doi: 10.1111/j.1467-9469.2007.00587.x

Likelihood Ratio Testing for Hidden Markov Models Under Non-standard Conditions

JÖRN DANNEMANN

Institut für Mathematische Stochastik, University of Göttingen

HAJO HOLZMANN

Institut für Stochastik, University of Karlsruhe

ABSTRACT. In practical applications, when testing parametric restrictions for hidden Markov models (HMMs), one frequently encounters non-standard situations such as testing for zero entries in the transition matrix, one-sided tests for the parameters of the transition matrix or for the components of the stationary distribution of the underlying Markov chain, or testing boundary restrictions on the parameters of the state-dependent distributions. In this paper, we briefly discuss how the relevant asymptotic distribution theory for the likelihood ratio test (LRT) when the true parameter is on the boundary extends from the independent and identically distributed situation to HMMs. Then we concentrate on discussing a number of relevant examples. The finite-sample performance of the LRT in such situations is investigated in a simulation study. An application to series of epileptic seizure counts concludes the paper.

Key words: bimodality, boundary, hidden Markov model, likelihood ratio test, marginal distribution, maximum-likelihood estimation, overdispersion

1. Introduction

Hidden Markov models (HMM) are a class of discrete-time stochastic processes that have been successfully applied in various fields of dependent data analysis, including speech recognition (Rabiner, 1989), econometrics (Rydén *et al.*, 1998), biology (Holzmann *et al.*, 2006), medical statistics (Albert, 1991) or biological sequence alignment (Arribas-Gil *et al.*, 2006). See the monograph by MacDonald & Zucchini (1997) for further examples.

An HMM consists of two ingredients, an unobservable finite-state Markov chain (X_k) and an observable stochastic process (Y_k) such that: (i) the (Y_k) are conditionally independent, given the (X_k); and (ii) given the (X_k), the distribution of Y_j depends on X_j only. As the Markov chain (X_k), which is sometimes called the regime or the latent process, is unobservable, inference has to be based on the (Y_k) alone. Typically, it is assumed that the state-dependent distributions, i.e. the distributions of Y_k given that $X_k = a$, $a \in \mathcal{M}$, come from a parametric family $(f_\theta)_{\theta \in \Phi}$ of densities or discrete distributions, e.g. the normal or the Poisson distribution. Thus, the unknown parameters in an HMM involve both the transition probabilities of the Markov chain and the parameters of the state-dependent distributions. The major approach to estimate the parameters in an HMM is by using likelihood-based methods. For general HMMs, strong consistency of the maximum-likelihood estimator (MLE) was proved by Leroux (1992). 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 *et al.*, 2004). Once these major results are obtained, the standard likelihood theory, such as asymptotic normality of the MLE with limit covariance $\mathcal{J}_0^{-1}$ (cf. Bickel *et al.*, 1998) and the asymptotic Chi-squared approximation to the distribution of the likelihood ratio test (LRT) under regularity conditions (cf. Giudici *et al.*, 2000), follows as in the i.i.d. setting.

In this paper, we observe that also the likelihood theory under non-standard conditions with parameters on the boundary, as developed by Chernoff (1954) and Self & 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 give 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.

Indeed, such testing situations are frequently encountered in practice when using HMMs. For example, if one wishes to test whether the state a is always left immediately, or whether the underlying Markov chain tends to stay in the state a, or whether the state a is on average more frequently visited than the state b, one requires testing for zero entries of the transition matrix, 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. Hence, all these testing problems and several others involve the extended asymptotic distribution of the LRT with true parameter on the boundary of the hypothesis.

The paper is organized as follows. In section 2, after formally introducing HMMs, we briefly discuss how the asymptotic distribution theory for the LRT under non-standard conditions can be extended from the i.i.d. case to HMMs. An extensive list of examples is given in section 3. In section 4, we present simulation results and illustrate the application of the tests for a series of epileptic seizure count data, previously analysed by Le *et al.* (1992). Some proofs are given in the Appendix.

2. Likelihood inference for HMMs

Let us start by introducing some further notation and definitions. Let $(X_k)_{k\geq 1}$ be a stationary, finite-state Markov chain with state space $\mathcal{M}=\{1,\dots,m\}$, transition probabilities $\alpha_{ab}=P(X_{k+1}=b\,|\,X_k=a)$ and unique stationary distribution $\pi=(\pi_1,\dots,\pi_m)$. Further, let $(Y_k)_{k\geq 1}$ be a stochastic process taking values in a Borel-measurable subset $\mathcal{Y}$ of Euclidean space, such that given $(X_k)_{k\geq 1}$, the $(Y_k)_{k\geq 1}$ are independent and the distribution of Y_k depends on X_k only. These conditional distributions are called the state-dependent distributions, and we assume that they come from a parametric family $\{f(y;\theta)\,|\,\theta\in\Phi\}$ of densities w.r.t. a σ-finite measure v on $\mathcal{Y}$, so that the distribution of Y_k, given that $X_k=a$, has density $f(\cdot;\theta_a)$. We assume that both the parameters of the transition matrix $\{\alpha_{ab}\}=\{\alpha_{ab}(\vartheta)\}$ and the parameters of the state-dependent densities $\theta_a=\theta_a(\vartheta)$ depend on a parameter $\vartheta\in\Theta\subset\mathbb{R}^d$. The standard parametrization is given by $\vartheta=(\alpha_{11},\dots,\alpha_{1,m-1},\alpha_{21},\dots,\alpha_{m,m-1},\theta_1,\dots,\theta_m)$. The subindex 0 indicates the true value ϑ_0 and the true distribution P_0 of the bivariate process $(X_k, Y_k)_{k\geq 1}$. Note that as the parameters of the transition matrix $\alpha_{ab}(\vartheta)$ depend on ϑ, so do the components of the unique stationary distribution $\pi_a=\pi_a(\vartheta)$.

The joint density of $(X_1,\dots,X_n,Y_1,\dots,Y_n)$ [w.r.t. (counting measure)$^n\times v^n$] is given by

$$p_n(x_1,\dots,x_n,y_1,\dots,y_n;\vartheta)=\pi_{x_1}(\vartheta)\prod_{k=1}^{n-1}\alpha_{x_k,x_{k+1}}(\vartheta)\prod_{k=1}^{n}f(y_k;\theta_{x_k}(\vartheta)),$$

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

$$p_n(y_1,\dots,y_n;\vartheta)=\sum_{x_1=1}^{m}\cdots\sum_{x_n=1}^{m}p_n(x_1,\dots,x_n,y_1,\dots,y_n;\vartheta),$$

and the log-likelihood is denoted by $L_n(\vartheta) = \log p_n(y_1, \ldots, y_n; \vartheta)$. An MLE $\hat{\vartheta}$ is any value of $\vartheta \in \Theta$ which maximizes $L_n(\vartheta)$.

Leroux (1992) gives conditions for HMMs with finite-state space and general observational space $\mathcal{Y}$ under which the MLE is strongly consistent, i.e. $\hat{\vartheta} \to \vartheta_0$ P_0 – almost surely as $n \to \infty$. Denote the matrix of second derivatives of $L_n(\vartheta)$ by $D^2_\vartheta L_n(\vartheta)$. Bickel *et al.* (1998) show that if $\tilde{\vartheta}_n$ is any strongly consistent sequence of estimates, under their assumptions 1–4,

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

as $n \to \infty$. The matrix $\mathcal{J}_0$ is called the Fisher information matrix. Douc *et al.* (2004) extend (1) to almost sure convergence. Bickel *et al.* (1998) further show under their assumptions 1–4 a central limit theorem for the score,

$$\sqrt{n} D_\vartheta L_n(\vartheta_0) \xrightarrow{\mathcal{L}} N(0, \mathcal{J}_0).$$

Once these fundamental results are established, the standard likelihood theory can be derived in complete analogy to the i.i.d. case. Bickel *et al.* (1998) conclude that if ϑ_0 is an interior point of Θ and if $\mathcal{J}_0$ is non-singular, $\sqrt{n}(\hat{\vartheta} - \vartheta_0)$ is asymptotically normally distributed with mean zero and covariance matrix $\mathcal{J}_0^{-1}$. Giudici *et al.* (2000) extend the Chi-squared approximation for the LRT for regular hypotheses from the i.i.d. setting to HMMs.

However, we observe that this strategy also applies when treating the LRT in case the true parameter is on the boundary of the hypothesis, cf. Chernoff (1954) and Self & Liang (1987) for the i.i.d. setting. To formulate the results, for $\Theta_0 \subset \Theta$ let

$$\lambda_n = \frac{\sup_{\vartheta \in \Theta_0} p_n(y_1, \ldots, y_n; \vartheta)}{\sup_{\vartheta \in \Theta} p_n(y_1, \ldots, y_n; \vartheta)}$$

be the likelihood ratio. If assumptions 1–4 in Bickel *et al.* (1998) hold true, if both Θ and Θ_0 satisfy assumption 1 (cf. the Appendix), if $\mathcal{J}_0$ is non-singular and if Θ can be approximated by a cone C_Θ at ϑ_0, and if Θ_0 can be approximated by a cone C_{Θ_0} at ϑ_0 (cf. the Appendix for the definition), then

$$T_n = -2 \log \lambda_n \xrightarrow{\mathcal{L}} \min_{z \in C_{\Theta_0} - \vartheta_0} (Z - z)^T \mathcal{J}_0 (Z - z) - \min_{z \in C_\Theta - \vartheta_0} (Z - z)^T \mathcal{J}_0 (Z - z), \tag{2}$$

where $Z \sim N(0, \mathcal{J}_0^{-1})$ and $\mathcal{J}_0$ is the Fisher information matrix in an HMM as given in (1). The asymptotic distribution in (2) in general depends on the Fisher information matrix. In some examples (cf. example 3 in section 3), 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 & Hughes, 2002). Alternatively, if one uses direct numerical maximization for computation of the MLE $\hat{\vartheta}$, most algorithms also give an estimate of the Hessian matrix at $\hat{\vartheta}$. From (1) 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) works best if the right-hand side does not depend on $\mathcal{J}_0$.

3. Examples

Example 1 (Zero entries of the transition matrix). 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 (9)). Therefore, testing for zeros in the transition matrix is evidently of some practical interest. Such boundary cases for LR testing were also studied by Bartolucci (2006) for latent

Markov models, and here we briefly discuss them in a general HMM framework. Consider the hypothesis that a certain entry α_{ab} of the transition matrix is zero:

$$H_{ab}^{0}:\alpha_{ab}=0 \quad \text{against } K_{ab}^{0}:\alpha_{ab}>0,$$

where it is assumed that under H_{ab}^{0}, the transition matrix still is ergodic, and the parameters of the state-dependent distributions are allowed to vary. In the case of two states, if $\alpha_{21}\neq 0$ and $\alpha_{22}\neq 0$ this is true for $\alpha_{11}=0$ but evidently not for $\alpha_{12}=0$. Under H_{ab}^{0}, from case 5 in Self & Liang (1987) it follows that

$$T_n \xrightarrow{\mathcal{L}} \tfrac{1}{2}\chi_0^2+\tfrac{1}{2}\chi_1^2, \tag{3}$$

where χ_0^2 is the point measure at 0. If one combines several of the H_{ab}^{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 (cf. Silvapulle & Sen, 2005, for a simulation procedure). Thus, a joint test would involve estimation of the Fisher information matrix. A simpler, although less efficient method would be to test several of the H_{ab}^{0}-type hypotheses by using some multiple testing procedure.

Example 2 (Boundary cases for parameters of the state-dependent distributions). Another possibility for model reduction is to test whether certain parameters of the state-dependent distributions are on the boundary of their parameter spaces.

As a first particular case, suppose that the state-dependent distributions are Bernoulli $B(p)$, and that the underlying Markov chain has two states. MacDonald & Zucchini (1997, pp. 140–144) consider the situation that in one of the states of the Markov chain (e.g. state 1) the outcome is just deterministic. This can be formulated by testing $H:p_1=0$ against $K:p_1>0$. The corresponding LRT again has the asymptotic distribution (3).

Another relevant case arises in the analysis of count data with overdispersion relative to the Poisson distribution (Leroux & Puterman, 1992). Such data are often modelled by finite mixtures of Poisson distributions, and HMMs then provide a natural generalization to the time-series context. Often, overdispersion mainly arises as there are too many zero observations (relative to Poisson), which is then modelled by zero inflation of the Poisson distribution (van den Broek, 1995). This can be interpreted as a two-component Poisson mixture, where $\lambda=0$ for one of the components. Thus, in the context of overdispersed count series, testing for zero inflation against general overdispersion structure can be accomplished by testing $H:\lambda_1=0$ against $K:\lambda_1>0$, with two-state underlying Markov chain and $\lambda_1<\lambda_2$. The LRT again has the asymptotic distribution (3).

Example 3 (One-sided tests for the transition probabilities). In this example, we consider one-sided hypotheses for the entries of the transition matrix. First consider

$$H:\alpha_{ac}\geq\alpha_{bc} \quad \text{against} \quad K:\alpha_{ac}<\alpha_{bc},$$

where $a,b,c\in\mathcal{M}$, i.e. that it is more probable under the alternative to have reached state c coming from state b than coming from state a. On the boundary of H, i.e. for $\alpha_{ac}=\alpha_{bc}$, T_n has the asymptotic distribution (3), and if $\alpha_{ac}>\alpha_{bc}$, by strong consistency of the MLE, $T_n\to 0$ in probability. Similarly, one can test $H:\alpha_{ab}\geq\alpha_{ac}$ against $K:\alpha_{ab}<\alpha_{ac}$, that under the alternative it is more likely to go from a to c than to b, or $H_{ab}^{q}:\alpha_{ab}\leq q$ against $K_{ab}^{q}:\alpha_{ab}>q$ for some $q\in(0,1)$ and $a,b\in\mathcal{M}$.

For the H_{ab}^{q}-type hypothesis, a relevant special case is when $a=b$ and $q=1/2$ because, in this case, under the alternative the HMM tends to stay in state $a\in\mathcal{M}$. We shall call such a state stable. Let us consider joint tests on two states $a,b\in\mathcal{M}$. First consider the testing problem

$$H_{a\wedge b,1/2} : \alpha_{aa} = 1/2 \wedge \alpha_{bb} = 1/2 \quad \text{against} \quad K_{a\wedge b} : \alpha_{aa} > 1/2 \wedge \alpha_{bb} > 1/2. \tag{4}$$

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. Using case 7 in Self & Liang (1987), one shows that under $H_{a\wedge b,1/2}$,

$$T_n \xrightarrow{\mathcal{L}} (\tfrac{1}{2} - p)\chi_0^2 + \tfrac{1}{2}\chi_1^2 + p\chi_2^2, \tag{5}$$

where the mixing quantity p can be evaluated as

$$p = \left(\cos^{-1}\rho\right)/(2\pi), \tag{6}$$

with ρ given in (12) in the Appendix, which also contains the proof.

The testing problem (4) 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

$$H_{a\wedge b} : \alpha_{aa} \le 1/2 \wedge \alpha_{bb} \le 1/2 \quad \text{against} \quad K_{a\wedge b} : \alpha_{aa} > 1/2 \wedge \alpha_{bb} > 1/2. \tag{7}$$

It turns out that in this testing problem, the asymptotic distribution given in (2) is no longer a finite mixture of Chi-squared distributions with different degrees of freedom (cf. Self & Liang, 1987, for an example where a nuisance parameter is on the boundary of the hypothesis). Here, the reason is that the whole parameter space under investigation, $\Theta = K_{a\wedge b} \cup H_{a\wedge b}$ is not convex. The asymptotic distribution of T_n for the testing problem (7) in the case of two states with known state-dependent distributions for $\alpha_{aa} = \alpha_{bb} = 1/2$ turns out to be

$$T_n \xrightarrow{\mathcal{L}} \frac{1}{2}\chi_0^2 + \frac{\pi - \phi}{2\pi}\chi_2^2 + \frac{\pi - \phi}{2\pi}P_1(\phi) + \frac{2\phi - \pi}{2\pi}P_2(\phi), \tag{8}$$

where $\phi = \cos^{-1}\rho \in [\pi/2, \pi)$, ρ is given in (12) and the densities of $P_1(\phi)$ and $P_2(\phi)$ are specified in (14) and (15) in the Appendix, which also contains the proof. For other parameter constellations under $H_{a\wedge b}$, the limit distribution is stochastically smaller. The asymptotic distributions (5) and (8) of the closely related testing problems (4) and (7) differ surprisingly strongly. As 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.

Example 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 (2) in principle allows to test one-sided hypotheses such as $\pi_a \ge \pi_b$ for $a, b \in \mathcal{M}$. However, the formulas for π in terms of the α_{ab} 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 the case of three states.

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 state-dependent parameters 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. $H : \pi_1 \ge \pi_2$ against $K : \pi_1 < \pi_2$. Evidently, H is equivalent to the linear restriction $\beta \ge \alpha$, and on the boundary of H, i.e. for $\alpha = \beta \ne 1$, one has the asymptotic distribution given in (3). Similarly, general restrictions $H_p : \pi_1 \le p$ for some $p \in (0, 1)$ can be formulated into linear restrictions $H_p : (1 - p)\beta \le p\alpha$, and on the boundary of the hypothesis, (3) applies as well.

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

$$\begin{pmatrix} 1-\alpha & \alpha & 0 \\ \beta & 1-\beta-\gamma & \gamma \\ 0 & \delta & 1-\delta \end{pmatrix}. \tag{9}$$

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 on the parameters of the transition matrix. For example, consider the hypothesis $H_{1,3}: \pi_1 \geq \pi_3$, which is equivalent to $H_{1,3}: \delta\beta \geq \gamma\alpha$, and one again obtains the asymptotic (3) for T_n. Further, if one wishes to test $H_{1,2,3}: \pi_1 = \pi_3, \pi_1 > \pi_2$, which is equivalent to $H_{1,2,3}: \delta\beta = \gamma\alpha, \beta > \alpha$, then (cf. case 6 in Self & Liang, 1987)

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

Example 5 (Testing for bimodality). Suppose that we have a two-state Gaussian HMM. Then the marginal distribution will be a two-component mixture of normals, where the mixing proportions are given by $\pi_1 = \beta/(\alpha+\beta)$ and $\pi_2 = 1 - \pi_1$, using the notation in example 4. A two-component mixture of normals is either unimodal or bimodal, and explicit characterizations for unimodal parameter regions are available (Robertson & Fryer, 1969). For example, for $\pi_1 = 1/2$, the mixture is unimodal if and only if $|\mu_1 - \mu_2| \leq 2\sigma$. Using these facts, one can construct an LRT for bimodality of the marginal distribution of a two-state normal HMM by using theorem 1. To this end, the model has to be reparameterized in (π_1, β) instead of (α, β), and the restriction of unimodality can then be formulated in terms of π_1 and the parameters of the state-dependent distributions. Let Θ_0 be the unimodal parameter region. Then, for all $\vartheta \in \Theta_0$, the asymptotic distribution of T_n is stochastically smaller than $(\chi_0^2 + \chi_1^2)/2$, and for some parameter values on the boundary of Θ_0 there is equality. Therefore, a test based on the critical value of $(\chi_0^2 + \chi_1^2)/2$ asymptotically keeps the level for all $\vartheta \in \Theta_0$. The test can be extended to other families of state-dependent distributions for which a characterization of the modality in two-component mixtures is available (e.g. the von Mises distribution, cf. Mardia & Sutton, 1975, or the multivariate normal distribution, cf. Ray & Lindsay, 2005).

4. Simulations and empirical illustration

4.1. Quality of asymptotic approximation for the LRT

As advocated by MacDonald & Zucchini (1997), for the numerical computation of the maximum-likelihood estimates, we use direct maximization by using a Newton-type algorithm. In each setting, we reparametrize the problem so that unconstrained maximization is possible.

(a) Consider testing the hypothesis $H: \alpha_{11} = 0$ in a stationary two-state normal HMM, as described in example 1, where the asymptotic distribution for T_n is given in (3). The transition matrix is taken as

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

and for the parameters of the state-dependent distributions we choose $\sigma^2 = 1$ and mean values $\mu_1 = 0$, $\mu_2 = 2$ in the first setting, which corresponds to sufficiently well-separated state-dependent distributions, and $\mu_1 = 0$, $\mu_2 = 1$ in the second setting, where the state-dependent distributions strongly overlap. In the simulation, we fix the parameters of the state-dependent distributions at their true values, and let only the parameters of the transition matrix vary. We generate $N = 10{,}000$ samples of various sizes, and for visualization 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 are displayed in Figs 1 and 2. It turns out that the asymptotic approximation for well-separated state-dependent distributions is relatively poor, even for large sample sizes such as $n = 500$ and in this simple situation with fixed parameters for the state-dependent

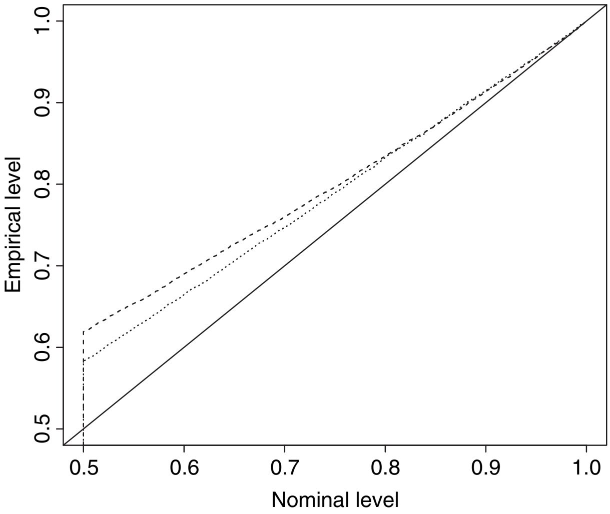

Fig. 1. PP plot of distribution of T_n for $n = 100$ (dashed line) and $n = 500$ (dotted line) for hypothesis $H: \alpha_{11} = 0$ in the case $\mu_1 = 0$, $\mu_2 = 2$.

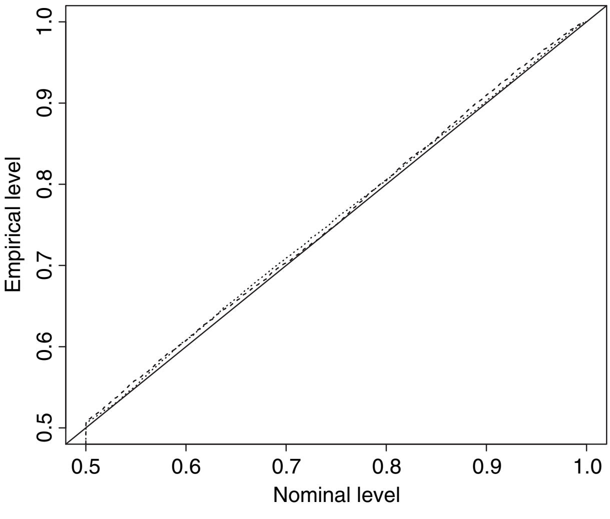

Fig. 2. PP plot of distribution of T_n for $n = 50$ (dashed line) and $n = 100$ (dotted line) for hypothesis $H: \alpha_{11} = 0$ in the case $\mu_1 = 0$, $\mu_2 = 1$.

distributions, while for strongly overlapping state-dependent distributions the approximation is quite good already for $n=50$.

(b) Further, we examine testing the hypothesis $H_{1,3}: \pi_1 \geq \pi_3$ as described in example 4 for a stationary three-state Poisson HMM. We use a transition matrix as in (9), with $\alpha=0.4$, $\beta=0.2$, $\gamma=0.3$ and $\delta=0.6$, yielding for the state-dependent distribution $\pi_1=\pi_3=0.25$, $\pi_2=0.5$. The parameters of the state-dependent distributions 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 (3).

First, we consider the described model for fixed and known values of the λs. Secondly, the λs are considered as unknown parameters one has to estimate. For both cases, we generate $N=5000$ samples of sizes $n=100$ and 500. The results are displayed in Figs 3 and 4 using PP plots. For fixed λs, the approximation is quite satisfactory even for small sample sizes ($n=100$). Naturally, estimation of the λs increases variation of the model. But for large sample sizes ($n=500$) the approximation is quite good in this case too.

4.2. 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 (Expectation Maximization) algorithm as suggested in Baum *et al.* (1970), Le *et al.* (1992) fit such models to the series of daily counts of epileptic seizures in one patient participating in a clinical trial at British Columbia's Children's Hospital. The originally published series consists of 225 observations; however, as indicated in MacDonald & Zucchini (1997, p. 147), observations 92–112 should be deleted; thus, we use the corrected data set of 204 observations.

In the neurology literature, Hopkins *et al.* (1985) proposed that the variation in seizure occurrences and its dependency structure could be modelled by a Markov chain. This is 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

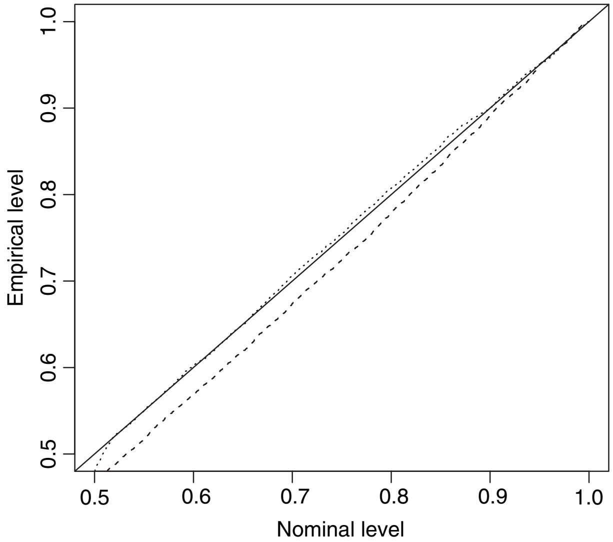

Fig. 3. PP plot of distribution of T_n for $n=100$ (dashed line) and $n=500$ (dotted line) for hypothesis $H_{1,3}$ in the case of fixed and known λs.

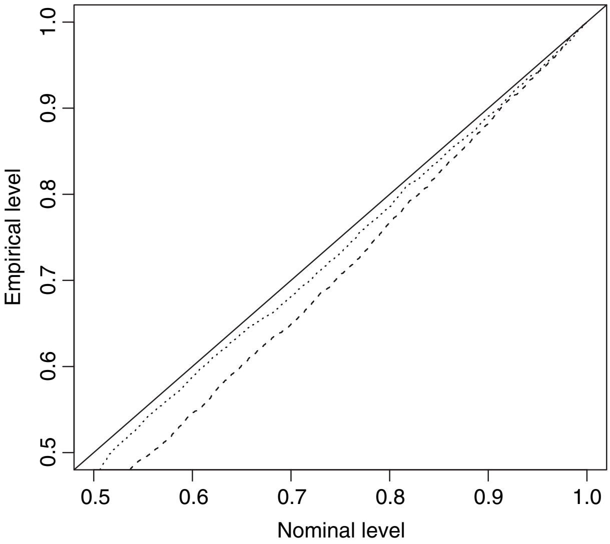

Fig. 4. PP plot of distribution of T_n for $n=100$ (dashed line) and $n=500$ (dotted line) for hypothesis $H_{1,3}:\alpha_{11}=0$ in the case of estimated λs.

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 & 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. Here, the asymptotic distribution of the likelihood ratio statistic T_n under the hypothesis H follows (3). 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 for the stationary distribution $\hat{\pi}=(0.567, 0.433)$. Therefore, the estimate indicates that state 1 with low seizure susceptibility is, on average, more frequently visited than state 2. 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 (3). For this test, the likelihood ratio statistic is $T_n=0.111$ with a corresponding p-value of 0.369. Hence, we cannot reject the hypothesis H at the 5% level; thus, there is not enough evidence that state 1 is more often visited than state 2.

Acknowledgements

Jörn Dannemann acknowledges financial support from the Georg Lichtenberg programme 'Applied Statistics & Empirical Methods'. Hajo Holzmann gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft, grant MU 1230/8-2. The authors would like to thank Axel Munk and Walter Zucchini as well as the editor, the associate editor and two referees for helpful comments.

References

Albert, P. S. (1991). A two-state Markov mixture model for time series of epileptic seizure counts. *Biometrics* **47**, 1371–1381.

Arribas-Gil, A., Gassiat, E. & Matias, C. (2006). Parameter estimation in pair-hidden Markov models. *Scand. J. Statist.* **33**, 651–671.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypothesis on the transition probabilities. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **68**, 155–178.

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41**, 164–171.

Bickel, P. J., Ritov, Y. & 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. & 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.

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

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.* **25**, 573–578.

Douc, R., Moulines, E. & Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.* **32**, 2254–2304.

Giudici, P., Rydén, T. & Vandekerkhove, P. (2000). Likelihood-ratio tests for hidden Markov models. *Biometrics* **56**, 742–747.

Haut, S. R. (2006). Seizure clustering. *Epilepsy Behav.* **8**, 50–55.

Holzmann, H., Munk, A., Suster, M. & Zucchini, W. (2006). Hidden Markov models for circular and linear–circular time series. *Environ. Ecol. Stat.* **13**, 325–347.

Hopkins, A., Davies, P. & Dobson, C. (1985). Mathematical models of patterns of seizures: their use in the evaluation of drugs. *Arch. Neurol.* **42**, 463–467.

Le, N. D., Leroux, B. G. & 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. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.* **40**, 127–143.

Leroux, B. G. & Puterman, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48**, 545–558.

Lystig, T. C. & Hughes, J. P. (2002). Exact computation of the observed information matrix for hidden Markov models. *J. Comput. Graph. Statist.* **11**, 678–689.

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

Mardia, K. V. & Sutton, T. W. (1975). On the modes of a mixture of two von Mises distributions. *Biometrika* **62**, 699–701.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286.

Ray, S. & Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *Ann. Statist.* **33**, 2042–2065.

Robertson, C. A. & Fryer, J. G. (1969). Some descriptive properties of normal mixtures. *Skandinavisk Aktuarietidskrift* **52**, 137–146.

Rydén, T., Terasvirta, T. & Asbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *J. Appl. Econometrics* **13**, 217–244.

Self, S. G. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82**, 605–610.

Silvapulle, M. J. & Sen, P. K. (2005). *Constrained statistical inference*. Wiley, Hoboken, NJ.

Van den Broek, J. (1995). Score test for zero inflation in a Poisson distribution. *Biometrics* **51**, 738–743.

Received May 2007, in final form October 2007

Hajo Holzmann, Institut für Stochastik, Universität Karlsruhe (TH), Englerstr. 2, 76128 Karlsruhe, Germany.
E-mail: holzmann@stoch.uni-karlsruhe.de

Appendix

Definition 1

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

$$\inf_{z \in C_\Theta} \|z - y\| = o(\|y - \vartheta_0\|)$$

for all $y \in \Theta$ and if

$$\inf_{y \in \Theta} \|z - y\| = o(\|z - \vartheta_0\|)$$

for all $z \in C_\Theta$.

A cone C with vertex at ϑ_0 is such that if $z \in C$, we also have $a(z - \vartheta_0) + \vartheta_0 \in C$ for all $a > 0$.

Assumption 1

We have $\vartheta_0 \in \Theta$, and the MLE *over Θ is strongly consistent.*

Proof of (6). Set $a = 1, b = 2$. From (10.29), p. 362 in Cappé *et al.* (2005), we have to compute the asymptotic covariance matrix of the score vector

$$E_0\left(\nabla_\vartheta \log(\pi_{X_1}(\vartheta)) \mid Y_{1:n}\right) + \sum_{k=1}^{n-1} E_0\left(\nabla_\vartheta \log(\alpha_{X_k, X_{k+1}}(\vartheta) \mid Y_{1:n}\right), \tag{10}$$

where $Y_{1:n} = (Y_1, \dots, Y_n)$. Let $\alpha_{12} = \alpha$ and $\alpha_{21} = \beta$, so that $\vartheta = (\alpha, \beta)$. Neglect the first term in (10) for the moment. The derivative $\partial/\partial\alpha$ ($\partial/\partial\beta$ is similar) of the second term are computed as

$$\frac{\partial}{\partial\alpha} \log\left(\alpha_{x,x'}(\vartheta)\right) = \frac{1}{\alpha}\delta_{(1,2)}(x, x') - \frac{1}{1-\alpha}\delta_{(1,1)}(x, x'),$$

so that the relevant sum in (10) (the derivative w.r.t. α in the second term) is

$$\sum_{k=1}^{n-1} E_0\left(\frac{1}{\alpha}\delta_{(1,2)}(X_k, X_{k+1}) - \frac{1}{1-\alpha}\delta_{(1,1)}(X_k, X_{k+1}) \mid Y_{1:n}\right), \tag{11}$$

where

$$\delta_x(y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise.} \end{cases}$$

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

$$\delta_{(i,j)}(X_k, X_{k+1}) = \delta_i(X_k)\delta_j(X_{k+1}), \quad i,j \in \{1, 2\},$$

(11) can be rearranged as

$$2\sum_{k=1}^{n-1} E_0\left(\delta_1(X_k) \mid Y_k\right)\left(E_0\left(\delta_2(X_{k+1}) \mid Y_{k+1}\right) - E_0\left(\delta_1(X_{k+1}) \mid Y_{k+1}\right)\right),$$

and a similar expression can be obtained for $\partial/\partial\beta$. Now

$$E_0\left(\delta_i(X_k) \mid Y_k\right) = \frac{f(Y_k; \theta_i)}{f(Y_k; \theta_1) + f(Y_k; \theta_2)}, \quad i \in \{1, 2\}.$$

Introducing the random variables

$$Z_{k,1}=\frac{f(Y_k;\theta_1)}{f(Y_k;\theta_1)+f(Y_k;\theta_2)}\frac{f(Y_{k+1};\theta_2)-f(Y_{k+1};\theta_1)}{f(Y_{k+1};\theta_1)+f(Y_{k+1};\theta_2)},$$
$$Z_{k,2}=\frac{f(Y_k;\theta_2)}{f(Y_k;\theta_1)+f(Y_k;\theta_2)}\frac{f(Y_{k+1};\theta_1)-f(Y_{k+1};\theta_2)}{f(Y_{k+1};\theta_1)+f(Y_{k+1};\theta_2)},$$

the components of the score are given by

$$S_{n,1}=2\sum_{k=1}^{n-1}Z_{k,1},\quad S_{n,2}=2\sum_{k=1}^{n-1}Z_{k,2}.$$

Although the $Z_{k,1}$s are not i.i.d., the covariances for different ks are 0, and the correlation can be computed as

$$\rho=\frac{E_0Z_{k,1}Z_{k,2}}{(E_0Z_{k,1}^2E_0Z_{k,2}^2)^{1/2}}. \tag{12}$$

For the first term in (10), one has that

$$E_0\left(\nabla_\vartheta\log(\pi_{X_1}(\vartheta))\mid Y_{1:n}\right)=E_0\left(\nabla_\vartheta\log(\pi_{X_1}(\vartheta))\mid Y_1\right),$$

of which the contribution to the asymptotic covariance matrix is zero. This proves (12), and (6) follows from case 7 in Self & Liang (1987). Note that this example generalizes example 4.3 in Bickel *et al.* (2002), who consider reversible Markov chains and thus only have a single parameter.

Proof of (8). We use (2) and a coordinate transformation to obtain

$$\begin{aligned}T_n\xrightarrow{\mathcal{L}}T&:=\min_{z\in H}(Z-z)^T(Z-z)-\min_{z\in H\cup K}(Z-z)^T(Z-z)\\&=\|Z-Z_H\|^2-\min\left(\|Z-Z_H\|^2,\|Z-Z_K\|^2\right)\end{aligned} \tag{13}$$

where Z is a bivariate standard normal random variable,

$$H=\mathcal{J}_0^{1/2}(\mathbb{R}^-\times\mathbb{R}^-)\quad\text{and}\quad K=\mathcal{J}_0^{1/2}(\mathbb{R}^+\times\mathbb{R}^+),$$

as shown in Fig. 5, Z_H and Z_K denote the orthogonal projections of Z onto H and K, respectively, and $\phi=\cos^{-1}\rho$.

First, we calculate the distribution of T conditional on the event $\{Z\in\text{region 1 or }1a\}$ and the weight $P(Z\in\text{region 1 or }1a)$. If Z is in region 1, then obviously both terms in (13) are zero. For region $1a$, the difference is zero as well, because $\|Z-Z_H\|\leq\|Z-Z_K\|$. Hence, the conditional distribution of T in region 1 and $1a$ is χ_0^2. 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 conditional on $\{Z\in\text{region 2}\}$ is determined by the first term, which gives a χ_2^2 distribution, as $Z_H=0$. The weight is given by $(\pi-\phi)/2\pi$. Observe that 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

$$\|Z-Z_H\|^2=r^2\sin^2(\pi-\psi)\quad\text{and}\quad\|Z-Z_K\|^2=r^2\sin^2(\psi-\phi).$$

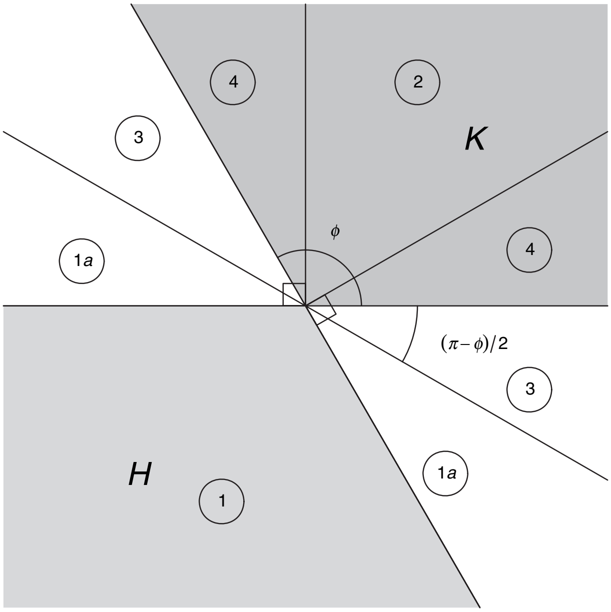

Fig. 5. Diagram of the parameter space of example 3.

Setting

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

for fixed ψ the difference

$$T(\psi)=\|Z-Z_H\|^2-\|Z-Z_K\|^2=r^2a_1(\psi,\phi)$$

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

$$h_1(t;\phi)=\frac{2}{\pi-\phi}\int_{\phi}^{(\phi+\pi)/2}\frac{1}{2a_1(\psi,\phi)}\exp\left(-\frac{t}{2a_1(\psi,\phi)}\right)\mathrm{d}\psi. \tag{14}$$

The weight of region 3 is given by $(\pi-\phi)/2\pi$.

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

$$h_2(t;\phi)=\frac{2}{2\phi-\pi}\int_{\pi/2}^{\phi}\frac{1}{2a_2(\psi)}\exp\left(-\frac{t}{2a_2(\psi)}\right)\mathrm{d}\psi \tag{15}$$

with weight

$$2\frac{\phi-(\pi/2)}{2\pi}=\frac{2\phi-\pi}{2\pi}.$$