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

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