The likelihood ratio test for hidden Markov models in two-sample problems

Jörn Dannemann, Hajo Holzmann *

Institute for Mathematical Stochastics, Georg-August-University of Göttingen, Maschmühlenweg 8–10, 37073 Göttingen, Germany

Abstract

The asymptotic distribution of the likelihood ratio test statistic in two-sample testing problems for hidden Markov models is derived when allowing for unequal sample sizes as well as for different families of state-dependent distributions. In both cases under regularity conditions the limit distribution is a standard χ^2–distribution, and in particular does not depend on the ratio of the distinct sample sizes. In a simulation study, the finite sample properties are investigated, and the methodology is illustrated in an application to modeling the movement of Drosophila larvae.

Key words: circular data, hidden Markov model, two-sample problem, likelihood ratio test

1 Introduction

Hidden Markov models (HMMs) are a class of stochastic processes applied in various fields of dependent data analysis (see Cappé et al. (2005) and MacDonald and Zucchini (1997) for descriptions of diverse applications). An HMM consists of two ingredients, an unobservable finite-state Markov chain (X_k) with state space $\mathcal{M}$ 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. Typically it is assumed that the state dependent distributions, i.e. the distributions of Y_k given that $X_k = i$, $i \in \mathcal{M}$, come from a parametric family $(f_\theta)_{\theta\in\Phi}$ of densities or discrete distributions. Therefore, the unknown parameters in an HMM involve both the transition probabilities of the Markov chain and the parameters of the state dependent distributions. Note that since the Markov chain (X_k) is unobservable, inference has to be based on the (Y_k) alone.

The major approach to estimate the parameters in an HMM is via likelihood-based methods. Strong consistency of the maximum likelihood estimator (MLE) was

* Corresponding author. Tel.: +49 551 39 13511; fax: +49 551 39 13505.
Email address: holzmann@math.uni-goettingen.de (Hajo Holzmann).

proved by Leroux (1992). Bickel et al. (1998) established the asymptotic normality of the MLE under Cramér-type conditions. The likelihood-ratio test (LRT) for HMMs in the one-sample case was studied by Giudici et al. (2000). They show that under regularity conditions both on the HMM and on the structure of the hypothesis, the standard asymptotic χ^2–distribution occurs under the hypothesis.

In this paper we study the LRT for HMMs in two-sample problems. Such testing problems often arise in practical applications when using HMMs. For example, MacDonald and Zucchini (1997) model the behavior of 24 locusts (locusta migratoria) via multivariate HMMs. These locusts were separated into two groups, according to whether they were recently fed or not, and a multivariate HMM was fitted to each group consisting of 12 animals. MacDonald and Zucchini (1997) concluded from the parameter estimates that the behavior of fed and unfed subjects differs; such a statement can be formally tested by a two-sample LRT for equality of the parameters. This situation occurs if one models two (or more) time series via HMMs, where the time series are related from the context (e.g. movement of two animals from the same species), but where the actual parameter values are expected to differ (e.g. fed and unfed animals).

On the other hand, if one models several time series via HMMs, where it is reasonable to assume that the HMMs, or at least some relevant parameters coincide (e.g. movement of two animals from the same species under equal conditions), the LRT should not reject the corresponding hypothesis if the HMMs provide a good fit. Here we consider the circular time series of direction changes of several Drosophila larvae, which were previously analyzed by means of HMMs by Holzmann et al. (2006). It turns out that at least the test for equality of the transition parameters of two distinct series can typically not be rejected.

Rydén et al. (1998) divide a long time series of the S&P 500 US stock price index into 10 subseries, and fit HMMs to each of these subseries. Using pairwise tests, one can now test for structural changes between series for different periods. Although independence between the samples will in general not hold exactly in this context, if the two subseries to be compared are sufficiently separated, they will typically be approximately independent. While Rydén et al. (1998) use subseries of equal length, in general it might also be reasonable to use breaks which lead to subseries of different lengths.

When discussing the LRT for HMMs in two-sample problems, we will allow for different sample sizes as well as for different parametric families of state-dependent distributions, and study the LRT for joint multi-dimensional restrictions on the parameters of both HMMs. It turns out that as long as both sample sizes are of the same order, the asymptotic distribution is still a standard χ^2–distribution and does not depend on the ratio of the sample sizes. Our results apply in particular to situations with equal state spaces and equal parametric families for the state-dependent distributions, if it is the purpose to test for equality of some of the parameters. In principle, all of our results could be extended to the multi-sample case, however, for simplicity of presentation we restrict ourselves to the two-sample case.

Let us mention that Michalek et al. (2001) considered the LRT in the two-sample case for two-state Gaussian HMMs with equal sample sizes. They reduced the problem to the one-sample case by constructing a superimposed HMM via adding the two time series. However, this method only works in very special situations, in particular it requires equal sample sizes and linear data, while we treat the problem in full generality.

The paper is organized as follows. In Section 2 we introduce some further notation and definitions. Section 3 contains the asymptotic distribution theory for the LRT in the two-sample case. The finite sample properties are investigated in Section 4 by means of a simulation study. In Section 5, we give an application to modeling the circular time series of direction changes of Drosophila larval movement. The data set can be obtained from
http://www.stochastik.math.uni-goettingen.de/pub/
Formal assumptions and proofs are deferred to an appendix.

2 Notation

Here we introduce some further notation. Let $(X_k)_{k\geq 1}$ be a stationary, finite-state Markov chain with state space $\mathcal{M} = \{1, \ldots, m\}$, transition probabilities $\alpha_{ab} = P(X_{k+1} = b|X_k = a)$, and unique stationary distribution $\pi = (\pi_1, \ldots, \pi_m)$. 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, we assume that they come from a parametric family $\{f(y; \theta)|\, \theta \in \Phi\}$ of densities w.r.t. a σ-finite measure ν 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

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

The subindex 0 indicates the true value ϑ_0 and the true distribution P_0 of the bivariate process $(X_k, Y_k)_{k\geq 1}$.

In the following we will consider two independent HMMs, so that the bivariate processes $(X_k^1, Y_k^1)_{k\geq 1}$ and $(X_k^2, Y_k^2)_{k\geq 1}$ are independent. In general we allow different image spaces $\mathcal{Y}_j$, state spaces $\mathcal{M}_j$, different parametric families $\left\{f^j(y; \theta)|\, \theta \in \Phi^j\right\}$, $j = 1, 2$, for the state-dependent densities and also different parameterizations $\vartheta^j \in \Theta^j \subset \mathbb{R}^{d^j}$. If we assume equality of these objects, we call the two HMMs *homogeneous* and drop the j-index from the notation. In this case we also assume that they have the same parametrization. Let us stress that homogeneous does not mean that the two HMMs have the same distributions, since the actual parameter values are allowed to differ.

3 Two-sample likelihood ratio tests for HMMs

In this section we study the asymptotic distribution of the LRT statistic for HMMs in two-sample problems.

Let $\left(X_k^j, Y_k^j\right)_{k\geq 1}$, $j = 1, 2$, be two independent HMMs. One observes $Y_1^1, \ldots, Y_{n_1}^1$

and $Y_1^2, \ldots, Y_{n_2}^2$, and we assume that for $n = n_1 + n_2$ there is a $c \in (0,1)$ such that as $n_1, n_2 \to \infty$,

$$\frac{n_1}{n} \to c, \qquad \frac{n_2}{n} \to 1 - c, \tag{1}$$

thus, n_1 and n_2 are of the same order of magnitude, i.e. $n_1/n_2 \to c/(1-c) \in (0, \infty)$. Let

$$L_{n_j}^j(\vartheta^j) := \log p_{n_j}(Y_1^j, \ldots, Y_{n_j}^j; \vartheta^j), \quad j = 1, 2,$$

denote the log-likelihood functions, as given in Eq. (1) in Bickel et al. (1998). If both HMMs satisfy Assumptions (A1)-(A4) (see the appendix), the normalized second derivatives of the log-likelihood functions, evaluated at sequences $\vartheta_n \to \vartheta_0$, converge P_0^j-almost surely (a.s.) to constant matrices $-\mathcal{J}_0^j$, where the $\mathcal{J}_0^j$ are called the Fisher information matrices, cf. Lemma 2 in Bickel et al. (1998). Further let

$$L_n^c(\vartheta^1, \vartheta^2) = L_{n_1}(\vartheta^1) + L_{n_2}(\vartheta^2) = \log p_{n_1}(Y_1^1, \ldots, Y_{n_1}^1; \vartheta^1) + \log p_{n_2}(Y_1^2, \ldots, Y_{n_2}^2; \vartheta^2)$$

be the joint log-likelihood function. We want to test joint hypotheses on the parameters ϑ^j, $j = 1, 2$. To this end assume that the joint hypothesis H_0 can be expressed as a regular r-dimensional restriction of the form $H_0 : R(\vartheta^1, \vartheta^2) = 0$, where R is a two-times differentiable map into $\mathbb{R}^r$ with continuous derivatives from some neighborhood around the true parameter $(\vartheta_0^1, \vartheta_0^2)$, i.e.

$$R \in C^2(B, \mathbb{R}^r), \qquad B := B_\delta(\vartheta_0^1, \vartheta_0^2) \subset \Theta^1 \times \Theta^2 \quad \text{for some } \delta > 0.$$

Denote by $\hat{\vartheta}_n^j$, $j = 1, 2$, the unrestricted MLEs and by $\tilde{\vartheta}_n^j$, $j = 1, 2$ the MLEs restricted to the parameter values under the hypothesis, i.e.

$$(\hat{\vartheta}_n^1, \hat{\vartheta}_n^2) := \arg \max_{(\vartheta^1, \vartheta^2) \in \Theta^1 \times \Theta^2} L_n^c(\vartheta^1, \vartheta^2), \qquad (\tilde{\vartheta}_n^1, \tilde{\vartheta}_n^2) := \arg \max_{\{R(\vartheta^1, \vartheta^2) = 0\}} L_n^c(\vartheta^1, \vartheta^2).$$

In the following we will assume that if the hypothesis is true, i.e. if $R(\vartheta_0^1, \vartheta_0^2) = 0$, both the unrestricted and the restricted MLEs are strongly consistent. In case of a single HMM Leroux (1992) proved the consistency of the MLE under regularity conditions. For the restricted MLE consistency can be shown in a similar fashion, for details see Dannemann (2006). The asymptotic distribution of the LRT is described in the following result.

Theorem 1. *Assume that (A1) - (A6) (cf. the appendix) hold for two independent HMMs $(X_k^j, Y_k^j)_{k \geq 1}$, $j = 1, 2$, respectively and that the Fisher information matrices $\mathcal{J}_0^j$, $j = 1, 2$, are non-singular. Further assume that under the hypothesis given by a regular, r-dimensional restriction $H_0 : R(\vartheta^1, \vartheta^2) = 0$, the restricted MLE is strongly consistent. Then under H_0, the likelihood-ratio test statistic λ_n is asymptotically χ_r^2–distributed, i.e.*

$$\lambda_n = 2\left(L_n^c(\hat{\vartheta}_{n_1}^1, \hat{\vartheta}_{n_2}^2) - L_n^c(\tilde{\vartheta}_{n_1}^1, \tilde{\vartheta}_{n_2}^2)\right) \xrightarrow{\mathcal{D}} \chi_r^2,$$

under $P_0^c = P_0^1 \otimes P_0^2$ as $n_1, n_2 \to \infty$ such that (1) holds.

Thus, the asymptotic distribution does not depend on the ratio of the sample sizes. The formal proof of the theorem is given in the appendix.

Remark 2. For equal sample sizes $n_1 = n_2$ the theorem can easily be reduced to the one-sample case by considering a superimposed HMM. In fact, the bivariate process $\mathbb{X} = (X_k^1, X_k^2)_{k\geq 1}$ is a stationary, finite-state Markov chain and further, $\mathbb{Y} = (Y_k^1, Y_k^2)_{k\geq 1}$ satisfies the properties of an HMM w.r.t. $\mathbb{X}$. Therefore, in this situation the result follows from the one-sample version (Giudici et al., 2000). Note that this way of constructing a superimposed HMM differs from the one suggested by Michalek et al. (2001).

Remark 3. Suppose that the two HMMs are homogeneous (i.e. have equal state spaces and families of state-dependent distributions). Using Theorem 1 we can construct a test for equality of some or all components of the parameters $H_0 : \vartheta^{1,i} = \vartheta^{2,i}$, $i = 1, \ldots, r$ with $r \leq d$. In this case, the LRT is asymptotically χ^2–distributed with r degrees of freedom (d.f.). In particular, for $r = d$,

$$\lambda_n = 2\left(\sup_{\vartheta^1,\vartheta^2} L_n^c(\vartheta^1, \vartheta^2) - \sup_{\vartheta} L_n^c(\vartheta, \vartheta)\right) \xrightarrow{\mathcal{D}} \chi_d^2.$$

To obtain a test with nominal level α we reject H_0 for $\lambda_n > \chi^2_{d,1-\alpha}$.

Remark 4. Even if the two HMMs are non-homogeneous, there are relevant problems leading to joint tests. For example, if the families of state-dependent distributions differ but if the state-spaces are the same, using Theorem 1 one can construct tests on the parameters of the transition matrices of the Markov chains, e.g. for equality of these parameters.

Remark 5. As Giudici et al. (2000) point out, the standard asymptotic χ^2–distribution for the LRT does not apply when testing for the number of states of an HMM in the one-sample case. Similarly, for the two-sample situation, standard asymptotics do not hold when testing for equality of the number of states.

4 Simulation experiments

We now investigate the finite sample properties of the proposed tests in a simulation study.

As advocated by MacDonald and Zucchini (1997) and Altman (2007), for the numerical computation of the maximum likelihood estimates we use direct maximization via a Newton-type algorithm, specifically the function *nlm* of the software package *R*. This is somewhat faster and simpler than the EM algorithm, although the results depend more sensitively on the starting values, and in practice one should use several distinct starting values in order to find the global maximum. However, in simulations this causes no problem since we can simply choose the true values as starting values. Furthermore, for the simulation the Markov chains are assumed to be in equilibrium, i.e. the stationary distribution π is used as initial distribution.

We consider two two-state Gaussian HMMs with transition matrices $\alpha_{12}^j = 0.4$, $\alpha_{21}^j = 0.2$, $j = 1, 2$, mean values $\mu_1^j = 0$, $\mu_2^j = 2$, $j = 1, 2$ and variances $(\sigma_i^j)^2 = 1$, $i, j = 1, 2$. The standard parametrization $\vartheta = (\vartheta^1, \vartheta^2)$ with $\vartheta^1 = (a_{12}^1, a_{21}^1, \mu_1^1, \mu_2^1)$ and $\vartheta^2 = (\alpha_{12}^2, \alpha_{21}^2, \mu_1^2, \mu_2^2)$ and fixed, known variance $\sigma^2 = 1$ is used.

First we use equal sample sizes n. Under the hypothesis $H_0 : \vartheta^1 = \vartheta^2$, the LRT statistic is χ^2_4–distributed, as shown in Theorem 1. We simulate $N = 5000$ times the statistic for sample sizes $n = 50, 100, 200$. For visualization we use P–P plots, which show for each $\alpha \in [0, 1]$ the empirical probability that the LRT statistic λ_n is $\leq q_\alpha$, where q_α is the α-quantile of the corresponding limit χ^2–distribution. Figure

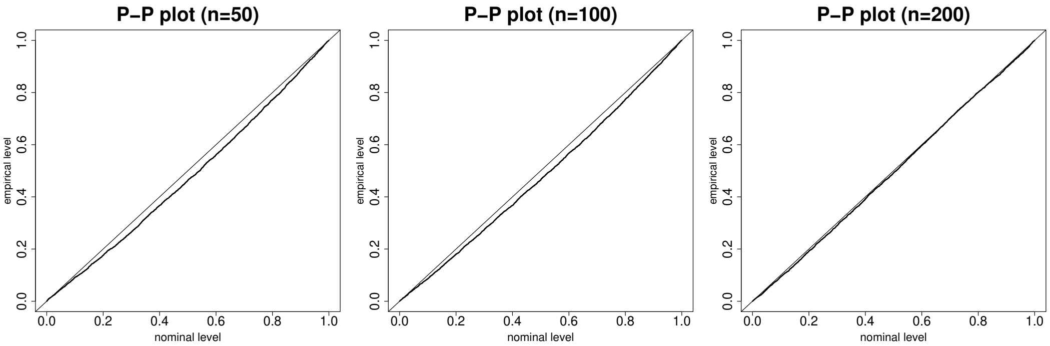

Fig. 1. P-P plots of the LRT statistic testing for equality of parameters of two HMMs with sample sizes $n = 50$, $n = 100$ and $n = 200$. (Number of replications: $N = 5000$)

1 shows the results. The fit is quite good, only for small sample sizes the LRT is somewhat anti-conservative.

Next we consider a test for equality of some of the parameters, where the additional parameters are treated as nuisance parameters. We test for equality of the regime, i.e. $H_0 : (\alpha^1_{12}, \alpha^1_{21}) = (\alpha^2_{12}, \alpha^2_{21})$. Under the hypothesis H_0 the LRT statistic is χ^2_2–distributed. Sample sizes and replications are chosen as in the first simulation. Figure 2 shows P-P plots of the results. The fit is even better than for the test of

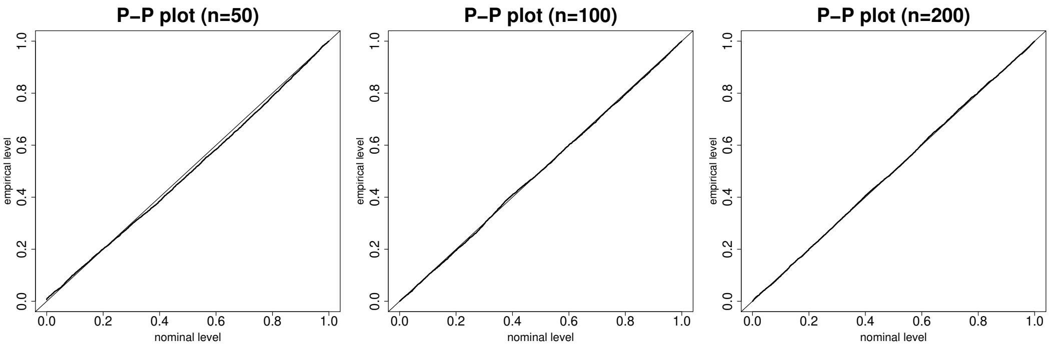

Fig. 2. P-P plots of the LRT statistic testing for equality of the transition probabilities of two HMMs with sample sizes $n = 50$, $n = 100$ and $n = 200$. (Number of replications: $N = 5000$)

equality of all parameters.

Finally, we construct a test in case of unequal sample sizes. In the above model we test for equality of all parameters, and use sample sizes $n_1 = 50$ and $n_2 = 50, 100, 500$. Figure 3 shows P-P plots for this testing problem. The fit is also quite good, and the asymptotic distribution evidently does not depend on the ratio of the sample sizes.

We also considered several other values for the transition probability matrices. Firstly, the situation of rather persistent states, $\alpha^j_{12} = 0.1$, $\alpha^j_{21} = 0.05$, $j = 1, 2$, was

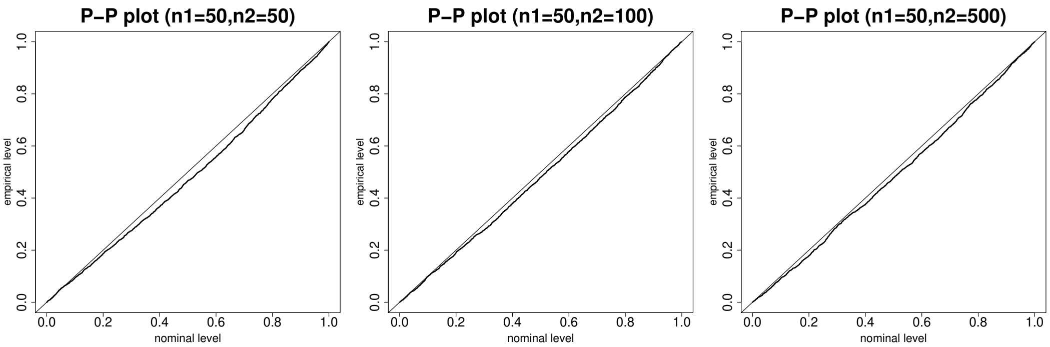

Fig. 3. P-P plots of the LRT statistic testing for equality of parameters of two HMMs with unequal sample sizes for $n_1 = 50$ and $n_2 = 50, 100, 500$ (Number of replications: N=2500).

investigated, second the opposite situation with $\alpha_{12}^j = 0.9$, $\alpha_{21}^j = 0.8$, $j = 1, 2$ and third $\alpha_{12}^j = 0.9$, $\alpha_{21}^j = 0.2$, $j = 1, 2$.

We simulated the distribution of the LRT for both hypotheses considered above, and observed good convergence properties in most cases. Only for the settings with few transitions between the states, when the smaller sample size $n = 50$ is used the fit is not yet satisfactory.

Furthermore we considered a two-state Poisson HMM with state-dependent intensities $\lambda_1^j = 1$ and $\lambda_2^j = 4$, $j = 1, 2$. Again, four different settings of transition probabilities were examined. Similarly to the results presented for the Gaussian HMM the simulations show a rather good fit even for the sample size of $n = 50$.

We would like to point out that for both HMMs, Gaussian as well as Poisson, the fit to the asymptotic distribution becomes less good if the difference between the state-dependent distributions is small. For example, if $(\mu_1^j - \mu_2^j)/\sigma = 1$, $j = 1, 2$, for the two-state Gaussian HMM or if $\lambda_1^j = 1, \lambda_2^j = 2$, $j = 1, 2$ for two-state Poisson HMM, the fit is not yet satisfactory for $n = 50$, but can be used for the larger sample size $n = 200$. In these cases estimation (and hence also testing) is hard since the resulting mixtures are quite similar to single components from these families, so that the parameters in the mixture can only be estimated in very large samples.

5 Application to modeling Drosophila larval movement

Holzmann et al. (2006) analyzed times series of Drosophila larval movement by means of HMMs. Here we apply our methodology to testing for equality of some of the parameters of times series of different larvae from this data set.

Let us start by briefly describing the data. A more complete discussion can be found in Holzmann et al. (2006). Locomotion of Drosophila larvae alternately features repeated episodes of linear movement and brief episodes of head swinging and turning. It can be summarized by measurements of speed and direction changes of the larvae over time. For each larva, the measurements are in the form of a bivariate time series of speed and direction changes, recorded once per second over three minutes. The paths taken by three larvae of the sample are displayed in Figure 4. In addition, scatterplots of the data sets are shown in Figure 5. We restrict ourselves to the univariate circular time series of direction changes, and fit two-

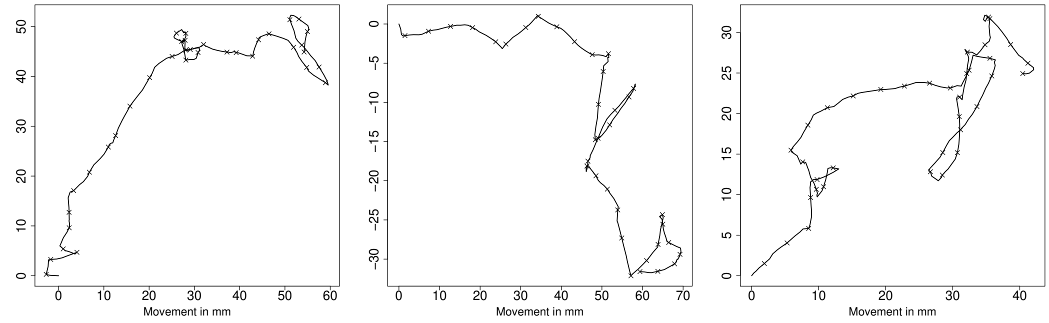

Fig. 4. Plots of the movements of three larvae of the fly Drosophila.

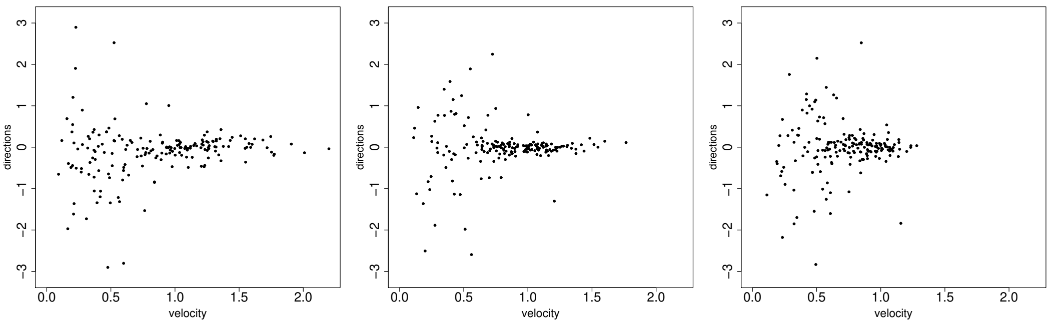

Fig. 5. Scatterplots of velocity and direction change of the movements of the three larvae.

state HMMs to these series, where the state-dependent densities are von Mises with location parameter $\mu \in \mathbb{R}$ and concentration parameter $\kappa > 0$. Furthermore, we assume that the location parameters in both states coincide. Hence the model has five parameters $\vartheta = (\alpha_{12}, \alpha_{21}, \mu, \kappa_1, \kappa_2)$. The MLEs were computed using direct maximization via a Newton-type algorithm as in Section 4. To ensure convergence to the global maximum several starting values were used.

In Table 1 the MLEs $\hat{\vartheta}_n$ for three larvae are displayed. Next we compare the times

no.	$\hat{\alpha}_{12}$	$\hat{\alpha}_{21}$	$\hat{\mu}$	$\hat{\kappa}_1$	$\hat{\kappa}_2$	$\hat{\pi}_1$
1	0.34	0.24	-0.0058	1.726	36.38	0.42
2	0.27	0.12	-0.0056	1.506	111.39	0.31
3	0.27	0.17	0.0029	1.559	43.99	0.40

Table 1
MLEs for a two-state von Mises-HMM.

series models for the different larvae in pairwise tests. First, we test the hypothesis of equality in all parameters $H_{hom} : \vartheta^1 = \vartheta^2$. A level-$\alpha$-test rejects H_{hom} if for the LRT statistic $\lambda_n^{hom} > \chi^2_{5:1-\alpha}$, where $\chi^2_{5:1-\alpha}$ denotes the $1-\alpha$-quantile of the χ^2–distribution with 5 d.f.s. Second, we test the hypothesis of equality in the regime parameters only:

$$H_{reg} : \alpha^1_{12} = \alpha^2_{12} \quad \text{and} \quad \alpha^1_{21} = \alpha^2_{21}.$$

In this case, the corresponding LRT statistic is χ^2–distributed with 2 d.f.s, and H_{reg} is rejected with level α if $\lambda_n^{reg} > \chi^2_{2:1-\alpha}$. The results of the pairwise tests are displayed in Table 2. It turns out that the hypothesis of equality in all parameters

no.	no.	λ_n^{hom}	p-value	λ_n^{reg}	p-value
1	2	24.06	0.000	3.26	0.196
1	3	2.17	0.826	0.98	0.612
2	3	18.04	0.002	1.00	0.606

Table 2
Results of the pairwise tests.

is rejected in two of the pairwise comparisons, while the hypothesis of equality in the regime parameters cannot be rejected on a 5%-level. Note that the test for the movement of larvae no. 1 against no. 3 has a comparable p-value as the test of no. 2 against no. 3. The reason is that while the regime parameters of no. 2 and no. 3 look more similar, they in fact give rise to stationary probabilities which differ more strongly than those of larvae 1 and 3.

The rejection of H_{hom} seems to be caused by the high value of κ_2 for the second larva. On the other hand, the estimates for α_{21} between the first and the second larva also differ strongly, which alone is however not yet significant for the LRT. Thus, the illustration shows once more that the LRT in HMMs is more sensitive w.r.t. the parameters of the state dependent distributions than w.r.t. the parameters of the regime.

Acknowledgements

Jörn Dannemann acknowledges financial support from the Georg Lichtenberg program "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 two anonymous referees, Axel Munk and Walter Zucchini for helpful comments and suggestions, and Max Suster for providing the data set which was published in Holzmann et al. (2006).

A Appendix

Here we give the technical assumptions as well as the proof of Theorem 1. Bickel et al. (1998) give assumptions (their Assumptions (A1)-(A6)) under which they establish the asymptotic normality of the MLE. It is shown that these assumptions are also sufficient to derive the asymptotics of the LRT statistic (Giudici et al., 2000).

Assumption 6 (A1)**.** *The transition probability matrix $\{\alpha_{ab}(\vartheta_0)\}$ is ergodic, i.e. irreducible and aperiodic.*

Assumption 7 (A2)**.** *There exists a $\delta > 0$ such that the maps $\vartheta \mapsto \alpha_{ab}(\vartheta)$ and $\vartheta \mapsto \pi_a(\vartheta)$ for $a, b \in \mathcal{M}$ have two continuous derivatives in the neighborhood $B_\delta(\vartheta_0) :=$*

$\{\vartheta | \, \|\vartheta - \vartheta_0\| < \delta\}$ *and the maps* $\vartheta \mapsto f(y;\ \theta_a(\vartheta))$ *for* $a \in \mathcal{M}$ *and* $y \in \mathcal{Y}$ *has two continuous derivatives in* $B_\delta(\vartheta_0)$.

Assumption 8 (A3)**.** *Let* $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$. *There exists a* $\delta > 0$ *such that*

1.) for all $i \in \{1, \ldots, d\}$ *and for all* $a \in \mathcal{M}$

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

2.) for all $i, j \in \{1, \ldots, d\}$ *and for all* $a \in \mathcal{M}$

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

3.) for $j = 1, 2$, *all* $i_l \in \{1, \ldots, d\}$, $l = 1, \ldots, j$ *and for all* $a \in \mathcal{M}$

$$\int \sup_{\vartheta \in B_\delta(\vartheta_0)} \left| \frac{d^2}{d\vartheta_{i_1} \cdots d\vartheta_{i_j}} f(y;\ \theta_a(\vartheta)) \right| d\nu(y) < \infty.$$

Assumption 9 (A4)**.** *There exists a* $\delta > 0$ *such that with*

$$\rho_0(y) = \sup_{\vartheta \in B_\delta(\vartheta_0)} \max_{1 \le a, b \le m} \frac{f(y;\ \theta_a(\vartheta))}{f(y;\ \theta_a(\vartheta))},$$

$P_0(\rho_0(Y_1) = \infty | X_1 = a) < 1$ *for all* $a \in \mathcal{M}$.

Assumption 10 (A5)**.** *The true parameter value* ϑ_0 *is an interior point of* Θ.

Assumption 11 (A6)**.** *The MLE* $\hat{\vartheta}_n$ *is strongly consistent.*

Proof of Theorem 1. The proof proceeds along similar lines as the proof for the one-sample case in Pruscha (2000). Since R is a differentiable map with 0 as regular value, there exists an injective C^2 mapping (chart) $\varphi : \Delta \to \Theta^1 \times \Theta^2$ such that $R(\varphi(t)) = 0$, where Δ is some neighborhood around t_0, with t_0 defined as the unique preimage of $(\vartheta_0^1, \vartheta_0^2)$, i.e. $\varphi(t_0) = (\vartheta_0^1, \vartheta_0^2)$. The Jacobian of φ is denoted by $C(t)$, it has rank $d^1 + d^2 - r$ for all t. Set $C_0 := C(t_0)$. Further let $H_j(t)$ denote the Hessian of the j^{th} component of $\varphi(t)$. Under the hypothesis the likelihood function can be written as $L_n^\varphi(t) = L_n^c(\varphi(t))$, and the score, the Hessian and the Fisher information matrix are given by

$$\mathrm{D}_t L_n^\varphi(t) = C^t(t)\, \mathrm{D}_{(\vartheta^1, \vartheta^2)} L_n^c(\vartheta^1, \vartheta^2),$$

$$\mathrm{D}_t^2 L_n^\varphi(t) = C^t(t)\, \mathrm{D}^2_{(\vartheta^1, \vartheta^2)} L_n^c(\vartheta^1, \vartheta^2) C(t) + \sum_{j=1}^{d^1+d^2} H_j(t) \left(\mathrm{D}_{(\vartheta^1, \vartheta^2)} L_n^c(\vartheta^1, \vartheta^2) \right)_j \text{ and}$$

$$\mathcal{J}_0^\varphi = C_0^t \mathcal{J}_0^c C_0 = C_0^t \begin{pmatrix} c\mathcal{J}_0^1 & 0 \\ 0 & (1-c)\mathcal{J}_0^2 \end{pmatrix} C_0. \tag{A.1}$$

Here $\mathrm{D}_{(\vartheta^1, \vartheta^2)} L_n^c(\vartheta^1, \vartheta^2) = \left(\mathrm{D}_{\vartheta^1} L_{n_1}(\vartheta^1), \mathrm{D}_{\vartheta^2} L_{n_2}(\vartheta^2) \right)^t$ denotes the joint score, and the index j indicates its j^{th} component. The joint Hessian is denoted as

$\mathrm{D}^2_{(\vartheta^1,\vartheta^2)} L_n^c(\vartheta^1,\vartheta^2) = \mathrm{Diag}\left(\mathrm{D}^2_{\vartheta^1} L_{n_1}(\vartheta^1),\ \mathrm{D}^2_{\vartheta^2} L_{n_2}(\vartheta^2)\right)$. Furthermore, $\mathcal{J}_0^c$ denotes the joint Fisher information, given explicitly in (A.1). $\mathcal{J}_0^1$ denotes the Fisher information of ϑ^1 and $\mathcal{J}_0^2$ the Fisher information of ϑ^2.

The key idea of the proof is to split the statistic of the composite hypothesis into statistics of two simple hypotheses

$$\begin{aligned}\lambda_n^c &= 2\left(L_n^c(\hat\vartheta^1_{n_1},\hat\vartheta^2_{n_2}) - L_n^c(\tilde\vartheta^1_n,\tilde\vartheta^2_n)\right)\\ &= 2\left(L_n^c(\hat\vartheta^1_{n_1},\hat\vartheta^2_{n_2}) - L_n^c(\vartheta^1_0,\vartheta^2_0)\right) - 2\left(L_n^c(\tilde\vartheta_n,\tilde\vartheta_n) - L_n^c(\vartheta^1_0,\vartheta^2_0)\right) = \lambda_n^0 - \lambda_n^{0,c}.\end{aligned}$$

A standard argument using Lemmas 1 and 2 as well as Theorem 1 in Bickel et al. (1998) shows that

$$\lambda_n^0 = 2\left(L_n^c(\hat\vartheta^1_{n_1},\hat\vartheta^2_{n_2}) - L_n^c(\vartheta^1_0,\vartheta^2_0)\right) = \frac{1}{n}\,\mathrm{D}L_n^c(\vartheta^1_0,\vartheta^2_0)^t\,(\mathcal{J}_0^c)^{-1}\,\mathrm{D}L_n^c(\vartheta^1_0,\vartheta^2_0) + o_P(1).$$

To investigate $\lambda_n^{0,c}$ we first use Lemmas 1 and 2 in Bickel et al. (1998) to establish

$$\frac{1}{n} C_0^t\,\mathrm{D}^2 L_n^c\left(\vartheta^{*1}_{n_1},\vartheta^{*2}_{n_2}\right) C_0 + \frac{1}{n}\sum_{j=1}^{d^1+d^2} H_j(t_n^*)\left(\mathrm{D}_{(\vartheta^1,\vartheta^2)} L_n^c\left(\vartheta^{*1}_{n_1},\vartheta^{*2}_{n_2}\right)\right)_j \xrightarrow{\mathcal{P}} -C_0^t\mathcal{J}_0^c C_0 \quad (\mathrm{A.2})$$

for $\varphi(t_n^*) = \left(\vartheta^{*1}_{n_1},\vartheta^{*2}_{n_2}\right) \xrightarrow{a.s.} (\vartheta^1_0,\vartheta^2_0)$, since

$$\frac{1}{n}\sum_{j=1}^{d^1+d^2} H_j(t_n^*)\left(\mathrm{D}_{(\vartheta^1,\vartheta^2)} L_n^c\left(\vartheta^{*1}_{n_1},\vartheta^{*2}_{n_2}\right)\right)_j = O_p(\sqrt{n}).$$

From (A.2) and (A.1), for the restricted MLE $\tilde t_n$ ($\varphi(\tilde t_n)) = (\tilde\vartheta^1_{n_1},\tilde\vartheta^2_{n_2})$) we obtain from a Taylor expansion of $D_t L_n^\varphi(t)$ around $\tilde t_n$ that

$$\sqrt{n}\left(\tilde t_n - t_0\right) = (\mathcal{J}_0^\varphi)^{-1}\frac{1}{\sqrt{n}} C_0^t\,\mathrm{D}L_n^c(\vartheta^1_0,\vartheta^2_0) + o_p(1). \tag{A.3}$$

Here we use the assumption that $\tilde t_n$ is strongly consistent, i.e. $\tilde t_n \xrightarrow{a.s.} t_0$. Using (A.3), a Taylor expansion of $L_n(\varphi(t))$ around $\tilde t_n$ yields

$$\begin{aligned}\lambda_n^{0,c} &= 2\left(L_n^\varphi(\tilde t_n) - L_n^\varphi(t_0)\right) = (\tilde t_n - t_0)^t\,\mathrm{D}_t^2 L_n^\varphi(\bar t_n)(\tilde t_n - t_0)\\ &= \frac{1}{n}\left(C_0^t\,\mathrm{D}L_n^c(\vartheta^1_0,\vartheta^2_0)\right)^t\left(C_0^t\mathcal{J}_0^c C_0\right)^{-1} C_0^t\,\mathrm{D}L_n^c(\vartheta^1_0,\vartheta^2_0) + o_p(1),\end{aligned}$$

where $\bar t_n$ denotes a point on the line segment $\overline{\tilde t_n t_0}$. In summary,

$$\begin{aligned}\lambda_n^c &= \frac{1}{n}\,\mathrm{D}L_n^c(\vartheta^1_0,\vartheta^2_0)^t\left((\mathcal{J}_0^c)^{-1} - C_0\left(C_0^t\mathcal{J}_0^c C_0\right)^{-1} C_0^t\right)\mathrm{D}L_n^c(\vartheta^1_0,\vartheta^2_0)\\ &= \frac{1}{\sqrt{n}}\left((\mathcal{J}_0^c)^{-1/2}\,\mathrm{D}L_n^c(\vartheta^1_0,\vartheta^2_0)\right)^t P\,\frac{1}{\sqrt{n}}(\mathcal{J}_0^c)^{-1/2}\,\mathrm{D}L_n^c(\vartheta^1_0,\vartheta^2_0),\end{aligned}$$

where $P := I_{(d^1+d^2)\times(d^1+d^2)} - (\mathcal{J}_0^c)^{1/2} C_0\left(C_0^t\mathcal{J}_0^c C_0\right)^{-1} C_0^t(\mathcal{J}_0^c)^{1/2}$.
Since $\frac{1}{\sqrt{n}}(\mathcal{J}_0^c)^{-1/2}\,\mathrm{D}L_n^c(\vartheta_0,\vartheta_0)$ is asymptotically standard normal and P is a projection of rank

$$
\begin{aligned}
\mathrm{rk}(P) &= \mathrm{Tr}(P) \\
&= \mathrm{Tr}(I_{(d^1+d^2)\times(d^1+d^2)}) - \mathrm{Tr}\left((\mathcal{J}_0^c)^{1/2} C_0 \left(C_0^t \mathcal{J}_0 C_0\right)^{-1} C_0^t (\mathcal{J}_0^c)^{1/2}\right) \\
&= \mathrm{Tr}(I_{(d^1+d^2)\times(d^1+d^2)}) - \mathrm{Tr}\left(\left(C_0^t \mathcal{J}_0^c C_0\right) \left(C_0^t \mathcal{J}_0^c C_0\right)^{-1}\right) \\
&= \mathrm{Tr}(I_{(d^1+d^2)\times(d^1+d^2)}) - \mathrm{Tr}\left(I_{(d^1+d^2-r)\times(d^1+d^2-r)}\right) = d^1 + d^2 - (d^1 + d^2 - r) = r,
\end{aligned}
$$

Theorem 1 follows. □

References

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 (407), 201–210.

Bickel, P. J., Ritov, Y., Rydén, T., 1998. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. Ann. Statist. 26 (4), 1614–1635.

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

Dannemann, J., 2006. Maximum-Likelihood-Inferenz für Hidden Markow Modelle. Diploma thesis, Inst. for Math. Stochastics, Georg-August-University of Göttingen, Göttingen.

Giudici, P., Rydén, T., Vandekerkhove, P., 2000. Likelihood-ratio tests for hidden Markov models. Biometrics 56 (3), 742–747.

Holzmann, H., Munk, A., Suster, M., Zucchini, W., 2006. Hidden Markov models for circular and linear-circular time series. Environmental and Ecological Statistics 13 (3), 325–347.

Leroux, B. G., 1992. Maximum-likelihood estimation for hidden Markov models. Stochastic Process. Appl. 40 (1), 127–143.

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

Michalek, S., Wagner, M., Timmer, J., Vach, W., 2001. Finite sample properties of the maximum likelihood estimator and of likelihood ratio tests in hidden Markov models. Biom. J. 43 (7), 863–879.

Pruscha, H., 2000. Vorlesungen über mathematische Statistik. B. G. Teubner, Stuttgart.

Rydén, T., Terasvirta, T., Asbrink, S., 1998. Stylized facts of daily return series and the hidden markov model. J. Appl. Econometrics. 13 (3), 217–244.