Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 1850–1859

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

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

Received 4 October 2006; received in revised form 7 May 2007; accepted 1 June 2007
Available online 7 June 2007

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.

Keywords: 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 $\mathscr{M}$ 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 $X_k = i$, $i \in \mathscr{M}$, 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 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

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

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.06.002

(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 $(X_k)_{k\geqslant 1}$ be a stationary, finite-state Markov chain with state space $\mathscr{M}=\{1,\ldots,m\}$, transition probabilities $\alpha_{ab}=P(X_{k+1}=b|X_k=a)$, and unique stationary distribution $\pi=(\pi_1,\ldots,\pi_m)$. Further let $(Y_k)_{k\geqslant 1}$ be a stochastic process taking values in a Borel-measurable subset $\mathscr{Y}$ of Euclidean space, such that given $(X_k)_{k\geqslant 1}$, the $(Y_k)_{k\geqslant 1}$ 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 v on $\mathscr{Y}$, 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 $(X_k,Y_k)_{k\geqslant 1}$.

In the following we will consider two independent HMMs, so that the bivariate processes $(X_k^1,Y_k^1)_{k\geqslant 1}$ and $(X_k^2,Y_k^2)_{k\geqslant 1}$ are independent. In general we allow different image spaces $\mathscr{Y}_j$, state spaces $\mathscr{M}_j$, different parametric