.10: PP Plot of distribution of T_n for $n = 100$ (solid line), $n = 200$ (dashed line) and $n = 500$ (dotted line) for hypothesis $H : \alpha_{11} = 0$ in case $\mu_1 = 0$, $\mu_2 = 3$ estimated.

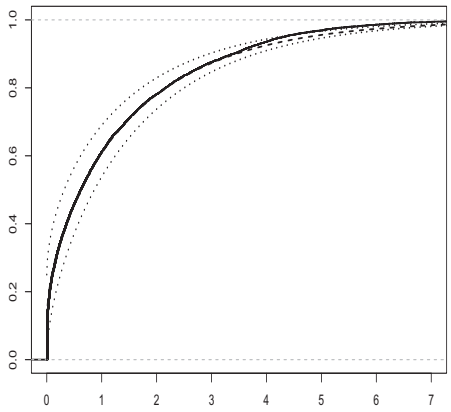


Figure 2.11: Empirical distribution function of T_n for $n = 100$ (solid line) for the hypothesis (2.11) in case $\lambda_1 = 1$, $\lambda_2 = 2$, together with the limit distribution (dashed line). The dotted lines indicate upper and lower stochastic bound independent of p.

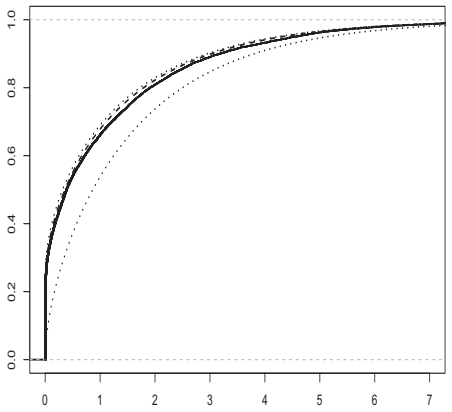


Figure 2.12: Empirical distribution function of T_n for $n = 100$ (solid line) for the hypothesis (2.11) in case $\lambda_1 = 1$, $\lambda_2 = 6$, together with the limit distribution (dashed line). The dotted lines indicate upper and lower stochastic bound independent of p.

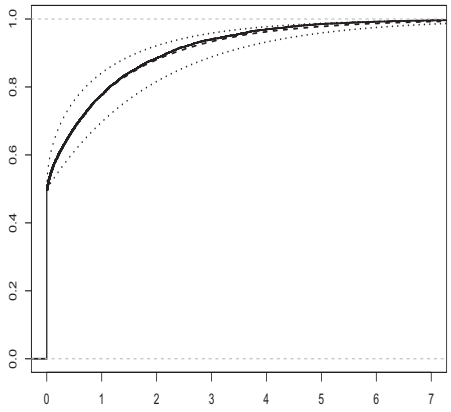


Figure 2.13: Empirical distribution function of T_n for $n = 100$ (solid line) for the hypothesis (2.15) in case $\lambda_1 = 1$, $\lambda_2 = 2$, together with the limit distribution (dashes line) and upper and lower bounds independent of p (dotted lines).

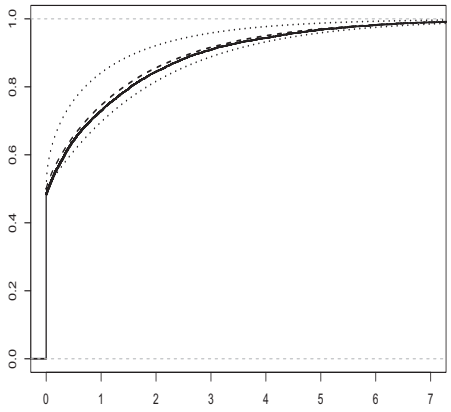


Figure 2.14: Empirical distribution function of T_n for $n = 100$ (solid line) for the hypothesis (2.15) in case $\lambda_1 = 1$, $\lambda_2 = 6$, together with the limit distribution (dashed line) and upper and lower bounds independent of p (dotted lines).

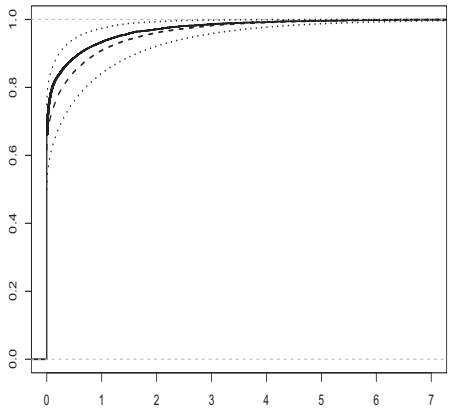


Figure 2.15: Empirical distribution function of T_n for $n = 100$ (solid line) for the hypothesis (2.17) in case $\lambda_1 = 1$, $\lambda_2 = 2$, together with the limit distribution (dashes line) and upper and lower bounds independent of p (dotted lines).

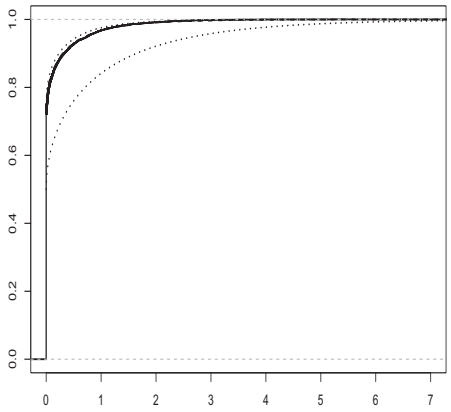


Figure 2.16: Empirical distribution function of T_n for $n = 100$ (solid line) for the hypothesis (2.17) in case $\lambda_1 = 1$, $\lambda_2 = 6$, together with the limit distribution (dashed line) and upper and lower bounds independent of p (dotted lines).

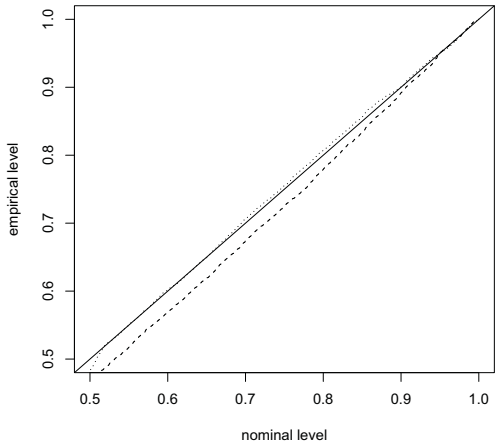


Figure 2.17: PP Plot of distribution of T_n for $n = 100$ (dashed line) and $n = 500$ (dotted line) for hypothesis $H_{1,3}$ in case of fixed and known λ's.

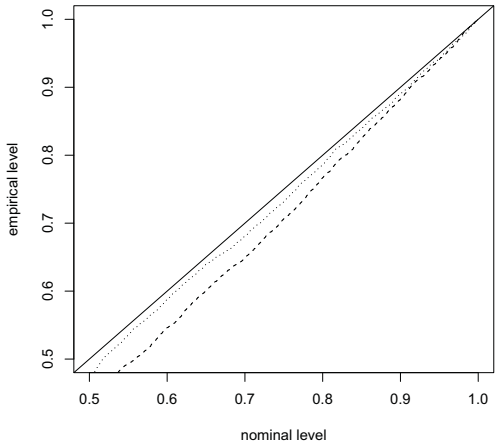


Figure 2.18: PP Plot of distribution of T_n for $n = 100$ (dashed line) and $n = 500$ (dotted line) for hypothesis $H_{1,3}$ in case of estimated λ's.

- 2.18 using PP-plots. For fixed λ's the approximation is quite satisfactory even for small sample sizes (n = 100). Naturally, estimation of the λ's increases variation of the model. But for large sample sizes (n=500) the approximation is quite good in this case, too.

2.3.3 Series of epileptic seizure counts

Albert (1991) proposed the use of two-state Poisson HMMs for series of daily seizure counts of epileptics. Using the implementation of the EM algorithm as suggested in Baum and Petrie (1966), Le et al. (1992) fit such models to the series of daily counts of epileptic seizures in one patient participating in a clinical trail at British Columbia's Children's Hospital. The originally published series consists of 225 observations, however, as indicated in MacDonald and Zucchini (1997, p.147), observations 92-112 should be deleted, thus we use the corrected data set of 204 observations.

Accounting for this the data are displayed in Figure 2.19 in the same way as in Le et al. (1992). From the figure one can observe that the data seems to be dependent over time, the empirical correlation between two consecutive observations (0.236) differs clearly from zero. As the variance (0.924) exceeds the mean (0.662), we see that overdispersion relative to Poisson is also present in the data. In the neurology literature, Hopkins et al. (1985) proposed that the variation of seizure occurrences and its dependency structure could be

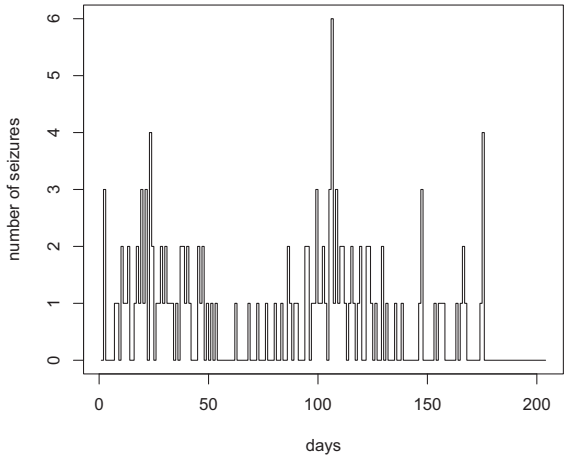


Figure 2.19: Daily recorded seizure counts from a single patient over 204 days.
(Source: The data set is published in MacDonald and Zucchini (1997, p.208, Table B.2).)

modeled by a Markov chain. Both features, dependency as well as overdispersion, can be incorporated naturally in the two-state Poisson HMM, where the two states of the chain represent two states of seizure susceptibility. Haut (2006) pointed out that such Markovian dependence of seizure susceptibility allows estimates for the expected incidence of subsequent seizure days, which might be useful for recognition of seizure clusters.

A relevant question in this model is whether seizures actually occur in both states of the HMM, or whether there is only a single "seizure state", whereas, in the other state, no seizures occur.

For the above mentioned data set MacDonald and Zucchini (1997) fitted a stationary two-state Poisson HMM with estimated transition matrix

$$\begin{pmatrix} 0.973 & 0.027 \\ 0.035 & 0.965 \end{pmatrix}$$

and seizure frequencies $\lambda_1 = 0.262$ and $\lambda_2 = 1.167$. Thus, we intend to test whether a model with $\lambda_1 = 0$ could be used instead, and therefore propose to test $H : \lambda_1 = 0$ as described in Example 2.2. Here, the asymptotic distribution of the likelihood ratio statistic T_n under the hypothesis H follows (2.9). However, the likelihood ratio test yields a value of $T_n = 10.25$, which corresponds to a p-value of nearly 0. Hence, the hypothesis H is rejected, and seizures occur in both states of the HMM.

The estimate of the transition matrix yields an estimate for the stationary distribution $\hat{\pi} = (0.567, 0.433)$. Therefore, the estimate indicates that state 1 with low seizure suscep-

tibility is on average more frequently visited than state 2. In order to test whether this observation is statistically significant, we test whether the hypothesis $H : \pi_2 \geq \pi_1$ can be rejected. Again, the asymptotic distribution of the likelihood ratio statistic T_n on the boundary of the hypothesis is (2.9). For this test the likelihood ratio statistic is $T_n = 0.111$ with corresponding p-value 0.369. Hence, we cannot reject the hypothesis H at a 5%-level, thus, there is not enough evidence that state 1 is more often visited than state 2.

2.4 Proofs

Proof of Theorem 2.1. As mentioned, the proof follows the arguments in Self and Liang (1987) respectively Chernoff (1954) using (2.3) and (2.4) from Bickel et al. (1998). At first $\sqrt{n}$-consistency of $\vartheta - \vartheta_0$ needs to be verified. As argued by Chernoff (1954), Lemma 1, the expansion of the log-likelihood around ϑ_0 gives

$$0 \leq \frac{1}{n}L_n(\hat{\vartheta}) - \frac{1}{n}L_n(\vartheta_0) = \frac{1}{n}DL_n(\vartheta_0)^T(\hat{\vartheta} - \vartheta_0) + (\hat{\vartheta} - \vartheta_0)^T D^2 L_n(\vartheta_0)(\hat{\vartheta} - \vartheta_0) + o_p(1).$$

By assumption and by (2.3,2.4) we have

$$\big\|\hat{\vartheta} - \vartheta_0\big\| = o_p(1), \qquad \big\|\frac{1}{n}D^2L_n(\vartheta_0) + \mathcal{J}_0\big\| = o_p(1) \text{ and } \big\|\frac{1}{n}DL_n(\vartheta_0)\big\| = O_p(\sqrt{n}).$$

in appropriate norms, and hence by positive definiteness of $\mathcal{J}_0$

$$0 \leq (\hat{\vartheta} - \vartheta_0)^T \mathcal{J}_0 (\hat{\vartheta} - \vartheta_0) \leq \big\|\hat{\vartheta} - \vartheta_0\big\| O_p(\sqrt{n}) + o_p(1),$$

which implies $\hat{\vartheta} - \vartheta_0 = O_p(n^{-1/2})$.

Secondly, as in Self and Liang (1987), Lemma 1, we may expand the likelihood for ϑ with $\hat{\vartheta} - \vartheta_0 = O_p(n^{-1/2})$ as follows

$$\begin{aligned}
\frac{2}{n}L_n(\vartheta) - \frac{2}{n}L_n(\vartheta_0) &= \frac{2}{n}DL_n(\vartheta_0)^T(\vartheta - \vartheta_0) + \frac{1}{n}(\vartheta - \vartheta_0)^T D^2 L_n(\vartheta_0)(\vartheta - \vartheta_0) + o_p(1) \\
&= \frac{1}{n}DL_n(\vartheta_0)^T(\vartheta - \vartheta_0) + \frac{1}{n}(\vartheta - \vartheta_0)^T DL_n(\vartheta_0) - (\vartheta - \vartheta_0)^T \mathcal{J}_0(\vartheta - \vartheta_0) + o_p(1) \\
&= Z_n^T \mathcal{J}_0(\vartheta - \vartheta_0) + (\vartheta - \vartheta_0)^T \mathcal{J}_0 Z_n - (\vartheta - \vartheta_0)^T \mathcal{J}_0(\vartheta - \vartheta_0) \\
&= Z_n^T \mathcal{J}_0 Z_n - (Z_n - (\vartheta - \vartheta_0))^T \mathcal{J}_0 (Z_n - (\vartheta - \vartheta_0)) + o_p(1)
\end{aligned}$$

with $Z_n = 1/n\mathcal{J}_0^{-1}DL_n(\vartheta_0)$, where the first summand does not depend on ϑ. Note that the second equality holds by (2.4). Hence

$$\begin{aligned}
\arg\max_{\vartheta \in \bar{\Theta}} L_n(\vartheta) &= \arg\min_{\vartheta \in \bar{\Theta}} (\sqrt{n}Z_n - \sqrt{n}(\vartheta - \vartheta_0))^T \mathcal{J}_0 (\sqrt{n}Z_n - \sqrt{n}(\vartheta - \vartheta_0)) + o_p(1) \\
&= \arg\min_{\vartheta \in C_{\bar{\Theta}}} (\sqrt{n}Z_n - \sqrt{n}(\vartheta - \vartheta_0))^T \mathcal{J}_0 (\sqrt{n}Z_n - \sqrt{n}(\vartheta - \vartheta_0)) + o_p(1) \\
&= \arg\min_{\vartheta \in C_{\bar{\Theta}} - \vartheta_0} (\sqrt{n}Z_n - \vartheta)^T \mathcal{J}_0 (\sqrt{n}Z_n - \vartheta) + o_p(1)
\end{aligned}$$

where the second equality holds by the definition of approximating cones (cf. Chernoff, 1954, p.578) and the fact that cones are homogeneous. Since $\sqrt{n}Z_n$ is centered asymptotic normal with covariance matrix $\mathcal{J}_0^{-1}$ by (2.3) this concludes the proof. □

Proof of Theorem 2.2. As mentioned, Theorem 2.2 is a straight forward extension of Theorem 3 by Self and Liang (1987), the proof follows the arguments therein. Denoting $Z_n = 1/n\mathcal{J}_0^{-1}DL_n(\vartheta_0)$ we apply the same expansion for $\sqrt{n}$-consistent sequences as in the proof of Theorem 2.1 and get

$$\begin{aligned}
T_n &= 2\big(\sup_{\vartheta\in\bar{\Theta}} L_n(\vartheta) - \sup_{\vartheta\in\bar{\Theta}_0} L_n(\vartheta)\big) = 2\big(\sup_{\vartheta\in\bar{\Theta}} L_n(\vartheta) - L_n(\vartheta_0)\big) - 2\big(\sup_{\vartheta\in\bar{\Theta}_0} L_n(\vartheta) - L_n(\vartheta_0)\big) \\
&= \inf_{\vartheta\in\bar{\Theta}_0} n(Z_n - (\vartheta-\vartheta_0))^T\mathcal{J}_0(Z_n - (\vartheta-\vartheta_0)) \\
&\quad - \inf_{\vartheta\in\bar{\Theta}} n(Z_n - (\vartheta-\vartheta_0))^T\mathcal{J}_0(Z_n - (\vartheta-\vartheta_0)) + o_p(1) \\
&= \inf_{\vartheta\in C_{\bar{\Theta}_0}} n(Z_n - (\vartheta-\vartheta_0))^T\mathcal{J}_0(Z_n - (\vartheta-\vartheta_0)) \\
&\quad - \inf_{\vartheta\in C_{\bar{\Theta}}} n(Z_n - (\vartheta-\vartheta_0))^T\mathcal{J}_0(Z_n - (\vartheta-\vartheta_0)) + o_p(1) \\
&= \inf_{\vartheta\in C_{\bar{\Theta}_0}-\vartheta_0} (\sqrt{n}Z_n - \vartheta)^T\mathcal{J}_0(\sqrt{n}Z_n - \vartheta) - \inf_{\vartheta\in C_{\bar{\Theta}}-\vartheta_0} (\sqrt{n}Z_n - \vartheta)^T\mathcal{J}_0(\sqrt{n}Z_n - \vartheta) + o_p(1).
\end{aligned}$$

The third equality follows from this expansion, the subsequent equalities use the approximation property and the homogeneity of the cones (cf. Chernoff, 1954). Since $\sqrt{n}Z_n$ is centered asymptotic normal with covariance matrix $\mathcal{J}_0^{-1}$ by (2.3), the statement follows. □

Proof of (2.14). As we consider a two-state HMM we choose w.l.o.g. $j = 1, k = 2$. We have to compute the asymptotic covariance matrix of the score vector

$$\begin{aligned}
\nabla_\vartheta L_n(\vartheta) &= \nabla_\vartheta \log p_n(Y_1,\ldots,Y_n;\vartheta) \\
&= E_\vartheta\left[\nabla_\vartheta \log p_n(U_1,\ldots,U_n,Y_1,\ldots,Y_n;\vartheta)|Y_{1:n}\right] \\
&= E_\vartheta\left[\nabla_\vartheta \log \pi_{U_1}(\vartheta)|Y_{1:n}\right] + \sum_{i=1}^{n-1} E_\vartheta\left[\nabla_\vartheta \log(\alpha_{U_i,U_{i+1}}(\vartheta))|Y_{1:n}\right], \qquad (2.20)
\end{aligned}$$

where $Y_{1:n} = \big(Y_1,\ldots,Y_n\big)$, following from Fisher's identity (cf. Cappé et al., 2005, (10.12.), p.354 and (10.29), p.362).

Let $\alpha_{12} = \alpha$ and $\alpha_{21} = \beta$, so that $\vartheta = (\alpha,\beta)$. Neglect the first term in (2.20) for the moment, corresponding to the initial distribution. The derivatives $\partial/\partial\alpha$, $\partial/\partial\beta$ of the

second term are computed as

$$\begin{aligned}\frac{\partial}{\partial\alpha}\log\big(\alpha_{u,u'}(\vartheta)\big) &= \frac{1}{\alpha}\delta_{(1,2)}(u,u') - \frac{1}{1-\alpha}\delta_{(1,1)}(u,u'),\\ \frac{\partial}{\partial\beta}\log\big(\beta_{x,x'}(\vartheta)\big) &= \frac{1}{\beta}\delta_{(1,2)}(u,u') - \frac{1}{1-\beta}\delta_{(1,1)}(u,u'),\end{aligned}$$

where $\delta_x(y) = 1$ if $x = y$ and $= 0$ otherwise. So the first component (the derivative w.r.t. α) of the second term in (2.20) is the sum

$$\sum_{i=1}^{n-1} E_\vartheta\Big[\frac{1}{\alpha}\delta_{(1,2)}(U_i,U_{i+1}) - \frac{1}{1-\alpha}\delta_{(1,1)}(U_i,U_{i+1})|Y_{1:n}\Big]. \tag{2.21}$$

The key observation is that for $\alpha_0 = \beta_0 = 1/2$, the (U_i) are independent Bernoulli distributed with success probability $1/2$, and the (Y_i) are also i.i.d. following a two-component mixture distribution with density

$$\frac{1}{2}f_{\theta_1}(y) + \frac{1}{2}f_{\theta_2}(y).$$

Using

$$\delta_{(j,k)}(U_i,U_{i+1}) = \delta_j(U_i)\delta_k(U_{i+1}), \qquad j,k\in\{1,2\},$$

(2.21) evaluated at ϑ_0 can be rearranged as

$$\begin{aligned}&\sum_{i=1}^{n-1} 2E_0\big[\delta_1(U_i)|Y_i\big]E_0\big[\delta_2(U_{i+1})|Y_{i+1}\big] - \sum_{i=1}^{n-1} 2E_0\big[\delta_1(U_i)|Y_i\big]E_0\big[\delta_1(U_{i+1})|Y_{i+1}\big],\\ &\qquad = 2\sum_{i=1}^{n-1} E_0\big[\delta_1(U_i)|Y_i\big]\Big(E_0\big[\delta_2(U_{i+1})|Y_{i+1}\big] - E_0\big[\delta_1(U_{i+1})|Y_{i+1}\big]\Big),\end{aligned}$$

and a similar expression can be obtained for the second component of the score (the derivative w.r.t. β). Now

$$E_0\big[\delta_j(U_i)|Y_i\big] = \frac{\frac{1}{2}f_{\theta_j}(Y_i)}{\frac{1}{2}f_{\theta_1}(Y_i)+\frac{1}{2}f_{\theta_2}(Y_i)} = \frac{f_{\theta_j}(Y_i)}{f_{\theta_1}(Y_i)+f_{\theta_2}(Y_i)}, \quad j\in\{1,2\}.$$

Introducing the random variables

$$\begin{aligned}Z_{i,1} &= \frac{f_{\theta_1}(Y_i)}{f_{\theta_1}(Y_i)+f_{\theta_2}(Y_i)}\frac{f_{\theta_2}(Y_{i+1})-f_{\theta_1}(Y_{i+1})}{f_{\theta_1}(Y_{i+1})+f_{\theta_2}(Y_{i+1})},\\ Z_{i,2} &= \frac{f_{\theta_2}(Y_i)}{f_{\theta_1}(Y_i)+f_{\theta_2}(Y_i)}\frac{f_{\theta_1}(Y_{i+1})-f_{\theta_2}(Y_{i+1})}{f_{\theta_1}(Y_{i+1})+f_{\theta_2}(Y_{i+1})},\end{aligned}$$

the components of the score are given by

$$S_{n,1} = 2\sum_{i=1}^{n-1} Z_{i,1}, \qquad S_{n,2} = 2\sum_{i=1}^{n-1} Z_{i,2}.$$

Although the $Z_{i,1}$'s are not i.i.d., the covariances for different i's are 0, since the Y_i's are independent and

$$\begin{aligned} & E_0[(f_{\theta_2}(Y_i) - f_{\theta_1}(Y_i))/(f_{\theta_1}(Y_i) + f_{\theta_2}(Y_i))] \\ &= \int (f_{\theta_2}(y) - f_{\theta_1}(y))/(f_{\theta_1}(y) + f_{\theta_2}(y))\,(1/2 f_{\theta_1}(y) + 1/2 f_{\theta_2}(y))dy \\ &= \int f_{\theta_2}(y) - f_{\theta_1}(y)\,dy = 1 - 1 = 0 \end{aligned}$$

Hence the correlation can be computed independently of i as

$$\begin{aligned} \rho &= \frac{E_0 Z_{i,1} Z_{i,2}}{(E_0 Z_{i,1}^2 E_0 Z_{i,2}^2)^{1/2}} \\ &= -\frac{\int f_{\theta_1}(y) f_{\theta_2}(y)/\big(f_{\theta_1}(y) + f_{\theta_2}(y)\big)dy}{\big(\int f_{\theta_1}^2(y)/\big(f_{\theta_1}(y) + f_{\theta_2}(y)\big)\,dy \int f_{\theta_2}^2(y)/\big(f_{\theta_1}(y) + f_{\theta_2}(y)\big)\,dy\big)^{1/2}}, \end{aligned}$$

For the first term in (2.20), one has that

$$E_0\big[\nabla_\vartheta \log(\pi_{U_1}(\vartheta))|Y_{1:n}\big] = E_0\big[\nabla_\vartheta \log(\pi_{U_1}(\vartheta))|Y_1\big],$$

of which the contribution to the asymptotic covariance matrix is zero , which concludes the proof. Note that this example generalizes the Example 4.3 in Bickel et al. (2002), who consider reversible Markov chains and thus only have a single parameter. □

Proof of (2.16) . To proof the asymptotic distribution of T_n under (2.15) we can apply Theorem 2.2 and a coordinate transformation to obtain

$$\begin{aligned} T_n \quad \xrightarrow{\mathcal{L}} \quad T &:= \inf_{z\in H} \big(Z - z\big)^T \big(Z - z\big) - \inf_{z\in H\cup K} \big(Z - z\big)^T \big(Z - z\big) \\ &= \|Z - Z_H\|^2 - \min\big(\|Z - Z_H\|^2, \|Z - Z_K\|^2\big) \end{aligned} \tag{2.22}$$

where Z is a bivariate standard normal random variable, $H = \mathcal{J}_0^{1/2}\big(\mathbb{R}^- \times \mathbb{R}^-\big)$ and $K = \mathcal{J}_0^{1/2}\big(\mathbb{R}^+ \times \mathbb{R}^+\big)$, as shown in Figure 2.20, Z_H and Z_K denote the orthogonal projections of Z onto H and K respectively. The spaces H and K are determined by the angle $\pi/2 \le \phi \le \pi$ which depends on $\mathcal{J}_0$, i.e. $\phi = \cos^{-1}\rho$. Firstly, we calculate the distribution of T conditional on the event $\{Z \in$ region 1 or $1a\}$ and the weight $P(Z \in$ region 1 or $1a)$. If Z is in region 1, then obviously both terms in (2.22) are zero. For region $1a$ the

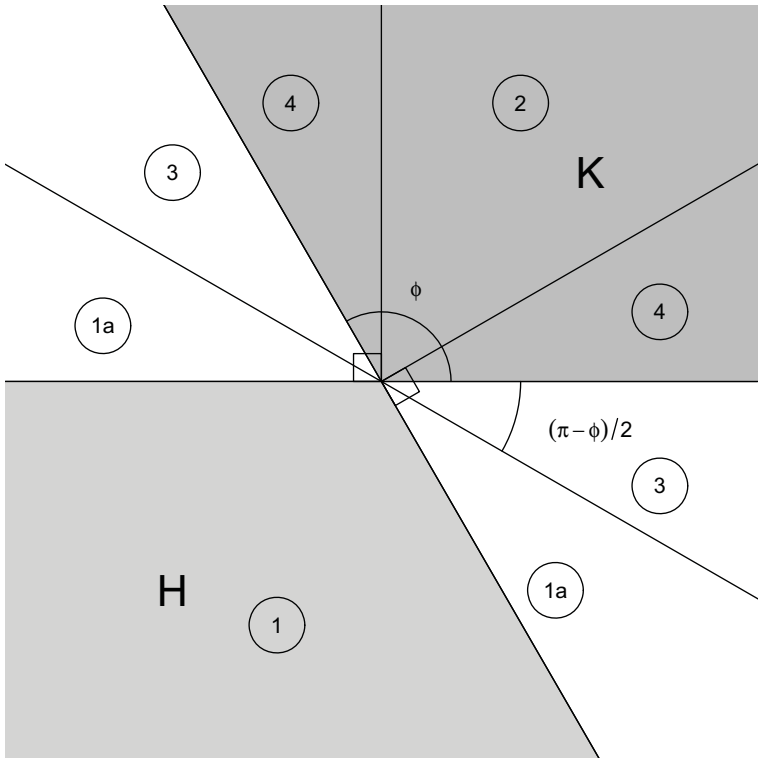


Figure 2.20: Diagram of the parameter space of Example 2.3.

difference is zero as well, since $\|Z - Z_H\| \leq \|Z - Z_K\|$. Hence the conditional distribution of T in region 1 and $1a$ is χ^2_0. As displayed in the figure its weight is $\frac{1}{2\pi}\left(\phi + 2\frac{\pi-\phi}{2}\right) = \frac{1}{2}$.
If Z is in region 2 and therefore in K the term $\|Z - Z_K\|$ is zero. Hence the distribution of T conditioned on $\{Z \in \text{ region } 2\}$ is determined by the first term, which gives a χ^2_2-distribution, since $Z_H = 0$. The weight is given by $\frac{\pi-\phi}{2\pi}$.
Observe that the r.v. Z is determined by its argument $\psi = \arg(Z)$ and its length $r(Z)$ by $Z = (r\cos\psi, r\sin\psi)$, where r^2 is χ^2_2-distributed and ψ is uniformly distributed on $[0, 2\pi)$, and r^2 and ψ are independent. If Z is now in region 3, i.e. $\psi \in [\phi, (\phi+\pi)/2) \cup [-(\pi-\phi)/2, 0)$, one has $\|Z - Z_H\|^2 = r^2\sin^2(\pi - \psi)$ and $\|Z - Z_K\|^2 = r^2\sin^2(\psi - \phi)$. Setting

$$a_1(\psi, \phi) = \sin^2(\pi - \psi) - \sin^2(\psi - \phi),$$

for fixed ψ the difference $T(\psi) = \|Z - Z_H\|^2 - \|Z - Z_K\|^2 = r^2 a_1(\psi, \phi)$ has a rescaled χ^2_2-distribution with density $p(t; \phi, \psi) = \frac{1}{2a_1(\psi,\phi)}\exp\left(-\frac{t}{2a_1(\psi,\phi)}\right)$. Hence the density of T is an averaged rescaled χ^2_2-distribution $P_1(\phi)$ with density

$$h_1(t; \phi) = \frac{2}{\pi - \phi}\int_\phi^{(\phi+\pi)/2} \frac{1}{2a_1(\psi, \phi)}\exp\left(-\frac{t}{2a_1(\psi, \phi)}\right) d\psi.$$

The weight of region 3 is given by $\frac{\pi-\phi}{2\pi}$.
For region 4 one proceeds similarly. For $\psi \in [0, \phi - \pi/2) \cup [\pi/2, \phi)$ one has $\|Z - Z_H\|^2 = r^2 \cos^2 \psi$ and trivially $\|Z - Z_K\| = 0$. Setting

$$a_2(\psi) = \cos^2\left(\psi - \frac{\pi}{2}\right),$$

this yields a conditional distribution $P_2(\phi)$ of T with density

$$h_2(t;\phi) = \frac{2}{2\phi - \pi} \int_{\pi/2}^{\phi} \frac{1}{2a_2(\psi)} \exp\left(-\frac{t}{2a_2(\psi)}\right) d\psi$$

with weight $2\frac{\phi - \frac{\pi}{2}}{2\pi} = \frac{2\phi-\pi}{2\pi}$. □

Proof of (2.18) . The proof follows the same arguments as the proof of (2.16). Again after coordinate transformation we have (2.22), where now $H = \mathcal{J}_0^{1/2}\big(\mathbb{R}^2 \backslash (\mathbb{R}^+ \times \mathbb{R}^+)\big)$ and $K = \mathcal{J}_0^{1/2}\big(\mathbb{R}^+ \times \mathbb{R}^+\big)$. So H is formed by the regions 1, 1a and 3 in Figure 2.20. Clearly for $Z \in H$ both terms in (2.22) vanish, leading to a χ_0^2-distribution with weight $P(Z \in H) = (2\pi - \phi)/2\pi$. Now we assume $Z \in K$ and $\psi = \arg(Z) \in [0, \phi/2)$, then $\|Z - Z_K\|^2$ is obviously zero and $\|Z - Z_H\|^2 = r^2 \sin^2(\psi)$ with r^2 being χ_2^2-distributed and ψ is uniformly distributed on $[0, 2\pi)$, mutually independent. Hence conditioned on $\{Z \in K, \arg(Z) \in [0, \phi/2)\}$ T follows a distribution P_3 with density

$$h_3(t;\phi) = \frac{2}{\phi} \int_0^{\phi/2} \frac{1}{2 \sin^2 \psi} \exp\Big(-\frac{t}{2 \sin^2 \psi}\Big) d\psi.$$

By symmetry T follows the same distribution if conditioned on $\{Z \in K, \arg(Z) \in [\phi/2, \phi)\}$ such that the weight of P_3 is given by $2\frac{1}{2\pi}\frac{\phi}{2} = \frac{\phi}{2\pi}$, which yields (2.18). □

Chapter 3

Testing for the number of states

In this chapter we discuss testing for the number of components or states in finite mixtures, HMMs as well as switching regression models with independent or Markov-dependent regime. We briefly introduce into the general problem of selecting the correct model in the framework of mixtures and HMMs and discuss the testing problem for homogeneity, i.e. $m = 1$ against $m > 1$, as considered by Ghosh and Sen (1985), Chen and Chen (2001) and many others for mixtures and Gassiat and Keribin (2000) for HMMs. Some new results of this chapter are published in Dannemann and Holzmann (2008c) and Dannemann and Holzmann (2010).

In general, we may encounter the situation, where we have two nested models $\mathcal{M}_0 \subset \mathcal{M}_1$ and must decide if the true model P_0 falls into $\mathcal{M}_0$ or $\mathcal{M}_1 \setminus \mathcal{M}_0$. It is a standard approach for various statistical models to treat this problem in terms of model selection criteria, e.g. AIC, BIC, ... (for a comprehensive overview see the monograph by Claeskens and Hjort, 2008). Based on the notion of penalized likelihood one defines penalties $pen(\mathcal{M}_0, n) < pen(\mathcal{M}_1, n)$ growing with the complexity of the model and one compares

$$\sup_{P \in \mathcal{M}_0} L_n(P) - pen(\mathcal{M}_0, n) \begin{array}{c} > \\ < \end{array} \sup_{P \in \mathcal{M}_1} L_n(P) - pen(\mathcal{M}_1, n)$$

with L_n denoting the log-likelihood function of the particular models. The AIC results from this notion by setting $pen(\mathcal{M}, n) = \dim(\mathcal{M})$, while the BIC is expressed through $pen(\mathcal{M}, n) = log(n) \dim(\mathcal{M})/2$. If one knows that $\sup_{P \in \mathcal{M}_1} L_n(P) - \sup_{P \in \mathcal{M}_0} L_n(P)$ has some asymptotic distribution P_{asym} the LRT w.r.t. a level α can also be interpreted as model selector with $pen(\mathcal{M}_1, n) - pen(\mathcal{M}_0, n) = q_{1-\alpha}(P_{asym})$, where $q_{1-\alpha}(P)$ denotes the $1 - \alpha$ quantile of a distribution P.

Selecting the number of states in latent variable models

For the models introduced in Chapter 1 the correct selection of the number of states m is very crucial for theoretical analysis, e.g. the asymptotic normality results for HMMs, and hence the Theorem 2.1 and 2.2 are only valid for the correct choice of m. For practical applications and interpretation of these models it is also of major importance to get m correctly.

For mixture models the log-likelihood $L_n^{(m)}$ is given by (1.2) and the dimension of an m-component mixture is $\dim(\mathcal{M}_m) = m - 1 + dm$ forming the mentioned criteria AIC and BIC, which are both frequently applied in practice (Frühwirth-Schnatter, 2006). Leroux (1992a) proves that these criteria do not underestimate the number of components of a finite mixture. For the BIC Keribin (2000) shows consistency, i.e. BIC asymptotically selects the true model. Other information criteria in the context of mixtures are discussed in McLachlan and Peel (2000).

For HMMs much less is known about model selection criteria and its behavior, when choosing the number of states m. As pointed out by Cappé et al. (2005, Chp. 15) this problem is closely related to order estimation for Markov chains. In practice model selection criteria are frequently used to determine m. Either one forms the criteria by considering the full-model log-likelihood defined in (2.1) (e.g. MacDonald and Zucchini, 1997) with $\dim(\mathcal{M}) = m(m-1) + dm$ or one may reduce the problem to selecting the number of components in the marginal mixture distribution (e.g. Poskitt and Zhang, 2005). Cappé et al. (2005) analyze order estimation in their monograph in detail and define a consistent criterion with a comparably strong penalty based on information theoretical analysis, however they focus on HMMs with finite sample space. Other suggestions are made by Celeux and Durand (2008), who consider the application of cross validation techniques to determine m, MacKay (2002), who gives a consistent order estimator based on minimum-distance methods and Rydén (1995), who investigates order estimation based on a split data likelihood function.

For switching regression models the classical model selection criteria are widely used to determine the number of components (e.g. Skrondal and Rabe-Hesketh, 2004). In addition to the choice of the number of components one needs to select the relevant covariates. Khalili and Chen (2007) consider the latter problem and argue that their approach can also be applied when m is not known aprior, but estimated consistently. Naik et al. (2007) propose a mixture regression criteria (MRC) of AIC-type dealing jointly with both problems.

Testing for the number of states in latent variable models

Although the LRT can be interpreted as model selection procedure we should point out that testing in general, and especially in this particular setting follows a different philosophy. Statistical tests allow for a decision between a hypothesis and an alternative, where the level $\alpha = P_H(\text{"reject } H\text{"})$ is at least asymptotically controlled. Naturally, testing procedures are not consistent, in the model selection sense, i.e. that asymptotically the true model is picked almost surely. A major advantage of tests is that in contrast to model selection criteria, they produce a p-value, which quantifies the confidence in the test result (cf. McLachlan and Peel, 2000).

The main problem of the LRT on the number of states is that the usual regularity conditions on Θ_H and Θ_K do not apply. Typically elements of Θ_H lie on the boundary of the parameter set and more unpleasant Θ_H is a non-identifiable subset of Θ_K. Hence the usual χ^2- or $\bar{\chi}^2-$asymptotic of twice the log-likelihood ratio does not hold. Let us, for example, consider $H : m = 1$ against $K : m = 2$ in a mixture model, then we have

$$\Theta_H = \{\vartheta = (\pi_1, \theta_1, \theta_2) | \pi_1(1 - \pi_1)(\theta_1 - \theta_2) = 0\}$$

such that H is fulfilled if $\pi_1 \in \{0, 1\}$, which is on the boundary, or if $\theta_1 = \theta_2$. Especially for $\vartheta_1 = (1, \theta_1, \theta)$, $\vartheta_2 = (0, \theta, \theta_1)$, $\vartheta_3 = (\pi, \theta_1, \theta_1)$ the law P_{ϑ_i} is the same for $i = 1, 2, 3$ for all $\theta \in \Theta$, $\pi \in [0, 1]$. Also, the geometry of

$$\Theta_H = \big(\{0, 1\} \times \Theta^2 \big) \cup \big([0, 1] \times \{(\theta, \theta) | \theta \in \Theta\} \big)$$

indicates the nonregularity of the problem. However, as illustrated by Chen (1995) a main reason for the nonregular behavior of the LRT, when testing for example $m = 1$, is that the Fisher information is typically degenerated at $\vartheta \in \Theta_H$. Chen (1995) analyzes the one-parametric model

$$g_\theta(x) = \frac{2}{3} f_{-\theta}(x) + \frac{1}{3} f_{2\theta}(x)$$

with $f_\theta \in \{\theta \in \Theta \subset \mathbb{R}\}$. He showed that for $\theta_0 = 0$, which corresponds to homogeneity, the MLE exhibits only a rate of $n^{-1/4}$, because the Fisher information is zero at $\theta_0 = 0$, although $\{g_\theta\}_{\theta \in \Theta}$ forms an identifiable, smooth family. In general, Chen (1995) deduced that $n^{-1/4}$ is the typical rate when estimating in overfitted mixture models, i.e. $m > m_0$.

Despite these difficulties, the asymptotic behavior of the LRT for testing for the number of states, especially for homogeneity, was investigated among others by Chen and Chen (2001), Azaïs et al. (2009) for mixtures, Gassiat and Keribin (2000) for HMMs, Zhu and Zhang (2004) for switching regression models and Cho and White (2007) for autoregressive models

with Markovian switch. All these results show fairly complicated asymptotic distributions of the LRT usually given by the distribution of the supremum of a Gaussian process, which may depend on the true parameter value ϑ_0. In some cases the LRT even diverges to infinity. Hence evaluation of the asymptotic distribution and its quantiles is demanding (if possible). It is frequently advocated to use some bootstrap procedures to access this problem (cf. Zhu and Zhang, 2004; Cho and White, 2007).
A more accessible alternative to the LRT is proposed by Chen et al. (2001, 2004) defining a modified likelihood for a model with m components by

$$\tilde{L}_n^{(m)}(\vartheta) = L_n^{(m)}(\vartheta) + C_m \sum_{k=1}^{m} \log \pi_k(\vartheta).$$

The additional penalty term ensures that the maximizer of the modified log-likelihood (MMLE) has weights $\tilde{\pi}_k$ that are bounded away from zero, such that the estimated mixture has no degenerated components. Chen et al. (2001, 2004) show that when testing $m = 1$ or $m = 2$ in mixture models against bigger models, the likelihood ratio evaluated at the MMLE rather then at the MLE exhibits an asymptotic distribution of the simple $\bar{\chi}^2-$type, if $\Theta \subset \mathbb{R}$.
In the next section we will present the results for testing based on the modified log-likelihood by Chen et al. (2001, 2004). In the subsequent sections we will show how this results can be used to tackle the similar testing problems for HMMs and switching regression models.

Remark 3.1. As pointed out by Frühwirth-Schnatter (2006) model selection and testing relies on the correct specification of the parametric class $\{f_\theta\}$ for the sdfs. In our view the question of robustness against misspecifications is although of practical relevance somehow misleading, since the specification of the correct number of components m makes sense only relative to the prespecified family. In real data examples one observes that the choice of the parametric family can influence the number of chosen components drastically, as an extreme case one may imagine that the family itself is formed by mixtures with m' components.

3.1 Testing for the number of components in a finite mixture model

In this section we discuss the results for testing based on the modified log-likelihood by Chen et al. (2001, 2004) for the hypothesis of homogeneity ($m = 1$) and the hypothesis of two components in mixture models.

3.1.1 Testing for homogeneity in a finite mixture model

As already mentioned the testing problem for homogeneity in a finite mixture model has been received much interest from many authors, especially the behavior of the LRT, which is shown to have a nonstandard asymptotic distribution, not $\bar{\chi}^2$-type.
Exemplarily, we cite the result from Chen and Chen (2001), who consider testing

$$H : \pi_1(1-\pi_1)(\theta_1-\theta_2) = 0 \quad \text{against} \quad K : \pi_1 \in [0,1], \theta_1, \theta_2 \in \Theta \tag{3.1}$$

and derive the asymptotic distribution of the LRT in the particular case of a one-parametric family $\Theta \subset \mathbb{R}^1$, with Θ compact and θ_0 interior point of Θ.
Chen and Chen (2001) and also Chen et al. (2001) require the following regularity conditions on the family $\{f_\theta\}_\theta$ and on the quantities

$$Z_i^1(\theta) = \frac{f_\theta(Y_i) - f_{\theta_0}(Y_i)}{(\theta-\theta_0) f_{\theta_0}(Y_i)}, \quad \theta \neq \theta_0 \quad \text{and} \quad Z_i^1(\theta_0) = \frac{f'_{\theta_0}(Y_i)}{f_{\theta_0}(Y_i)},$$
$$Z_i^2(\theta) = \frac{Z_i^1(\theta) - Z_i^1(\theta_0)}{(\theta-\theta_0)}, \quad \theta \neq \theta_0 \quad \text{and} \quad Z_i^2(\theta_0) = Z_i^{1\prime}(\theta_0)$$

where the prime indicates the derivative w.r.t. θ.

Assumption 3.1 (Wald-type integrability condition)**.** Let $E_0\left[|\log f_{\theta_0}(Y_i)|\right] < \infty$, and there exists $\varepsilon > 0$ such that, for each θ, $f(y;\theta,\varepsilon) := 1 + \sup_{|\theta-\theta'|\leq\varepsilon} f_{\theta'}(y)$ is measurable and $E_0\left[\log f(Y_i;\theta,\varepsilon)\right] < \infty$.

Assumption 3.2 (Smoothness)**.** The support of $f_\theta(y)$ does not depend on θ and $f_\theta(y)$ is two times continuously differentiable w.r.t. $\theta \subset \Theta$. The derivatives are jointly continuous in y and θ.

Assumption 3.3 (Strong identifiability)**.** The family $\{f(y;\theta) \,|\, \theta \in \Theta\}$ is strong identifiable, i.e. for $\theta_1 \neq \theta_2$

$$\sum_{k=1}^{2} \left(a_k f_{\theta_k}(y) + b_k f'_{\theta_k}(y) + c_k f''_{\theta_k}(y)\right) = 0$$

for all y implies $a_k = b_k = c_k = 0$ for $k = 1, 2$.

Assumption 3.4 (Uniform boundedness)**.** There exists an integrable function g and $\delta > 0$ such that $|Z_i^1(\theta)|^{4+\delta} \leq g(Y_i)$ and $\left|Z_i^{1\prime}(\theta)\right|^3 \leq g(Y_i)$ for all θ.

Assumption 3.5 (Tightness)**.** The processes $n^{-1/2}\sum_i Z_i^1(\theta)$, $n^{-1/2}\sum_i Z_i^{1\prime}(\theta)$ as well as $n^{-1/2}\sum_i Z_i^2(\theta)$ and $n^{-1/2}\sum_i Z_i^{2\prime}(\theta)$ are tight.

For a discussion of the assumptions see Chen and Chen (2001) and the related working paper. Note that for commonly used families as the binomial or the Poisson family the conditions are fulfilled, also for normal density with fixed variance σ^2, while the family of all normal densities is not strongly identifiable.
Chen and Chen (2001) derive the following theorem:

Theorem 3.1. *Suppose that Assumptions 3.1 – 3.5 hold and that $f_{\theta_0}(y)$ is the density of the true null distribution of $(Y_i)_i$. Then*

$$2\big(\sup_{\vartheta\in\bar{\Theta}_K} L_n^{(2)}(\vartheta) - \sup_{\theta\in\Theta} L_n^{(1)}(\theta)\big) \xrightarrow{\mathcal{L}} (\sup_{\theta\in\Theta} W^+(\theta))^2$$

where $\bar{\Theta}_K = [0,1]\times\Theta^2$ and W^+ is the positive part of a standard Gaussian process with correlation function

$$\rho(\theta,\theta') = Corr(W_1(\theta), W_1(\theta'))$$

with $W_1(\theta) = Z_1^2(\theta) - (E[Z_1^1(\theta_0)Z_1^2(\theta)]/E[(Z_1^1(\theta_0))^2])Z_1^1(\theta_0)$.

For the proof see Chen and Chen (2001). Although this theorem is of primary theoretical interest, the evaluation of the asymptotic distributions and hence the application of the LRT is rather difficult. Moreover, the asymptotic distribution depends on the true parameter θ_0 (if $Z_1^2(\theta)$ and $Z_1^1(\theta_0)$ are correlated) and on the parameter space Θ, especially on its magnitude. Clearly, Theorem 3.1 implies that $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ serves as an asymptotic lower bound for the LRT, but applying that lower bound would lead to a liberal test, which is usually undesirable.

A modified LRT to test for homogeneity

As an alternative, Chen et al. (2001) show how one can test (3.1) via a modified log-likelihood function:

$$\tilde{L}_n^{(2)}(\vartheta) = L_n^{(2)}(\vartheta) - pen(\pi_1(\vartheta))$$

where $L_n^{(2)}$ denotes the ordinary log-likelihood for a mixture with two components (see (1.2)) and a penalty $pen(\pi_1)$. The penalty function should vanish for $\pi_1 = 1/2$ and tend to infinity, if π_1 approaches the bounds of $[0,1]$ to penalize degenerated components. A common choice is

$$pen(\pi_1(\vartheta)) = -C\log 4\pi_1(\vartheta)(1-\pi_1(\vartheta))$$

with some positive constant C. Based $\tilde{L}_n$ the modified likelihood ratio test (MLRT) to test (3.1) with the penalty function chosen as suggested above is then formed as

$$T_n = 2\big(\sup_{\vartheta\in\bar{\Theta}_K} \tilde{L}_n^{(2)}(\vartheta) - \sup_{\theta\in\Theta} L_n^{(1)}(\theta)\big).$$

Chen et al. (2001) show that the MLRT in contrast to the ordinary LRT follows a simple $\bar{\chi}^2$-distribution under the null hypothesis:

Theorem 3.2. *Suppose that Assumptions 3.1 – 3.5 hold and that the density of the true null distribution of $(Y_i)_i$ is given by $f_{\theta_0}(y)$, i.e. just one component. Then*

$$T_n \xrightarrow{\mathcal{L}} \frac{1}{2}\,\chi_0^2 + \frac{1}{2}\,\chi_1^2. \tag{3.2}$$

For the proof see Chen et al. (2001) and the related technical report. A key observation in the proof of Theorem 3.2 is that for the MMLE $\hat{\vartheta} = (\hat{\pi}_1, \hat{\theta}_1, \hat{\theta}_2) = \arg\max_\vartheta \tilde{L}_n^{(2)}(\vartheta)$ the quantities

$$\hat{\pi}_1(\hat{\theta}_1 - \theta_0) + (1 - \hat{\pi}_1)(\hat{\theta}_2 - \theta_0) \quad \text{and} \quad \hat{\pi}_1(\hat{\theta}_1 - \theta_0)^2 + (1 - \hat{\pi}_1)(\hat{\theta}_2 - \theta_0)^2$$

are under the hypothesis both of the order $n^{-1/2}$, which implies $\hat{\theta}_k - \theta_0 = O_p(n^{-1/4})$ for $k = 1, 2$. This coincides with the results of Chen (1995) where ML-estimation in mixtures with too many components is considered. Chen et al. (2001) also show some local optimality result for the MLRT by considering local alternatives in a $n^{-1/4}$-neighborhood of θ_0.

Remark 3.2. Simulations show that the asymptotic distribution from Theorem 3.2 approximates the finite sample distribution of T_n under the null hypothesis moderately well. Note that for the testing problem

$$H : \pi_1 = 1/2, \theta_1 = \theta_2 \quad \text{against} \quad K' : \pi_1 = 1/2,\ \theta_1, \theta_2 \in \Theta,$$

where the weight π_1 is fixed at $1/2$ the LRT (which coincides with the MLRT here) also follows asymptotically the distribution (3.2). For this hypothesis the asymptotic distribution approximates the finite sample distribution satisfactorily. However, when testing H against K' rather than against K the power decreases. By proposing a testing procedure based on the EM-principle Li et al. (2009) develop an *EM-test* that has the same asymptotic distribution under the hypothesis. It provides a good approximation under the hypothesis and possesses a power comparable to the MLRT.

Remark 3.3. A serious lack of the results formulated in Theorem 3.1 and 3.2 is that the explicit asymptotic distributions are only valid for one parametric families, i.e. $\Theta \subset \mathbb{R}$. For many applications this assumption is met, for example for Poisson, binomial or exponential distributions. But for many other applications it appears to restrictive, e.g. location-scale models with unknown mean and variance. For finite mixtures the notion of an additional structural parameter, which is unknown, but equal in all components, appears natural

and useful. Chen et al. (2008) provide some results also for this model, which is also closely related to switching regression models with switching intercept (cf. Section 3.3). More general results of the behavior of the LRT for mixture with and without structural parameter are given by Azaïs et al. (2009) and also the paper by Zhu and Zhang (2004) implies asymptotic results for both, the LRT and the MLRT.

3.1.2 Testing for two components in a finite mixture model

Although testing for homogeneity is the most relevant testing problem for finite mixture models, it is further of interest to test whether the latent distribution only has two components or more than two components, for example, when population heterogeneity is evident or has already been established. Indeed, two components often correspond to two contrast levels, which may have a straight interpretation, whereas more than two components express a whole range of possibilities and indicate a more complex structure.
For notational convenience we introduce the following notation of distributions with m support points on $\Theta \subset \mathbb{R}$:

$$\mathfrak{M}_m = \Big\{ G(\theta) = \sum_{k=1}^{m} \pi_k I_{\{\theta_k \leq \theta\}} : \ \theta_1 \leq \ldots \leq \theta_m, \ \sum_{k=1}^{m} \pi_k = 1, \pi_k \geq 0 \Big\}$$

denote the set of all m-point distributions on Θ, and let $\mathfrak{M} = \cup_{m \geq 2} \mathfrak{M}_m$. For $G \in \mathfrak{M}_m$ with parameters $(\pi_1(G), \ldots, \pi_m(G), \theta_1(G), \ldots, \theta_m(G))$ we let $f_{\text{mix}}(y; G)$ denote the mixing density

$$f_{\text{mix}}(y; G) = \int f(y; \theta) dG(\theta) \ = \ \pi_1(G) f(y; \theta_1(G)) + \ldots + \pi_m(G) f(y; \theta_m(G)).$$

In the following we may write π_k, θ_k instead of $\pi_k(G), \theta_k(G)$ to keep the notation handy. Chen et al. (2004) investigate testing for two components, i.e. testing

$$H : \ G_0 \in \mathfrak{M}_2 \qquad \text{against} \quad K : \ G_0 \in \mathfrak{M} \setminus \mathfrak{M}_2. \tag{3.3}$$

A modified LRT to test for two components

Similarly to the test for homogeneity (see above) Chen et al. (2004) bypass the analysis of LRT by proposing an MLRT based on a modified likelihood function

$$\tilde{L}_n^{(m)}(G) = L_n^{(m)}(G) - pen(G)$$

with the ordinary log-likelihood function

$$L_n^{(m)}(G) = \sum_{i=1}^{n} \log\ f_{\text{mix}}\big(Y_i; G\big)$$

and a penalty

$$pen(G) = -C_m \sum_{k=1}^{m} \log\big(\pi_k(G)\big),$$

for $G \in \mathfrak{M}_m$ with a positive constant $C_m > 0$. Again the penalty $pen(G)$ tends to infinity, if $\pi_k(G)$ approaches zero for some k such that degenerated components are penalized.
We denote the MMLE, i.e. the maximizer of $\tilde{L}_n^{(m)}(\,\cdot\,)$, as $\hat{G}^{(m)}$. Based on the MMLEs the MLRT statistic for testing (3.3) is

$$T_n^{\text{mod}} = 2\big(L_n^{(m)}(\hat{G}^{(m)}) - L_n^{(2)}(\hat{G}^{(2)})\big), \tag{3.4}$$

for some choice of m, where $L_n^{(m)}$ and $L_n^{(2)}$ denote the ordinary likelihood functions. At a first glance this is in contrast to the MLRT for homogeneity, where the modified log-likelihood is not only used to determine $\hat{G}^{(2)}$ but also is part of the test statistic. In case of Theorem 3.2 one may replace $\tilde{L}_n$ by L_n, since the penalty tends to zero under the homogeneity hypothesis. This is not the case for the penalty introduced above, since $pen(G^{(m)}) \geq C_m m \log m$.
In the remainder of this section we present the asymptotic results on T_n^{mod} by Chen et al. (2004) who established

$$T_n^{\text{mod}} \xrightarrow{\mathcal{L}} (\frac{1}{2} - p)\,\chi_0^2 + \frac{1}{2}\chi_1^2 + p\,\chi_2^2,$$

for some $p \in [0, 1/2]$.

Regularity conditions

In the same spirit as in the previous section we define the following quantities

$$Z_{ik}^1(\theta) = \frac{f_\theta(Y_i) - f_{\theta_{0k}}(Y_i)}{f_{\text{mix}}\big(Y_i; G_0\big)}, \quad \text{for } k = 1, 2$$

$$Z_i^l(\theta) = \frac{\frac{d^l}{d\theta^l} f_\theta(Y_i)}{f_{\text{mix}}\big(Y_i; G_0\big)}, \quad \text{for } l = 1, 2, 3$$

where G_0 denotes the true two component mixing distribution

$$G_0(\theta) = \pi_0\, I_{\{\theta_{01} \leq \theta\}} + (1 - \pi_0) I_{\{\theta_{02} \leq \theta\}}.$$

For $G \in \mathfrak{M}_m$ we define $Z_{ik}^1(G) = \int Z_{ik}^1(\theta) dG(\theta)$. Similar regularity conditions as above are used by Chen et al. (2004). Essentially, the assumptions reformulate Assumptions 3.1-3.5 in terms of a true underlying distribution $f_{\text{mix}}\big(y; G_0\big)$ rather than $f_{\theta_0}(y)$. In addition, one requires the third derivative of the kernel function $f_\theta(y)$.

Assumption 3.1' (Wald-type integrability condition) Let $E_0\left[|\log f_{\text{mix}}(Y_i; G_0)|\right] < \infty$ and there exists $\varepsilon > 0$ such that, for each G, $f_{\text{mix}}(y; G, \varepsilon) := 1 + \sup_{|Q-G|\leq\varepsilon} f_{\text{mix}}(y; Q)$ is measurable and $E_0\left[\log f_{\text{mix}}(Y_i; G, \varepsilon)\right] < \infty$.

Assumption 3.2' (Smoothness) The support of $f_\theta(y)$ does not depend on θ and $f_\theta(y)$ is three times continuously differentiable w.r.t. $\theta \subset \Theta$. The derivatives are jointly continuous in y and θ.

Assumption 3.3' (Strong identifiability) Same as Assumption 3.3.

Assumption 3.4' (Uniform boundedness) There exists an integrable function g and $\delta > 0$ such that $|Z_{ik}^1(\theta)|^{4+\delta} \leq g(Y_i)$ and $\left|Z_i^l(\theta)\right|^3 \leq g(Y_i)$ for all θ, for $k = 1, 2; l = 1, 2, 3$.

Assumption 3.5' (Tightness) The processes $n^{-1/2}\sum_i Z_{ik}^1(\theta)$, $n^{-1/2}\sum_i Z_i^l(\theta)$ are tight for $k = 1, 2; l = 1, 2, 3$.

Asymptotic analysis

For the analysis of $G^{(m)}$ we introduce a representation of the form

$$G^{(m)} = \hat{\pi}\hat{G}_1^{(m)} + (1 - \hat{\pi})\hat{G}_2^{(m)}$$

with $\hat{\pi} = \hat{G}^{(m)}(\theta_{mid})$ and $\hat{G}_1^{(m)}(\theta_{mid}) = 1$, $\hat{G}_2^{(m)}(\theta_{mid}) = 0$, $\theta_{mid} = 1/2(\theta_{01} + \theta_{02})$, i.e. we collect all support points θ of $\hat{G}^{(m)}$ with $\theta \leq \theta_{mid}$ and the corresponding weights in $\hat{G}_1^{(m)}$ and the remaining ones in $\hat{G}_2^{(m)}$.
Using this representation consistency of the MMLE $\hat{G}$ can be formulated as follows

$$\hat{\pi} \xrightarrow{\mathcal{P}} \pi_0 := \pi_1(G_0), \qquad |\hat{G}_k(\theta)^{(m)} - I_{\{\theta_{0k}\leq\theta\}}| \xrightarrow{\mathcal{P}} 0.$$

In the working paper by Chen et al. (2004) consistency of $\hat{G}$ is proved based on the fact that the probability of $\hat{G}$ to possess degenerated components tends to zero due to the structure of the penalty.

Before starting the analysis of the MLRT we note that test statistic T_n^{mod} is defined for a specified number of components m under the alternative. Chen et al. (2004) show that if one chooses $m \geq m^* := \max\left\{\lfloor 1.5/\pi_0 \rfloor, \lfloor 1.5/(1-\pi_0) \rfloor, 4\right\}$ this ensures that $\hat{G}_k^{(m)}$ for

$k = 1, 2$ is not degenerated to a single point distribution, which is necessary to obtain the asymptotic $\bar{\chi}^2$- distribution. Hence we assume that in the definition of T_n^{mod} m is chosen appropriately.
Since the asymptotic analysis of T_n^{mod} is also the cornerstone for the results of the Sections 3.2.3 and 3.3.2 we discuss it in some detail here. As standard for the LRT we start with the decompose

$$T_n^{\text{mod}} = T_{1n}^{\text{mod}} - T_{0n}^{\text{mod}} = 2\big(L_n^{(m)}(\hat{G}^{(m)}) - L_n^{(2)}(G_0)\big) - 2\big(L_n^{(2)}(\hat{G}^{(2)}) - L_n^{(2)}(G_0)\big).$$

We begin our analysis with T_{1n}^{mod} and obtain

$$T_{1n}^{\text{mod}} = 2\sum_i \log(1+\delta_i) \leq 2\sum_i \delta_i - \sum_i \delta_i^2 + 2/3 \sum_i \delta_i^3$$

for $\delta_i = \big(f_{\text{mix}}\big(Y_i; \hat{G}^{(m)}\big) - f_{\text{mix}}\big(Y_i; G_0\big)\big)/f_{\text{mix}}\big(Y_i; G_0\big)$. We may omit the index m in this part, i.e. $\hat{G}^{(m)} = \hat{G}$. The crucial observation is that

$$\delta_i = (\hat{\pi} - \pi_0)\Delta_i + \hat{\pi} Z_{i1}^1(\hat{G}_1) + (1-\hat{\pi}) Z_{i2}^1(\hat{G}_2) \quad \text{ with } \Delta_i = Z_{i2}^1(\theta) - Z_{i1}^1(\theta),$$

such that one can basically apply similar arguments on $Z_{ik}^1(\hat{G}_k)$ for each $k = 1, 2$ as those from the homogeneity case (Chen et al., 2001).
Expanding $Z_{ik}^1(\theta)$ yields for $k = 1, 2$ and $i = 1, \ldots, n$

$$Z_{ik}^1(\theta) = (\theta - \theta_{0k}) Z_i^1(\theta_{0k}) + (\theta - \theta_{0k})^2/2 \; Z_i^2(\theta_{0k}) + \varepsilon_{ik} \tag{3.5}$$

and

$$Z_{ik}^1(\hat{G}_k) = m_{1k}(\hat{G}_k) Z_i^1(\theta_{0k}) + m_{2k}(\hat{G}_k)/2 \; Z_i^2(\theta_{0k}) + \tilde{\varepsilon}_{ik} \tag{3.6}$$

with $m_{lk}(G) = \int (\theta - \theta_{0k})^l dG(\theta)$. Hence

$$\begin{aligned} \sum_i \delta_i &= \sum_i (\hat{\pi} - \pi_0)\Delta_i + \hat{\pi} m_{11}(\hat{G}_1) Z_i^1(\theta_{01}) + (1-\hat{\pi}) m_{12}(\hat{G}_2) Z_i^1(\theta_{02}) \\ &\quad + \hat{\pi} m_{21}(\hat{G}_1)/2 \; Z_i^2(\theta_{01}) + (1-\hat{\pi}) m_{22}(\hat{G}_2)/2 \; Z_i^2(\theta_{02}) + \tilde{\varepsilon}_{1n} \\ &= t(\hat{G})\, \mathbf{b} + \tilde{\varepsilon}_{1n} \end{aligned}$$

for

$$\mathbf{b} = \sum_i b_i = \sum_i (\Delta_i, Z_i^1(\theta_{01}), Z_i^1(\theta_{02}), Z_i^2(\theta_{01}), Z_i^2(\theta_{02}))^T. \tag{3.7}$$

and

$$t(G) = \big(\pi(G) - \pi_0, \pi_0 m_{11}(G_1), (1-\pi_0) m_{12}(G_2), \pi_0 m_{21}(G_1)/2, (1-\pi_0) m_{22}(G_2)/2\big)^T$$

with $G = \pi(G)G_1 + (1 - \pi(G))G_2$ partitioned as above. Similarly we get

$$\sum \delta_i^2 = t(\hat{G})^T \, \mathbf{B} \, t(\hat{G}) + \tilde{\varepsilon}_{2n}$$

with $\mathbf{B} = \sum_i b_i b_i^T$. Chen et al. (2004) prove under the Assumptions 3.1' – 3.5' the asymptotic neglectability $\tilde{\varepsilon}_{1n}$, $\tilde{\varepsilon}_{2n}$ and other higher order terms, and conclude

$$T_{1n}^{\text{mod}} \leq \sup_{G \in \mathfrak{M}_m} \big(2\, t(G)^T \mathbf{b} - t(G)^T \, \mathbf{B} \, t(G) \big) + o_P(1). \tag{3.8}$$

Since

$$\sup_{G \in \mathfrak{M}_m} \big(2\, t(G)^T \mathbf{b} - t(G)^T \, \mathbf{B} \, t(G) \big) = \sup_{t \in C} \big(2\, t^T \mathbf{b} - t^T \, \mathbf{B} \, t \big)$$

with $C = \mathbb{R}^3 \times [0, \infty) \times [0, \infty)$, one has simply the supremum of a quadratic form over a closed cone C, which is attained by some value $t^* \in C$. Examine t as a function of G yields that for all $t^* \in C$ there exists G^* such that $t^* = t(G^*)$. In particular, $G^* = \pi(G^*)G_1^* + (1 - \pi(G^*))G_2^*$ must satisfy

$$\pi(G^*) = t_1^* + \pi_0, \;\; m_{1k}(G_k^*) = t_{k+1}^*/\pi_0, \;\; m_{2k}(G_k^*) = 2t_{k+3}^*/\pi_0 \geq 0$$

for $k = 1, 2$, which is possible for all $t^* \in C$, if one chooses G_1^* and G_2^* in the way that they possess at least two support points. Hence we have

$$\sup_{G \in \mathfrak{M}_m} \big(2\, t(G)^T \mathbf{b} - t(G)^T \, \mathbf{B} \, t(G) \big) = \; 2\, t(G^*)^T \mathbf{b} - t(G^*)^T \, \mathbf{B} \, t(G^*) \tag{3.9}$$

and by the expansion above

$$T_{1n}^{\text{mod}} \geq 2\big(L_n^{(m)}(G^*) - L_n^{(2)}(G_0) \big) = \; 2\, t(G^*)^T \mathbf{b} - t(G^*)^T \, \mathbf{B} \, t(G^*) + o_P(1)$$

Combining this with (3.8) results in

$$T_{1n}^{\text{mod}} = \sup_{G \in \mathfrak{M}_m} \big(2\, t(G)^T \mathbf{b} - t(G)^T \, \mathbf{B} \, t(G) \big) + o_P(1). \tag{3.10}$$

Note that, this implies that $m_{1k}(G^*)$ and $m_{2k}(G^*)$ are of the same order (by the tightness assumption this is $n^{-1/2}$), and hence $|G^* - G_0| = O_P(n^{-1/4})$. This corresponds the rates in the homogeneity case discussed Section 3.1.1. For the estimation under the hypothesis, $\hat{G}_1^{(2)}$ and $\hat{G}_2^{(2)}$ are both single point distributions which implies that $(m_{1k}(\hat{G}_k^{(2)}))^2 = m_{2k}(\hat{G}_k^{(2)})$, yielding $m_{2k}(\hat{G}_k^{(2)}) = o_p(m_{1k}(\hat{G}_k^{(2)}))$ for $k = 1, 2$ and hence

$$\begin{aligned} T_{0n}^{\text{mod}} &= \sup_{G \in \mathfrak{M}_2} \big(2\, t(G)^T \mathbf{b} - t(G)^T \, \mathbf{B} \, t(G) \big) + o_P(1) \\ &= \sup_{t_1 \in \mathbb{R}^3} \big(2\, t_1^T \mathbf{b_1} - t_1^T \, B_{11} \, t_1 \big) + o_P(1) = \mathbf{b}_1^T {B_{11}}^{-1} \mathbf{b}_1^T + o_P(1) \end{aligned}$$

for $\mathbf{b}^T = (\mathbf{b}_1^T, \mathbf{b}_2^T)$, $t^T = (t_1^T, t_2^T)$ with $\mathbf{b}_1, t_1 \in \mathbb{R}^3$ and

$$\mathbf{B} = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array} \right), \qquad B_{11} \in \mathbb{R}^{3\times 3}.$$

Assumption 3.3' ensures that $\mathbf{B}$ is invertible (at least for large n), hence orthogonalization leads to

$$T_{1n}^{\text{mod}} = \mathbf{b}_1^T {B_{11}}^{-1} \mathbf{b}_1^T + \sup_{t_2 \in [0,\infty)\times[0,\infty)} \left(2\, t_2^T \tilde{\mathbf{b}}_2 - t_2^T\, \tilde{\mathbf{B}}_{22}\, t_2 \right) + o_P(1)$$

with $\tilde{\mathbf{b}}_2^T = \mathbf{b}_2^T - \mathbf{b}_1^T {B_{11}}^{-1} B_{12}$ and $\tilde{\mathbf{B}}_{22} = B_{22} - B_{21} {B_{11}}^{-1} B_{12}$, where t_2 is restricted to positive values.
Therefore we have

$$T_n^{\text{mod}} \quad = \quad \sup_{t_2 \in [0,\infty)\times[0,\infty)} \left(2\, t_2^T \tilde{\mathbf{b}}_2 - t_2^T\, \tilde{\mathbf{B}}_{22}\, t_2 \right) + o_P(1) \tag{3.11}$$

and the following proposition completes the analysis.

Proposition 3.1. *Suppose that Assumptions 3.1' – 3.5' hold and that true distribution of $(Y_i)_i$ is a two component finite mixture, i.e. $G_0 \in \mathfrak{M}_2$. Then*

$$\sup_{t_2 \in [0,\infty)\times[0,\infty)} \left(2\, t_2^T \tilde{\mathbf{b}}_2 - t_2^T\, \tilde{\mathbf{B}}_{22}\, t_2 \right) \xrightarrow{\mathcal{L}} (\frac{1}{2} - p)\, \chi_0^2 + \frac{1}{2}\chi_1^2 + p\, \chi_2^2,$$

where $p = \left(\cos^{-1} \rho \right)/(2\pi)$ and ρ is the correlation coefficient in the asymptotic covariance matrix $\tilde{\Sigma}$ of $n^{-1/2}\tilde{\mathbf{b}}_2$.

The proposition is proved slightly different than in Chen et al. (2004) in Sec. 3.4. We can now derive the asymptotic distribution of T_n^{mod} and formulate the theorem by Chen et al. (2004).

Theorem 3.3. *Suppose that Assumptions 3.1' – 3.5' hold and that true distribution of $(Y_i)_i$ is a two component finite mixture. Further assume that m in the definition of T_n^{mod} in (3.4) satisfies $m \geq m^* := \max\left\{ \lfloor 1.5/\pi_0 \rfloor, \lfloor 1.5/(1-\pi_0) \rfloor, 4 \right\}$. Then*

$$T_n^{mod} \xrightarrow{\mathcal{L}} (\frac{1}{2} - p)\, \chi_0^2 + \frac{1}{2}\chi_1^2 + p\, \chi_2^2,$$

where $p = \left(\cos^{-1} \rho \right)/(2\pi)$ and ρ is the correlation coefficient in the covariance matrix $\tilde{\Sigma}$.

Remark 3.4 (Choice of m and C_2, C_m)**.** Chen et al. (2001, 2004) make several suggestions, how to choose m and the constants C_2, C_m to form the MLRT. Especially, Chen et al. (2004) advocate an data-adaptive choice of

$$m \geq \hat{m} = \max\left\{ \left\lfloor 1.5/\pi_1(\hat{G}^{(2)}) \right\rfloor, \left\lfloor 1.5/(1-\pi_1(\hat{G}^{(2)})) \right\rfloor, 4 \right\}.$$

Intuitively, a greater value of m should increase the value of T_n^{mod}, but it neither affects the asymptotic behavior of T_n^{mod} (by the previous theorem) nor simulation studies indicate a big difference, as long as m is not too small. For $m < m^*$, $\hat{G}_k^{(m)}$ for $k = 1, 2$ may degenerate to a single point which does affect the asymptotic distribution of T_n^{mod} such that the resulting test is conservative. Concerning the choice of C_2, C_m simulation studies do not indicate a strong influence of those constants. In general, increasing C yields a stronger effect of the penalization on the estimates and pushes $\pi_k^{(m)}$ towards uniform weights $1/m$. One may expect that this should result in better approximations under the hypothesis, but in a decrease of power. A common choice is $C_m = C_2 = 1$, but one use the constants to adjust the level of the test under the hypothesis for a particular model verified by means of simulation studies.

Remark 3.5. To apply the asymptotic distribution in Theorem 3.3, i.e. to determine the critical region of the test, one needs the correlation coefficient ρ. The standard estimator $\hat{\rho}$ is the correlation coefficient of $\tilde{\mathbf{B}}_{22}$. As it is consistent, a test based on the $\bar{\chi}^2$-mixture with $\hat{p} = \big(\cos^{-1}\hat{\rho}\big)/(2\pi)$ keeps the prespecified level asymptotically.

3.2 Testing for the number of states in a hidden Markov model

Testing problems concerning the number of components in a finite mixture model are similarly present in the HMM context, where one wishes to specify the number of hidden states correctly. For testing m = 1 against m = 2 for an HMM, Gassiat and Keribin (2000) show that the LRT statistic diverges to infinity. Their result is based on studying the subproblem, where the parameters of the sdfs are fixed, and investigates a family of Gaussian processes V_η, $\eta \in (0, 1)$, which are smaller than the LRT, but unbounded if η approaches zero. This result is rather remarkable, since the divergence of the LRT takes place in a finite-dimensional setup although the standard regularity and compactness assumptions are fulfilled. However, simulations indicate that the speed of convergence is pretty slow. Although the work by Gassiat and Keribin (2000) benefits from the fact that for $m = 1$, the $(Y_i)_i$ are simply an i.i.d. sequence from $f_{\theta_1}(y)$, their result requires a lot of technical effort.

The simplest non-trivial (i.e. dependent) HMM has to have at least two states. Therefore, testing for $m = 2$ against $m \geq 3$ states for an HMM is the problem of primary practical interest. For this problem the LRT has not been investigated so far. The results of Gassiat and Keribin (2000) are not very encouraging, but they indicate the severity of the

problem. Due to the lack of asymptotic theory Rydén et al. (1998) used a bootstrap version of the LRT for this problem. However, bootstrapping in this context is computationally demanding, since it requires repeated maximization of the full log-likelihood function of an HMM for more than two states.

In this section, we investigate how the MLRT of Chen et al. (2001, 2004) can be extended to HMMs. For mixture models the MLRT is proved to have a relatively simple limit theory, and is computationally easy to handle, since it does no require bootstrapping of the asymptotic distribution. We apply these methods to HMMs hoping (and proving) that the desirable properties are preserved.

Note, that the marginal distribution of observations $(Y_i)_i$ from an HMM is given by the finite mixture $f_{\text{mix}}(y; G)$ where the support points of G correspond to the parameters of the sdfs and the weights to the stationary distribution π. In particular, when one assumes that the parameters of the sdfs are distinct, the number of hidden states m coincides naturally with the number of components of the marginal mixture distribution. Hence, testing for the number of components of the marginal mixture is equivalent to testing for the number of states of the underlying latent process.

In general, estimation based on a quasi likelihood function formed by the marginal mixture distribution, also called likelihood function under independence assumption, was proposed by Lindgren (1978) for HMMs and switching regression models. As an illustration we first discuss based on Lindgren's results how to test regular hypotheses on parameters of the marginal mixture distribution of the HMM via a quasi LRT in Section 3.2.1.

It turns out that this test statistic is not asymptotically χ^2-distributed in general, but rather requires an adjustment for the dependence structure of the HMM. This is not the case when applying the MLRT for testing for homogeneity in an HMM (see Section 3.2.2) for the simple reason that under the hypothesis $m = 1$ no dependency is present. Surprisingly, as shown in Section 3.2.3, the MLRT for $m = 2$ against $m \geq 3$ in an HMM does also not require an adjustment. Although the observations exhibit dependency under the hypothesis $m = 2$, the limit distribution of the MLRT is the same as for independent mixtures. We may point out that this makes its use for HMMs particularly simple and attractive. Since the evaluation of the MMLE is much simpler and hence much faster than for the MLE based on the full log-likelihood function of the HMM, inference based on the marginal mixture distribution of the HMM is also computationally attractive.

The section is concluded by simulation studies, where the theoretical results are illustrated. In particular, we verify empirically that the performance of the MLRT for testing $m = 2$ is hardly influenced by different forms of the transition matrix, as long as its stationary distribution remains the same. Finally we give two empirical illustrations, one for the series

of fetal lamb movements analyzed in Leroux and Puterman (1992), and the other to the series of log-returns of the S&P 500 (cf. Rydén et al., 1998).

3.2.1 The LRT under independence assumption

The density of marginal distribution of the observations from an HMM $(Y_i)_i$ is given by the density of the finite mixture

$$f_{\text{mix}}(y; \pi_1, \ldots, \pi_m, \theta_1, \ldots, \theta_m) = \pi_1 f_{\theta_1}(y) + \cdots + \pi_m f_{\theta_m}(y)$$

where π is the stationary distribution of the transition matrix. For estimation based on $f_{\text{mix}}(y; \pi_1, \ldots, \pi_m, \theta_1, \ldots, \theta_m)$ the parameter of interest contains the entries of the stationary distribution rather than the entries of the transition matrix. We denote a suitable parametrization by ω with $\pi_k(\omega)$ and $\theta_k(\omega)$ for $1 \leq k \leq m$ and assume $\omega \in \Omega \subset \mathbb{R}^{\bar{d}}$, and write $f_{\text{mix}}(y; \omega) = f_{\text{mix}}(y; \pi_1, \ldots, \pi_m, \theta_1, \ldots, \theta_m)$. Based on the marginal density one forms the log-likelihood function under independence assumption

$$L_n^I(\omega) = \sum_{i=1}^{n} \log\, f_{\text{mix}}(Y_i; \omega), \tag{3.12}$$

and defines the maximum likelihood estimator under independence assumption (MLEI) by

$$\hat{\omega} = \arg\max_{\omega \in \Omega} L_n^I(\omega).$$

Lindgren (1978) proposes this estimator for the stationary distribution and the parameters of the sdfs for HMMs and phrases a CLT for those. With the matrices

$$\begin{aligned} \Sigma_0 &= E\big[h(Y_1; \omega_0)\, h(Y_1; \omega_0)^T\big] \\ \text{Cov}_0 &= \Sigma_0 + \sum_{j \geq 2} E\big[h(Y_1; \omega_0)\, h(Y_j; \omega_0)^T + h(Y_j; \omega_0)\, h(Y_1; \omega_0)^T\big], \end{aligned}$$

where $h(y; \omega) = \big(D_\omega \log f_{\text{mix}}(y; \omega)\big)^T$, we can formulate the following theorem in the spirit of Lindgren (1978).

Theorem 3.4. *Suppose that we have an HMM with ergodic regime fulfilling Assumptions 2.1', 2.2', 2.3, rephrased for the parametrization ω. Assume that ω_0 is an interior point of Ω compact. Then, if the MLEI is strongly consistent and Σ_0 nonsingular, we have*

$$\sqrt{n}\big(\hat{\omega} - \omega_0\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \Sigma_0^{-1} \text{Cov}_0 \Sigma_0^{-1}\big). \tag{3.13}$$

An outline of the proof is given in Section 3.4. As we require the MLEI to be consistent, we may note that this holds under Wald-type conditions, as a repetition of the arguments in Rydén (1994) shows. In this paper Rydén (1994) extends Lindgren's approach by assuming independence only between blocks of observations.

Based on the MLEIs one can test hypotheses about ω via a LRT under independence assumption (LRTI). We now briefly discuss the LRTI for regular hypotheses in order to illustrate that its asymptotic distribution is not given by a simple χ^2-distribution but is in general significantly influenced by the dependence structure of the HMM. Note that since π is uniquely determined by the transition matrix $\big(\alpha_{jk}\big)_{1\le j,k\le m}$, hypotheses on ω can in principle be reformulated into hypotheses on the original parameters of the HMM, and hence be tested by the usual LRT for HMMs (cf. Sections 2.1.1, 2.1.2 and Example 2.4). However, as seen in Example 2.4 the expression of π in terms of the entries of the transition matrix is highly nonlinear for $m \ge 3$, and thus the ordinary LRT becomes difficult to use in such situations. Hence, for $m \ge 3$ the LRTI is also an attractive procedure to test hypotheses on the stationary distribution of an HMM, e.g. testing for $\pi_1 = \ldots = \pi_m = 1/m$. More precisely, suppose that we want to test a regular r-dimensional restriction

$$H_s:\ s(\omega_0) = 0 \qquad \text{against} \qquad K_s:\ s(\omega_0) \neq 0,$$

where $s : \mathbb{R}^{\bar{d}} \to \mathbb{R}^r$, $r \le \bar{d}$, is a differentiable map with Jacobian $D_\omega s(\omega_0)$ of full rank r at ω_0. Let

$$T_n^I = 2\big(\sup_{\omega\in\Omega} L_n^I(\omega) - \sup_{\{\omega|s(\omega)=0\}} L_n^I(\omega)\big)$$

be the LRTI statistic. In order to derive the asymptotic distribution of T_n^I, reparametrize H_s (at least locally around ω_0) as the image of a differentiable mapping $\varphi : \mathbb{R}^{\bar{d}-r} \supset U \to \mathbb{R}^{\bar{d}}$, i.e. $s(\varphi(t)) = 0$, and these are the only solutions locally around ω_0. Let $S_0 = D_t\varphi(t_0)$, where $\varphi(t_0) = \omega_0$.

Theorem 3.5. *Suppose that we have an HMM with ergodic regime fulfilling Assumptions 2.1', 2.2', 2.3, rephrased for the parametrization ω and t. Assume that $\omega_0 = \varphi(t_0)$ is an interior point of Ω compact. Then, if the restricted MLEI as well as the unrestricted MLEI are strongly consistent and Σ_0 nonsingular and S_0 of full rank we have*

$$T_n^I \xrightarrow{\mathcal{L}} Z^T \mathrm{Cov}_0^{1/2}\Big(\Sigma_0^{-1} - S_0(S_0^T\Sigma_0 S_0)^{-1}S_0^T\Big)\mathrm{Cov}_0^{1/2} Z, \tag{3.14}$$

where $Z \sim N(0, I_{\bar{d}\times\bar{d}})$.

Once Theorem 3.4 and simultaneously an CLT for the score and an LLN for the Fisher information are established, deriving the asymptotic distribution of T_n^I follows from a standard argument. For details see Sec. 3.4.
In principle, one may also use the LRTI to test hypotheses on the stationary distribution involving the boundary of the parameter space as in Example 2.4.

Remark 3.6. The quadratic form which occurs as asymptotic distribution in (3.14) is a linear combination of independent χ_1^2 distributed variables, where the weights are given by the eigenvalues of the matrix $\mathrm{Cov}_0^{1/2}\Big(\Sigma_0^{-1} - S_0(S_0^T\Sigma_0 S_0)^{-1}S_0^T\Big)\mathrm{Cov}_0^{1/2}$. If the observation were independently drawn from a finite mixture, this matrix would be an orthogonal projection, since in that case $\mathrm{Cov}_0 = \Sigma_0$. In general, Σ_0 and Cov_0 differ due to the dependence structure of an HMM and hence the matrix is no orthogonal projection. Hence, the asymptotic distribution of the LRTI will in general not be a simple χ^2-distribution.
For an application of (3.14), these eigenvalues have to be estimated, by consistently estimating all component matrices Cov_0, Σ_0 and S_0 and using the fact that the eigenvalues depend continuously on the entries.

Remark 3.7. As an alternative to the LRTI one can also use a Wald-type statistic as follows. Suppose that Σ_0 and Cov_0 are non-singular, and let Σ_n and $\widehat{\mathrm{Cov}}_n$ be consistent estimates of Σ_0 and Cov_0, respectively. Then, under H_s and non-singularity of Σ_0, one shows by using the δ-method that

$$W_n^I = ns(\hat{\omega})^T\Big(D_\omega s(\hat{\omega})\,\Sigma_n^{-1}\widehat{\mathrm{Cov}}_n\Sigma_n^{-1}\left(D_\omega s(\hat{\omega})\right)^T\Big)^{-1}s(\hat{\omega}) \xrightarrow{\mathcal{L}} \chi_r^2. \tag{3.15}$$

The Wald test and the LRTI will be less efficient than the LRT (based on the full-model MLEs), and the usual LRT should thus be employed if possible. However, in the simulation section we illustrate that the loss in power is pretty small, and hence that the tests under independence assumption offer a reasonable, simple alternative.

3.2.2 Testing for homogeneity in an HMM

We briefly discuss testing for homogeneity using the MLRT approach (see Sec. 3.1.1) in an HMM. This section is mainly motivated by questions raised at the GOCPS (Aachen, 2008) and the WCPS (Singapore, 2008) when presenting the paper by Dannemann and Holzmann (2008c). The question is whether one can formulate the MLRT based on the log-likelihood under independence assumption to test $H : m = 1$ against $K : m = 2$ in an HMM, i.e.

$$H : \alpha_{12}\alpha_{21}(\theta_1 - \theta_2) = 0 \quad \text{against} \quad K : \alpha_{12}, \alpha_{21} \in [0,1], \theta_1, \theta_2 \in \Theta. \tag{3.16}$$

The answer is positive, the MLRT from Chen et al. (2001) can be applied. Since under the hypotheses in (3.1) and (3.16) the observations are simply i.i.d. with $Y_i \sim f_{\theta_0}$, Theorem 3.2 gives the asymptotic behavior of the MLRT under the null hypothesis. However, in the case of HMMs one compares a very simple model with independent observations with a much more complex model with a dependency structure. The result by Gassiat and Keribin (2000) indicates theses model classes are quite different.
In the end the simple answer given above refers to the behavior of the MLRT under H only. For a complete picture one has to analyze the behavior of of the MLRT under an HMM belonging to the alternative, to see whether the dependency structure of the HMM influences the test. In principle, one can also test (3.16) by statistical tools detecting dependency, rather than using a method which ignores dependency by construction.

3.2.3 Testing for two states in an HMM

As argued at the beginning of this section testing for two states in an HMM is of major interest, since a two-state HMM is the smallest non-trivial model. We now propose a test for

$$H : m = 2 \quad \text{against} \quad K : m \geq 3$$

in an HMM. To apply the MLRT from Chen et al. (2004) we assume that the parameters of the sdfs are all distinct and one dimensional, i.e. $\Theta \subset \mathbb{R}$. Adopting the notation introduced in Sec. 3.1.2 by denoting the true regime distribution of an HMM as G_0 the testing problem for two states is given by

$$H : \ G_0 \in \mathfrak{M}_2 \qquad \text{against} \quad K : \ G_0 \in \mathfrak{M} \setminus \mathfrak{M}_2. \tag{3.17}$$

With respect to this notation the ordinary log-likelihood under independence assumption is given by $L_n^{I(m)}(G) = \sum_{i=1}^n \log f_{\text{mix}}\big(Y_i; G\big)$ and the modified log-likelihood under independence assumption is defined as

$$\tilde{L}_n^{I(m)}(G) = L_n^{I(m)}(G) - pen(G)$$

with a penalty

$$pen(G) = -C_m \sum_{k=1}^{m} \log\big(\pi_k(G)\big).$$

Analogous to the notion in Sec. 3.1.2 we denote the MMLE under independence assumption, i.e. the maximizer of $\tilde{L}_n^{I(m)}(\,\cdot\,)$, as $\hat{G}^{(m)}$. The MLRT statistic under independence assumption for testing (3.17) is then formed by

$$T_n^{\text{mod}} = 2\big(L_n^{I(m)}(\hat{G}^{(m)}) - L_n^{I(2)}(\hat{G}^{(2)})\big), \tag{3.18}$$

Chen et al. (2004) show that under the hypothesis of a two-component mixture T_n^{mod} follows asymptotically a $\bar{\chi}^2$-mixture (cf. Theorem 3.3). Now we must investigate the asymptotic behavior of T_n^{mod} under the hypothesis of a two-state (i.e. dependent) HMM.
Following the analysis by Chen et al. (2004) discussed in Sec. 3.1.2 we see that under the Assumptions 3.1' – 3.5' the expansion

$$T_n^{\text{mod}} = \sup_{t_2 \in [0,\infty)\times[0,\infty)} \left(2\, t_2^T \tilde{\mathbf{b}}_2 - t_2^T\, \tilde{\mathbf{B}}_{22}\, t_2 \right) + o_P(1)$$

remains valid for the quantities $\tilde{\mathbf{b}}_2$, $\tilde{\mathbf{B}}_{22}$ defined in Sec. 3.1.2. For the final step Chen et al. (2004) exploit the fact that the covariance matrix of the asymptotic normal distribution of $n^{-1/2}\tilde{\mathbf{b}}_2$ coincides with the limit of $n^{-1}\tilde{\mathbf{B}}_{22}$ denoted as $\tilde{\Sigma}$. For independent data this is obviously true, since $\mathbf{b}$ is in this case the sum of independent increments, i.e. $E\big[b_i b_j^T\big] = 0$ if $1 \le i \ne j \le n$. In our case these increments are dependent and hence $n^{-1/2}\tilde{\mathbf{b}}_2$ is asymptotically normally distributed with mean zero and covariance matrix

$$\widetilde{\text{Cov}} = \tilde{\Sigma} + \sum_{i=2}^{\infty} E\big[\tilde{b}_{21}\tilde{b}_{2i}^T + \tilde{b}_{2i}\tilde{b}_{21}^T\big].$$

with $\tilde{b}_{2i}^T = b_{2i}^T - b_{1i}^T \bar{B}_{11}^{-1} \bar{B}_{12}$, $b_i^T = (b_{1i}^T, b_{2i}^T)$, $b_1 \in \mathbb{R}^3$ and $\bar{B}_{jk} = E\big[b_{ji} b_{ki}\big] = \lim_n n^{-1} B_{jk}$ for $1 \le j, k \le 2$.
Surprisingly, for the asymptotic distribution of $n^{-1/2}\tilde{\mathbf{b}}_2$ we indeed have $\tilde{\Sigma} = \widetilde{\text{Cov}}$, as stated in the next proposition.

Proposition 3.2. *Suppose that Assumptions 3.1' – 3.5' hold and that the true marginal distribution of the HMM $(Y_i)_i$ is a two-component finite mixture. Then we have*

$$E\big[\tilde{b}_{21}\tilde{b}_{2i}^T\big] = E\big[\tilde{b}_{2i}\tilde{b}_{21}^T\big] = 0 \qquad \textit{for all } i \ge 2.$$

The proof is given in Sec. 3.4. The result is much in contrast to the relation of the matrices Σ_0 and Cov_0 introduced in Section 3.2.1, as we shall illustrate in the simulation study in Section 3.2.4. Proposition 3.2 implies that T_n^{mod} will have the same limit distribution as for independent mixtures. In particular, analogously to Theorem 3.3 we have

Theorem 3.6. *Suppose that Assumptions 3.1' – 3.5' hold and that the true marginal distribution of the HMM $(Y_i)_i$ is a two-component finite mixture. Further assume that m in the definition of T_n^{mod} in (3.18) satisfies $m \ge m^* := \max\{\lfloor 1.5/\pi_1^0 \rfloor, \lfloor 1.5/\pi_2^0 \rfloor, 4\}$. Then*

$$T_n^{mod} \xrightarrow{\mathcal{L}} \left(\frac{1}{2} - p\right)\chi_0^2 + \frac{1}{2}\chi_1^2 + p\,\chi_2^2, \tag{3.19}$$

where $p = \big(\cos^{-1}\rho\big)/(2\pi)$ and ρ is the correlation coefficient in the covariance matrix $\tilde{\Sigma}$.

After Proposition 3.2 is established the proof of the theorem follows the proof in Chen et al. (2004), sketched in Sec. 3.1.2. Comments on the validity of the assumptions are passed in a subsequent remark (and proved in Sec. 3.4). Clearly, the Remarks 3.4, 3.5 on the choice of m and the constants C_2, C_m as well as on estimation of the correlation coefficient ρ also apply when the MLRT is used for testing for two states. In addition we should comment how the Assumptions 3.1' – 3.5' could be verified in the context of HMMs

Remark 3.8. The Assumptions 3.1' – 3.4' are mainly concerned with the kernel functions $f_\theta(\,\cdot\,)$, so that their validity remains in the context of HMMs. Concerning the tightness of the processes $n^{-1/2}\sum_i Z^1_{ik}(\theta)$, $n^{-1/2}\sum_i Z^l_i(\theta)$ (Assumption 3.5') Chen et al. (2004) argue that for independent mixtures this is implied by Assumption 3.4' by applying Theorem 12.3 in (Billingsley, 1968, p.95). This is also true for the HMM setup, which is proved in Sec. 3.4.

3.2.4 Simulation experiments

Here we present some results of an simulation study of the tests proposed in the previous sections in the HMM setup. For the maximization of the log-likelihood function (under independence assumption) we use direct maximization via a Newton-type algorithm (cf. Sec. 2.3.1).

The LRT under independence assumption

In this section we shall illustrate two aspects about the LRT under independence assumption discussed in Sec. 3.2.1. First, the difference between Σ_0 and Cov_0 can be quite large and the distribution of T_n^I can be quite far from a χ^2-distribution, even in a simple setting. Second, we show that (at least in a particular example), the LRTI and the Wald test under independence assumption have little loss in power when compared to the LRT based on full-model MLEs discussed in Sec. 2.4. Thus, ignoring the dependence structure in the test statistic need not result in a significant loss of power.
We start by suggesting estimators for the matrices Σ_0 and Cov_0, where Σ_0 is estimated by $\Sigma_n = \frac{1}{n}\sum_{i=1}^n h(Y_i;\hat\omega)h(Y_i;\hat\omega)^T$, and Cov_0 by

$$\widehat{\mathrm{Cov}}_n = \Sigma_n + \sum_{j=1}^{J} \frac{n-j}{n}\,\Sigma_{n,j}, \quad \Sigma_{n,j} = \frac{1}{n-j}\sum_{i=1}^{n-j}\Big(h(Y_i;\hat\omega)h(Y_{i+j};\hat\omega)^T + h(Y_{i+j};\hat\omega)h(Y_i;\hat\omega)^T\Big),$$

where J is small compared to n. Typically, the covariances decrease exponentially fast, so a small number for J will suffice. In practice one can simply check for each j whether the entries of $\Sigma_{n,j}$ are small compared to Σ_n.

We simulate from a stationary three-state Poisson HMM, where the means of the Poisson sdfs are given by $\theta_1 = 1$, $\theta_2 = 5$ and $\theta_3 = 9$, and the transition matrix of the underlying Markov chain is of the form (2.19). We examine testing the hypothesis $H : \pi_1 = \pi_3$. Under H, we choose the entries in (2.19) as $\alpha = 0.4$, $\beta = 0.2$, $\gamma = 0.3$ and $\delta = 0.6$, yielding the stationary distribution $\pi_1 = \pi_3 = 0.25$, $\pi_2 = 0.5$. In the following, for simplicity we fix the θs at their true values, and estimate the parameter $\omega = (\pi_1, \pi_3)$ only. First, we generate estimates of Σ_0 and Cov_0 from a single sample of size 10^6, yielding for $J = 8$

$$\Sigma_n = \begin{pmatrix} 3.56 & 0.16 \\ 0.16 & 2.12 \end{pmatrix}, \qquad \widehat{\mathrm{Cov}}_n = \begin{pmatrix} 8.13 & -1.61 \\ -1.61 & 2.74 \end{pmatrix}, \qquad P_n = \begin{pmatrix} 1.34 & -1.18 \\ -1.18 & 1.03 \end{pmatrix}$$

where $P_n = \widehat{\mathrm{Cov}}_n^{1/2}\Big(\Sigma_n^{-1} - S_0(S_0^T\Sigma_n S_0)^{-1}S_0^T\Big)\widehat{\mathrm{Cov}}_n^{1/2}$ is an estimate of the matrix in the quadratic form in (3.14) (here, S_0 does not depend on ω). Thus, the matrices Σ_0 and Cov_0 apparently differ significantly. The matrix P_n is singular, its non-zero eigenvalue is equal to 2.38. Hence, the asymptotic distribution of the LRTI is a scaled χ^2_1–distribution with scaling factor 2.38.

The distribution of the LRTI-statistic and the Wald-statistic was investigated for sample size $n = 500$ with $N = 10000$ replications. Figure 3.1 shows the empirical cumulative distribution functions. In both cases one can hardly visually distinguish between the sample and the asymptotic distribution functions. However, one clearly observes that the distribution of LRTI differs strongly from the standard χ^2_1-distribution.

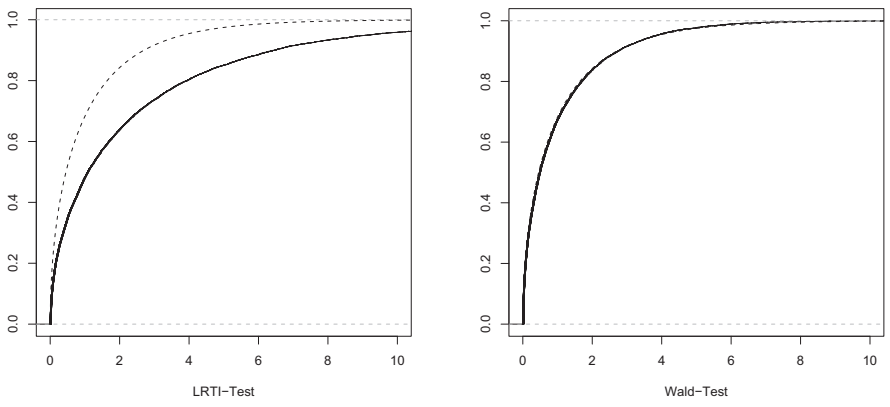


Figure 3.1: Distribution of the LRTI-Statistic and of the Wald-Statistic (solid), the dotted line (hardly visible) indicates the asymptotic distribution of the LRTI-Statistic and the dashed line the χ^2_1–distribution.

Finally, we conduct a power comparison between the LRTI, the Wald test under independence assumption and the LRT based on the full model MLEs. We again test the hypothesis $H : \pi_1 = \pi_3$, and the parameters in (2.19) are taken as $\alpha_1 = \alpha_0 + \kappa$, $\beta_1 = \beta_0 + \kappa$, $\gamma_1 = \gamma_0 - \kappa$ and $\delta_1 = \delta_0 - \kappa$, where $\alpha_0, \ldots, \delta_0$ are chosen as above, and for κ we use $\kappa = 0, 0.05, 0.1, 0.15, 0.25$. For all tests, the asymptotic critical values are employed (in case of the LRTI the critical value is estimated for each sample). The sample size was taken as $n = 500$, and $N = 10000$ samples were used to estimate the power in each setting. The results are displayed in Table 3.1. It turns out that at least in this specific scenario, there is little loss in power when using the tests based on the MLEI.

Table 3.1: Simulated rejection rates of the LRT based on the MLE, LRTI and Wald test based on MLEIs under the hypothesis ($\kappa = 0$) and under the alternative ($\kappa > 0$). The model is a three-state HMM with means 1, 5 and 8. The transition matrix is of the form (2.19) with $\alpha = 0.4 + \kappa, \beta = 0.2\kappa, \gamma = 0.3 - \kappa$ and $\delta = 0.6 - \kappa$.

κ	0	0.05	0.1	0.15	0.25
π_1 (true value)	0.25	0.276	0.300	0.323	0.377
π_3 (true value)	0.25	0.226	0.200	0.169	0.078
Power LRT	0.052	0.135	0.381	0.717	0.994
Power LRTI	0.050	0.131	0.373	0.709	0.999
Power W_n^I	0.047	0.121	0.351	0.683	0.997

Testing for homogeneity in HMMs

Here, we investigate the finite sample behavior of the test for homogeneity, i.e. $m = 1$ against $m \geq 2$. As discussed in Section 3.2.2 we are mainly interest in the behavior under an non-i.i.d. alternative. We consider the alternative a two-state HMM with sdfs from the Poisson family. We examine the empirical levels under the hypothesis ($\theta_1 = \theta_2 = 3$) and under two-state HMM with Poisson intensities $\theta_1 = 3, \theta_2 = 5$ and five different transition matrices T1 - T5 (Table 3.2). We set $C = 1$.

Table 3.3 shows that the different transition matrices T1-T5 under the alternative do not seem to have much influence of the power of the test. Especially for T1 - T3, where the stationary distributions are the same the results every similar to each other. For T4 and T5, where the weight π_1 for one component is small the power of the test decreases, which appears natural and coincides with simulation results by Chen et al. (2001) for the i.i.d. setup.

Table 3.2: Transitions for two-state HMMs.

	α_{12}	α_{21}	π_1
T1	0.50	0.50	0.50
T2	0.25	0.25	0.50
T3	0.75	0.75	0.50
T4	0.80	0.40	0.33
T5	0.90	0.30	0.25

Table 3.3: Simulated rejection rates of the modified LRT for testing homogeneity under the hypothesis and under alternative models (two-state HMMs with transition probabilities T1 - T5 given in Table 3.2) for sample size $n = 100$ and $n = 200$ with $N = 10000$ replications.

	Hyp.	Alt. (HMM), n = 100				
Level	H	T1	T2	T3	T4	T5
0.025	0.029	0.398	0.380	0.381	0.301	0.228
0.05	0.052	0.512	0.484	0.491	0.403	0.323
0.1	0.097	0.632	0.610	0.622	0.522	0.446

	Hyp.	Alt. (HMM), n = 200				
Level	H	T1	T2	T3	T4	T5
0.025	0.027	0.634	0.627	0.641	0.516	0.391
0.05	0.053	0.733	0.726	0.740	0.626	0.502
0.1	0.100	0.830	0.826	0.832	0.744	0.628

Testing for two states

In the following we investigate the finite-sample performance of the MLRT for $m = 2$ against $m \geq 3$ states as suggested in Section 3.2.3. We both consider the standard example of HMMs with Poisson sdfs, as well as with zero-mean Gaussian sdfs which are used to model financial times series (cf. Rydén et al., 1998; Robert et al., 2000).

First, we examine the empirical levels under the hypothesis and consider two-state HMMs with Gaussian sdfs (N1, N2) and Poisson sdfs (P1) and five different transition matrices T1 - T5 (see Table 3.2). The specific parameter combinations of N1, N2 and P1 are displayed in Table 3.4.

To perform the test we need to specify the number of states m for the evaluation of

Table 3.4: Parameter values of the Gaussian sdfs under the hypothesis (N1, N2) and the alternative (A1, A2) as well as parameter values of the Poisson sdfs under the hypothesis (P1) and the alternative (A3, A4).

Gaussian ($\mu = 0$)			
	σ_1^2	σ_2^2	σ_3^2
N1	1	2.5	
N2	1	4	
A1	1	2.5	4
A2	1	3	6

Poisson			
	θ_1	θ_2	θ_3
P1	3	12	
A3	3	8	1
A4	3	12	7

$L_n^{I(m)}(\hat{G}^{(m)})$ and the constants C_2, C_m. Under the hypothesis we choose the minimal $m = m^*$, i.e. $m = 4$ for T1-T4 and $m = 6$ for T5. Under the alternative we always take $m = 4$. We set $C_2 = C_4 = C_6 = 1$ and choose the starting values as suggested by Chen et al. (2004).

Tables 3.5 - 3.6 show the simulated rejection rates for sample sizes $n = 200$ and $n = 1000$ for different levels. Note that models N1 and N2 are often used for financial time series analysis where large data sets are available Rydén et al. (1998).

In general, the simulated rejection rates correspond to the specified levels under the hypothesis in a satisfactory manner. Only for sample size $n = 200$ and for N1 and N2, the test is somewhat conservative. The simulations also show better results for N2, where the components differ clearly, than for N1. Note that as expected from the asymptotic theory, the different transition matrices T1-T5 do not seem to have much influence on the results. Indeed, the finite sample behavior for different transition matrices with equal stationary distribution hardly differs, at least as long as transitions are not made too rarely or too frequently (i.e. the diagonal entries are not too close to 0 or 1).

Second, we examine the power of the tests under alternative models. We consider three-state HMMs with Gaussian sdfs (A1, A2) and Poisson sdfs (A3, A4) and four different transition matrices T6 - T9, where T6, T7 and T8 are matrices of the form (2.19) and T9 the corresponding i.i.d. mixture model. The specific parameter combinations of A1 -A4 are displayed in Table 3.4 and transition probabilities for T6 - T9 are given in Table 3.7. The results for sample sizes $n = 200$ and $n = 500$ and additionally for $n = 1000$ for A1 and A2 are displayed in Tables 3.8 - 3.9.

Generally speaking, the simulations show that one should expect only a slight loss of power when introducing dependence. In fact, the influence of the different transition matrices

Table 3.5: Simulated rejection rates of the modified LRT for the models under the hypothesis N1, N2 and P1 in Table 3.4 with transition probabilities T1 - T5 given in Table 3.2 for sample size $n = 200$ with $N = 10000$ replications.

N1 (Gaussian), n = 200					
Level	T1	T2	T3	T4	T5
0.025	0.010	0.009	0.012	0.008	0.010
0.05	0.021	0.021	0.022	0.016	0.018
0.1	0.045	0.045	0.044	0.034	0.037

N2 (Gaussian), n = 200					
Level	T1	T2	T3	T4	T5
0.025	0.020	0.022	0.019	0.013	0.013
0.05	0.039	0.040	0.039	0.028	0.029
0.1	0.073	0.074	0.072	0.061	0.058

P1 (Poisson), n = 200					
Level	T1	T2	T3	T4	T5
0.025	0.032	0.032	0.030	0.031	0.032
0.05	0.056	0.056	0.056	0.054	0.060
0.1	0.101	0.101	0.098	0.098	0.109

Table 3.6: Simulated rejection rates of the modified LRT for the models under the hypothesis N1, N2 and P1 in Table 3.4 with transition probabilities T1 - T5 given in Table 3.2 for sample size $n = 1000$ with $N = 10000$ replications.

N1 (Gaussian), n = 1000					
Level	T1	T2	T3	T4	T5
0.025	0.022	0.023	0.020	0.016	0.018
0.05	0.044	0.044	0.039	0.035	0.034
0.1	0.082	0.080	0.076	0.067	0.070

N2 (Gaussian), n = 1000					
	T1	T2	T3	T4	T5
0.025	0.033	0.031	0.030	0.027	0.027
0.05	0.063	0.057	0.055	0.053	0.054
0.1	0.110	0.104	0.103	0.103	0.107

P1 (Poisson), n = 1000					
Level	T1	T2	T3	T4	T5
0.025	0.030	0.030	0.034	0.034	0.036
0.05	0.055	0.055	0.059	0.061	0.063
0.1	0.096	0.102	0.104	0.111	0.116

Table 3.7: Transitions probabilities for models under the alternative. The transition matrix is of the form (2.19).

	α	β	γ	δ	π_1	π_2	π_3
T6	0.60	0.60	0.35	0.70	0.40	0.40	0.20
T7	0.10	0.10	0.20	0.40	0.40	0.40	0.20
T8	0.05	0.05	0.05	0.10	0.40	0.40	0.20
T9			i.i.d.		0.40	0.40	0.20

Table 3.8: Simulated rejection rates of the modified LRT for the models under the alternative A1 - A4 in Table 3.4 with transition probabilities T6 - T9 given in Table 3.7 for sample size $n = 200$ and $n = 500$ with $N = 10000$ replications.

A1 (Gaussian), n = 200				
Level	T6	T7	T8	T9
0.025	0.048	0.049	0.040	0.046
0.05	0.090	0.086	0.072	0.083
0.1	0.157	0.153	0.129	0.155

A1 (Gaussian), n = 500				
Level	T6	T7	T8	T9
0.025	0.146	0.145	0.137	0.149
0.05	0.227	0.219	0.213	0.227
0.1	0.343	0.333	0.315	0.342

A2 (Gaussian), n = 200				
Level	T6	T7	T8	T9
0.025	0.214	0.192	0.152	0.216
0.05	0.313	0.288	0.234	0.313
0.1	0.441	0.413	0.351	0.443

A2 (Gaussian), n = 500				
Level	T6	T7	T8	T9
0.025	0.570	0.552	0.540	0.582
0.05	0.684	0.664	0.649	0.695
0.1	0.793	0.772	0.759	0.799

A3 (Poisson), n = 200				
Level	T6	T7	T8	T9
0.025	0.327	0.293	0.239	0.326
0.05	0.437	0.399	0.334	0.446
0.1	0.567	0.529	0.451	0.573

A3 (Poisson), n = 500				
Level	T6	T7	T8	T9
0.025	0.700	0.680	0.622	0.714
0.05	0.791	0.774	0.713	0.807
0.1	0.878	0.857	0.804	0.881

A4 (Poisson), n = 200				
Level	T6	T7	T8	T9
0.025	0.247	0.233	0.216	0.231
0.05	0.348	0.329	0.305	0.330
0.1	0.476	0.463	0.427	0.466

A4 (Poisson), n = 500				
Level	T6	T7	T8	T9
0.025	0.532	0.529	0.506	0.531
0.05	0.646	0.640	0.609	0.646
0.1	0.761	0.754	0.722	0.769

Table 3.9: Simulated rejection rates of the modified LRT for the models under the alternative A1 - A4 in Table 3.4 with transition probabilities T6 - T9 given in Table 3.7 for sample size $n = 1000$ with $N = 10000$ replications.

A1 (Gaussian), n = 1000				
Level	T6	T7	T8	T9
0.025	0.313	0.310	0.304	0.307
0.05	0.422	0.419	0.410	0.417
0.1	0.556	0.550	0.546	0.551

A2 (Gaussian), n = 1000				
Level	T6	T7	T8	T9
0.025	0.882	0.884	0.868	0.884
0.05	0.929	0.930	0.916	0.933
0.1	0.965	0.965	0.957	0.966

on the resulting power is small. Only, for models where transitions are sparse as for the models with transition matrix T8 one observes a slight loss of power, as might be expected. Furthermore, one observes that the test is more powerful against A2 than against A1. Similarly, for the Poisson case there is a higher power against A4 than against A3. Note that Poisson-mixtures were also investigated in the simulations by Chen et al. (2004), our results are rather close to those obtained in that paper.

3.2.5 Empirical illustrations

Fetal lamb movements

As a first illustration, let us revisit the fetal movement data set which is displayed and analyzed in Leroux and Puterman (1992) and reanalyzed by Chen et al. (2004). Leroux and Puterman (1992) fit both two- and three component independent Poisson mixtures as well as two- and three-state Poisson HMMs. They find for these data that while independent mixtures are only marginally better than a negative binomial model, the fits provided by the HMMs are much superior and should be used. In fact, there is strong evidence for autocorrelation in these data (cf. Figure 3.2). For a two-state Poisson HMM, ordinary maximum likelihood yields the following estimates: $\hat{\alpha}_{12} = 0.011$, $\hat{\alpha}_{21} = 0.310$, $\hat{\theta}_1 = 0.256$ and $\hat{\theta}_2 = 3.115$. Assuming $m = 2$, the ordinary likelihood ratio test rejects the hypothesis

of independence, i.e. $H: \alpha_{12} = 1 - \alpha_{21}$ with a p-value nearly zero. The comparison of the autocorrelation functions of the sample and the two-state Poisson HMM with parameters $\left(\hat{\alpha}_{12}, \hat{\alpha}_{21}, \hat{\theta}_1, \hat{\theta}_2\right)$ displayed in Figure 3.2 indicates that a two-state Poisson HMM is an appropriate model for the given data.

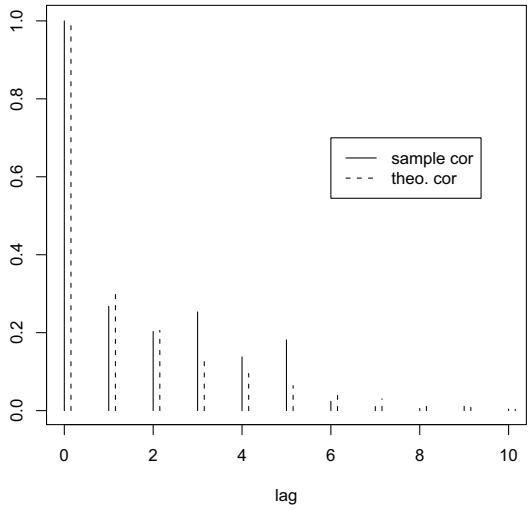


Figure 3.2: Autocorrelation function for the series of fetal lamb movements.
(Source: The data set is published in Leroux and Puterman (1992, p.547, Table 1).)

However, using formal model selection criteria one cannot decide between the two-state HMM (selected by BIC) and the three-state HMM (selected by the AIC). Using the modified LRT for two components in independent mixtures, Chen et al. (2004) test the hypothesis of two components which, yielding a p-value of 0.085, cannot be rejected. From Theorem 1, it follows that their analysis remains valid for the marginal mixture distribution even if the model of choice is an HMM.

Series of log-returns of the S&P 500

Rydén et al. (1998) use HMMs with zero-mean Gaussian state-dependent distributions to analyze the series of log-returns of daily values of the S&P 500 index (formerly called S&P 90). Specifically, they consider the series of log-returns of ten subseries of length 1700 of the S&P 90/500 from 3 January 1928 to 30 April 1991. We shall examine the same ten subseries $A, \ldots, J$, with outlier replacement and centering of each subseries being conducted as in their paper.

In order to determine the number of states of the HMM, Rydén et al. (1998) use an M- out-of N ($M = 800$) bootstrap for the full-model LRT for two against three components. This procedure requires repeated maximization of the full log-likelihood function of an HMM with three states. We found this procedure extremely computationally expensive, since proper maximization also require the choice of several starting value combinations, and we were not able to investigate the properties in an adequate simulation. In fact, in their analysis Rydén et al. (1998) only used very small bootstrap samples for the distribution of the LRT of size 50, and rejected the hypothesis if the LRT statistic from the first M observations of the sample exceeded 48 (or more) values of the bootstrap distribution. Also, the choice of M in the M- out-of N bootstrap is a somewhat subjective manner, and may (at least in practice) significantly influence the results.
Therefore, we apply the modified LRT for two against more states to this data set, where we use $m = m^*$ and set $C_m = 1$ for all m. As illustration, we present the estimates of the fitted models $\hat{G}^{(2)}$ and $\hat{G}^{(m)}$ for the subseries $H\,(m = 4)$, $I\,(m = 4)$ and $J\,(m = 5)$ in Table 3.10.

Table 3.10: Estimates $\hat{G}^{(2)}$ and $\hat{G}^{(m)}$ for the subseries H, I and J of the series of log-returns of the S&P 500 index, each of length 1700.

	$\hat{\pi}_1$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
H	0.679	0.0064	0.0125
I	0.562	0.0062	0.0115
J	0.704	0.0063	0.0154

	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\pi}_5$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$	$\hat{\sigma}_5$
H	0.181	0.308	0.308	0.203		0.0043	0.0077	0.0077	0.0136	
I	0.232	0.287	0.287	0.193		0.0049	0.0084	0.0084	0.0131	
J	0.173	0.242	0.252	0.252	0.081	0.0032	0.0063	0.0101	0.0101	0.0210

One observes that for H and I the mixture distribution $\hat{G}^{(m)}$ exhibits only three different components, while for series J four distinct components are present. The values of the LRT statistic, the estimated $\hat{p}$ in the limit distribution in (3.19) as well as the p-values of the test are displayed in Table 3.11.
While for the series J, the hypothesis of two states can be rejected at a level of $\alpha < 0.001$, for series H and I rejection is only possible at a nominal level of 0.1. Note that from the

Table 3.11: Test results of the hypothesis $m = 2$ for the subseries H, I and J of the series of log-returns of the S&P 500 index, each of length 1700.

	LRT	$\hat{p}$	p-value
H	2.68	0.09	0.074
I	2.16	0.08	0.099
J	21.72	0.12	0.000

simulations in Section 3.2.4 we may expect that the test is somewhat conservative in such settings, so that a test decision on a nominal level of 0.1 appears to be reasonable.

3.3 Testing for the number of components in a switching regression model

The class of switching regression models (SRMs) extends the class of finite mixture models in another direction than HMMs, namely it enables the integration of covariates. Here, the specification of the number of components is also an important issue. As described in Sec. 1.3 we assume that the joint density of (Y_i, X_i) is of the form

$$f_{\text{switch}}(y_i, x_i; \beta, G) = (\pi_1(G) f(y_i, x_i; \beta, \theta_1) + \ldots + \pi_m(G) f(y_i, x_i; \beta, \theta_m))\, h(x_i), \quad (3.20)$$

where β denotes the structural parameter, which is equal in all components, and G denotes an m-point distribution on Θ. SRMs are applied in various settings and for different families $\{f(y, x; \beta, \theta) | \beta \in \mathcal{B}, \theta \in \Theta\}$, of which we present some commonly used examples from the context of generalized linear models.

Example 3.1 (*switching logistic regression*)**.** Let U_i be independent copies of a latent variable with values in Θ following an m-point distribution G on Θ. If $(Y_i, X_i)_i$ are conditionally independent given U_i, satisfy $P(Y_i \in \{0, 1\}) = 1$ and

$$\text{logit}\, P\big(Y_i = 1 | X_i = x_i, U_i = \theta_k\big) = x_i^T \beta + w_i^T \theta_k,$$

where w_i is an l-dimensional vector of covariates, then the model is called an m-component switching logistic regression model. The density of (Y_i, X_i) is then given by

$$\begin{aligned} & f_{\text{switch}}(y_i, x_i; \beta, G) \\ &= \sum_{k=1}^{m} \pi_k(G) \left(\text{logit}^{-1}(x_i^T \beta + w_i^T \theta_k)\right)^{y_i} \left(1 - \text{logit}^{-1}(x_i^T \beta + w_i^T \theta_k)\right)^{(1-y_i)} h(x_i). \end{aligned}$$

This model can be extended in a straightforward fashion to switching binomial regression models, in which case we denote the number of successes by **n** (cf. Sec. 3.3.3).

Example 3.2 (*switching Poisson regression*)**.** Again let U_i be independent copies of a latent variable with values in Θ following the distribution G. If $(Y_i, X_i)_i$ are conditionally independent given $U_i$and satisfy for $y_i \in \mathbb{N}_0$

$$P\big(Y_i = y_i | X_i = x_i, U_i = \theta_k\big) = \frac{1}{y_i!}\,\lambda_{i;k}^{y_i}\,\exp(-\lambda_{i;k}),$$

where $\lambda_{i;k} = \exp\big(x_i^T\beta + w_i^T\theta_k\big)$, then the model is called an m-component switching Poisson regression model. For (Y_i, X_i) we then have that

$$f_{\text{switch}}(y_i, x_i, \beta, G) \;=\; \sum_{k=1}^{m} \pi_k(G)\,\frac{1}{y_i!}\,\lambda_{i;k}^{y_i}\,\exp(-\lambda_{i;k})\,h(x_i).$$

Example 3.3 (*linear switching regression*)**.** Here,

$$Y_i = x_i^T\beta + w_i^T U_i + \varepsilon_i,$$

where the ε_i are independently distributed with $E\,\varepsilon_i = 0$ and $Var\,\varepsilon_i = \sigma^2 < \infty$., e.g. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

In all three examples choosing the number of components m is of primary interest in every application. The testing problem of homogeneity is investigated by Zhu and Zhang (2004) by means of the LRT as well as the MLRT. Again our focus is also on situations where heterogeneity is evident and it is of interest to test whether the latent distribution only has two states or more than two states.
The remainder of the section is organized in the way that we first present the results by Zhu and Zhang (2004) for testing homogeneity, we then derive the MLRT for testing for two components in SRMs. As Zhu and Zhang (2004) we also discuss extensions of the described model to a longitudinal setup and to SRMs with not i.i.d., but Markov-dependent regime (MSRMs), which are closely related to HMMs. In a simulation study we investigate the finite-sample behavior of the test for two components. Finally, we apply the methodology to data of a dental health trial. Here, the model selection criteria AIC and BIC favor distinct binomial regression models with switching intercept (AIC three components, BIC two components). The MLRT allows us to reject the hypothesis of two components in favor of three components.

3.3.1 Testing for homogeneity in a switching regression model

Zhu and Zhang (2004) investigate the LRT for testing for homogeneity in a switching regression model, i.e.

$$H : \pi_1(1-\pi_1)(\theta_1-\theta_2) = 0, \beta \in \mathcal{B} \quad \text{against} \quad K : \pi_1 \in [0,1], \beta \in \mathcal{B}, \theta_1, \theta_2 \in \Theta. \quad (3.21)$$

Similarly to the result by Chen and Chen (2001) presented in Sec. 3.1.1 they show that the LRT converges in distribution to the maxima of a stochastic process (cf. Zhu and Zhang, 2004, Thm. 1).

Zhu and Zhang (2004) also consider the MLRT for testing (3.21) following Chen et al. (2001). As Zhu and Zhang (2004) do not restrict themselves as Chen et al. (2001) to the case of a one-dimensional switching parameter, i.e. $\Theta \subset \mathbb{R}$, their result is more general, but the derived asymptotic distribution is in general not of a simple form as in Theorem 3.2. Rather then presenting Zhu and Zhang's result in full generality we stick to the assumption $\Theta \subset \mathbb{R}$. In particular, this covers the important class of SRMs with switching intercept.

The modified log-likelihood function for an m-component SRM is given by

$$\tilde{L}_n^{(m)}(\beta, G) = L_n^{(m)}(\beta, G) - pen(G) \quad (3.22)$$

with the ordinary log-likelihood function for an m-component SRM

$$L_n^{(m)}(\beta, G) = \sum_i \log f_{\text{switch}}(y_i, x_i; \beta, G)$$

for $G \in \mathfrak{M}_m$ (the m-point distributions on Θ) and a penalty $pen(G)$. Zhu and Zhang (2004) consider as Chen et al. (2001) the penalty

$$pen(G) = -C \log 4\pi_1(G)(1-\pi_1(G))$$

for $G \in \mathfrak{M}_2$. Similarly the MLRT for the testing problem (3.21) is given by

$$T_n^{\text{mod}} = 2\big(\sup_{\beta \in \mathcal{B}, G \in \mathfrak{M}_2} \tilde{L}_n^{(2)}(\beta, G) - \sup_{\beta \in \mathcal{B}, \theta \in \Theta} L_n^{(1)}(\beta, \theta) \big).$$

Zhu and Zhang (2004) show that Chen's result for the MLRT in finite mixtures (see Theorem 3.2) transfers to SRMs:

Theorem 3.7. *Suppose that the Assumptions A.1-A.3 in Zhu and Zhang (2004) are satisfied. Further assume that $f(y, x; \beta_0, \theta_0)h(x)$ is the density of the true regression model $(Y_i, X_i)_i$. Then*

$$T_n^{mod} \xrightarrow{\mathcal{L}} \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2. \quad (3.23)$$

For the proof see Zhu and Zhang (2003, Proof of Thm. 2). Note that, Assumptions A.1-A.3 in Zhu and Zhang (2004) apply to a very general, possibly longitudinal setting. For our purposes an alternative set of conditions might be formulated by adopting Assumptions 3.1 - 3.5 to the switching regression context.

3.3.2 Testing for two components in a switching regression model

We consider the testing problem for two components in a switching regression model with independent regime as well as with Markov-dependent regime.

Testing for two components in an SRM with independent regime

Suppose that for different i, the observations (Y_i, X_i) are independent. We propose a test for two components in an SRM with one-dimensional switch, i.e.

$$H: \; G_0 \in \mathfrak{M}_2 \qquad \text{against} \quad K: \; G_0 \in \mathfrak{M} \setminus \mathfrak{M}_2, \tag{3.24}$$

in an SRM with G_0 denoting the true two component distribution of the switching parameter on $\Theta \subset \mathbb{R}$, i.e.

$$G_0(\theta) = \pi_0 \, I_{\{\theta_{01} \leq \theta\}} + (1 - \pi_0) I_{\{\theta_{02} \leq \theta\}}.$$

Following Chen et al. (2004) the MLRT is based on the modified log-likelihood function $\tilde{L}_n^{(m)}(\beta, G)$ as displayed in (3.22) with penalty function

$$pen(G) = -C_m \sum_{k=1}^{m} \log\big(\pi_k(G)\big),$$

for $G \in \mathfrak{M}_m$ on Θ with a positive constant $C_m > 0$. We denote the MMLE, i.e. the maximizer of $\tilde{L}_n^{(m)}(\,\cdot\,)$ by $(\hat{\beta}^{(m)}, \hat{G}^{(m)})$, or more explicitly

$$(\hat{\beta}^{(m)}, \hat{\pi}_1^{(m)}, \ldots, \hat{\pi}_m^{(m)}, \hat{\theta}_1^{(m)}, \ldots, \hat{\theta}_m^{(m)}).$$

The MLRT statistic for the hypothesis (3.24) in SRMs is then formed by

$$T_n^{\text{mod}} = 2\big(L_n^{(m)}(\hat{\beta}^{(m)}, \hat{G}^{(m)}) - L_n^{(2)}(\hat{\beta}^{(2)}, \hat{G}^{(2)})\big), \tag{3.25}$$

Regularity conditions

Following Chen et al. (2004) we require the following regularity conditions. Essentially, the assumptions reformulate Assumptions 3.2'-3.5' in terms the two-component switching model $f_{\text{switch}}\big(y, x; \beta_0, G_0\big)$ rather than $f_{\text{mix}}\big(y; G_0\big)$. Note, that only θ is assumed to be

one-dimensional, while $\beta \in \mathbb{R}^q$ such that the derivatives w.r.t. β should be understood in terms of the gradient etc. As in Sec. 3.1.2 we define similar quantities as for corresponding testing problem in finite mixtures:

$$\begin{aligned} Z^1_{i0}(\beta,\theta) &= \frac{f(Y_i,X_i;\beta,\theta)-f(Y_i,X_i;\beta_0,\theta)}{f_{\text{switch}}\big(Y_i,X_i;\beta_0,G_0\big)}, \\ Z^1_{ik}(\beta,\theta) &= \frac{f(Y_i,X_i;\beta,\theta)-f(Y_i,X_i;\beta,\theta_{0k})}{f_{\text{switch}}\big(Y_i,X_i;\beta_0,G_0\big)}, \quad \text{for } k=1,2 \\ Z^{l_1l_2}_i(\beta,\theta) &= \frac{\frac{d^{l_1+l_2}}{d\theta^{l_1}d\beta^{l_2}}f(Y_i,X_i;\beta,\theta)}{f_{\text{switch}}\big(Y_i,X_i;\beta_0,G_0\big)}, \quad \text{for } l_1=0,1,2,3, l_2=0,1,2, l_1+l_2\neq 0. \end{aligned}$$

where G_0 denotes again the true two component switching distribution. As above we define $Z^1_{ik}(\beta,G)=\int Z^1_{ik}(\beta,\theta)dG(\theta)$ for $G\in\mathfrak{M}_m$.

Assumption 3.1" (Wald-type integrability condition) Let

$$E_0\left[|\log f_{\text{switch}}(Y_i,X_i;\beta,G_0)|\right]<\infty$$

and there exists $\varepsilon>0$ such that $f_{\text{switch}}(y,x;Q,\varepsilon) := 1+\sup_{\|Q'-Q\|\leq\varepsilon} f_{\text{switch}}(y,x;Q')$ is measurable and $E_0\left[\log f_{\text{switch}}(Y_i,X_i;Q,\varepsilon)\right]<\infty$ for each $Q:=(\beta,G)$.

Assumption 3.2" (Smoothness) The support of each function $f(y,x;\beta,\theta)$ does not depend on (β,θ), and the derivatives

$$\frac{d^{l_1+l_2}}{d\theta^{l_1}d\beta^{l_2}}f(y,x;\beta,\theta)$$

with $l_1=0,1,2,3$, $l_2=0,1,2$, $l_1+l_2\neq 0$ exist and are jointly continuous in (y,x) and (β,θ).

Assumption 3.3" (Strong identifiability) The family $\{f(y,x;\beta,\theta)\}$ is strong identifiable, in the sense that for $\theta_1\neq\theta_2$

$$\sum_{k=1}^{2}\Big(a_kf(y,x;\beta,\theta_k)+\sum_{l=1}^{q}b_{kl}\frac{d}{d\beta_l}f(y,x;\beta,\theta_k)+c_k\frac{d}{d\theta}f(y,x;\beta,\theta_k)+d_k\frac{d^2}{d\theta^2}f(y,x;\beta,\theta_k)\Big)=0$$

for all (y,x) implies $a_k=b_{k1}=\ldots=b_{kq}=c_k=d_k=0$ for $k=1,2$.

This assumption implies that the asymptotic covariance matrix $\lim_{n\to\infty}1/n\sum_{i=1}^n b_ib_i^T$ with b_i defined below is positive definite.

Assumption 3.4" (Uniform boundedness) There exists an integrable function g and $\delta>0$ such that $|Z^1_{ik}(\beta,\theta)|^{4+\delta}\leq g(Y_i)$ and $\left|Z^{l_1l_2}_i(\beta,\theta)\right|^3\leq g(Y_i)$ for all (β,θ), for $k=0,1,2$; $l_1=0,1,2,3$; $l_2=0,1,2$; $l_1+l_2\neq 0$.

Assumption 3.5" (Tightness) The processes $n^{-1/2}\sum_i Z^1_{ik}(\beta,\theta)$, $n^{-1/2}\sum_i Z^{l_1 l_2}_i(\beta,\theta)$ are tight for $k=0,1,2; l_1=0,1,2,3; l_2=0,1,2; l_1+l_2\neq 0$.

Asymptotic analysis

In the following we derive the asymptotic distribution of the MLRT under the hypothesis. We follow the approach by Chen et al. (2004) presented in Sec. 3.1.2 and start with the standard decomposition

$$T^{\text{mod}}_n = T^{\text{mod}}_{1n} - T^{\text{mod}}_{0n} = 2\big(L^{(m)}_n(\hat\beta^{(m)},\hat G^{(m)}) - L^{(2)}_n(\beta_0,G_0)\big) - 2\big(L^{(2)}_n(\hat\beta^{(2)},\hat G^{(2)}) - L^{(2)}_n(\beta_0,G_0)\big).$$

Beginning with T^{mod}_{1n} we observe $T^{\text{mod}}_{1n} = 2\sum_i \log(1+\delta_i)$ with

$$\delta_i = \big(f_{\text{switch}}\big(Y_i,X_i;\hat\beta^{(m)},\hat G^{(m)}\big) - f_{\text{switch}}\big(Y_i,X_i;\beta_0,G_0\big)\big)/f_{\text{switch}}\big(Y_i,X_i;\beta_0,G_0\big).$$

For the further analysis of T^{mod}_{1n} we omit the index m, i.e. $\hat\beta := \hat\beta^{(m)}$ and $\hat G := \hat G^{(m)}$. Following Chen et al. (2004) we define $\hat\pi = \hat G(\theta_0)$ for $\theta_0 := (\theta_{01}+\theta_{02})/2$ and $\hat G_k$ for $k=1,2$ such that $\hat G(\theta) = \hat\pi\hat G_1(\theta) + (1-\hat\pi)\hat G_2(\theta)$ (cf. Sec. 3.1.2) and observe

$$\delta_i \;=\; (\hat\pi-\pi_0)\Delta_i(\hat\beta) + Z^1_{i0}(\hat\beta,G_0) + \hat\pi Z^1_{i1}(\hat\beta,\hat G_1) + (1-\hat\pi)Z^1_{i2}(\hat\beta,\hat G_2) \tag{3.26}$$

with $\Delta_i(\hat\beta) = Z^1_{i2}(\beta,\theta) - Z^1_{i1}(\beta,\theta)$. Expanding the functions $\Delta_i, Z^1_{i0}, Z^1_{i1}, Z^1_{i2}$ as in (3.5) and (3.6) yield

$$\begin{aligned}\delta_i \;=\;& (\hat\pi-\pi_0)\Delta_i(\beta_0) + (\hat\beta-\beta_0)Z^{01}_i(\beta_0,G_0) + \hat\pi\hat m_{11}Z^{10}_i(\beta_0,\theta_{01}) + (1-\hat\pi)\hat m_{12}Z^{10}_i(\beta_0,\theta_{02})\\ &+\hat\pi\hat m_{21}/2\; Z^{20}_i(\beta_0,\theta_{01}) + (1-\hat\pi)\hat m_{22}/2\; Z^{20}_i(\beta_0,\theta_{02}) + o_P(1).\end{aligned}$$

We may now define similar quantities as in (3.7) by

$$b_i = \big(\Delta_i(\beta_0), Z^{01}_i(\beta_0,G_0), Z^{10}_i(\beta_0,\theta_{01}), Z^{10}_i(\beta_0,\theta_{02}), Z^{20}_i(\beta_0,\theta_{01}), Z^{20}_i(\beta_0,\theta_{02})\big)^T \tag{3.27}$$

and

$$t(\beta,G) = \big(\pi(G)-\pi_0, \beta-\beta_0\pi_0 m_{11}(G_1), (1-\pi_0)m_{12}(G_2), \pi_0 m_{21}(G_1)/2, (1-\pi_0)m_{22}(G_2)/2\big)^T$$

with $m_{lk}(G) = \int(\theta-\theta_{0k})^l dG(\theta)$ and $\mathbf{b} = \sum_i b_i \in \mathbb{R}^{q+5}$, $\mathbf{B} = \sum_i b_i b_i^T \in \mathbb{R}^{(q+5)\times(q+5)}$.
This yields

$$\sum_i \delta_i \;=\; t(\hat\beta,\hat G)\,\mathbf{b} + o_P(1) \quad \text{and} \quad \sum_i \delta_i^2 = t(\hat\beta,\hat G)^T\,\mathbf{B}\,t(\hat\beta,\hat G) + o_P(1)$$

where $\mathbf{b}, \mathbf{B}$ are defined in terms of the SRM as follows: Since $\log(1+\delta_i) \leq \delta_i - \delta_i^2/2 + \delta_i^3/3$, this yields

$$T_{1n}^{\text{mod}} \leq \sup_{\beta \in \mathcal{B}, G \in \mathfrak{M}_m} \left(2\, t(\beta, G)^T \mathbf{b} - t(\beta, G)^T\, \mathbf{B}\, t(\beta, G) \right) + o_P(1). \tag{3.28}$$

Similarly as in (3.9) there exists (β^*, G^*) such that for $t^* = t(\beta^*, G^*)$ the supremum on the right hand side of (3.28) is attained. This implies

$$\begin{aligned} T_{1n}^{\text{mod}} &\geq \left(L_n^{(m)}(\beta^*, G^*) - L_n^{(2)}(\beta_0, G_0) \right) \\ &= 2\, t^{*T} \mathbf{b} - t^{*T}\, \mathbf{B}\, t^* + o_P(1) = \sup_{\beta \in \mathcal{B}, G \in \mathfrak{M}_m} \left(2\, t(\beta, G)^T \mathbf{b} - t(\beta, G)^T\, \mathbf{B}\, t(\beta, G) \right) + o_P(1) \end{aligned} \tag{3.29}$$

and hence

$$T_{1n}^{\text{mod}} = \sup_{\beta \in \mathcal{B}, G \in \mathfrak{M}_m} \left(2\, t(\beta, G)^T \mathbf{b} - t(\beta, G)^T\, \mathbf{B}\, t(\beta, G) \right) + o_P(1) \tag{3.30}$$

For T_{0n}^{mod} following Chen et al. (2004) shows

$$T_{0n}^{\text{mod}} \;=\; \mathbf{b}_1^T {B_{11}}^{-1} \mathbf{b}_1^T + o_P(1)$$

for $\mathbf{b}^T = (\mathbf{b}_1^T, \mathbf{b}_2^T)$ with $\mathbf{b}_2 \in \mathbb{R}^2$ and

$$\mathbf{B} = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array} \right), \qquad B_{22} \in \mathbb{R}^{2\times 2}.$$

Finally, we arrive via orthogonalization of the right hand side of (3.30) at

$$T_n^{\text{mod}} \;=\; \sup_{t_2 \in [0,\infty)\times[0,\infty)} \left(2\, t_2^T \tilde{\mathbf{b}}_2 - t_2^T\, \tilde{\mathbf{B}}_{22}\, t_2 \right) + o_P(1) \tag{3.31}$$

for $\tilde{\mathbf{b}}_2 = \mathbf{b}_2 - B_{21}\,(B_{11})^{-1}\,\mathbf{b}_1$ and $\tilde{\mathbf{B}}_{22} = B_{22} - B_{21}\,(B_{11})^{-1}\,B_{12}$, such that we can formulate the following theorem.

Theorem 3.8. *Suppose that Assumptions 3.1" - 3.5" hold and that m in the definition of T_n^{mod} in (3.25) satisfies $m \geq m^* := \max\left([1.5/\pi_0], [1.5/(1-\pi_0)], 4\right)$. Then under H, the modified likelihood ratio test statistic T_n^{mod} converges in distribution to a mixture of χ^2-distributions,*

$$T_n^{mod} \xrightarrow{\mathcal{L}} (\frac{1}{2} - p)\, \chi_0^2 + \frac{1}{2}\chi_1^2 + p\, \chi_2^2, \tag{3.32}$$

where $p = \left(\cos^{-1}\rho\right)/(2\pi)$ and ρ is the correlation coefficient in the covariance matrix $\tilde{\mathbf{B}}_{22}$.

The proof follows Chen et al. (2004) and is given in more detail manner in Sec. 3.4.

Remark 3.9. Zhu and Zhang (2004) derive Theorem 3.8 not only for homogeneous SRMs, i.e. $f_{i,\text{switch}}(y_i, x_i; \beta, \theta) = f_{\text{switch}}(y_i, x_i; \beta, \theta)$ for all i. They also examine the longitudinal setup, where one considers n_i observations $(Y_{i,j}, X_{i,j})$, $j = 1, \ldots, n_i$ per unit "i". If one assumes that the observations within one unit are also independent conditioned on U_i, one may interpret this settings as blocks of observations of which one has the additional information that switching is only apparent between but not within blocks.
Zhu and Zhang (2004) show that their result remains valid under general conditions on the families $\{f_{i,\text{switch}}(y_i, x_i; \beta, \theta)\}$. The same is true for Theorem 3.8. However, for the extensions of the Theorem 3.8 to the longitudinal setup, one must carefully redefine the modified log-likelihood of the model and the quantities Z_{ik}^1, $Z_i^{l_1 l_2}$ and modify the Assumptions 3.1" - 3.5" appropriately. Zhu and Zhang (2003) illustrate, for example, how consistency of the MMLE can be derived from uniform laws of large number on the log-likelihood function of the model. They also give insights concerning the tightness assumptions. For example, they show that for the logistic regression model the boundedness of n_i, $||x_i||$ as well as positive definitiveness of $\text{var}[(X_i^T, 1)]$ is relevant. For further details we defer to Dannemann and Holzmann (2010).

Testing for two components in an SRM with Markov-dependent regime

An interesting extension of the model is to allow some dependence structure in the underlying latent variables. We shall consider the case of a Markov-dependent regime, i.e. Markov-switching regression models (MSRMs) as introduced in Sec. 1.3. Following the notation from Sec. 1.5 we denote the regime $U_1, \ldots, U_n$ as a stationary, ergodic Markov chain on a finite set $\{\theta_1, \ldots, \theta_m\} \subset \Theta$ with transition matrix P and stationary distribution G. The marginal density of (Y_i, X_i) is of the form (3.20), where the weights $\pi_1(G), \ldots, \pi_m(G)$ are determined by G.
Similar as for HMMs discussed in Sec. 3.2.3, the testing problem (3.24) simply translates to testing whether the number of states m in the MSRM equals two, i.e.

$$H: \; G_0 \in \mathfrak{M}_2 \qquad \text{against} \quad K: \; G_0 \in \mathfrak{M} \setminus \mathfrak{M}_2, \tag{3.33}$$

where G_0 is the stationary distribution of the regime.The modified log-likelihood function (3.22) neglect the introduced dependence structure and should be called modified log-likelihood function under independence assumption for MSRMs. But they can still be used to estimate the parameters β, $\pi_1(G), \ldots, \pi_m(G)$ and $\theta_1, \ldots, \theta_m$, and to form the test statistic (3.25) in the present situation. However, one should expect that the asymptotic distribution (3.32) in Theorem 3.8 must be modified due to the dependence structure of

the regime. Surprisingly, the asymptotic distribution remains the same for the switching regression model with independent regime as similarly observed for HMMs in Sec. 3.2.3. To see this, we must again examine the covariance of the asymptotic normal distribution of $n^{-1/2}\tilde{\mathbf{b}}_2$, now formed by means of an SRM,

$$\widetilde{\mathrm{Cov}} = \tilde{\Sigma} + \sum_{i=2}^{\infty} E\big[\tilde{b}_{21}\tilde{b}_{2i}^T + \tilde{b}_{2i}\tilde{b}_{21}^T\big].$$

One needs to show that the asymptotic covariance of $\tilde{\mathbf{b}}_2$ remains the same as for independent switching.

Proposition 3.3. *Suppose that for a Markov-switching regression model Assumptions 3.1" - 3.5" hold true. Then, under the hypothesis H of a two-state Markov regime we have*

$$E\left[\tilde{b}_{21}\tilde{b}_{2i}^T + \tilde{b}_{2i}\tilde{b}_{21}^T\right] = 0 \quad \text{for all } i \geq 2.$$

The proof of the proposition is similar to the proof of Proposition 3.2 and is given in Sec. 3.4. Proposition 3.3 implies that the asymptotic distribution (3.32) remains true for Markov-switching regression models, where the weight p is determined as in Theorem 3.8 from the covariance matrix $\tilde{B}_{22}$. We state this as a corollary of Theorem 3.8.

Corollary 3.1. *Suppose that for a Markov-switching regression model Assumptions 3.1" - 3.5" hold and that m in the definition of the test statistic T_n^{mod} in (3.25) satisfies $m \geq m^* := \max\big([1.5/\pi_0], [1.5/(1-\pi_0)], 4\big)$. Then under the hypothesis H of a two-state Markov regime the asymptotic distribution of T_n^{mod} is as in (3.32).*

Note that for the more general longitudinal setup described in Remark 3.9 the corollary only holds true under additional assumptions on the sequence $(n_i)_i$, e.g. if these are chosen at random according to a bounded, stationary process.
One may also think about relaxing the i.i.d. assumption on the regressors. Our simulations indicate that the effect of dependent regressors, for example if X_i are univariate and follow an AR(1) process, on the asymptotic covariance matrices and hence on the asymptotic distribution of the test statistic under the hypothesis is small. So we may conclude that the described testing procedure is quite robust against violations of the independence assumptions.

3.3.3 Simulation experiments

We examine the finite sample behavior of the testing procedures for testing for two components in an SRM for both independent and Markov dependent regime. A simulation study for testing homogeneity in SRMs is available in Zhu and Zhang (2004).

Independent regime

We consider two specific switching Poisson regression models and two switching Binomial regression models. To apply the testing procedure we use $C_2 = C_m = 1$ and choose m, the number of components in the alternative, by m^*, where we replace π_0 by its estimate under the hypothesis.

The switching Poisson regression models are specified by $q = 2, l = 1$ with covariates $X_i = (1, X_{i1})^T$, X_{i1} independent r.v.s uniformly distributed on the unit interval. For the first model P1 we choose $W_i = 1$ which leads to the Poisson regression model with switching intercept:

$$P\big(Y_i = y_i | X_i = x_i, U_i = \theta_k\big) = \frac{1}{y_i!}\lambda_{i;k}^{y_i} \exp(-\lambda_{i;k}),$$

with $\lambda_{i;k} = \exp\big(x_i\beta + \theta_k\big)$. Choosing $W_i = X_i$ leads to the Poisson regression model with switching regression coefficient (P2) where intensities is given by $\lambda_{i;k} = \exp\big(x_i\theta_k + \beta\big)$.

The two Binomial regression models with switching intercept (B1, B2) are given by

$$P\big(Y_i = y_i | X_i = x_i, U_i = \theta_k\big) = \binom{\mathbf{n}}{y_i} p_{i;k}^{y_i}\big(1 - p_{i;k}\big)^{\mathbf{n}-y_i}$$

with

$$\operatorname{logit} p_{i;k} = x_i^T\beta + \theta_k.$$

For both models we specify $\mathbf{n} = 8$ and for B1 we consider the covariates as in P1, i.e., $q = 2, l = 1$, $X_i = (1, X_i)^T$, X_i independent r.v.s uniformly distributed on the unit interval and $W_i = 1$. For B2 the covariates are categorical variables as in the application we discuss below, i.e., $q = 9, l = 1$, $X_i = (1, X_{i1}, \ldots, X_{i8})^T$, $X_{ij} \in \{0, 1\}$ eight independent copies of Bernoulli r.v.s with success probability 1/2 and $W_i = 1$. The specific parameter combinations for models of P1, P2, B1, B2 under the hypothesis ($m = 2$) and under the alternative ($m = 3$) are given in Table 3.12. For B2 β is given by $(-0.8, -0.5, -0.3, -0.4, -0.2, -0.1, 0.1, 0.2)^T$.

In general, the simulated rejection rates correspond to the specified levels under the hypothesis in a satisfactory manner (see Table 3.13). For small sample size and for the model B2.H, the test is somewhat conservative. Note, that for small sample sizes the estimation of ρ as correlation coefficient might fail (if the empirical version of B_{11} is not invertible), in this case one may use $p = 0.5$ leading to a conservative test decision.

Dependent regime

Next we investigate the behavior of the proposed test when the regime is Markov-dependent. As discussed in Section 3.3.2 if the switching process is a Markov chain rather than an i.i.d.

Table 3.12: Parameter values of the switching Poisson regression models and the switching Binomial regression models under the hypothesis (P1.H, P2.H, B1.H, B2.H) and the alternative (P1.A, P2.A, B1.A, B2.A). For the model B2 the parameter β is given in the text.

	β	θ_1	θ_2	θ_3	π_1	π_2	π_3
P1.H	0.5	1	2		0.6	0.4	
P1.A	0.5	0	1	2	0.33	0.33	0.33
P2.H	1	1	2		0.6	0.4	
P2.A	1	0	1	2	0.33	0.33	0.33
B1.H	-2	-2	0		0.6	0.4	
B1.A	-2	-2	0	1	0.33	0.33	0.33
B2.H		-2	0		0.6	0.4	
B2.A		-2	0	1	0.33	0.33	0.33

process, the asymptotic behavior of the test statistic remains the same.
In addition, we also investigate the case of dependent covariates. Although we have no formal theory, our simulation indicates that also in this case there is no a significant change in the asymptotic behavior of the test statistic.
For our simulations we consider the models P1.H and B1.H (see Table 3.12) and choose the switching process as a Markov chain with transition matrix and P and stationary distribution π specified as

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix} \quad \text{and} \quad \pi = \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}.$$

In addition we construct the covariate $X_i = (1, X_i)^T$ based on an autoregressive process

$$\tilde{X}_i = \rho_0 \tilde{X}_{i-1} + \varepsilon_i$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d., with $\sigma^2 = 1$. Based on $\tilde{X}_i$ we construct a process with uniform marginals by

$$X_i = \phi^{-1}\left(\tilde{X}_i / \sqrt{\sigma^2/(1-\rho_0^2)}\right).$$

with ϕ denoting the distribution function for the standard normal distribution. The regression models with Markov-switching and independent covariates are denoted by P1.MC.H and B1.MC.H, whereas the models with Markov-switching and dependent covariates (with

Table 3.13: Simulated rejection rates of the MLRT for the models under the hypothesis P1.H, P2.H, B1.H, and B2.H in Table 3.12 for various sample sizes with $N = 5000$ replications.

	P1.H (Poisson)			
Level	$n = 50$	100	200	500
0.025	0.017	0.018	0.022	0.027
0.05	0.031	0.038	0.040	0.047
0.1	0.057	0.068	0.073	0.085

	B1.H (Binomial)			
Level	$n = 100$	200	500	1000
0.025	0.010	0.013	0.020	0.019
0.05	0.018	0.028	0.039	0.039
0.1	0.040	0.060	0.073	0.081

	P2.H (Poisson)			
Level	$n = 50$	100	200	500
0.025	0.011	0.015	0.023	0.027
0.05	0.021	0.030	0.046	0.052
0.1	0.044	0.060	0.083	0.091

	B2.H (Binomial)			
Level	$n = 100$	200	500	1000
0.025	0.011	0.014	0.017	0.016
0.05	0.017	0.025	0.034	0.033
0.1	0.031	0.048	0.062	0.066

$\rho_0 = 0.5$) are denoted by P1.MCAR.H and B1.MCAR.H. The results displayed in Table 3.15 confirm the small effect of the dependency structure on the asymptotic distribution of the test statistic.

3.3.4 Empirical illustration: Application to dental health trial

We discuss an application of the switching regression model to the dental data set analyzed by Böhning et al. (1999). In a dental health trail 797 children were exposed to different treatments for the improvement of their dental health. This was measured by the number of decayed, missing or filled teeth (DMFT - Index). The index provides counting data, which cannot exceed 8 in our case, since only eight molars were under examination in the trail. As covariates there are the six different treatment groups, sex and three ethnic groups.

The collected data set as displayed in Table 3.18 exhibits a large fraction of zero outcomes. Since the common generalized linear models, e.g. with Poisson or binomial distributed response, typically do not capture this feature, it is a standard method in the GLM framework to introduce a zero-inflating component to the model. Indeed, the zero-inflated regression model is a special case of a switching regression model.

Table 3.14: Simulated rejection rates of the MLRT for the models under the alternative P1.A, P2.A, B1.A and B2.A in Table 3.12 for various sample sizes with $N = 5000$ replications.

	P1.A (Poisson)			
Level	$n = 50$	100	200	500
0.025	0.113	0.248	0.502	0.909
0.05	0.180	0.356	0.622	0.947
0.1	0.281	0.485	0.743	0.976

	B1.A (Binomial)			
Level	$n = 100$	200	500	1000
0.025	0.097	0.208	0.507	0.813
0.05	0.161	0.292	0.630	0.883
0.1	0.256	0.419	0.749	0.937

	P2.A (Poisson)			
Level	$n = 50$	100	200	500
0.025	0.085	0.220	0.450	0.864
0.05	0.144	0.320	0.564	0.917
0.1	0.233	0.440	0.686	0.956

	B2.A (Binomial)			
Level	$n = 100$	200	500	1000
0.025	0.058	0.137	0.394	0.763
0.05	0.094	0.202	0.510	0.838
0.1	0.141	0.301	0.636	0.908

Skrondal and Rabe-Hesketh (2004, p. 349) fit several latent Poisson and binomial regression models to the data. They consider zero-inflation, two - and three component switching regression models with switching intercept as well as models with a normally-distributed intercept (normal intercept models).
In the notation of the introduction, we have $n = 797$, $q = 9$ and $l = 1$ for $i = 1, \ldots, n$ with covariate $X_i = (1, X_{i1}, \ldots, X_{i8})^T$, $X_{ij} \in \{0, 1\}$ for $j = 1, \ldots, 8$ and $W_i = 1$. The zero-inflated models are defined by two switching components, where one component is a point mass at zero. Further, in the normal intercept models, the latent variable U_i is distributed $N(\mu, \sigma^2)$. Specifically, the density of $Y_i|X_i = x_i, U_i = \theta_k$ is, in the Poisson case,

$$f(y_i, x_i; \beta, \theta_k) = \exp\big(- \exp\left(x_i^T\beta + \theta_k\right)\big) \frac{\exp\big(y_i\left(x_i^T\beta + \theta_k\right)\big)}{y_i!}$$

and in the binomial case with $\mathbf{n} = 8$,

$$f(y_i, x_i; \beta, \theta_k) = \binom{8}{y_i} \big(\text{logit}^{-1}(x_i^T\beta + \theta_k)\big)^{y_i} \big(1 - \text{logit}^{-1}(x_i^T\beta + \theta_k)\big)^{8-y_i}.$$

A point mass at zero arises as a limiting case for $\theta_1 = -\infty$, thus, the zero-inflated model can be thought of as a sub-model of the two-component switching model (with one parameter less).

Table 3.15: Simulated rejection rates of the modified LRT for the models with dependency structure over time P1.MC.H, P1.MCAR.H, B1.MC.H and B1.MCAR.H under the hypothesis for various sample sizes with $N = 5000$ replications.

	P1.MC.H		
Level	$n = 100$	200	500
0.025	0.018	0.024	0.024
0.05	0.032	0.041	0.046
0.1	0.064	0.074	0.081

	B1.MC.H		
Level	$n = 200$	500	1000
0.025	0.014	0.018	0.019
0.05	0.026	0.037	0.040
0.1	0.059	0.078	0.085

	P1.MCAR.H		
Level	$n = 100$	200	500
0.025	0.023	0.025	0.027
0.05	0.043	0.047	0.045
0.1	0.071	0.085	0.079

	B1.MCAR.H		
Level	$n = 200$	500	1000
0.025	0.013	0.018	0.027
0.05	0.027	0.036	0.049
0.1	0.060	0.070	0.086

Table 3.16 contains the results of the model selection criteria AIC and BIC for the above mentioned latent regression models (we omit the simple Poisson and binomial regression model without latent variable). We note that except for the zero-inflated variant, the models based on the binomial distribution perform better than the corresponding model based on the Poisson distribution. Further, the overall best model in terms of AIC is the two-state binomial model, and in terms of BIC the three-state binomial model. Thus, in this problem

Table 3.16: Log-Likelihood, AIC and BIC for the latent regression models. ZI: zero-inflated model, NI: normal intercept model.

	Poisson				Binomial			
	ZI	m=2	m=3	NI	ZI	m=2	m=3	NI
Log-like	−1410	−1406	−1406	−1433	−1431	−1400	−1397	−1409
AIC	2841	2834	2838	2886	2882	2822	**2821**	2838
BIC	2887	2886	2899	2932	2929	**2874**	2882	2885

model selection criteria give a clear indication of population heterogeneity, but do not allow to decide between a two- and a three component model. Therefore, we test for two against

three states by using the above methodology. Note that since m^* is at least 4, the test will be asymptotically conservative. Table 3.17 provides the penalized maximum likelihood estimators for the two- and three-component binomial switching regression model. When fitting a model with four potential components, two components (i.e. parameters of the binomial components) were equal, thus, it reduced to the three-component model.
We use $C_2 = C_4 = 1$ and obtain $T_n^{\text{mod}} = 5.71$, with the estimate $\hat{p} = 0.40$ this yields a P-value of 1.4%. Thus, the test clearly rejects two components in favor of three components. Finally, in Table 3.18 we display the observed and expected frequencies under the fitted models (estimated without penalization). We also see here that the three component model provides quite a good fit to the data.

Table 3.17: Penalized ML-estimators for the two- and three-component model.

	π_1	π_2	π_3	θ_1	θ_2	θ_3	Log-like
m=2	0.49	0.51		−2.30	−0.28		−1400.19
m=3	0.23	0.44	0.33	−3.47	−1.27	−0.01	−1397.33

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
m=2	−0.42	−0.17	−0.46	−0.28	−0.74	0.17	0.13	−0.14
m=3	−0.41	−0.16	−0.50	−0.38	−0.81	0.16	0.13	−0.18

Table 3.18: Observed and expected frequencies under the fitted binomial models.

	obs	m=1	ZI	m=2	m=3
0	231	107.21	227.88	221.72	226.84
1	163	230.81	120.50	180.94	171.36
2	140	233.84	174.21	130.26	141.90
3	116	145.04	150.21	113.99	109.83
4	70	59.90	84.21	85.11	77.62
5	55	16.76	31.33	45.26	45.10
6	22	3.08	7.53	16.00	18.86
7	0	0.34	1.06	3.40	4.90
8	0	0.02	0.07	0.33	0.59

3.4 Proofs

Proof of Proposition 3.1 . By the boundedness of $b_i b_i^T$ (Assumption 3.4') the classical strong law of large numbers (e.g. Shiryaev, 1996, Chp. IV, § 3, Thm. 3) yields

$$\frac{1}{n}\tilde{\mathbf{B}}_{22} \xrightarrow{a.s.} \tilde{\Sigma}$$

with $\tilde{\Sigma}$ positive definite 2×2-matrix. The classical central limit theorem (e.g. Shiryaev, 1996, Chp. III, § 3; Thm. 3) gives

$$\frac{1}{\sqrt{n}}\tilde{\mathbf{b}}_2 \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tilde{\Sigma})$$

since b_i and hence $\tilde{\mathbf{b}}_2$ is centered under G_0. Therefore we have

$$\sup_{t_2 \in C} \left(2\, t_2^T \tilde{\mathbf{b}}_2 - t_2^T\, \tilde{\mathbf{B}}_{22}\, t_2 \right) = \sup_{t_2 \in C} \left(2\, t_2^T Z - t_2^T\, \tilde{\Sigma}\, t_2 \right) + o_P(1)$$

for the cone $C = [0, \infty) \times [0, \infty)$ and $Z \sim \mathcal{N}(0, \tilde{\Sigma})$. Since

$$\sup_{t_2 \in C} \left(2\, t_2^T Z - t_2^T\, \tilde{\Sigma}\, t_2 \right) = \sup_{\tilde{t} \in \tilde{C}} \left(2\, \tilde{t}^T\, \tilde{\Sigma}^{-1/2} Z - \tilde{t}^T\, \tilde{t} \right) = \tilde{Z}^T \tilde{Z} - \inf_{\tilde{t} \in \tilde{C}} (\tilde{Z} - \tilde{t})^T\, (\tilde{Z} - \tilde{t}),$$

for $\tilde{C} = \tilde{\Sigma}^{1/2} C$ and $\tilde{Z}$ bivariate standard normal r.v., we can apply the arguments from Self and Liang (1987, case 7) as in Example 2.3 to establish the proposition. □

Proof of Theorem 3.4 . Under the stated assumptions a standard expansion gives

$$0 = D_\omega L_n^I(\hat{\omega}) = D_\omega L_n^I(\omega_0) + D_\omega^2 L_n^I(\omega_0)(\hat{\omega} - \omega_0) + o(\|\hat{\omega} - \omega_0\|). \tag{3.34}$$

Ergodicity and stationarity of the process $(Y_i)_i$ (cf. Leroux, 1992b) yield

$$\lim_{n \to \infty} D_\omega^2 L_n^I(\omega_0) = E_0 \left[D_\omega^2 \log\, f_{\text{mix}}(Y_1; \omega_0) \right] = \Sigma_0. \tag{3.35}$$

Lindgren (1978) showed that the process $(Y_i)_i$ is strongly mixing with exponentially decaying mixing coefficient. This allows us to apply Theorem 18.5.3 from Ibragimov and Linnik (1971) to see

$$\frac{1}{\sqrt{n}} D_\omega L_n^I(\omega_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \text{Cov}_0). \tag{3.36}$$

In fact, one requires a multivariate extension of Ibragimov and Linnik's theorem, which can be obtained by adopting the Cramér-Wold device. The well-definition of the matrix

$$\text{Cov}_0 \;=\; E\big[h(Y_1; \omega_0)\, h(Y_1; \omega_0)^T\big] + \sum_{j \geq 2} E\big[h(Y_1; \omega_0)\, h(Y_j; \omega_0)^T + h(Y_j; \omega_0)\, h(Y_1; \omega_0)^T\big],$$

with $h(y;\omega) = \big(D_\omega \log f_{\text{mix}}(y;\omega)\big)$ is a consequence of Ibragimov and Linnik's theorem, however for HMMs this as well as the mixing property just follows from the fact that for ergodic Markov chains the n-step transition probabilities approach the stationary probabilities exponentially fast. Combining the first three displayed equations concludes the proof. □

Proof of Theorem 3.5 . The proof follows the same lines as for i.i.d. models, as for example discussed in Pruscha (2000, pp. 251–256). We start with the decomposition

$$T_n^I = T_{1n}^I - T_{0n}^I = 2\big(L_n^I(\hat\omega) - L_n^I(\omega_0)\big) - 2\big(L_n^I(\varphi(\hat t)) - L_n^I(\varphi(t_0))\big).$$

where $\hat\omega$ and $\hat t$ denote the unrestricted and restricted MLEIs. Using the consistency of $\hat\omega$ and (3.34) and (3.35) standard expansion technique shows

$$\begin{aligned} T_{1n}^I &= (\hat\omega-\omega_0)^T\big(D_\omega^2 L_n^I(\omega_0)\big)(\hat\omega-\omega_0) + o_p(1) = n(\hat\omega-\omega_0)^T\Sigma_0(\hat\omega-\omega_0) + o_p(1) \\ &= \frac{1}{n} D_\omega L_n^I(\omega_0)^T\Sigma_0^{-1} D_\omega L_n^I(\omega_0) + o_p(1) \end{aligned}$$

Since the proof of Theorem 3.4 and hence the equations (3.34) and (3.35) can be rephrased in the parametrization t a similar expansion applies to T_{0n}^I, in particular we note that

$$\begin{aligned} &D_t L_n^I(\varphi(t_0)) = S_0^T D_\omega L_n^I(\omega_0) + o_p(\sqrt{n}) \\ &\frac{1}{n} D_t^2 L_n^I(\varphi(t_0)) = S_0^T\Sigma_0 S_0 + o_p(1) \\ &S_0^T\Sigma_0 S_0(\hat t - t_0) = \frac{1}{n} D_t L_n^I(\varphi(t_0)) + o_p\big(\frac{1}{\sqrt{n}}\big) \end{aligned}$$

with $S_0 = D_t\varphi(t_0)$ such that one has for an consistent MLEI $\hat t$

$$\begin{aligned} T_{0n}^I &= (\hat t - t_0)^T D_t^2 L_n^I(\varphi(t_0))(\hat t - t_0) + o_p(1) = n(\hat t - t_0)^T S_0^T\Sigma_0 S_0(\hat t - t_0) + o_p(1) \\ &= \frac{1}{n} D_t L_n^I(\varphi(t_0))^T (S_0^T\Sigma_0 S_0)^{-1} D_t L_n^I(\varphi(t_0)) + o_p(1) \\ &= \frac{1}{n} D_\omega L_n^I(\omega_0)^T S_0 (S_0^T\Sigma_0 S_0)^{-1} S_0^T D_\omega L_n^I(\omega_0) + o_p(1) \end{aligned}$$

Combining these calculations concludes the proof

$$\begin{aligned} T_n^I &= \frac{1}{n} D_\omega L_n^I(\omega_0)^T\Sigma_0^{-1} D_\omega L_n^I(\omega_0) - \frac{1}{n} D_\omega L_n^I(\omega_0)^T S_0 (S_0^T\Sigma_0 S_0)^{-1} S_0^T D_\omega L_n^I(\omega_0) + o_p(1) \\ &= \frac{1}{n} D_\omega L_n^I(\omega_0)^T\big(\Sigma_0^{-1} - S_0(S_0^T\Sigma_0 S_0)^{-1} S_0^T\big) D_\omega L_n^I(\omega_0) + o_p(1) \\ &= Z^T \mathrm{Cov}_0^{1/2}\Big(\Sigma_0^{-1} - S_0(S_0^T\Sigma_0 S_0)^{-1} S_0^T\Big)\mathrm{Cov}_0^{1/2} Z \end{aligned}$$

with $Z \sim N(0, I_{\bar d\times\bar d})$, where the last equation follows from (3.36). □

Proof of Proposition 3.2 . Set $B^i = E\big[b_1 b_i^T\big]$, and partition B^i into

$$B^i = \left(\begin{array}{c|c} B^i_{11} & B^i_{12} \\ \hline B^i_{21} & B^i_{22} \end{array}\right), \qquad B^i_{11} \in \mathbb{R}^{3\times 3}.$$

Let for $k = 1, 2$

$$\lambda_k = E\left[b_1 | U_1 = k\right] = \int b_1(y) f_{\theta_{0k}}(y)\, dy \in \mathbb{R}^5.$$

From $E[b_1] = 0$ it easily follows that $\lambda_2 = c_1\lambda_1$, where $c_1 = -\frac{\alpha_{21}}{\alpha_{12}} \neq 0$. Using this and

$$\begin{aligned} E\big[\Delta_1 b_1\big] &= E\big[\big(Z^1_{12}(\theta) - Z^1_{11}(\theta)\big) b_1\big] = \int \frac{f_{\theta_{01}}(y) - f_{\theta_{02}}(y)}{f_{\text{mix}}\big(y; G_0\big)} b_1(y) f_{\text{mix}}\big(y; G_0\big)\, dy \\ &= \int \big(b_1(y) f_{\theta_{01}}(y) - b_1(y) f_{\theta_{02}}(y)\big)\, dy = E\big[b_1 | U_1 = 1\big] - E\big[b_1 | U_1 = 2\big] = \lambda_1 - \lambda_2 \end{aligned}$$

we arrive at

$$B^1\, \bar{1} = \lambda_1 - \lambda_2 = (1 - c_1)\lambda_1 \tag{3.37}$$

where $\bar{1} = (1, 0, 0, 0, 0)^T$. Further, using $\lambda_2 = c_1\lambda_1$ and $E\big[b_1 b_i^T | U_1, U_i\big] = E\big[b_1 | U_1\big] E\big[b_i | U_i\big]^T$ one shows that

$$B^i = E\big[b_1 b_i^T\big] = c_i \lambda_1 \lambda_1^T, \qquad i \geq 2, \tag{3.38}$$

where $c_i = \frac{\alpha_{21}}{\alpha_{12}} \Big(1 - \alpha_{12}^{(i-1)} - \alpha_{21}^{(i-1)}\Big)$, and $\alpha_{jk}^{(i)} = P(U_{i+1} = k | U_1 = j)$ denotes the i-step transition probability. Note, that $c_i = 0$ for all i if and only if $a_{12} + a_{21} = 1$, which leads to independence of the (Y_i). Furthermore, note that (3.38) implies the symmetry of B^i. In order to show $\tilde{B}^i = 0$ for $i \geq 2$, we compute

$$\begin{aligned} \tilde{B}^i &= E\big[\tilde{b}_{21} \tilde{b}_{2i}^T\big] \\ &= B^i_{22} - B^i_{21} \left(B^1_{11}\right)^{-1} B^1_{12} - B^1_{21} \left(B^1_{11}\right)^{-1} B^i_{12} + B^1_{21} \left(B^1_{11}\right)^{-1} B^i_{11} \left(B^1_{11}\right)^{-1} B^1_{12} \end{aligned}$$

To establish our claim, we show that all four summands in this expansion coincide. From (3.37), one observes

$$B^1_{11}\, \bar{1} = (1 - c_1)\, (\lambda_{11}, \lambda_{12}, \lambda_{13})^T$$

and hence

$$\left(B^1_{11}\right)^{-1} (\lambda_{11}, \lambda_{12}, \lambda_{13})^T = \frac{1}{1 - c_1}\, \bar{1},$$

where $\bar{1} = (1, 0, 0)^T$ and λ_{1j} denotes the j^{th} component of λ. Using this, (3.37) and (3.38) give

$$\begin{aligned} B^1_{21} \left(B^1_{11}\right)^{-1} B^i_{12} &= c_i B^1_{21} \left(B^1_{11}\right)^{-1} (\lambda_{11}, \lambda_{12}, \lambda_{13})^T\, (\lambda_{14}, \lambda_{15}) \\ &= \frac{c_i}{1 - c_1} B^1_{21}\, \bar{1}\, (\lambda_{14}, \lambda_{15}) = c_i\, (\lambda_{14}, \lambda_{15})^T\, (\lambda_{14}, \lambda_{15}) = B^i_{22}. \end{aligned}$$

Since B^1 and B^i are symmetric, one also has $B^i_{21}\left(B^1_{11}\right)^{-1}B^1_{12} = B^i_{22}$. The same argument applies to the last matrix

$$\begin{aligned} & B^1_{21}\left(B^1_{11}\right)^{-1}B^i_{11}\left(B^1_{11}\right)^{-1}B^1_{12} \\ &= c_i B^1_{21}\left(B^1_{11}\right)^{-1}(\lambda_{11},\lambda_{12},\lambda_{13})^T(\lambda_{11},\lambda_{12},\lambda_{13})\left(B^1_{11}\right)^{-1}B^1_{12} \\ &= \frac{c_i}{(1-c_1)^2}B^1_{21}\,\bar{1}\,\bar{1}^T B^1_{12} = c_i\,(\lambda_{14},\lambda_{15})^T(\lambda_{14},\lambda_{15}) = B^i_{22}. \end{aligned}$$

This concludes the proof. □

Proof of the fact that Ass. 3.4' implies Ass. 3.5' (Remark 3.8). Chen et al. (2004) show that Assumption 3.4' implies Assumption 3.5' by using Theorem 12.3 in (Billingsley, 1968, p.95). To apply this theorem they show

$$E\big[\big(n^{-1/2}\sum_i Z^1_{ik}(\theta_1) - n^{-1/2}\sum_i Z^1_{ik}(\theta_2)\big)^2\big] = E\big[\big(Z^1_{1k}(\theta_1) - Z^1_{1k}(\theta_2)\big)^2\big] \le E\big[g^{2/3}(Y_1)\big]\,|\theta_1-\theta_2|^2$$

Since in the case of dependent data the first equality fails, we rephrase the calculation for the HMM setup:

$$\begin{aligned} & E\big[\big(n^{-1/2}\sum_i Z^1_{ik}(\theta_1) - n^{-1/2}\sum_i Z^1_{ik}(\theta_2)\big)^2\big] \\ &= n^{-1}\sum_i E\big[\big(Z^1_{ik}(\theta_1) - Z^1_{ik}(\theta_2)\big)^2\big] + n^{-1}\sum_{i\neq j} E\big[\big(Z^1_{ik}(\theta_1) - Z^1_{ik}(\theta_2)\big)\big(Z^1_{jk}(\theta_1) - Z^1_{jk}(\theta_2)\big)\big] \\ &= E\big[\big(Z^1_{1k}(\theta_1) - Z^1_{1k}(\theta_2)\big)^2\big] + 2\sum_{i=2}^{n} E\big[\big(Z^1_{1k}(\theta_1) - Z^1_{1k}(\theta_2)\big)\big(Z^1_{ik}(\theta_1) - Z^1_{ik}(\theta_2)\big)\big] \\ &\le \Big(E\big[g^{2/3}(Y_1)\big] + 2\sum_{i=2}^{\infty} E\big[g^{1/3}(Y_1)g^{1/3}(Y_i)\big]\Big)\;|\theta_1-\theta_2|^2 \end{aligned}$$

Since $(Y_i)_i$ is strongly mixing with exponentially decaying coefficients (cf. Lindgren, 1978) the sum in the last line is finite and hence Thm. 12.3 from Billingsley (1968) can also be applied here. □

Proof of Theorem 3.8. We follow Chen et al. (2004) and provide some details to the sketch given above. Starting with (3.26) we consider the expansions of Z^1_{i0}, Z^1_{i1} and Z^1_{i2} for $k = 1, 2$ and $i = 1, \ldots, n$

$$\begin{aligned} Z^1_{i0}(\beta,\theta) &= (\beta-\beta_0)Z^{01}_i(\beta_0,\theta) + \varepsilon_{i0} \\ Z^1_{ik}(\beta,\theta) &= (\theta-\theta_{0k})Z^{10}_i(\beta,\theta_{0k}) + (\theta-\theta_{0k})^2/2\; Z^{20}_i(\beta,\theta_{0k}) + \varepsilon_{ik} \end{aligned}$$

and

$$\begin{aligned} Z_i^1(\beta, \hat{G}_k) &= (\beta - \beta_0) Z_i^{01}(\beta_0, G_0) + \tilde{\varepsilon}_{i0} \\ Z_{ik}^1(\beta, \hat{G}_k) &= m_{1k}(\hat{G}_k) Z_i^{10}(\beta_0, \theta_{0k}) + m_{2k}(\hat{G}_k)/2 \; Z_i^{20}(\beta_0, \theta_{0k}) + \tilde{\varepsilon}_{ik}. \end{aligned}$$

Hence,

$$\begin{aligned} \delta_i \; = (\hat{\pi} - \pi_0)\Delta_i(\beta_0) + (\hat{\beta} - \beta_0) Z_i^{01}(\beta_0, G_0) + \hat{\pi}\hat{m}_{11} Z_i^{10}(\beta_0, \theta_{01}) + (1 - \hat{\pi})\hat{m}_{12} Z_i^{10}(\beta_0, \theta_{02}) \\ + \hat{\pi}\hat{m}_{21}/2 \; Z_i^{20}(\beta_0, \theta_{01}) + (1 - \hat{\pi})\hat{m}_{22}/2 \; Z_i^{20}(\beta_0, \theta_{02}) + \varepsilon_{in}, \end{aligned}$$

Chen et al. (2008) show that the tightness condition (see Assumption 3.5") ensures that

$$\varepsilon_n := \sum_{i=1}^{n} \varepsilon_{in} = o_p(1).$$

Using $\log(1+\delta_i) \leq \delta_i - \delta_i^2/2 + \delta_i^3/3$ and the fact that the remainder of the square and cubic part is at least of the same order as the linear part (cf. Chen et al., 2008), we obtain (3.28). Repeating the arguments yielding (3.10) shows that the upper bound can be attained and (3.30) holds. For the expansion of T_{0n}^{mod} we again follow Chen et al. (2004) and observe

$$T_{0n}^{\text{mod}} = \mathbf{b}_1^T B_{11}^{-1} \mathbf{b}_1 + o_p(1),$$

for $\mathbf{b}_1 \in \mathbb{R}^{q+3}$ as defined above, since $\hat{G}_k^{(2)}$ are single point distributions on Θ. This implies $m_{2k}(\hat{G}_k^{(2)}) = m_{1k}(\hat{G}_k^{(2)})^2$ for $k = 1, 2$.
Since $\mathbf{B}$ is invertible (at least for large n by Assumption 3.3") we can decompose $\mathbf{b}$ via orthogonalization, which yields

$$T_{1n}^{\text{mod}} \;=\; \mathbf{b}_1^T {B_{11}}^{-1} \mathbf{b}_1^T + \sup_{t_2} \left(2\, t_2^T \tilde{\mathbf{b}}_2 - t_2^T\, \tilde{\mathbf{B}}_{22}\, t_2 \right) + o_P(1)$$

and hence

$$T_n^{\text{mod}} \;=\; \sup_{t_2 \in [0,\infty)\times[0,\infty)} \left(2\, t_2^T \tilde{\mathbf{b}}_2 - t_2^T\, \tilde{\mathbf{B}}_{22}\, t_2 \right) + o_P(1).$$

Since Proposition 3.1 also holds for the corresponding quantities in the SRM setup, namely the b_is are defined by (3.27), we have

$$\sup_{t_2 \in [0,\infty)\times[0,\infty)} \left(2\, t_2^T \tilde{\mathbf{b}}_2 - t_2^T\, \tilde{\mathbf{B}}_{22}\, t_2 \right) \xrightarrow{\mathcal{L}} \left(\frac{1}{2} - p \right) \chi_0^2 + \frac{1}{2}\chi_1^2 + p\, \chi_2^2,$$

where $p = \left(\cos^{-1} \rho \right)/(2\pi)$ and ρ is the correlation coefficient in the matrix $\tilde{\mathbf{B}}_{22}$, which completes the analysis. □

Proof of Proposition 3.3. The proof follows the one of Proposition 3.2, where the quantities b_i, $\tilde{b}_{2i}$, B^i are now defined in terms of the SRM, in particular $b_i \in \mathbb{R}^{q+5}$, $\tilde{b}_{2i} \in \mathbb{R}^2$, $B^i \in \mathbb{R}^{(q+5)\times(q+5)}$. Analogously we define for $k = 1, 2$

$$\lambda_k = E\big[b_1 | U_1 = k\big] = \int b_1(y,x) f(y,x;\beta,\theta_{0k}) h(x)\, dy\, dx \in \mathbb{R}^{q+5}.$$

An important step in the proof of Prop. 3.2 is (3.37), therefore we need to verify

$$\begin{aligned} E\big[\Delta_1(\beta_0) b_1\big] &= E\big[\big(Z^1_{12}(\beta,\theta) - Z^1_{11}(\beta,\theta)\big) b_1\big] \\ &= \int \frac{f(y,x;\beta,\theta_{01}) - f(y,x;\beta,\theta_{01})}{f_{\text{switch}}\big(y,x;\beta_0,G_0\big)} b_1(y,x) f_{\text{switch}}\big(y,x;\beta_0,G_0\big) h(x)\, dy\, dx \\ &= \int \big(b_1(y,x) f(y,x;\beta,\theta_{01}) - b_1(y,x) f(y,x;\beta,\theta_{02})\big) h(x)\, dy\, dx \\ &= E\big[b_1 | U_1 = 1\big] - E\big[b_1 | U_1 = 2\big] = \lambda_1 - \lambda_2. \end{aligned}$$

The remainder of the proof follows from conditional independence of $(Y_i, X_i)_i$ given $(U_i)_i$, which holds for HMMs as well as for SRMs, and straight forward calculations as in the proof of Prop. 3.2. □

Chapter 4

Modeling HMMs with flexible state-dependent distributions

In the previous chapters we discuss statistical inference in HMMs with a given parametric family of state-dependent distributions. From a data analysis point of view this means that one chooses a parametric family and further statistical inference is conducted based on this choice. For many applications this approach is very common and frequently used. However, estimation and test results are clearly dependent on the chosen parametric family. In particular, one expects a systematic bias in the estimates in case of misspecifications. To avoid systemic errors due to the parametric assumptions one may consider HMMs with more flexible state-dependent distributions (sdfs), for example by applying nonparametric methods. Let us, for example, consider fitting a two-state Gaussian HMM while the true sdf of the second state F_2 is in fact a skew gamma distribution. For the specific setting of our simulation study (see Sec. 4.3) we observe that especially the estimator of the transition probability α_{21} is strongly influenced by the parametric assumptions:

F_2	bias of $\hat{\alpha}_{21}$	std.dev. of $\hat{\alpha}_{21}$
normal	-0.044	0.064
gamma	0.012	0.057
nonparam.	0.013	0.078

Excerpt of Table 4.2 for the estimator $\hat{\alpha}_{21}$.

We see that the model with flexible nonparametric sdfs exhibits a bias comparable to the true parametric model. As expected the estimator is more variable than in the parametric models.

In this chapter we investigate methods of modeling HMMs with flexible sdfs and a prespecified number of states. In the following, we restrict ourselves to two-state models for simplicity. Two-state HMMs are reasonably implemented for many applications, for example one state might represent the normal status of a system while the other one might stand for an abnormal situation, e.g. a crisis in financial markets, a major break down in a transport network etc.
Clearly, the a priori assumption of a two-state model inherits the danger of misspecifications, when the true model in fact exhibits more states. In principle, this problem occurs similarly in purely parametric HMMs. But due to the additional flexibility of the proposed model one should expect that such misspecifications are less transparent and more difficult to detect, since models with flexible sdfs may capture some of the resulting artifacts. Such an phenomenon is called masking effect and should not be neglected in the application of flexible sdfs.
Firstly in the chapter, we propose a two-state HMM with flexible sdfs by assuming that only for the first state the corresponding sdf comes from a prespecified parametric family while the sdf corresponding to the second state is given by a finite mixture with components from the same parametric family. We will see that the standard methods and analytic tools for HMMs in the classical framework are also applicable for the proposed model.
Secondly, we investigate how nonparametric methods can be used to provide HMMs with flexible sdfs. In the recent literature semi- and nonparametric techniques in finite mixture models are widely discussed under various constraints, e.g. shape constraints (Chang and Walther, 2007), symmetry (Bordes et al., 2006a), smoothness (Ma, Gudlaugsdottir and Wood, 2008). For many applications such quality constraints are plausible and appear less restrictive than parametric assumptions. We will discuss how these concepts of semiparametric mixtures can be used for modeling of HMMs.

4.1 An HMM with flexible sdfs: a parametric approach

We propose a two-state HMM which allows us to model the sdfs or one of the sdfs in a rather flexible way. We assume that the distribution of one state belongs to some parametric family while the other distribution is a finite mixture of distributions from the same family. Such a model gives a more flexible framework compared with the usual parametric two-state HMM.
We define a two-state HMM with flexible sdfs $\mathcal{M}_{\text{flex}}$. Following the notation in Sec. 1.5

let $(U_i)_i$ be an unobserved stationary and ergodic two-state Markov chain and $(Y_i)_i$ an observed process fulfilling the HMM dependency assumptions 1., 2. in Section 1.2. Let $(F_\theta)_\theta$ be a parametric family of distributions on the observation space $\mathcal{Y}$. In contrast to the usual assumption that the sdfs $F_k = P(Y_i \leq y_i | U_i = k)$ for $k = 1, 2$ belong to some parametric family, we consider the HMM with

$$\begin{aligned} F_1(y) &= F_{\theta_1}(y) \\ F_2(y) &= \bar{\pi}_1 F_{\theta_2}(y) + \ldots + \bar{\pi}_{\bar{m}} F_{\theta_{\bar{m}+1}}(y) \end{aligned}$$

with $\bar{\pi}_k > 0$, $\sum_{k=1}^{\bar{m}} \bar{\pi}_k = 1$ and $(F_{\theta_i})_{1 \leq i \leq \bar{m}+1} \subset (F_\theta)_\theta$ some parametric family. The parameter of the model is therefore

$$\vartheta = (\alpha_{12}, \alpha_{21}, \bar{\pi}_1, \ldots, \bar{\pi}_{\bar{m}-1}, \theta_1, \ldots, \theta_{\bar{m}+1}).$$

Let us briefly consider the marginal distribution of the model $\mathcal{M}_{\text{flex}}$

$$F_{\text{mix}}(y) = \pi_1 F_1(y) + \pi_2 F_2(y) = \pi_1 F_{\theta_1}(y) + \pi_2 \bar{\pi}_1 F_{\theta_2}(y) + \ldots + \pi_2 \bar{\pi}_{\bar{m}} F_{\theta_{\bar{m}+1}}(y)$$

with (π_1, π_2) being the stationary distribution of $(U_i)_i$. Clearly, the marginal distribution is just a $\bar{m}+1$-component mixture w.r.t. the family $(F_\theta)_\theta$ and weights $\tilde{\pi}_1 := \pi_1$, $\tilde{\pi}_2 := \pi_2 \bar{\pi}_1, \ldots, \tilde{\pi}_{\bar{m}+1} := \pi_2 \bar{\pi}_{\bar{m}}$. Since the mapping $(\pi_1, \bar{\pi}_1, \ldots, \bar{\pi}_{m-1}) \to (\tilde{\pi}_1, \ldots, \tilde{\pi}_m)$ is one-to-one (if one excludes $\pi_1 = 1$), we deduce that from the viewpoint of the marginal distribution the model $\mathcal{M}_{\text{flex}}$ just gives a nonstandard parametrization for the $\bar{m}+1$-component mixture w.r.t. $(F_\theta)_\theta$. Analogously, we will see that the proposed model $\mathcal{M}_{\text{flex}}$ can be interpreted as an HMM with sdfs from some parametric family with a specific nonstandard parametrization.

Proposition 4.1. *Let $(Y_i)_i$ be observations from the model $\mathcal{M}_{flex}$ defined above and let $(\tilde{Y}_i)_i$ be observations from a $(\bar{m}+1)$-state HMM with sdfs $F_{\theta_1}, \ldots, F_{\theta_{\bar{m}+1}}$ and transition matrix*

$$\begin{pmatrix} \alpha_{11} & \alpha_{12}\bar{\pi}_1 & \alpha_{12}\bar{\pi}_2 & \cdots & \alpha_{12}\bar{\pi}_{\bar{m}} \\ \alpha_{21} & \alpha_{22}\bar{\pi}_1 & \alpha_{22}\bar{\pi}_2 & \cdots & \alpha_{22}\bar{\pi}_{\bar{m}} \\ \alpha_{21} & \alpha_{22}\bar{\pi}_1 & \alpha_{22}\bar{\pi}_2 & \cdots & \alpha_{22}\bar{\pi}_{\bar{m}} \\ \vdots & \vdots & \vdots & & \vdots \\ \alpha_{21} & \alpha_{22}\bar{\pi}_1 & \alpha_{22}\bar{\pi}_2 & \cdots & \alpha_{22}\bar{\pi}_{\bar{m}} \end{pmatrix} \tag{4.1}$$

Then, the sequences $(Y_i)_i$ and $(\tilde{Y}_i)_i$ follow the same law.

The proof of the proposition is straightforward and outlined in Section 4.4. The proposition shows that the models of $\mathcal{M}_{\text{flex}}$-type form a subset of $(\bar{m}+1)$-state HMMs with sdfs from a parametric family, which is defined by the restrictions on the transition matrix displayed above. Note that a standard $(\bar{m}+1)$-state HMM has $(\bar{m}+1)^2$ parameters, while a model of $\mathcal{M}_{\text{flex}}$-type only has $2(\bar{m}+1)$ parameters.

Based on the fact that the proposed model can be represented as a $(\bar{m}+1)$-state HMM with nonstandard parametrization ϑ it is clear that the methods of ML-estimation as discussed above apply to the $\mathcal{M}_{\text{flex}}$ model. Consistency and asymptotic normality of the MLE can be derived by verifying the assumptions by Leroux (1992b) and Bickel et al. (1998) for the parametrization ϑ respectively. The sequence $\sqrt{n}(\hat{\vartheta} - \vartheta_0)$ is centered asymptotically normal, if Assumptions 2.1 - 2.3 are fulfilled, ϑ_0 lies in the interior of the parameter space and if the strong consistency of the MLE, the positive definiteness of the Fisher information matrix $\mathcal{J}_0$ and the ergodicity of the underlying Markov chain hold for the parametrization ϑ. Note that, the ergodicity property follows from the ergodicity of the two-state Markov chain and $\bar{\pi}_k$ for all $1 \leq k \leq \bar{m}$. Note that the number of components or states must be specified correctly to ensure asymptotic normality, i.e. $\bar{m}$ must be known. Alternatively $\bar{m}$ can be estimated for example based on the model selection criteria proposed by Poskitt and Zhang (2005). Numerical evaluation of the MLE is straightforward when general purpose methods are used, however for application of an EM algorithm the restrictions on the transition matrix need to be addressed appropriately.

In summary, it can be stated that the model $\mathcal{M}_{\text{flex}}$ is a nonstandard two-state HMM with quite flexible second sdfs (even for small $\bar{m}$) and a handy number of parameters, which is easy to handle and to interpret. In addition, it can be analyzed by means of the theoretical results available for standard HMMs with sdfs from some parametric family.

4.2 An HMM with flexible sdfs: a semiparametric approach

As recent publications show the benefit of nonparametric methods in mixture models, e.g. Chang and Walther (2007), we take semiparametric mixtures as a starting point. We discuss how these results can be extended to HMMs in terms of identifiability and estimation via an EM algorithm.

4.2.1 Semiparametric mixtures

The label of semiparametric mixtures has multiple usage, in particular one needs to distinguish whether nonparametric methods enter the model via the sdfs or through the mixing distribution. In the following we concentrate on the first case. Semiparametric mixtures build a vivid field of research as many recent publications show. Hall and Zhou (2003) and Hall et al. (2005) consider multivariate mixtures and observe a reverse "curse of dimensionality" (Hall and Zhou, 2003), i.e. the number of components m, for which a mixture is identifiable, increases with the dimension of the observations d. In particular, Hall and Zhou (2003) show identifiability for the two-component mixture for $d \geq 3$ under the key assumption of independent marginals. Under this assumption one may interpret multivariate observations as blocks of univariate observations for which the additional knowledge is available that they belong to the same component, i.e. increasing d amplifies the knowledge about the component membership. For univariate observations Bordes et al. (2006b) and Hunter et al. (2007) show identifiability of two-components mixtures with location-shift under the key assumption of symmetric components. Bordes et al. (2006a) and Bordes and Vandekerkhove (2008) consider two-component mixtures with one known and one symmetric component.
In addition to the mentioned papers, which prove identifiability of the models and provide very specific estimation procedures, other contributions motivate their approaches rather heuristically, but provide quite general estimation procedures based on the EM principle for models from the world of nonparametric maximum likelihood estimation (NPMLE), see Ma et al. (2008) for mixtures with components of some degree of smoothness, Chang and Walther (2007), Eilers and Borgdorff (2007) and Cule et al. (2008) for uni- and multivariate mixtures of log-concave densities, Bordes et al. (2007) for mixtures with symmetric components and Benaglia et al. (2009) for general multivariate mixtures.

4.2.2 Semiparametric HMMs

In the following we discuss how to make use of the mentioned results for HMMs, in particular we show that identifiability of the marginal mixture of a two-state HMM implies the identifiability of the HMM parameters and propose an EM algorithm with nonparametric assumptions on the sdfs.
We consider a stationary, ergodic two-state HMM with parameter $\vartheta = (\alpha_{12}, \alpha_{21}, F_1, F_2)$, $\alpha_{12}, \alpha_{21} \in [0,1], F_k \in \mathfrak{F}_k, k = 1,2$. If $\mathfrak{F} = \mathfrak{F}_1 = \mathfrak{F}_2$ denotes a parametric family for which two-component mixtures are identifiable, then the identifiability of the HMM follows from

Teicher (1967) as argued by Leroux (1992b). We will see that this also holds true if $\mathfrak{F}_k$ are general, possibly nonparametric families of distributions.

Proposition 4.2. *Suppose that we have a stationary two-state HMM with ergodic regime, such that the marginal mixture of the HMM is identifiable, i.e. for* $\pi_1, \pi_1' \in [0,1]$, $F_k, F_k' \in \mathfrak{F}_k$ *for* $k = 1, 2$ *with* $\mathfrak{F}_1, \mathfrak{F}_2$ *families of distributions on* $\mathbb{R}^d$

$$\pi_1 F_1(y) + (1-\pi_1)F_2(y) = \pi_1' F_1'(y) + (1-\pi_1')F_2'(y) \qquad \text{a.e.}$$

with $F_1 \not\equiv F_2$, $F_1' \not\equiv F_2'$ *implies* $\pi_1 = \pi_1'$ *and* $F_k = F_k'$ *for* $k = 1, 2$. *Then the HMM is also identifiable, i.e. if* $P_\vartheta^{(Y_i)} = P_{\vartheta'}^{(Y_i)}$ *for* $\vartheta = (\alpha_{12}, \alpha_{21}, F_1, F_2)$, $\vartheta' = (\alpha_{12}', \alpha_{21}', F_1', F_2')$ *then* $\vartheta = \vartheta'$ *holds.*

The proof uses the fact that every stationary, ergodic two-state Markov chain is reversible. Details are deferred to Sec. 4.4. The proposition shows that the identifiability results established by Hall and Zhou (2003) and Bordes et al. (2006a) for semiparametric two-component mixtures transfer to two-state HMMs, e.g. an HMM with one fixed and one symmetric sdf, whose marginal mixture fulfills the conditions of Proposition 2 in Bordes et al. (2006a), is identifiable.

A standard technique in nonparametric density estimation is NPMLE without and with penalization, i.e.

$$\hat{f} = \arg\max_{f \in \mathfrak{F}} \sum_i \log f(Y_i) \tag{4.2}$$

$$\hat{f} = \arg\max_{f \in \mathfrak{F}} \sum_i \log f(Y_i) - C\ Pen(f). \tag{4.3}$$

Clearly, if $\mathfrak{F}$ denotes the class of all densities these maximization problems do not have a solution. However, for the class of monotone densities on the positive real line the Grenander estimator is a prominent example of a unique solution of (4.2). Also for log-concave densities an NPMLE exists (cf. Rufibach, 2006). A commonly used penalty measuring the roughness of a function is $Pen(f) = \int (f''(x))^2 dx$ (e.g. Silverman, 1982).
We now consider a stationary, ergodic two-state HMM with flexible sdfs $\mathcal{M}_{\text{semi}}$, such that $F_1 \in \mathfrak{F}_1 = \{F_\theta | \theta \in \Theta\}$ belongs to some parametric family of distributions with densities w.r.t. the Lebesgue measure. Let $F_2 \in \mathfrak{F}_2$ denote a nonparametric class, e.g. $\mathfrak{F}_2 = \{\text{distributions with monotone densities on } [0, \infty)\}$. Our aim is to propose an EM algorithm to fit an HMM of the $\mathcal{M}_{\text{semi}}$-type. We will see that the essential requirement for the formulation of an EM algorithm for HMMs is the following condition.

Condition 4.1. Let $\mathfrak{F}_2$ be a class of distributions and $\mathcal{F}_2$ the corresponding class of densities w.r.t. the Lebesgue measure such that for observations $Y_1, \dots Y_n$ and for all weights $w_1, \dots, w_n \geq 0$, $\sum_i w_i = 1$ the unpenalized or penalized maximization problem

$$\arg\max_{f \in \mathcal{F}_2} \sum_i w_i \log f(Y_i)$$
$$\text{or} \qquad \arg\max_{f \in \mathcal{F}_2} \sum_i w_i \log f(Y_i) - C\, Pen(f)$$

have a unique solution $\hat{f} \in \mathcal{F}_2$.

The EM algorithm for HMMs of $\mathcal{M}_{\text{semi}}$-type

We recall that the log likelihood function of an HMM with m states is given by

$$L_n(\vartheta) = \log p_n(Y_1, \dots, Y_n; \vartheta) = \log \sum_{U_1=1}^{m} \cdots \sum_{U_n=1}^{m} p_n(U_1, \dots, U_n, Y_1, \dots, Y_n; \vartheta),$$

with the complete information likelihood function

$$p_n(u_1, \dots, u_n, y_1, \dots, y_n; \vartheta) = \pi_{u_1}(\vartheta) \prod_{i=1}^{n-1} \alpha_{u_i, u_{i+1}}(\vartheta) \prod_{i=1}^{n} f_{u_i}(y_i; \vartheta).$$

As standard for the EM algorithm we consider iterative maximization of the objective function

$$Q(\vartheta, \vartheta') = E_{\vartheta'} \left[\log p_n(U_1, \dots, U_n, Y_1, \dots, Y_n; \vartheta) | Y_1^n \right]$$

with $Y_1^n = (Y_1, \dots, Y_n)$. In the HMM framework $Q(\vartheta, \vartheta')$ can be expressed as follows (cf. Cappé et al., 2005).

$$\begin{aligned}
Q(\vartheta, \vartheta') &= E_{\vartheta'} \Big[\log \pi_{U_1}(\vartheta) \prod_{i=1}^{n-1} \alpha_{U_i, U_{i+1}}(\vartheta) \prod_{i=1}^{n} f_{U_i}(Y_i; \vartheta) \Big| Y_1^n \Big] \\
&= E_{\vartheta'} \left[\log \pi_{U_1}(\vartheta) | Y_1^n \right] + \sum_{i=1}^{n-1} E_{\vartheta'} \left[\log \alpha_{U_i, U_{i+1}}(\vartheta) | Y_1^n \right] + \sum_{i=1}^{n} E_{\vartheta'} \left[\log f_{U_i}(Y_i; \vartheta) | Y_1^n \right] \\
&= E_{\vartheta'} \Big[\sum_{k=1}^{m} \mathbf{1}_{\{U_1 = k\}} \log \pi_k(\vartheta) \Big| Y_1^n \Big] + \sum_{i=1}^{n-1} E_{\vartheta'} \Big[\sum_{j=1}^{m} \sum_{k=1}^{2} \mathbf{1}_{\{U_i = j, U_{i+1} = k\}} \log \alpha_{jk}(\vartheta) \Big| Y_1^n \Big] \\
&\quad + \sum_{i=1}^{n} E_{\vartheta'} \Big[\sum_{k=1}^{m} \mathbf{1}_{\{U_i = k\}} \log f_k(Y_i; \vartheta) \Big| Y_1^n \Big]
\end{aligned}$$

$$\begin{aligned} Q(\vartheta,\vartheta') &= \sum_{k=1}^{m} E_{\vartheta'}[\mathbf{1}_{\{U_1=k\}}|Y_1^n]\log\pi_k(\vartheta) + \sum_{i=1}^{n-1}\sum_{j=1}^{m}\sum_{k=1}^{m} E_{\vartheta'}[\mathbf{1}_{\{U_i=j,U_{i+1}=k\}}|Y_1^n]\log\alpha_{jk}(\vartheta) \\ &\quad + \sum_{i=1}^{n}\sum_{k=1}^{m} E_{\vartheta'}[\mathbf{1}_{\{U_i=k\}}|Y_1^n]\log f_k(Y_i;\vartheta). \end{aligned}$$

Note, that taking the conditional expectation of indicator functions just gives conditional probabilities. These can be computed based on the forward and backward probabilities (cf. Sec. 2.3.1) as follows

$$\begin{aligned} \phi_{i|n}(k;\vartheta') &:= P_{\vartheta'}(U_i=k|Y_1^n) = a_i(k;\vartheta')b_i(k;\vartheta')/L \\ \phi_{i,i+1|n}(j,k;\vartheta') &:= P_{\vartheta'}(U_i=j,U_{i+1}=k|Y_1^n) = a_i(j;\vartheta')\alpha_{jk}(\vartheta')f_k(y_{i+1};;\vartheta')b_{i+1}(k;\vartheta')/L(\vartheta') \end{aligned}$$

for $1\le j,k\le m$ with $L(\vartheta')=\sum_{k=1}^{m} a_i(k;\vartheta')b_i(k;\vartheta')$ for some $1\le i\le n$.
Let us now consider an HMM of $\mathcal{M}_{\text{semi}}$-type with $m=2$ states and the infinite-dimensional parameter $\vartheta=(\alpha_{12},\alpha_{21},\theta,f)$ where the sdfs are given by $f_1(y;\vartheta)=f_{\theta(\vartheta)}(y)\in(f_\theta)_\theta$ and $f_2(y;\vartheta)=f(y;\vartheta)\in\mathcal{F}_2$. Then the objective function Q simplifies to

$$\begin{aligned} Q(\vartheta,\vartheta') &= \phi_{1|n}(1;\vartheta')\log\pi_1(\vartheta) + \phi_{1|n}(2;\vartheta')\log(1-\pi_1(\vartheta)) \\ &\quad + \sum_{i=1}^{n-1}\phi_{i,i+1|n}(1,1;\vartheta')\log(1-\alpha_{12}(\vartheta)) + \sum_{i=1}^{n-1}\phi_{i,i+1|n}(1,2;\vartheta')\log\alpha_{12}(\vartheta) \\ &\quad + \sum_{i=1}^{n-1}\phi_{i,i+1|n}(2,1;\vartheta')\log\alpha_{21}(\vartheta) + \sum_{i=1}^{n-1}\phi_{i,i+1|n}(2,2;\vartheta')\log(1-\alpha_{21}(\vartheta)) \\ &\quad + \sum_{i=1}^{n}\phi_{i|n}(1;\vartheta')\log f_{\theta(\vartheta)}(Y_i) + \sum_{i=1}^{n}\phi_{i|n}(2;\vartheta')\log f(Y_i;\vartheta). \end{aligned}$$

From this representation it is clear that the l-th E-step consists of the computation of the probabilities $\phi_{i|n}(k;\vartheta^{(l)}),\phi_{i,i+1|n}(j,k;\vartheta^{(l)})$. The M-step is carried out by

$$\vartheta^{(l+1)} = \arg\max_{\vartheta} Q(\vartheta,\vartheta^{(l)})$$

If we assume the initial distribution as fixed, we can obtain explicit expressions for $\alpha_{12}^{(l+1)}$ and $\alpha_{21}^{(l+1)}$

$$\begin{aligned} \alpha_{12}^{(l+1)} &= \frac{\sum_{i=1}^{n-1}\phi_{i,i+1|n}(1,2;\vartheta^{(l)})}{\sum_{i=1}^{n-1}\phi_{i,i+1|n}(1,1;\vartheta^{(l)}) + \sum_{i=1}^{n-1}\phi_{i,i+1|n}(1,2;\vartheta^{(i)})} \\ \alpha_{21}^{(l+1)} &= \frac{\sum_{i=1}^{n-1}\phi_{i,i+1|n}(2,1;\vartheta^{(l)})}{\sum_{i=1}^{n-1}\phi_{i,i+1|n}(2,1;\vartheta^{(l)}) + \sum_{i=1}^{n-1}\phi_{i,i+1|n}(2,2;\vartheta^{(l)})}. \end{aligned}$$

If we consider the initial distribution as the stationary one, i.e. $\pi_1 = \alpha_{21}/(\alpha_{12} + \alpha_{21})$, one needs to maximize a quadratic function, which can easily be obtained by standard analytic or numerical methods. Secondly, we have

$$\theta^{(l+1)} = \arg \max_{\{\theta \in \Theta\}} \sum_{i=1}^{n} \phi_{i|n}(1; \vartheta^{(l)}) \log f_\theta(Y_i)$$

which is simply a weighted MLE with weights $\phi_{i|n}(1; \vartheta^{(l)}) / \sum_{i=1}^{n} \phi_{i|n}(1; \vartheta^{(l)})$, that can be computed easily for many parametric families. In particular, if $(f_\theta)_\theta$ is an exponential family and $\theta = E_\vartheta[Y_i | U_i = 1] = \int y f_\theta(y) dy$ one has

$$\theta^{(l+1)} = \frac{\sum_{i=1}^{n} \phi_{i|n}(1; \vartheta^{(l)}) Y_i}{\sum_{i=1}^{n} \phi_{i|n}(1; \vartheta^{(l)})}.$$

For the nonparametric component the representation of Q shows that we need to find a maximizer

$$f^{(l+1)} = \arg \max_{\{f \in \mathcal{F}_2\}} \sum_{i=1}^{n} \phi_{i|n}(2; \vartheta^{(l)}) \log f(Y_i), \tag{4.4}$$

i.e. we need to find a weighted NPMLE with weights $\phi_{i|n}(2; \vartheta^{(l)}) / \sum_{i=1}^{n} \phi_{i|n}(2; \vartheta^{(l)})$. Such an estimator exists and can often be computed as easily as the unweighted version when for example the class of monotone or log-concave densities is considered. In total the M-step gives

$$\vartheta^{(l+1)} = \left(\alpha_{12}^{(l+1)}, \alpha_{21}^{(l+1)}, \theta^{(l+1)}, f^{(l+1)}\right).$$

Starting with an initial guess $\vartheta^{(0)}$, e.g. provided by a parametric HMM, conducting iteratively the E- and M-step yields a sequence of estimators for the $\mathcal{M}_{\text{semi}}$-model $\vartheta^{(0)}, \vartheta^{(1)}, \vartheta^{(2)}, \ldots$. Our description of the EM algorithm is completed by proposing some determination rule, e.g. $L_n(\vartheta^{(l+1)}) - L_n(\vartheta^{(l)}) < \varepsilon$ or $\left\|\vartheta^{(l+1)} - \vartheta^{(l)}\right\| < \varepsilon$ for some appropriate norm.

Remark 4.1. Above, the proposed EM algorithm is based on the likelihood function without penalization. If one considers the log-likelihood function with penalization of a $\mathcal{M}_{\text{semi}}$ model

$$L_n(\vartheta) - C\, pen(f)$$

instead, the algorithm remains the same besides the fact that in the M-step in (4.4) $f^{(l+1)}$ is replaced by the penalized weighted NPMLE

$$f^{(l+1)} = \arg \max_{\{f \in \mathcal{F}_2\}} \sum_{i=1}^{n} \phi_{i|n}(2; \vartheta^{(l)}) \log f(Y_i) - C\, pen(f).$$

Remark 4.2. Other choices of $\mathfrak{F}_1, \mathfrak{F}_2$, e.g. $\mathfrak{F}_1 = \{F_0\}$ as in Bordes et al. (2006a) or $\mathfrak{F}_1 = \mathfrak{F}_2 = \{\text{distributions with log-concave densities}\}$ as in Chang and Walther (2007), or additional structure as in location-shift mixture models are also possible and the presented EM-techniques may extend to those models, but to us the proposed model $\mathcal{M}_{\text{semi}}$ with one parametric and one nonparametric sdf appears most useful in applications of the nature described above.

Concluding remark

The previous section illustrates how results on semiparametric mixtures can be translated to the HMM framework, especially to models of $\mathcal{M}_{\text{semi}}$-type. In particular, this enables the estimation in semiparametric HMMs. However, the section does not bridge the gap between theoretical results on quite specific models with serious drawbacks in application on one hand and a general class of models with practicable algorithms lacking of a theoretical analysis on the other hand. Filling this gap must be a subject of further research in the field of semiparametric mixtures as well as semiparametric HMMs.

Extensions and outlook

Extensions of the proposed model are possible in several directions. In particular, one may define models of $\mathcal{M}_{\text{semi}}$-type with m states with $m > 2$, of which $m - 1$ sdfs follow parametric models. In this case selecting m via model choice procedures as discussed in Chapter 3 becomes a relevant task. The theoretical analysis of these extensions seems to be a hard problem, also the practical applicability needs to be investigated. In particular, masking effects resulting in an underestimation of m are expected due to the high flexibility of the model.
It remains a challenging task to derive asymptotic results for the $\mathcal{M}_{\text{semi}}$-model in the HMM context as well as for corresponding finite mixture models. Bordes and Vandekerkhove (2008) take a first step in this direction, but their techniques appear to be quite specific. Based on the asymptotic analysis a test theory enabling testing hypotheses on the parameters would be a desirable goal as well as validation techniques for parametric assumptions.

4.3 Simulation experiments

In this simulation study we illustrate the previously introduced methods for two different scenarios both motivated by applications. We concentrate on univariate continuous models.

A waiting time model with contaminations

At first we consider a waiting time model with normal contaminations given by an HMM M1 with transitions $\alpha_{12} = 0.2, \alpha_{21} = 0.3$, an exponential sdf with $\lambda = 1$ in state one and a Gaussian sdf with $\mu = 4, \sigma = 1$ in state two. The marginal density of the model is displayed in Fig. 4.1. Analyzing the exon length distribution in the human genome, Ma et al. (2008) propose comparable models based on semiparametric mixtures.
In our study we simulate N=100 samples each of size $m = 500$ from M1 and perform estimation under several model assumptions on the families of densities of the sdfs $\mathcal{F}_1$, $\mathcal{F}_2$:

- M1par denotes the true parametric model, i.e.

$$\begin{aligned} \mathcal{F}_1 &= \mathcal{F}_{exp} = \left\{ f_\lambda(x) = \lambda \exp(-\lambda x) \mathbf{1}_{\{x>0\}} | \lambda > 0 \right\} \\ \mathcal{F}_2 &= \mathcal{F}_{norm} = \left\{ f_{\mu,\sigma^2}(x) = 1/(\sigma\sqrt{2\pi}) \exp(-(x-\mu)^2/(2\sigma^2)) | \mu \in \mathbb{R}, \sigma^2 > 0 \right\}. \end{aligned}$$

- M1LC denotes the semiparametric model with log-concave component, i.e.

$$\mathcal{F}_1 = \mathcal{F}_{exp},\ \mathcal{F}_2 = \{ f \text{ densities} \,|\, \log(f) \text{ concave} \}.$$

- M1Mon denotes the semiparametric model with monotone component, i.e.

$$\mathcal{F}_1 = \{ f \text{ monotone densities} \},\ \mathcal{F}_2 = \mathcal{F}_{norm}.$$

While for the parametric model the MLE is evaluated using numerical maximization based on the **R** (R Development Core Team, 2009) function *nlm*, this is not feasible for semiparametric models. Therefore we perform estimation for the semiparametric models using the EM algorithm proposed above. As one should investigate the sensitivity of the EM algorithm w.r.t. starting values, in particular for the nonparametric component, we consider in addition to the true values as starting values also the density of the uniform distribution on $[0, \max(Y_i)]$ as $f^{(0)}$. In our simulations the proposed EM does not appear very sensitive in our settings. For the evaluation of the weighted NPMLE which needs to be performed in each M-step (cf. Eq. (4.4)) the package *logcondens* by Rufibach and Dümbgen (2009) is used for the M1LC model with minor modifications. Note, that the weighted NPMLE over the class of log-concave densities exists uniquely and its logarithm is a linear spline (cf. Rufibach, 2006). Based on the function *isoMean* in the package *logcondens* one can also perform the M-step for the M1Mon model. Here a weighted Grenander estimator needs to be computed, which is given by the left derivative of the least concave majorant of the weighted distribution function $\sum_{i=1}^n w_i \mathbf{1}_{(-\infty,t]}(Y_i)$ with weights $w_i \geq 0$, $\sum_i w_i = 1$

(cf. van der Vaart, 1998, Lemma 24.5.). We determine the EM algorithm with the stopping rule

$$L_n(\vartheta^{(l+1)}) - L_n(\vartheta^{(l)}) < \varepsilon = 10^{-6} \text{ or } l \geq L := 100.$$

In Figures 4.1-4.2 we display the histogram of one sample from M1 and the marginal densities of the true and estimated models. In Figure 4.3 the nonparametric component

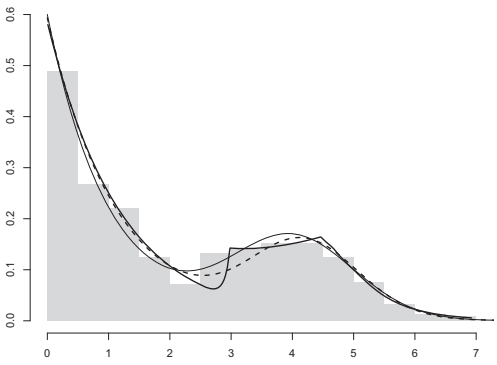


Figure 4.1: Histogram of a sample from M1 with the marginal densities of the M1LC model (solid), the M1par model (dashed) and the true model (thin line).

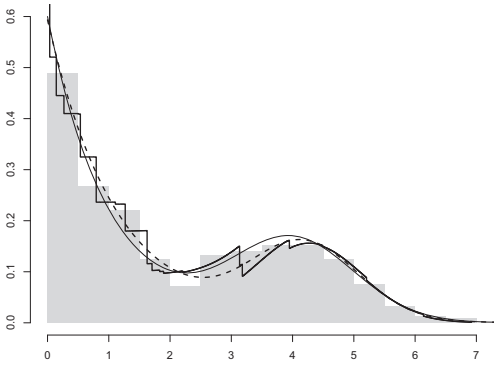


Figure 4.2: Histogram of a sample from M1 with the marginal densities of the M1Mon model (solid), the M1par model (dashed) and the true model (thin line).

of the model M1LC is displayed and Figure 4.4 shows the nonparametric component for the model M1mon. In both plots the parametrically estimated and the true densities are added.

The simulation results are displayed in Table 4.1. They show that the semiparametric

Table 4.1: Simulated estimators for the models M1par, M1LC and M1mon. The mean and the standard deviation (in brackets) of the estimators are displayed.

	α_{12}	α_{21}	λ	μ	σ
true	0.200	0.300	1.000	4.000	1.000
M1par	0.204	0.302	1.006	3.992	0.987
	(0.031)	(0.040)	(0.097)	(0.098)	(0.079)
M1LC	0.206	0.300	1.023	-	-
	(0.031)	(0.040)	(0.099)	-	-
M1mon	0.175	0.287	-	3.993	0.959
	(0.034)	(0.041)	-	(0.105)	(0.081)

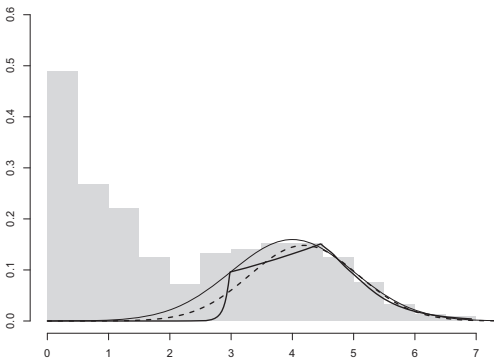


Figure 4.3: Histogram of a sample from M1 with the weighted density $(1-\hat{\pi})\hat{f}$ of the nonparametric sdf of the M1LC model (solid). The weighted densities of the M1par model (dashed) $(1-\hat{\pi})f_{\hat{\mu},\hat{\sigma}}$ and the true model (thin line) $(1-\pi_0)f_{\mu_0,\sigma_0}$ are added.

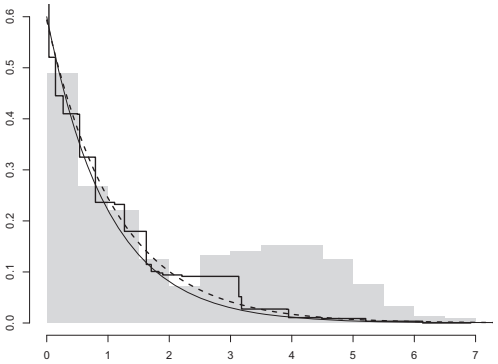


Figure 4.4: Histogram of a sample from M1 with the weighted density $\hat{\pi}\hat{f}$ of the nonparametric sdf of the M1Mon model (solid). The weighted densities of the M1par model (dashed) $\hat{\pi}f_{\hat{\lambda}}$ and the true model (thin line) $\pi_0 f_{\lambda_0}$ are added.

HMMs deliver in general results which are comparable to the fully parametric model, in terms of the transition probabilities and the remaining parametric component, i.e. λ in M1LC and (μ, σ^2) in M1Mon. We see that in the model M1Mon the transition probabilities are slightly underestimated. The variances exhibit a small increase for the semiparametric models.

An HMM with a skew component

Secondly, we consider an HMM with a skew component M2 given by the same transition probabilities as above $\alpha_{12} = 0.2, \alpha_{21} = 0.3$, a Gaussian sdf with $\mu = 7, \sigma = 1$ in state one and a gamma sdf with $\alpha = 3, \beta = 1/2$ in state two. Note that for $X \sim \Gamma(\alpha, \beta)$ with $\alpha = 3, \beta = 1/2$ one has $E[X] = \alpha/\beta = 6, V[X] = \alpha/\beta^2 = 12$ and skewness $\nu(X) = 2/\sqrt{\alpha} = \sqrt{6}/3 \approx 0.82$. The marginal density of the model is displayed in Fig. 4.5. The proposed model seems suitable to the data sets considered in Güttler (2006) where delay time differences from railway data collected for eight high speed tracks are under investigation. Capturing the dependence structure of such data sets as it is possible using HMMs could give insights to the nature of delay times and their development over time.
As above we simulate N=100 samples each of size $m = 500$ from M2 and perform estimation under several model assumptions on the families of densities of sdfs $\mathcal{F}_1$, $\mathcal{F}_2$:

- M2par denotes the true parametric model, i.e.

$$\mathcal{F}_1 = \mathcal{F}_{norm},\ \mathcal{F}_2 = \mathcal{F}_{gamma} = \left\{ f_{\alpha,\beta}(x) = x^{\alpha-1}\beta^{\alpha}/\Gamma(\alpha)\exp(-\beta x)\,\mathbf{1}_{\{x>0\}} \,|\, \alpha,\beta > 0 \right\}.$$

- M2norm denotes a parametric model under normality assumption (misspecified), i.e.

$$\mathcal{F}_1 = \mathcal{F}_2 = \mathfrak{F}_{norm}.$$

- M2LC denotes the semiparametric model with a log-concave component, i.e.

$$\mathcal{F}_1 = \mathcal{F}_{norm},\ \mathcal{F}_2 = \{ f \text{ densities} \,|\, \log(f) \text{ concave} \}.$$

- M2flex denotes the parametric model $\mathcal{M}_{\text{flex}}$ with $\bar{m} = 2$ under normality assumption (misspecified), i.e.

$$\mathcal{F}_1 = \mathcal{F}_{norm},\ \mathcal{F}_2 = \mathcal{F}_{norm}\,\{ \text{ densities of Gaussian two component mixtures } \}.$$

The simulation settings are the same as for M1. Again the **R** package *logcondens* is used to evaluate the weighted NPMLE for the EM algorithm of the semiparametric model M2LC. The sensitivity w.r.t. starting values was investigated and can be considered as minor. The starting values for the M2flex model we choose $\bar{\pi} = 1/2$, $\bar{\mu}_2 = 3$, $\bar{\mu}_1 = 9$, $\bar{\sigma}_1 = \bar{\sigma}_2 = 2$ leading to similar mean and variance as the true sdf of the state two.
In Figures 4.5-4.6 we display the histogram of one sample from M2 and the marginal densities of the true and estimated models. In Figure 4.7 the nonparametric component of the model M2LC is displayed and Figure 4.8 shows the nonparametric component for the models M2flex and M2norm. In both plots the parametrically estimated and the true densities are added.
The simulation results are displayed in Table 4.2. The results show that the semiparametric model M2LC gives almost as accurate results as the parametric model under the true parametric assumptions M2par, in particular the estimated variances for M2LC exceed the variances for M2par only slightly. The parametric model under the misspecified Gaussian assumption M2norm captures the expectation and variance of the sdf of the state two surprisingly well. Only the estimator $\hat{\alpha}_{21}$ exhibits a strong negative bias.
The sample displayed in Fig. 4.8 indicates for the $\mathcal{M}_{\text{flex}}$-model the tendency of the sdf of state two to be split into two components left and right of the sdf of state one. This seems to be the reason for the strong negative bias of $\hat{\alpha}_{12}$ and the strong positive bias of $\hat{\alpha}_{21}$.

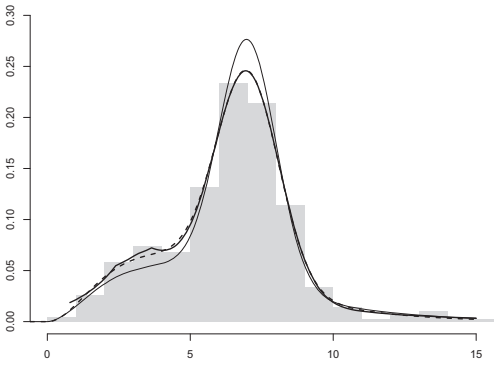


Figure 4.5: Histogram of a sample from M2 with the marginal densities of the M2LC model (solid), the M2par model (dashed) and the true model (thin line).

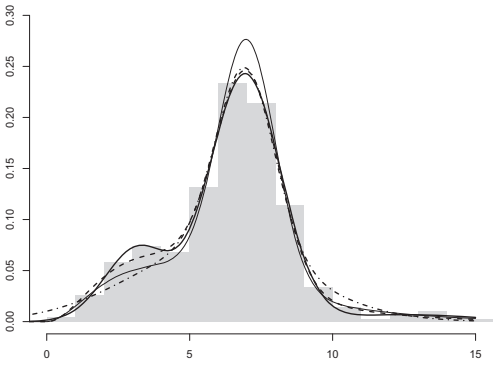


Figure 4.6: Histogram of a sample from M2 with the marginal densities of the M1flex model (solid), the M2par model (dashed), the M2norm model (dashed dotted) and the true model (thin line).

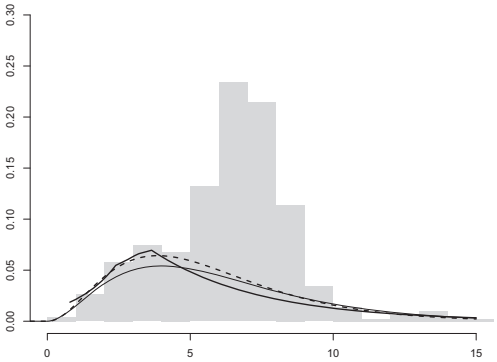


Figure 4.7: Histogram of a sample from M2 with the weighted density $(1-\hat{\pi})\hat{f}$ of the nonparametric sdf of the M2LC model (solid). The weighted densities of the M2par model (dashed) $(1-\hat{\pi})f_{\hat{\alpha},\hat{\beta}}$ and the true model (thin line) $(1-\pi_0)f_{\alpha_0,\beta_0}$ are added.

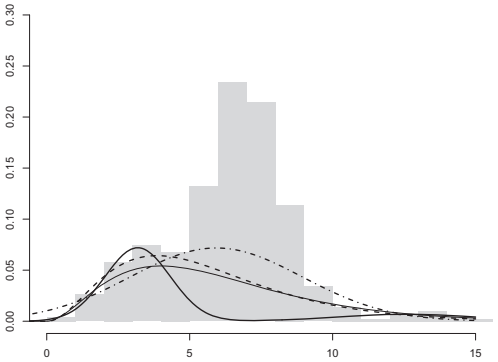


Figure 4.8: Histogram of a sample from M2 with the weighted densities $(1-\hat{\pi})\hat{f}_{\hat{\mu},\hat{\sigma}}$ of the sdf of the M2norm model (dashed dotted) and $(1-\hat{\pi})\hat{f}_{flex}$ of the sdf of the M2flex model (solid). The weighted densities of the M2par model (dashed) $(1-\hat{\pi})f_{\hat{\alpha},\hat{\beta}}$ and the true model (thin line) $(1-\pi_0)f_{\alpha_0,\beta_0}$ are added.

Table 4.2: Simulated estimators for the models M2par, M2norm, M2LC and M2flex. The mean and the standard deviation (in brackets) of the estimators are displayed. For the model M2norm the mean of the estimators μ_2, σ_2 are reported instead of α, β.

	α_{12}	α_{21}	μ	σ	α	β
true	0.200	0.300	7.000	1.000	3.000	0.500
M2par	0.206	0.312	7.011	0.974	3.031	0.509
	(0.043)	(0.064)	(0.073)	(0.066)	(0.345)	(0.062)
M2norm	0.217	0.256	6.972	0.933	6.164	3.295
	(0.049)	(0.057)	(0.081)	(0.069)	(0.231)	(0.327)
M2LC	0.202	0.313	7.011	0.976	-	-
	(0.041)	(0.078)	(0.080)	(0.073)	-	-
M2flex	0.186	0.414	6.987	1.031	-	-
	(0.033)	(0.121)	(0.089)	(0.104)	-	-

Concluding remark

In general, our simulations show that semiparametric HMMs are applicable via the proposed EM procedure. Although it is not known whether the considered models are identifiable, reasonable estimates have been obtained. In conclusion our study confirms the evident fact that the correctly specified parametric model is superior over the semiparametric model, while the semiparametric model seems superior over the incorrectly specified parametric models. Hence in situations where parametric assumptions are questionable, using the proposed methods might be advantageous.

However, we should point out that the simulation results should not be overgeneralized. Clearly, the performance of semiparametric methods does depend on the specific models, in particular the shape of the marginal density.

4.4 Proofs

Proof of Proposition 4.1. We denote the unobserved variables as follows: as above $(U_i)_i$ denotes the two-state Markov chain of the $\mathcal{M}_{\text{flex}}$ model, $(\bar{U}_i)_i$ denotes the unobserved component in the mixture F_2, i.e. $\bar{U}_i \sim Mult(\bar{\pi})$, and $(\tilde{U}_i)_i$ denotes the $(\bar{m}+1)$-state Markov chain of the HMM $(\tilde{Y}_i)_i$. We set $A_1 = \{U_i = 1\}$, $A_{k+1} = \left\{U_i = 2, \bar{U}_i = k\right\}$ for $k = 1, \ldots, \bar{m}$ and $B_k = \left\{\tilde{U}_i = k\right\}$ for $k = 1, \ldots, \bar{m}+1$. Clearly (Y_i) and $(\tilde{Y}_i)$ follow the

same law, if $P(B_k|B_j) = P(A_k|A_j)$ for all $j,k = 1,\dots,\bar{m}+1$. This holds, if

$$\begin{aligned}
&P(\tilde{U}_{i+1} = 1|\tilde{U}_i = 1) = P(U_{i+1} = 1|U_i = 1) = \alpha_{11}\\
&P(\tilde{U}_{i+1} = k+1|\tilde{U}_i = 1) = P(U_{i+1} = 2, \tilde{U}_{i+1} = k|U_i = 1) = \alpha_{12}\bar{\pi}_k\\
&P(\tilde{U}_{i+1} = 1|\tilde{U}_i = k+1) = P(U_{i+1} = 1|U_i = 2, \tilde{U}_i = k) = \alpha_{12}\\
&P(\tilde{U}_{i+1} = k+1|\tilde{U}_i = j+1) = P(U_{i+1} = 2, \tilde{U}_{i+1} = k|U_i = 2, \tilde{U}_i = j) = \alpha_{22}\bar{\pi}_k
\end{aligned}$$

for $1 \le j,k \le \bar{m}$, which leads to the transition matrix (4.1). □

Proof of Proposition 4.2. Let $\vartheta = (\alpha_{12}, \alpha_{21}, F_1, F_2)$, $\vartheta' = (\alpha'_{12}, \alpha'_{21}, F'_1, F'_2)$ two parameterizations for an HMM leading to the same law $P_\vartheta^{(Y_i)} = P_{\vartheta'}^{(Y_i)}$. Since the marginal mixture of the HMM is identifiable, $P_\vartheta(Y_1 \le y_1) = P_{\vartheta'}(Y_1 \le y_1)$ a.e. implies $\pi_1 = \pi'_1$ for the stationary distribution and for the sdfs $F_k = F'_k$ for $k = 1,2$. Hence it remains to show that the transition probabilities also coincide. Since stationary two-state HMMs are reversible one has $\pi_1\alpha_{12} = \pi_2\alpha_{21}$ such that it suffices to show $\alpha'_{12} = \alpha_{12}$. Since $F_1 \not\equiv F_2$ we find y_1, y_2 such that $F_1(y_i) \ne F_2(y_i)$ for $i = 1,2$ and compute

$$\begin{aligned}
&P_\vartheta(Y_1 \le y_1, Y_2 \le y_2)\\
&\quad= \pi_1\alpha_{11}F_1(y_1)F_1(y_2) + \pi_1\alpha_{12}F_1(y_1)F_2(y_2) + \pi_2\alpha_{21}F_2(y_1)F_1(y_2) + \pi_2\alpha_{22}F_2(y_1)F_2(y_2)\\
&\quad= \pi_1(1-\alpha_{12})F_1(y_1)F_1(y_2) + \pi_1\alpha_{12}F_1(y_1)F_2(y_2)\\
&\qquad+\pi_2\alpha_{21}F_2(y_1)F_1(y_2) + \pi_2(1-\alpha_{21})F_2(y_1)F_2(y_2)\\
&\quad= \pi_1F_1(y_1)F_1(y_2) + \pi_2F_2(y_1)F_2(y_2) + \pi_1\alpha_{12}\left(F_1(y_1)F_2(y_2) - F_1(y_1)F_1(y_2)\right)\\
&\qquad+\pi_2\alpha_{21}\left(F_2(y_1)F_1(y_2) - F_2(y_1)F_2(y_2)\right)\\
&\quad= \pi_1F_1(y_1)F_1(y_2) + \pi_2F_2(y_1)F_2(y_2)\\
&\qquad+\pi_1\alpha_{12}\left(F_1(y_1)F_2(y_2) - F_1(y_1)F_1(y_2) + F_2(y_1)F_1(y_2) - F_2(y_1)F_2(y_2)\right)\\
&\quad= \pi_1F_1(y_1)F_1(y_2) + \pi_2F_2(y_1)F_2(y_2) + \pi_1\alpha_{12}\left(F_1(y_1) - F_2(y_1)\right)\left(F_2(y_2) - F_1(y_2)\right).
\end{aligned}$$

Since $\left(F_1(y_1) - F_2(y_1)\right)\left(F_2(y_2) - F_1(y_2)\right) \ne 0$ this yields

$$\alpha_{12} = \frac{P_\vartheta(Y_1 \le y_1, Y_2 \le y_2) - \pi_1F_1(y_1)F_1(y_2) - \pi_2F_2(y_1)F_2(y_2)}{\pi_1\left(F_1(y_1) - F_2(y_1)\right)\left(F_2(y_2) - F_1(y_2)\right)}.$$

As the same calculation holds true for $P_{\vartheta'}(Y_1 \le y_1, Y_2 \le y_2)$ we obtain $\alpha_{12} = \alpha'_{12}$. □

Bibliography

Albert, P. S. (1991). A two-state Markov mixture model for time series of epileptic seizure counts. *Biometrics*, **47** 1371–1381.

Altman, R. M. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *J. Amer. Statist. Assoc.*, **102** 201–210.

Azaïs, J.-M., Gassiat, E. and Mercadier, C. (2009). The likelihood ratio test for general mixture models with or without structural parameter. *ESAIM: Probab. Stat.*, **13** 301–327.

Balabdaoui, F., Mielke, M. and Munk, A. (2009). The likelihood ratio test for non-standard hypotheses near the boundary of the null - with application to the assessment of non-inferiority. *Statist. Decisions*. To appear.

Barbu, V. and Limnios, N. (2006). Maximum likelihood estimation for hidden semi-Markov models. *C. R. Math. Acad. Sci. Paris*, **342** 201–205.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68** 155–178.

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

Benaglia, T., Chauveau, D. and Hunter, D. R. (2009). An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *J. Comput. Graph. Statist.*, **18** 505–526.

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L. and Kirchner, U. (1999). The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J. R. Stat. Soc. Ser. A*, **162** 195–209.

Böhning, D., Seidel, W., Alfó, M., Garel, B., Patilea, V. and Walther, G. (2007). Advances in mixture models. *Comput. Stat. Data Anal.*, **51** 5205 – 5210.

Bickel, P. J. and Ritov, Y. (1996). Inference in hidden Markov models. I. Local asymptotic normality in the stationary case. *Bernoulli*, **2** 199–228.

Bickel, P. J., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.*, **26** 1614–1635.

Bickel, P. J., Ritov, Y. and Rydén, T. (2002). Hidden Markov model likelihoods and their derivatives behave like i.i.d. ones. *Ann. Inst. H. Poincaré Probab. Statist.*, **38** 825–846.

Billingsley, P. (1968). *Convergence of probability measures*. John Wiley & Sons Inc., New York.

Böhning, D. (1999). *Computer-assisted analysis of mixtures and applications*. Chapman & Hall/CRC, Boca Raton.

Bordes, L., Chauveau, D. and Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Comput. Stat. Data Anal.*, **51** 5429–5443.

Bordes, L., Delmas, C. and Vandekerkhove, P. (2006a). Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Statist.*, **33** 733–752.

Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006b). Semiparametric estimation of a two components mixture model. *Ann. Statist.*, **34** 1204–1232.

Bordes, L. and Vandekerkhove, P. (2008). Semiparametric two-component mixture model when a component is known: a class of asymptotically normal estimators. *Preprint*.

Bulla, J. (2006). *Application of Hidden Markov Models and Hidden Semi-Markov Models to Financial Time Series*. PhD thesis, Inst. for Statistics and Econometrics, Georg-August-University of Göttingen, Göttingen.

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

Celeux, G. and Durand, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Comput. Statist.*, **23** 541–564.

Chang, G. T. and Walther, G. (2007). Clustering with mixtures of log-concave distributions. *Comput. Stat. Data Anal.*, **51** 6242–6251.

Chen, H. and Chen, J. (2001). The likelihood ratio test for homogeneity in finite mixture models. *Canad. J. Statist.*, **29** 201–215.

Chen, H. and Chen, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statist. Sinica*, **13** 351–365.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **63** 19–29.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **66** 95–115.

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.*, **23** 221–233.

Chen, J., Li, P. and Fu, Y. (2008). Testing homogeneity in a mixture of von Mises distributions with a structural parameter. *Canad. J. Statist.*, **36** 129–142.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.*, **25** 573–578.

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

Claeskens, G. and Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press, Cambridge.

Cule, M. L., Samworth, R. J. and Stewart, M. I. (2008). Maximum likelihood estimation of a multidimensional log-concave density. *Preprint*.

Dainotti, A., Pescape, A., Rossi, P. S., Palmieri, F. and Ventre, G. (2008). Internet traffic modeling by means of hidden Markov models. *Computer Networks*, **52** 2645–2662.

Dannemann, J. (2006). *Maximum-Likelihood-Inferenz für Hidden Markow Modelle*. Diploma thesis, Inst. for Math. Stochastics, Georg-August-University of Göttingen, Göttingen.

Dannemann, J. and Holzmann, H. (2008a). The likelihood ratio test for hidden Markov models in two-sample problems. *Comput. Stat. Data Anal.*, **52** 1850–1859.

Dannemann, J. and Holzmann, H. (2008b). Likelihood ratio testing for hidden Markov models under non-standard conditions. *Scand. J. Statist.*, **35** 309–321.

Dannemann, J. and Holzmann, H. (2008c). Testing for two states in a hidden Markov model. *Canad. J. Statist.*, **36** 505–520.

Dannemann, J. and Holzmann, H. (2010). Testing for two components in a switching regression model. *Comput. Stat. Data Anal.* Accepted for publication.

De Gunst, M. C. M., Künsch, H. R. and Schouten, J. G. (2001). Statistical analysis of ion channel data using hidden Markov models with correlated state-dependent noise and filtering. *J. Amer. Statist. Assoc.*, **96** 805–815.

Demos, A. and Sentana, E. (1998). Testing for GARCH effects: a one-sided approach. *J. Econometrics*, **86** 97 – 127.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **39** 1–38.

Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, **7** 381–420.

Douc, R., Moulines, É. and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, **32** 2254–2304.

Drton, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.*, **37** 979–1021.

Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G. (1999). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Eilers, P. H. C. and Borgdorff, M. W. (2007). Non-parametric log-concave mixtures. *Comput. Stat. Data Anal.*, **51** 5444–5451.

Everitt, B. S. and Hand, D. J. (1981). *Finite mixture distributions*. Chapman & Hall, London.

Feller, W. (1943). On general class of "contagious" distributions. *Ann. Math. Statist.*, **14** 389–400.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer, New York.

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

Ghosh, J. and Sen, P. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II*. Wadsworth, Monterey, 789–806.

Giudici, P., Rydén, T. and Vandekerkhove, P. (2000). Likelihood-ratio tests for hidden Markov models. *Biometrics*, **56** 742–747.

Güttler, S. (2006). *Statistical Modeling of Railway Data*. Diploma thesis, Inst. for Math. Stochastics, Georg-August-University of Göttingen, Göttingen.

Hall, P., Neeman, A., Pakyari, R. and Elmore, R. (2005). Nonparametric inference in multivariate mixtures. *Biometrika*, **92** 667–678.

Hall, P. and Zhou, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.*, **31** 201–224.

Haut, S. R. (2006). Seizure clustering. *Epilepsy & Behavior*, **8** 50 – 55.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *J. Classification*, **17** 273–296.

Holzmann, H., Munk, A. and Gneiting, T. (2006a). Identifiability of finite mixtures of elliptical distributions. *Scand. J. Statist.*, **33** 753–763.

Holzmann, H., Munk, A., Suster, M. and Zucchini, W. (2006b). Hidden Markov models for circular and linear-circular time series. *Environ. Ecol. Stat.*, **13** 325–347.

Hopkins, A., Davies, P. and Dobson, C. (1985). Mathematical models of patterns of seizures: Their use in the evaluation of drugs. *Arch. Neurol.*, **42** 463–467.

Hunter, D. R., Wang, S. and Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.*, **35** 224–251.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and stationary sequences of random variables*. Wolters-Noordhoff Publishing, Groningen.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, **62** 49–66.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.*, **102** 1025–1038.

Kiefer, N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica*, **46** 427–434.

Kitchens, L. J. (1998). *Exploring statistics: a modern introduction to data analysis and inference*. Brook/Cole, Pacific Grove.

Le, N. D., Leroux, B. G. and Puterman, M. L. (1992). Exact likelihood evaluation in a Markov mixture model for time series of seizure counts. *Biometrics*, **48** 317–323.

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

Leroux, B. G. (1992b). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.*, **40** 127–143.

Leroux, B. G. and Puterman, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, **48** 545—558.

Li, P., Chen, J. and Marriott, P. (2009). Non-finite Fisher information and homogeneity: an EM approach. *Biometrika*, **96** 411–426.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scand. J. Statist.*, **5** 81–91.

Lystig, T. C. and Huges, J. P. (2002). Exact computation of the observed information matrix for hidden Markov models. *J. Comput. Graph. Statist.*, **11** 678–689.

Ma, J., Gudlaugsdottir, S. and Wood, G. (2008). On semi-parametric mixture distributions and their generalized EM estimation. *Preprint*.

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

MacKay, R. J. (2002). Estimating the order of a hidden Markov model. *Canad. J. Statist.*, **30** 573–589.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley-Interscience, New York.

Naik, P. A., Shi, P. and Tsai, C.-L. (2007). Extending the Akaike information criterion to mixture regression models. *J. Amer. Statist. Assoc.*, **102** 244–254.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. London A*, **185** 71–110.

Petrie, T. (1969). Probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **40** 97–115.

Piger, J. (2009). Econometrics: Models of regime changes. In *Encyclopedia of Complexity and Systems Science* (R. A. Meyers, ed.). Springer, New York, 2744–2757.

Poskitt, D. S. and Zhang, J. (2005). Estimating components in finite mixtures and hidden Markov models. *Aust. N. Z. J. Stat.*, **47** 269–286.

Pruscha, H. (2000). *Vorlesungen über mathematische Statistik*. B.G. Teubner Verlag, Stuttgart.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc.*, **73** 730–752. With comments and a rejoinder by the authors.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabiner, L. (1989). A tutorial on hmm and selected applications in speech recognition. *Proceedings of the IEEE*, **77** 257–286.

Robert, C. P., Rydén, T. and Titterington, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **62** 57–75.

Rufibach, K. (2006). *Log-concave density estimation and bump hunting for i.i.d. observations*. PhD thesis, University of Bern/University of Göttingen, Bern/Göttingen.

Rufibach, K. and Dümbgen, L. (2009). *Logcondens: Estimate a log-concave probability density from iid observations*. R package version 1.3.3.

Rydén, T. (1994). Consistent and asymptotically normal parameter estimates for hidden Markov models. *Ann. Statist.*, **22** 1884–1895.

Rydén, T. (1995). Estimating the order of hidden Markov models. *Statistics*, **26** 345–354.

Rydén, T., Terasvirta, T. and Asbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *J. Appl. Econometrics.*, **13** 217–244.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.*, **82** 605–610.

Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, **72** 133–144.

Shapiro, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *Internat. Statist. Rev.*, **56** 49–62.

Shiryaev, A. N. (1996). *Probability*. 2nd ed. Springer-Verlag, New York.

Silvapulle, M. J. and Sen, P. K. (2005). *Constrained statistical inference*. Wiley-Interscience, Hoboken.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, **10** 795–810.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Chapman & Hall/CRC, Boca Raton.

Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.*, **34** 1265–1269.

Teicher, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist.*, **38** 1300–1302.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. John Wiley & Sons Ltd., Chichester.

Turner, R. (2008). Direct maximization of the likelihood of a hidden Markov model. *Comput. Stat. Data Anal.*, **52** 4147 – 4160.

van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, **51** 738–743.

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

Wang, P. and Puterman, M. L. (1998). Mixed logistic regression models. *J. Agric. Biol. Environ. Stat.*, **3** 175–200.

Wang, P. and Puterman, M. L. (1999). Markov Poisson regression models for discrete time series. I. Methodology. *J. Appl. Statist.*, **26** 855–869.

Wang, P. and Puterman, M. L. (2001). Analysis of longitudinal data of epileptic seizure counts—a two-state hidden Markov regression approach. *Biom. J.*, **43** 941–962.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist.*, **39** 209–214.

Young, D. S. (2008). An overview of mixture models. *Preprint on arxiv.org*.

Zhu, H.-T. and Zhang, H. (2003). Hypothesis testing in mixture regression models (mathematical details). Tech. rep., Yale University School of Medicine, New Haven.

Zhu, H.-T. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **66** 3–16.

Zucchini, W., Raubenheimer, D. and MacDonald, I. L. (2008). Modeling time series of animal behavior by means of a latent-state model with feedback. *Biometrics*, **64** 807–815.

Curriculum Vitae

Jörn Dannemann

born 8. December 1980 in Eutin
married, one child, German

1987 – 2000	Schooling *Abitur* at the *Johann-Heinrich-Voß-Gymnasium Eutin*
2000 – 2001	civil service
2001 – 2006	diploma studies in mathematics and economics (minor) *Faculty of Mathematics, University of Göttingen* diploma thesis: *Maximum-Likelihood-Inferenz für Hidden Markow Modelle* supervised by *Prof. Dr. Axel Munk*
2001 – 2006	scholar of the Evangelisches Studienwerk e.V. Villigst
2003 – 2004	exchange student at the Warwick Mathematics Institute (United Kingdom)
2006 – 2009	Ph.D. studies in mathematics *Institute for Mathematical Stochastics, University of Göttingen* supervised by *Prof. Dr. Axel Munk* and *Prof. Dr. Hajo Holzmann*
2006 – 2009	member of the Ph.D. Programm "Applied Statistics and Empirical Methods" *Centre for Statistics, University of Göttingen*
2007 – 2009	"Georg-Lichtenberg"-scholar