View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

ISSN 1440-771X

MONASH University

Australia

Department of Econometrics
and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Estimating Components in Finite Mixtures and Hidden Markov Models

D.S. Poskitt and Jing Zhang

Working Paper 10/04

Estimating Components in Finite Mixtures and Hidden Markov Models

D. S. Poskitt
Department of Econometrics and Business Statistics
Monash University
Victoria 3800
Australia
Tel: +61 3 9905 9378
Fax: +61 3 9905 5474
E-mail: don.poskitt@Buseco.monash.edu.au

Jing Zhang
KMV LLC
1620 Montgomery Street, Suite 40, San Francisco, CA 94111, USA.
Jing.Zhang@kmv.com

March 28, 2003
Australian & New Zealand Journal of Statistics submission.

Abstract

When the unobservable Markov chain in a hidden Markov model is stationary the marginal distribution of the observations is a finite mixture with the number of terms equal to the number of the states of the Markov chain. This suggests estimating the number of states of the unobservable Markov chain by determining the number of mixture components in the marginal distribution. We therefore present new methods for estimating the number of states in a hidden Markov model, and coincidentally the unknown number of components in a finite mixture, based on penalized quasi-likelihood and generalized quasi-likelihood ratio methods constructed from the marginal distribution. The procedures advocated are simple to calculate and results obtained in empirical applications indicate that they are as effective as current available methods based on the full likelihood. We show that, under fairly general regularity conditions, the methods proposed will generate strongly consistent estimates of the unknown number of states or components.

Some key words: finite mixture, hidden Markov process, model selection, number of states, penalized quasi-likelihood, generalized quasi-likelihood ratio, strong consistency.

JEL: C51, C52

1 Introduction

Hidden Markov models, also known as Markov regime switching models, have become a widely used tool for modelling sequences of dependent random variables. Given a model structure with a known number of regimes, efficient and sophisticated estimation and forecasting schemes have been successfully developed and applied in a variety of fields including speech recognition, Juang & Rabiner (1990), DNA composition, Churchill (1989), neurobiology, Chung, Moore, Xia, Premkumar & Gages (1990) and Fredkin & Rice (1992a,b), the analysis of business cycles, Hamilton (1989) and modelling stock market and asset returns, Turner, Startz & Nelson (1990). In these applications, the correct specification of the number of states of the Markov chain has both theoretical and practical significance. The construction of realistic models that describe the gating mechanism of ion channels requires knowledge of the number of elementary channels, or states, contributing to the observed current fluctuations in the cell membrane, for example, whilst in economics the unobserved Markov chain is used to model the underlying states of the economy, the number of states corresponding to the number of qualitative categories into which the economy is classified.

Techniques currently available for choosing the number of states of a Markov chain are, however, somewhat incomplete and often difficult to apply. The conventional likelihood ratio test breaks down if one tries to fit a k state model when the true process has $k_0 < k$ states since under the null hypothesis $k_0 = k' < k$ the parameters that describe the k state model are unidentified. Hansen (1992, 1996) proposed a procedure that avoids this problem but his technique only gives bounds for the likelihood ratio statistic and requires extensive computation involving large scale simulation and optimization over a three-dimensional grid. Using a similar approach Hamilton (1996) has also developed a variety of misspecification tests. Leroux & Puterman (1992) and Rydén (1995) have studied the use of model selection criteria and the latter established that criteria such as AIC and BIC will not underestimate the true number of states. Consistency of BIC for the number of states of a Markov chain has been recently established by Csiszar & Shields (2000). An alternative approach to the selection of the number of states exploits the relationship of the autocovariances of these processes to those of $ARMA$ structures. Poskitt & Chung (1996) showed that, given appropriate assumptions, a k-state Markov chain process embedded in noise will posses an autocovariance function corresponding to that of an $ARMA(k-1, k-1)$ process and Zhang & Stine (2001) have extended their results to the case of a vector switching autoregression. Both papers proposed estimating the number of states of the Markov model by determining the order of its $ARMA$ representation.

In this paper we exploit the fact that the marginal distribution of a hidden Markov model is a finite mixture and use the number of estimated mixture components to determine the order of the hidden Markov model. The following section of the paper reviews the techniques and ideas that form the background to the development of our methodology whilst introducing our notation. Section 3 discusses two estimators of the number of states of the Markov chain that are based on the maximization of a penalized quasi-likelihood in which the full likelihood is approximated using the finite mixture marginal distribution. In Section 4 we show that both estimators are strongly consistent under fairly general regularity conditions. Section 5 presents a two step procedure based on a generalized quasi-likelihood ratio. This generates two further estimators that are also shown to be strongly consistent. Illustrations of the application of the proposed methodology in the context of both finite mixture and hidden Markov models are presented in Section 6. We close in Section 7 with a brief conclusion. Proofs of the basic results are assembled in an appendix.

2 Notation and Background

To begin, let $f(\cdot \mid \theta)$ denote densities on $\mathcal{Y} \subseteq \mathcal{R}^d$ with respect to a measure μ, parameterized by $\theta \in \Theta \subseteq \mathcal{R}^q$, and suppose that $Y_t \sim f(y \mid \theta_{S_t})$, $t = 1, \ldots, T$, where $\theta_s = \vartheta(s)$, $\vartheta : \mathcal{S} \to \Theta$, $\mathcal{S} = \{\sigma_1, \ldots, \sigma_k\}$, and $\{S_t\}$ is a k-state Markov chain process with state space $\mathcal{S}$ and $k \times k$ transition matrix $P = [p_{ij}]$. Now set $\Psi_k = \text{Closure}\{\bigcup P\}$ where the union is taken over all stochastic matrices P of order k such that the Markov chain is irreducible and aperiodic. For each k this set is compact. Furthermore, let us assume Θ is compact and set $\Phi_k = \Psi_k \times \Theta^k$ so that Φ_k is the parameter space for the model with a k-state Markov chain. (If Θ is not compact the compactification technique suggested by Rydén (1995) can be used to show that the results presented in the sequel will still hold.) We call k the dimension or order of the process. Given k the model dimension $\varphi_k = \dim(\Phi_k)$.

If $\phi_k \in \Phi_k$ the joint density of $Y^T = (Y_1, Y_2, \ldots, Y_T)$ with respect to the product measure μ^T associated with μ is

$$\begin{aligned} p(Y^T \mid \phi_k) &= \sum_{s_1,\ldots,s_T} \pi_{s_1} f(Y_1 \mid \theta_{s_1}) \cdot p_{s_1 s_2} f(Y_2 \mid \theta_{s_2}) \\ &\qquad \cdots\cdots\cdots p_{s_{T-1} s_T} f(Y_T \mid \theta_{s_T}), \end{aligned} \tag{2.1}$$

where $\sum_{s_1,\ldots,s_T}$ denotes that the summation is taken over all possible paths $s_t,\ t = 1, \ldots, T$, of the Markov chain. The marginal distribution $p(y \mid \phi_k)$ of Y_t has a much simpler form however, namely,

$$p(y \mid \phi_k) = \sum_{i=1}^{k} \pi_i f(y \mid \theta_{\sigma_i}), \tag{2.2}$$

a k component finite mixture where the mixing distribution $(\pi_1, \pi_2, \ldots, \pi_k)$ is given by the stationary distribution of the Markov chain. In general the mapping $\vartheta : \mathcal{S} \to \Theta$ could be many to one: It is possible that ion channels could have different open-closed states with the same conductance levels and the number of components will be determined by the observed conductance levels and not the different physical states of the channel dynamics, see Fredkin & Rice (1992b). Thus the number of unique components in (2.2) can be less than the order of the chain: $k' = \inf\{\kappa : p(y \mid \phi_k) = \sum_{i=1}^{\kappa} \pi_i f(y \mid \theta_{\sigma_i})\} \leq k$. If $\vartheta(s)$ is one-to-one, however, then there is a direct correspondence between the number of components in the mixture and the number of states. Thus we can contemplate estimating k, or at least k', by ascertaining the number of components in the marginal distribution of Y_t.

Finite mixture models have been extensively studied in statistics and have found application in various fields. Comprehensive treatments of the subject can be found in the monographs by Titterington, Smith & Markov (1985), McLachlan & Basford (1988) and Lindsay (1995). Basing inference in the hidden Markov model on finite mixtures is not in itself new. Lindgren (1978) showed under mild regularity conditions that the estimator $\hat{\phi}_k$ obtained by maximizing the following quasi-likelihood function based on the marginal distributions

$$L^m(Y^T \mid \phi_k) = \prod_{t=1}^{T} p(Y_t \mid \phi_k) \tag{2.3}$$

would be consistent and asymptotically normal. Leroux & Puterman (1992) also suggested that fitting a finite mixture model was an effective strategy for obtaining parameter estimates suitable for the initiation of the iterative calculations required for the evaluation of the exact maximum likelihood estimates of a hidden Markov model.

From our perspective, the advantages of working with (2.3) are twofold. First, the combinatorial calculations associated with investigating (2.1) theoretically make its analysis extremely difficult. The structure of $L^m(Y^T \mid \phi_k)$, on the other hand, is far simpler and facilitates the application of appropriate limiting arguments. Second, there are substantial computational gains that derive from using (2.3). Not only does the efficient evaluation of (2.1) involve $O((2k^2+k)T+2k-1)$ operations compared to $O(2kT-1)$ for (2.3) but the maximization of (2.1) presents a far more complicated task than the optimization of (2.3) because of the presence of the unobserved states. See Lindgren, and Leroux and Puterman *op. cit.* as well as Böhning, Schlattmann & Lindsay (1992) for some discussion of the numerical issues raised here.

Generally speaking, there are three different approaches that have been taken to estimating the number of components of a finite mixture:

(i) The first involves the construction of estimators that minimize some charaterisation of the distance between the true distribution and the fitted or empirical distribution. Henna (1985) and Chen & Kalbfleisch (1996) show that this approach will yield consistent estimates under appropriate regularity conditions but the methods advocated are not readily implementable since, as noted by Chen and Kalbfleisch, issues associated with the choice of distance measure and the construction of effective algorithms remain unresolved.

(ii) The second considers the application of hypothesis testing procedures using the likelihood ratio principle, as in Feng & McCulloch (1996). It is widely known, however, see Hartigan (1985), that the null distribution of the likelihood ratio statistic for testing $k_0 = k-1$ components versus $k_0 = k$ components does not converge to that of the conventional Chi-squared variate and Feng and McCulloch propose a remedy based on using bootstrapped likelihood ratios. Heckman, Robb & Walker (1990) develop an alternative test statistic based upon the method of moments and a recent contribution that we include here, but which should perhaps be given a separate heading of its own, is that of Richardson & Green (1997), who consider a Bayesian analysis based on Markov chain Monte Carlo methods.

(iii) The third approach considers the use of model selection devices based on penalized likelihood methods and is exemplified by the work of Leroux (1992) and Dacunha-Castelle & Gassiat (1997).

It is the latter approach that we follow in this paper. Our work differs from that of previous authors, however, in that our methodology incorporates aspects of all three approaches and we develop four alternative estimators of k. We are also able to show that these estimators will yield strongly consistent estimates of the true order k_0 under regularity conditions that allow for both independent and Markov dependent processes.

3 Order Estimation Based on Finite Mixtures

We propose two types of estimate based on the quasi-likelihood $L^m(Y^T \mid \hat{\phi}_k)$ calculated from the finite mixture in (2.3). The first is defined in terms of the penalized quasi-likelihood function

$$\Delta_T(k) = \log L^m(Y^T \mid \hat{\phi}_k) - d_{kT} \qquad (3.1)$$

and is defined as

$$\hat{k_I} = \min\{\kappa : \Delta_T(\kappa) = \max_{k\in\{1,\ldots,K\}}\{\Delta_T(k)\}\}, \tag{3.2}$$

where K is prescribed by the practitioner and $d_{kT} > 0$ is a penalty term which depends on both the order of the fitted model and the sample size T and is such that $d_{(k+1)T} > d_{kT}$ given T. Two common choices for the penalty term are $d_{kT} = \varphi_k$ and $d_{kT} = (\varphi_k/2)\log T$. The model dimension $\varphi_k = \dim(\Phi_k) = k(k-1)+kq$ if the individual components of the transition matrix P are to be counted, as in the evaluation of the full likelihood, and $k-1+kq$ if only the marginal distributions are considered, as in (3.2).

The estimator $\hat{k}_I$ is similar in spirit to that proposed by Rydén (1995). He based his estimator on maximizing the *split data likelihood function* defined by

$$L^S(Y^T \mid \phi_k) = \prod_{j=1}^{N} p_M(Y_{M(j-1)+1}, \cdots, Y_{Mj} \mid \phi_k), \tag{3.3}$$

where $T = MN$ and $p_M(Y_{M(j-1)+1}, \cdots, Y_{Mj} \mid \phi_k)$ is the M-dimensional joint density of the variates $Y_{M(j-1)+1}, \cdots, Y_{Mj}$ defined as in (2.1). Rydén (1995) advocated using $M > 2k$ in (3.3) and showed that his estimator would not underestimate k_0 in the limit. Compared to the approach of Rydén (1995) or the use of penalized likelihood methods based on evaluation of the full likelihood, however, the estimator $\hat{k}_I$ is much easier to implement and it can be shown to be strongly consistent.

If $K < k_0$, so that the upper bound on the number of states considered is chosen too small, then the proof of consistency given below indicates that $\hat{k}_I = K$ as $T \to \infty$ if d_{kT}/T approaches zero as T increases. Suppose then that $K = K_T$, an increasing function of T. Eventually we must have $K_T > k_0$ for any $k_0 < \infty$ and $\hat{k}_I \geq k_0$ with probability one. Thus the possibility of underestimating the number of states is circumvented, but the rate of increase of K_T is unknown and the problem of selecting an upper bound for the true order is, to this extent, unresolved.

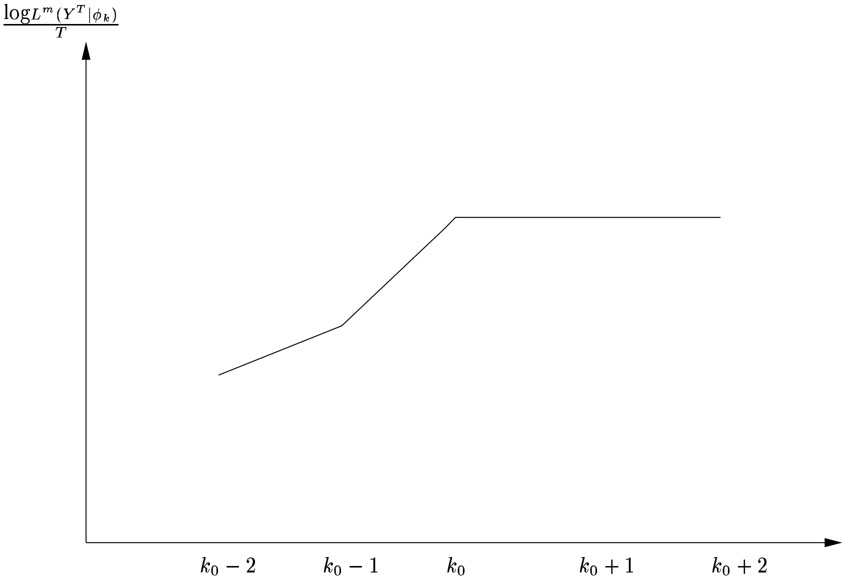
An alternative solution is to consider a procedure that avoids the need to make an *a-priori* selection of K altogether. The second estimator, $\hat{k}'_I$, is defined by the requirement that

$$\begin{aligned} \hat{k}'_I &= 1 \text{ if } \Delta_T(1) \geq \Delta_T(2) \text{ and} \\ \hat{k}'_I &= k' \text{ if } \Delta_T(k) < \Delta_T(k+1),\ 1 \leq k < k',\ \text{and} \\ &\qquad \Delta_T(k') \geq \Delta_T(k'+1),\ k' = 2, 3, \ldots. \end{aligned} \tag{3.4}$$

This estimator locates the right-hand end-point of the interval on which $\Delta_T(k)$ is strictly increasing. The rationale behind $\hat{k}'_I$ is illustrated in Figure 1. For T sufficiently large the average support, $T^{-1}\log L^m(Y^T \mid \hat{\phi}_k)$, will be an increasing function of k whenever $k < k_0$ but it will remain roughly constant once the fitted order exceeds the true one. This leads us to the conclusion that the first local maximum of $\Delta_T(k)$ in the range $1 \leq k < \infty$ will occur at $k = k_0$ and that $\hat{k}'_I \geq k_0$ if $d_{kT}/T \to 0$ as $T \to \infty$. In fact, $\hat{k}'_I$ will yield a consistent estimate of k_0 under the same conditions as for $\hat{k}_I$.

We have adopted the notation $\hat{k}'_I$ rather than the more obvious $\hat{k}_{II}$ for the second estimator so as to preserve the later notation for the two step estimator presented below. The estimator $\hat{k}_{II}$ shares a similar rationale to $\hat{k}'_I$ but is based on the incremental value in the quasi-likelihood ratio rather than the imposition of a penalty term. The details of $\hat{k}_{II}$ are given in Section 5, after some asymptotic properties of $\hat{k}_I$ and $\hat{k}'_I$ are presented in the next

Figure 1: Hypothetical Log Quasi-Likelihood



section.

4 Theoretical Bounds for $\hat{k}_I$ and $\hat{k}'_I$

First we will establish that the estimators $\hat{k}_I$ and $\hat{k}'_I$ will not underestimate the true order asymptotically if the penalty term is chosen so that $d_{kT}/T \to 0$ as $T \to \infty$. To achieve this we note that $\hat{k}_I$ coincides with Rydén's estimator if $M = 1$ in (3.3) and Rydén's derivations are independent of the value of M. Thus, a similar argument to that employed by Rydén (1995), which builds on the methodology of Leroux (1992), can be used to show that our estimators will behave as claimed.

We now introduce our regularity conditions. These conditions follow those commonly employed in the analysis of maximum likelihood estimation and correspond more or less to those given in Leroux (1992) and Rydén (1995). In what follows $\phi^0_{k_0} \in \Phi_{k_0}$ will denote the true parameter value where k_0, the true order, is assumed to be minimal, that is, there does not exist a parameter value $\phi_k \in \Phi_k$ with $k < k_0$ such that ϕ_k and $\phi^0_{k_0}$ generate identical probability laws for Y^T with $p(y^T \mid \phi_k) = p(y^T \mid \phi^0_{k_0})$ almost everywhere. Unless stated otherwise, all probability statements are assumed to obtain under the true probability measure $\nu_{\phi^0_{k_0}}$ induced by $\phi^0_{k_0}$.

C1. The process $\{S_t\}$ is a stationary and ergodic Markov chain.

C2. For each fixed k the family of finite mixtures $\sum_{i=1}^k \pi_i f(y \mid \theta_{\sigma_i})$ is identifiable up to a permutation of the indices.

C3. The density function $f(y \mid \theta)$ is continuous on $\mathcal{Y} \times \Theta$ and twice continuously differentiable with respect to θ.

C4. For each compact set $C \subseteq \mathcal{Y}$ and $\epsilon > 0$ there is a compact set $C_\epsilon \subseteq \Theta$ such that $f(y \mid \theta) \leq \epsilon$ on $C \times (\Theta \backslash C_\epsilon)$.

C5. There is a set $Z \subseteq \mathcal{Y}$ and a set $\Omega \subseteq \Theta$ such that $\mu(Z) > 0$, $f(y \mid \theta) = 0$ on $Z \times (\Theta \backslash \Omega)$, and $f(y \mid \theta) > 0$ on $\mathcal{Y} \times \Omega$.

C6. For all $\theta \in \Theta$ the integral $\int p(y \mid \phi^0_{k_0})(\log f(y \mid \theta))^- \mu(dy) < 0$ where $(\cdot)^-$ denotes the negative part of the argument and there is a continuous function $h : \mathcal{Y} \to \mathcal{R}$ such that $f(y \mid \theta) \leq h(y)$ and $\int p(y \mid \phi^0_{k_0}) |\log h(y)| \mu(dy) < \infty$.

We will also suppose:

C2' Let $\mathrm{Int}\{\Phi_k\}$ denote the interior of Φ_k. There exist $m < \varphi_k$ affinely independent, continuous functions $r_i(\phi_k)$, $i = 1, \ldots, m$, such that $r(\phi_k) = (r_1(\phi_k), \ldots, r_m(\phi_k))' = 0$ for all ϕ_k that lie on the boundary $\Phi_k \backslash \mathrm{Int}\{\Phi_k\}$.

C3'. The first and second order partial derivatives of $\log f(y \mid \theta)$ satisfy $|\partial \log f(y \mid \theta)/\partial\theta_u| < h_u(y)$ and $|\partial^2 \log f(y \mid \theta)/\partial\theta_u\partial\theta_v| < h_{u,v}(y)$, $1 \leq u, v \leq q$, where $\int p(y \mid \phi^0_{k_0}) h_u(y)\mu(dy) < \infty$ and $\int p(y \mid \phi^0_{k_0}) h_{u,v}(y)\mu(dy) < \infty$.

C6' There exits a $\delta > 0$ such that $\int \|y\|^{2+\delta} f(y \mid \theta)\mu(dy) < \infty$ for all $\theta \in \Theta$.

Leroux and Rydén *op. cit.* show that assumptions C2-C6 hold for various families of probability distributions, including Poisson and Gaussian mixtures of the type considered in the applications below. To motivate C2' consider a finite mixture process of order k with density as in (2.2) where $f(y \mid \theta)$ is chosen such that C2 is satisfied. Then $\phi_k = (\pi_1, \ldots, \pi_{k-1}, \theta_1, \ldots, \theta_k)'$ and Φ_k consists of the cartesian product of the simplex $0 \leq \pi_i \leq 1$, $i = 1 \ldots, k-1$, $\sum_{i=1}^{k-1} \pi_i \leq 1$, and the k fold reproduction of Θ. The boundary $\Phi_k \backslash \mathrm{Int}\{\Phi_k\}$ contains all those parameterizations corresponding to processes of order k', $1 \leq k' < k$. Modulo a permutation of the indices, these can be represented by the restrictions $r_i(\phi_k) = \pi_{k'+i} = 0$, $i = 1, \ldots, m-1$, $r_m(\phi_k) = \sum_{i=1}^{k'} \pi_i - 1 = 0$, $m = k - k'$. Bickel, Ritov & Rydén (1998), pages 1618-1619, illustrate the application of conditions essentially equivalent to C1-C6 and C3' to hidden Markov models such as those described in Fredkin & Rice (1992a,b). The application of C6' to finite mixtures and hidden Markov models involving Poisson, binomial, exponential or mean-translated Gaussian distributions is obvious.

In order to proceed let us define the distance between the true distribution $p(y \mid \phi^0_{k_0})$ and $p(y \mid \phi_k)$ as the Kullback-Leibler divergence

$$K(\phi^0_{k_0}, \phi_k) = \int p(y \mid \phi^0_{k_0}) \log \frac{p(y \mid \phi^0_{k_0})}{p(y \mid \phi_k)} \mu(dy)\,, \tag{4.1}$$

where $p(y \mid \phi_k)$ is the marginal distribution of Y_t as defined in (2.2). Now set

$$K(\phi^0_{k_0}, \Phi_k) = \inf_{\phi_k \in \Phi_k} \{K(\phi^0_{k_0}, \phi_k)\}\,. \tag{4.2}$$

and let $\Phi^0_k = \{\phi_k : K(\phi^0_{k_0}, \phi_k) = K(\phi^0_{k_0}, \Phi_k)\}$. If the fitted order $k < k_0$ then there is no parameter which is observationally equivalent to $\phi^0_{k_0}$ and therefore we must have $K(\phi^0_{k_0}, \Phi_k) > 0$. The following lemmas give formal content to this idea.

Lemma 1 *If conditions C3, C4, and C6 hold then for each k there exists a $\phi^0_k \in \Phi_k$ such that $K(\phi^0_{k_0}, \phi^0_k) = K(\phi^0_{k_0}, \Phi_k)$.*

Lemma 2 *Assume that conditions C2-C4 and C6 hold. Then $K(\phi^0_{k_0}, \Phi_k) > 0$ if $k < k_0$ and $K(\phi^0_{k_0}, \Phi_k) = 0$ if $k \geq k_0$. Moreover, $K(\phi^0_{k_0}, \Phi_{k+1}) < K(\phi^0_{k_0}, \Phi_k)$ for all $k < k_0$.*

At this point we note that both $\hat{k}_I$ and $\hat{k}'_I$ can be viewed as being derived via a succession of quasi-likelihood ratio tests. The test statistic is given by $\log\{L^m(Y^T \mid \hat{\phi}_{k'})/L^m(Y^T \mid \hat{\phi}_k)\}$

and the null hypothesis $H_N : k_0 = k$ is rejected in favor of the alternative hypothesis $H_A : k_0 = k'$ if the test statistic exceeds $(d_{k'T} - d_{kT})$. The estimator $\hat{k}_I$ corresponds to calculating all $\frac{1}{2}K(K-1)$ likelihood ratios for every pair of values k and k' in the set $\{1, \ldots, K\}$, $k \neq k'$, and choosing the value that is accepted against all others. The estimate $\hat{k}'_I$ is the first value of k for which the null hypothesis is accepted when testing in sequence $H_N : k_0 = k$ against $H_A : k_0 = k' = k+1$ starting at $k = 1$. The implied noncentrality parameter $T\{K(\phi^0_{k_0}, \phi^0_k) - K(\phi^0_{k_0}, \phi^0_{k'})\}$ is positive for $k < k' \leq k_0$ and consequently neither test procedure will select an order less than k_0 if the critical value $c_T = (d_{k'T} - d_{kT})$ is chosen such that $c_T/T \to 0$ as $T \to \infty$. This result is stated formally in Theorem 1 immediately below and is contingent on the following lemma.

Lemma 3 *If conditions C1-C6, C3' and C6' obtain then for all k*

$$\lim_{T\to\infty} \frac{1}{T} \log L^m(Y^T \mid \hat{\phi}_k) = \int p(y \mid \phi^0_{k_0}) \log p(y \mid \phi^0_k) \mu(dy) \tag{4.3}$$

and

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \frac{\partial^2 \log L^m(Y^T \mid \phi_k)}{\partial \phi_k \partial \phi'_k} &= -\frac{\partial^2 K(\phi^0_{k_0}, \phi_k)}{\partial \phi_k \partial \phi'_k} \\ &= \int p(y \mid \phi^0_{k_0}) \frac{\partial^2 \log p(y \mid \phi_k)}{\partial \phi_k \partial \phi'_k} \mu(dy) \end{aligned} \tag{4.4}$$

with probability one.

Theorem 1 *Assume conditions C1-C6, C3' and C6' hold and suppose that* $\overline{\lim}_{T\to\infty} d_{kT}/T = 0$*. Then* $\lim_{T\to\infty} \hat{k}_I = K$ *if* $K < k_0$ *and* $\underline{\lim}_{T\to\infty} \hat{k}_I \geq k_0$ *if* $K \geq k_0$*, furthermore,* $\underline{\lim}_{T\to\infty} \hat{k}'_I \geq k_0$ *with probability one.*

Whereas Theorem 1 establishes that k_0 provides the infimum of $\hat{k}_I$ and $\hat{k}'_I$, we now demonstrate additional constraints on the growth rate of d_{kT} as a function of T that will guarantee that the limit supremum of $\hat{k}_I$ and $\hat{k}'_I$ is also k_0. These constraints are dependent on the order of magnitude of the logarithm of the quasi-likelihood ratio $\log\{L^m(Y^T \mid \hat{\phi}_k)/L^m(Y^T \mid \hat{\phi}_{k_0})\}$. The next two lemmas serve to set a bound on the growth rate of this latter statistic.

Lemma 4 *Suppose that the regularity conditions C1-C6, C2', C3' and C6' hold, then for each k there exits a* $\phi^0_k \in \Phi^0_k$ *such that*

$$\log L^m(Y^T \mid \hat{\phi}_k) - \log L^m(Y^T \mid \phi^0_k) = O(\log \log T) \ .$$

Lemma 5 *Under the same assumptions as for Lemma 4 we have*

$$\overline{\lim}_{T\to\infty} \left\{ \log \frac{L^m(Y^T \mid \phi^0_k)}{L^m(Y^T \mid \phi^0_{k_0})} - 2M \log \log T - 2 \log T \right\} < 0$$

with probability one for all $k \neq k_0$ *and all* $M > 1$.

We are now in a position to give conditions on d_{kT} that will avoid overestimation in the limit.

Theorem 2 *Suppose assumptions C1-C6, C2', C3' and C6' obtain and that the penalty term d_{kT} can be written as $d_{kT} = g(T) \cdot h(k)$ where g and h are increasing functions of T and k that satisfy*

$$\begin{cases} \underline{\lim}_{T\to\infty} g(T)/\log T & \geq 1 \textit{ and} \\ h(k') - h(k) & \geq 2, \; k' > k \; . \end{cases} \tag{4.5}$$

Then $\overline{\lim}_{T\to\infty}\hat{k}_I = k_0$ *and* $\overline{\lim}_{T\to\infty}\hat{k}'_I = k_0$ *with probability one.*

At first sight Theorem 2 may seem curious. For the mixture likelihood (2.3) $\dim(\Phi_k) = \varphi_k = k(q+1) - 1$ and each increment in k increases the value of φ_k by $q+1$. Thus the theorem indicates that the value of the penalty term need not reflect the overall number of parameters to be estimated. This is explicable, however, because asymptotically it is the number of densities $f(\cdot \mid \theta)$ that have been fitted to the data that is critical in determining the behaviour of the criterion irrespective of the value of q and $q+1 \geq 2$ for all densities containing at least one parameter. Hence the restriction on $h(k)$.

Theorem 2 yields restrictions on d_{kT} which when taken in conjunction with Theorem 1 ensure that both $\hat{k}_I$ and $\hat{k}'_I$ are consistent. Thus, if in practice we set $d_{kT} = \varphi_k \log T$, then $d_{kT} = g(T) \cdot h(k)$ where $g(T) = \log T$ and $h(k) = \varphi_k$. Clearly $\lim_{T\to\infty} d_{kT}/T = 0$ and the penalty term satisfies (4.5) since $(\varphi_k - \varphi_{k_0}) = (k - k_0)(1+q) \geq 2$ for all $k > k_0$. Combining Theorems 1 and 2 we have the following result.

Corollary 1 *Assume the conditions of Theorems 1 and 2 hold. If the penalty term* $d_{kT} = \varphi_k \log T$ *then* $\hat{k}_I$ *and* $\hat{k}'_I$ *converge to* k_0 *almost surely.*

5 A Two Step Estimator

To simplify the presentation of the two step estimator let us first introduce the generalized quasi-likelihood ratio

$$R(Y^T, k) = \frac{1}{T}\{\log \frac{L^m(Y^T \mid \hat{\phi}_k)}{L^m(Y^T \mid \hat{\phi}_{k-1})} + \log \frac{L^m(Y^T \mid \hat{\phi}_{k+1})}{L^m(Y^T \mid \hat{\phi}_k)}\}, \; k = 1, 2, \ldots, \tag{5.1}$$

where, by definition, $L^m(Y^T \mid \hat{\phi}_1)/L^m(Y^T \mid \hat{\phi}_0) \equiv 1$. Let $\mathcal{K}_T(\hat{k}_I) = \{k \in \{1, \ldots, K\} : k \leq \hat{k}_I$ and $R(Y^T, k) > \eta_T\}$ where $\eta_T > 0$ is a non-increasing function of T, yet to be prescribed. The two step estimator $\hat{k}_{II}$ is defined as

$$\hat{k}_{II} = \max\{k \in \mathcal{K}_T(\hat{k}_I)\} \tag{5.2}$$

and $\hat{k}'_{II}$ is defined in a completely analogous way by simply replacing $\mathcal{K}_T(\hat{k}_I)$ by $\mathcal{K}_T(\hat{k}'_I) = \{k \in \{1, 2, \ldots\} : k \leq \hat{k}'_I$ and $R(Y^T, k) > \eta_T\}$.

The idea underlying the two step estimator is to recognize that if d_{kT} satisfies the conditions of Theorem 1 then ultimately both $\hat{k}_I$ and $\hat{k}'_I$ are likely to exceed k_0. The behaviour of $T^{-1} \log L^m(Y^T \mid \hat{\phi}_k)$, as exhibited in Figure 1, can then be exploited to ascertain the extent to which either $\hat{k}_I$ or $\hat{k}'_I$ needs to be reduced. This is done by examining the two step increment in $T^{-1} \log L^m(Y^T \mid \hat{\phi}_k)$ between $k-1$ and $k+1$. For T sufficiently large this increment will be positive whenever $k \leq k_0$ and arbitrarily small when $k > k_0$. For example, for the preschool health status data considered below the values of $R(Y^{602}, k)$, to two decimal places, are 1671.12, 1743.96, 92.58, 19.74, 0.0 and 0.0, for $k = 1, \ldots, 6$, providing reasonably strong evidence in favour of $k_0 = 4$ in this case.

Theorem 3 *Assume that the conditions of Theorem 1 obtain and suppose that* $K \geq k_0$ *where* $k_0 < \infty$*. If* $\eta_T \to 0$ *as* $T \to \infty$ *such that* $R(Y^T, k)/\eta_T \to 0$ *almost surely for all* $k > k_0$*, then the two step estimators* $\hat{k}_{II}$ *and* $\hat{k}'_{II}$ *are both strongly consistent for* k_0*.*

To construct $\hat{k}_{II}$ or $\hat{k}'_{II}$ the practitioner will have to select values for η_T. If $\eta_T = \eta$ where η is a predetermined positive constant chosen such that $\eta < K(\phi^0_{k_0}, \phi^0_{k_0-1})$ then $\hat{k}_{II}$ and $\hat{k}'_{II}$ will be strongly consistent because $R(Y^T, k)$ will be arbitrarily close to zero for T sufficiently large when $k > k_0$ whereas $R(Y^T, k_0)$ will converge to $K(\phi^0_{k_0}, \phi^0_{k_0-1})$ and hence eventually k_0 will be the largest k in either $\mathcal{K}_T(\hat{k}_I)$ or $\mathcal{K}_T(\hat{k}'_I)$. Since $K(\phi^0_{k_0}, \phi^0_{k_0-1})$ is unknown the inequality $\eta_T < K(\phi^0_{k_0}, \phi^0_{k_0-1})$ can be achieved asymptotically by choosing η_T such that $\eta_T \to 0$ as $T \to \infty$. But, as indicated in Theorem 3, the magnitude of η_T must not decrease too quickly otherwise the condition required to maintain consistency, $R(Y^T, k) < \eta_T$ for $k > k_0$, cannot be ensured. These arguments intimate that small values of η_T will decrease the chance of underestimation while larger values will decrease the chance of overestimation. Lemmas 4 and 5 shed some additional light on the choice of η_T since they imply that $R(Y^T, k)$ is at most $O((\log\log T + \log T)/T)$ when $k > k_0$ and in view of this result we could in practice set $\eta_T = \frac{1}{2}\{(\log T)^{5/4}/T\}$, for example, and achieve the conditions of Theorem 3. It is unlikely, however, that any one choice of the tuning parameter η_T will be optimal, in the sense of maximizing the probability of correctly selecting k_0, over all possible structures, parameterizations and sample sizes.

6 Empirical Illustrations

6.1 Finite mixture models

Böhning et al. (1992) have employed several data sets available in the literature to demonstrate the use of finite mixture models and here we will illustrate the application of the techniques described above to this data. Computer software and the data sets can be accessed at: http://ftp.ukbf.fu-berlin.de/sozmed/caman.html. Note that for finite mixture models $\log L^m(Y^T \mid \hat{\phi}_k)$ gives the exact likelihood function and in what follows $\Delta_T(k)$ based on the exact likelihood function with $d_{kT} = \varphi_k$ and $d_{kT} = (\varphi_k/2)\log T$ will be labelled AIC and BIC respectively, $\text{BIC}_2 = \log L^m(Y^T \mid \hat{\phi}_k) - \varphi_k \log T$.

The first example relates to a study from northeast Thailand in which the health status of 602 preschool children was checked every 2 weeks from June 1982 until September 1985. The data consists of the number of occurrences of fever and/or cough symptoms recorded during the study period and is described in more detail in Schelp, Vivatanasept, Sitaputra, Sormani, Pongpaew, Vudhivai, Egormaiphol & Böhning (1990). Fitting a finite Poisson mixture model with eight components produces the maximum likelihood estimate of the distribution as plotted in Figure 2. Böhning et al. (1992) argued that there is clear evidence of separation into three groups, those who were often sick, those who were frequently sick and those who were rarely, if ever, sick, these groups comprising about 5%, 30% and 65% of the children respectively. They also suggested that the last group might be separated further into those children who were seldom sick and those who were never sick.

To illustrate our methods we list in Table 6.1 the values of the log-likelihood, AIC, BIC and BIC$_2$ for up to 8 components. One can clearly see that $\log L^m(Y^T \mid \hat{\phi}_k)$ increases monotonically with k for $k < 4$ but becomes flat thereafter, *c.f.* Figure 1. AIC, BIC and BIC$_2$ all choose a four-component model and the second step modification $\hat{k}_{II}$ also yields $k = 4$. Both $\hat{k}_I$ and $\hat{k}'_I$ give the same result. Thus we would suggest that a four-component

Figure 2: Maximum likelihood estimates. Eight component Poisson mixture for preschool health status data

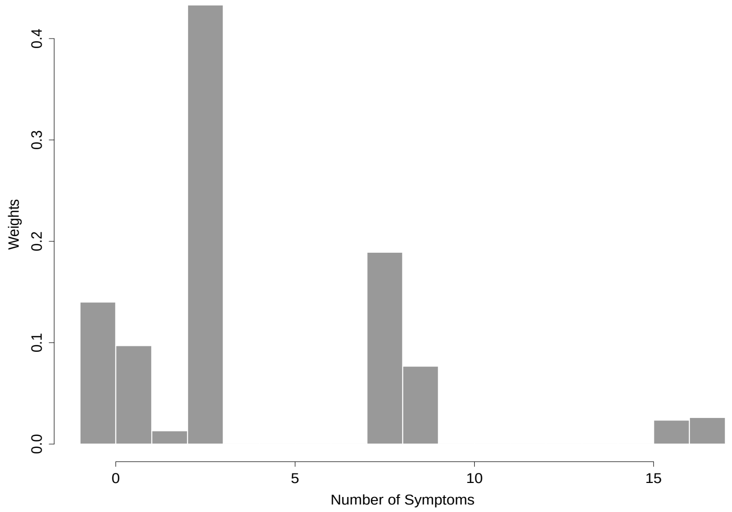


Table 1 Log likelihood, $LL_m = \log L^m(Y^T \mid \hat{\phi}_k)$, AIC, BIC and BIC$_2$ for preschool health status data

	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8
LL_m	-3317.39	-1646.27	-1573.43	-1553.69	-1553.69	-1553.69	-1553.69	-1553.68
AIC	-3318.39	-1649.27	-1578.43	-1560.69	-1562.69	-1564.69	-1566.69	-1568.68
BIC	-3320.66	-1656.10	-1589.81	-1576.62	-1583.18	-1589.73	-1596.29	-1602.83
BIC$_2$	-3323.79	-1665.47	-1605.43	-1598.49	-1611.29	-1624.09	-1636.89	-1649.29

model provides a good representation of this data.

Böhning et al. (1992) also fitted translated standard normal mixtures to the anthropometric measurements of 708 preschool children who were examined for subclinical malnourishment. They found a two-component model to be appropriate and this was confirmed by the results from $\hat{k}_I$, $\hat{k}'_I$, $\hat{k}_{II}$ and $\hat{k}'_{II}$.

The second example concerns a study of the number of occurrences of sudden infant death syndrome (cot deaths) in 100 North Carolina (U.S.A) counties over a 4-year period. Symons, Grimson & Yuan (1983) used a two-component Poisson mixture to model this data with $f(y \mid \theta_s) = \exp(-\theta_s)(\theta_s)^y/y!$ $s = 1, 2$ where y denotes the number of deaths and $\theta_s = \beta_s N$ with β_s representing the incident rate per birth with N the number of live births. The two components were interpreted as representing normal and high-risk categories. Böhning et al. (1992), however, suggested using a four-component model based on the nonparametric maximum likelihood estimate of the mixing distribution. The values of the fitted log likelihood for k from 1 to 5 are -255.57, -237.28, -234.40, -233.36 and -233.35 respectively. AIC with $K = 5$ chooses a four-component model, although the second step modification $\hat{k}_{II}$ reduces the order to two, which is the same order chosen by BIC and BIC$_2$. The parameter estimates for the two-component model are $\hat{\pi}_1 = 1 - \hat{\pi}_2 = 0.75$ with $\hat{\beta}_1 = 0.0016$ and $\hat{\beta}_2 = 0.0035$ while a four component model yields the mixing probabilities $(0.322, 0.515, 0.152, 0.011)$ with rates of incidence $(0.0012, 0.0021, 0.0037, 0.0090)$. These estimates indicate that it may be reasonable to combine the first two adjacent components and trim the last component in the four-component structure to form a two-component model, as is suggested by the selection criteria.

6.2 Hidden Markov models

Hidden Markov chains are one of the many tools used to analyze DNA sequences, see Churchill (1989). In these models the DNA is assumed to be composed of homogeneous segments belonging to a small number of distinct compositional classes and the probability of observing a base $y_t \in \{C, T, A, G\}$ at a given site t on the molecule depends on the type of segment in which it lies. An underlying organization of the DNA is supposed in which switching from one segment to the next follows an unobserved Markov chain, the states of the hidden process indicating the type of segment. Thus the number of states of the hidden Markov chain corresponds to the number of distinct segments and has an important influence on the probability of observing different bases. Churchill (1992) used BIC based on the full likelihood function to choose the number of states and here we will illustrate the use of penalized quasi-likelihood methods based on finite mixtures.

The first data set contains the simian virus 40 genome, which is a circular double-stranded DNA molecule of 5243 bass-pairs, and the second consists of the complete genome of the bacteriophage lambda, a double-stranded circular DNA molecule of 48,502 base pairs. (The data is available at http://www.ncbi.nlm.nih.gov/.) The results are listed in Table 2. In this table AICm and BICm indicate the number of states chosen by the criterion $\Delta_T(k)$ with $d_{kT} = \varphi_k$ and $d_{kT} = (\varphi_k/2) \log T$ respectively and, as previously, AIC and BIC indicate the values obtained using penalized likelihood methods based on the exact likelihood function. The final column gives the number of states determined by the pattern identification method described in Zhang & Stine (2001).

Table 2 Estimated number of homogeneous segments of DNA sequence

DNA	AIC	AICm($\hat{k}_{II}$)	BIC	BICm	Pattern
SV40	4	4(3)	2	2	2
Lambda	4	4(4)	3	4	4

For simian virus 40, the pattern method indicates that the temporal dependence of neighbouring bases is characterized by an $ARMA(1, 1)$ process, which implies that at least two states are needed to capture the structure of the sequence. BIC and BICm choose a two-state model, but AIC and AICm select a four-state model and the second step modification of AICm, AICm($\hat{k}_{II}$), gives a three-state model. From the theoretical results presented above we can anticipate that AIC and AICm will have a tendency to overestimate k_0 and although $\hat{k}_{II}$ can be used to reduce the chance of over estimation BIC and BICm seem likely to produce a more accurate estimate of the true number of states. In fact it is known, Reddy *et al.* (1978), that the expression of simian virus 40 genes is regulated by two major transcripts and, as shown by Churchill (1992), these transcripts are apparent in the data, the two states reflecting regions of distinct dinucleotide composition. Thus it seems sensible to believe in a simple two-state model for simian virus 40 as suggested by BIC and BICm.

For bacteriophage lambda the BIC criterion chooses a three-state model while the other criteria, AIC, AICm, $\hat{k}_{II}$ and BICm, choose a four-state model. The pattern method also suggests a hidden Markov chain of at least four states is needed to describe the temporal dependence between neighbouring bases. Based on other analyses, Churchill (1992) concluded that compositional variation in bacteriophage lambda does not fall into a small number of distinct states, which seems to be the conclusion reached by the methods considered here.

7 Conclusion

In this paper we have developed four estimators for the number of components in a finite mixture. By interpreting the mixture likelihood as a quasi-likelihood we have also suggested how these estimators can be employed to determine the number of states in a hidden Markov chain process. Strong consistency of the estimators under suitable regularity, including Markov dependency, has been established.

Empirical illustrations indicate that the application of our techniques to real world data sets produces outcomes that are both heuristically understandable and scientifically explicable. The observed behaviour of the methods is also in close accord with what might be expected from theoretical considerations, with all four estimates being coincident more often than not. The results suggest that little if anything may be lost in terms of statistical performance by using our methods whilst considerable gains may be achieved in terms of computational speed and simplicity, particularly when the sample size is large. This is especially relevant here because in the analysis of DNA sequences and ion channel records, two areas of application where hidden Markov models are becoming of increasing importance, data sets in excess of 10,000 observations are not uncommon.

Acknowledgments: The authors are extremely grateful to Professor Gary Churchill for sharing his C-code and data sets. The second author would also like to thank Professor Robert Stine for his support and guidance during the production of this work and Professor Peter Hall and his colleagues for the warm hospitality he received at The Australian National University where the major part of the research was carried out.

Appendix: Proofs

Proof of Lemma 1: See the proof of Lemma 2 in Rydén (1995). ∎

Proof of Lemma 2: By Lemma 1 $\Phi_k^0 \subset \Phi_k$ and $\Phi_k^0 \neq \emptyset$. When $k < k_0$, there is no $\phi_k \in \Phi_k$ such that $p(y \mid \phi_k) = p(y \mid \phi_{k_0}^0)$ and by a standard application of Jensen's inequality $K(\phi_{k_0}^0, \phi_k) \geq 0$ with equality holding if and only if $p(y \mid \phi_{k_0}^0) = p(y \mid \phi_k)$ and therefore $K(\phi_{k_0}^0, \Phi_k) > 0$. When $k \geq k_0$ there exists a $\phi_k \in \Phi_k$ such that $p(y \mid \phi_k) = p(y \mid \phi_{k_0}^0)$ and hence $K(\phi_{k_0}^0, \Phi_k) = 0$ for $k \geq k_0$ by the same argument. The monotone structure of $K(\phi_{k_0}^0, \Phi_k)$ for $k < k_0$ follows directly from Lemma 3 of Leroux (1992). ∎

Proof of Lemma 3: Both (4.3) and (4.4) are examples of the strong law of large numbers applied to mixing processes. In the case of (4.3), for example, it is clear that the expected value of $\log p(Y_t \mid \phi_k)$ equals $\int p(y \mid \phi_{k_0}^0) \log p(y \mid \phi_k) \mu(dy)$ and from Lindgren (1978) page 87 we know $\{Y_t\}$ is a stationary and ergodic process that is strongly mixing at a geometric rate. By Theorem 14.1 of Davidson (1994) $\log p(Y_t \mid \phi_k)$ is also α-mixing at a geometric rate and therefore, by Davidson (1994) Theorem 20.19 or Theorem 5 of Oodairia & Yoshihara (1971), $T^{-1} \log L^m(Y^T \mid \phi_k)$ converges to $\int p(y \mid \phi_{k_0}^0) \log p(y \mid \phi_k) \mu(dy)$ almost surely. Moreover, by Assumption C3 $T^{-1} \log L^m(Y^T \mid \phi_k)$ is continuously differentiable on Φ_k for all T and therefore by Theorem 21.10 of Davidson (1994) it follows that the function sequence $T^{-1} \log L^m(Y^T \mid \phi_k)$, $T = 1, 2, \ldots$, is stochastically equicontinuous and hence the convergence is uniform, see Davidson (1994), §21.4, for details.

Since Φ_k is compact every sequence of estimates $\{\hat{\phi}_k\}$ contains at least one cluster point $\phi_k^* \in \Phi_k$. Let $\{\hat{\phi}_{k,T}\} = \{\hat{\phi}_k : T = T_1, T_2, \ldots\}$ denote a subsequence converging to ϕ_k^*. By continuity and uniform convergence

$$\lim_{T\to\infty} T^{-1}|\log L^m(Y^T \mid \hat{\phi}_{k,T}) - \log L^m(Y^T \mid \phi_k^*)| = 0 \text{ and}$$

$$\lim_{T\to\infty} |T^{-1}\log L^m(Y^T \mid \phi_k^*) - \int p(y \mid \phi_{k_0}^0)\log p(y \mid \phi_k^*)\mu(dy)| = 0$$

with probability one. By definition, however, $\int p(y \mid \phi_{k_0}^0)\log\{p(y \mid \phi_k^0)/p(y \mid \phi_k^*)\}\mu(dy) \geq 0$ and $\log L^m(Y^T \mid \hat{\phi}_k) \geq \log L^m(Y^T \mid \phi_k^0)$ for all T, leading to the conclusion that $\int p(y \mid \phi_{k_0}^0)\log\{p(y \mid \phi_k^0)/p(y \mid \phi_k^*)\}\mu(dy) = 0$. It follows that ϕ_k^* belongs to the coset Φ_k^0, but this is true for all cluster points ϕ_k^*, giving (4.3).

Readers are referred to Leroux (1992), Lemma 1 and Theorem 1, and Rydén (1995), Lemma 3 for results whose content and proof is closely analogous to that of (4.3). The result in (4.4) follows directly from Theorem 20.19 of Davidson (1994) or Theorem 5 of Oodairia & Yoshihara (1971) since ϕ_k is fixed and $\partial^2\log p(Y_t \mid \phi_k)/\partial\phi_k\partial\phi_k'$ is α-mixing at a geometric rate, the interchange of the operations of differentiation and integration being allowed by virtue of the integrability conditions C3'. ∎

Proof of Theorem 1: First consider $\hat{k}_I'$. Clearly $\hat{k}_I' \geq 1$; suppose therefore that $k_0 > 1$. To show that $\hat{k}_I' \geq k_0$ as $T \to \infty$ if $\overline{\lim}_{T\to\infty} d_{kT}/T = 0$ observe that by Lemma 3

$$\lim_{T\to\infty} \frac{\log L^m(Y^T \mid \hat{\phi}_{k+1}) - \log L^m(Y^T \mid \hat{\phi}_k)}{T} = K(\phi_{k_0}^0, \phi_k^0) - K(\phi_{k_0}^0, \phi_{k+1}^0) \tag{A.1}$$

almost surely. In view of Lemma 2 the right hand side of equation (A.1) is positive for all $k < k_0$ and this implies that $\lim_{T\to\infty} T^{-1}\{\Delta_T(k) - \Delta_T(k+1)\} < 0$, $k = 1, \ldots, k_0 - 1$, with probability one since $\lim_{T\to\infty}(d_{(k+1)T} - d_{kT})/T = 0$. Hence $\hat{k}_I' \geq k_0$ almost surely. This result also leads to the conclusion that $\lim_{T\to\infty} \hat{k}_I = K$ if $K < k_0$ and that $\liminf_T \hat{k}_I \geq k_0$ when $k_0 \leq K$, as required. ∎

Proof of Lemma 4: From Lemma 3 we know that if ϕ_k^* is a cluster point of $\{\hat{\phi}_k\}$ then $\phi_k^* \in \Phi_k^0$ and since this is true for all subsequential limit points it follows that $\hat{\phi}_k \to \Phi_k^0$ almost surely as $T \to \infty$ where convergence is defined relative to the quotient norm $\inf\{\|\phi_k - \phi_k^0\| : \phi_k^0 \in \Phi_k^0\}$ on Φ_k^0. Let $\{\hat{\phi}_{k,T}\} = \{\hat{\phi}_k : T = T_1, T_2, \ldots\}$ denote a subsequence converging to the element ϕ_k^0 of Φ_k^0.

If $\phi_k^0 \in \text{Int}\{\Phi_k\}$ then there exists an $\epsilon > 0$ such that $\mathcal{B}_\epsilon(\phi_k^0) = \{\phi_k : \|\phi_k - \phi_k^0\| < \epsilon\} \subset \text{Int}\{\Phi_k\}$ and the event $\hat{\phi}_{k,T} \in \mathcal{B}_\epsilon(\phi_k^0)$ will occur infinitely often with probability one. Thus $\hat{\phi}_{k,T}$ will be a critical point of $\log L^m(Y^T \mid \phi_k)$ and a second order Taylor series expansion of $\log L^m(Y^T \mid \phi_k^0)$ about $\log L^m(Y^T \mid \hat{\phi}_{k,T})$ yields

$$\log L^m(Y^T \mid \hat{\phi}_{k,T}) - \log L^m(Y^T \mid \phi_k^0) = \frac{T}{2}(\hat{\phi}_{k,T} - \phi_k^0)' H_T(\hat{\phi}_{k,T})(\hat{\phi}_{k,T} - \phi_k^0) + R_{2,T}(\hat{\phi}_{k,T}, \phi_k^0) \tag{A.2}$$

where $H_T(\phi_k) = -T^{-1}\partial^2\log L^m(Y^T \mid \phi_k)/\partial\phi_k\partial\phi_k'$ and $|R_{2,T}(\hat{\phi}_{k,T}, \phi_k^0)| = o(\|\hat{\phi}_{k,T} - \phi_k^0\|^2)$, see Remark 1 following the proof of Theorem 6.10 of Marsden (1974). Similarly, using the Mean Value Theorem for vector-valued functions, see Theorem 6.7 of Marsden (1974) or Apostol (1974) Theorem 12.9, we can express the gradient of $\log L^m(Y^T \mid \phi_k)$ at ϕ_k^0,

$\nabla L_T^m(\phi_k^0) = \partial \log L^m(Y^T \mid \phi_k^0)/\partial\phi_k$, as

$$\nabla L_T^m(\phi_k^0) = T H_T(\bar{\phi}_{k,T}^0)(\hat{\phi}_{k,T} - \phi_k^0) \tag{A.3}$$

where $\bar{\phi}_{k,T}^0 = \phi_k^0 + \lambda \odot (\hat{\phi}_{k,T} - \phi_k^0)$, the sum of ϕ_k^0 and the Hadamard product of $\lambda = (\lambda_1, \ldots, \lambda_{\varphi_k})'$, $0 \leq \lambda_i \leq 1$, $i = 1, \ldots, \varphi_k$, with $\hat{\phi}_{k,T} - \phi_k^0$. The Hessian matrix $H_T(\bar{\phi}_{k,T}^0)$ converges to $H_k^0 = \partial^2 K(\phi_{k_0}^0, \phi_k^0)/\partial\phi_k\partial\phi_k'$ with probability one because $\|H_T(\bar{\phi}_{k,T}^0) - H_k^0\| \leq \|H_T(\bar{\phi}_{k,T}^0) - H_T(\phi_k^0)\| + \|H_T(\phi_k^0) - H_k^0\|$ and the first term on the right hand side approaches zero by continuity and the fact that $\|\bar{\phi}_{k,T}^0 - \phi_k^0\| \leq \|\hat{\phi}_{k,T} - \phi_k^0\| \to 0$ as $T \to \infty$ and the second converges to zero by Lemma 3. Substituting $H_T(\bar{\phi}_{k,T}^0) = H_k^0 + \Delta_0\hat{H}_k$ in (A.3) and using $H_k^{0\dagger}$ to denote the Moore-Penrose inverse of H_k^0 we find, after some straightforward algebra, that

$$\begin{aligned} \nabla L_T^m(\phi_k^0)' H_k^{0\dagger} \nabla L_T^m(\phi_k^0) &= T^2(\hat{\phi}_{k,T} - \phi_k^0)'[H_k^0 + \Delta_0\hat{H}_k]' H_k^{0\dagger}[H_k^0 + \Delta_0\hat{H}_k](\hat{\phi}_{k,T} - \phi_k^0) \\ &= T^2\left\{(\hat{\phi}_{k,T} - \phi_k^0)' H_k^0 (\hat{\phi}_{k,T} - \phi_k^0) + o(\|\hat{\phi}_k - \phi_k^0\|^2)\right\} \ . \end{aligned} \tag{A.4}$$

The expression in (A.4) derives first, from the equality $H_k^0 H_k^{0\dagger} H_k^0 = H_k^0$, and second, from the fact that $\Delta_0\hat{H}_k = o(1)$ implies that the quadratic form

$$(\hat{\phi}_{k,T} - \phi_k^0)'(H_k^0 H_k^{0\dagger}\Delta_0\hat{H}_k + \Delta_0\hat{H}_k H_k^{0\dagger} H_k^0 + \Delta_0\hat{H}_k H_k^{0\dagger}\Delta_0\hat{H}_k)(\hat{\phi}_{k,T} - \phi_k^0)$$

is $o(\|\hat{\phi}_k - \phi_k^0\|^2)$. Writing $\lambda_{max}(H_k^{0\dagger})$ for the largest eigenvalue of the matrix $H_k^{0\dagger}$ and applying the Rayleigh-Ritz theorem to the left hand side of (A.4) leads to the conclusion that

$$T^2\left\{(\hat{\phi}_{k,T} - \phi_k^0)' H_k^0 (\hat{\phi}_{k,T} - \phi_k^0) + o(\|\hat{\phi}_{k,T} - \phi_k^0\|^2)\right\} \leq \lambda_{max}(H_k^{0\dagger})\|\nabla L_T^m(\phi_k^0)\|^2 \ .$$

But $(\hat{\phi}_{k,T} - \phi_k^0)' H_k^0 (\hat{\phi}_{k,T} - \phi_k^0) = (\hat{\phi}_{k,T} - \phi_k^0)' H_T(\hat{\phi}_{k,T})(\hat{\phi}_{k,T} - \phi_k^0) + o(\|\hat{\phi}_{k,T} - \phi_k^0\|^2)$ because $H_T(\hat{\phi}_{k,T}) = H_k^0 + o(1)$. Therefore we can bound $\log\{L^m(Y^T \mid \hat{\phi}_{k,T})/L^m(Y^T \mid \phi_k^0)\} \geq 0$ by

$$\frac{1}{2T\lambda_{min}^+(H_k^0)}\|\nabla L_T^m(\phi_k^0)\|^2 + o(\|\hat{\phi}_{k,T} - \phi_k^0\|^2) \tag{A.5}$$

where $\lambda_{min}^+(H_k^0)$ denotes the smallest nonzero eigenvalue of H_k^0.

To establish the order of magnitude of (A.5) observe that

$$\|\nabla L_T^m(\phi_k^0)\|^2 = \sum_{r=1}^{\varphi_k} |\sum_{t=1}^{T} \frac{\partial \log p(Y_t \mid \phi_k^0)}{\partial\phi_{kr}}|^2 \ .$$

Now, as in the proof of Lemma 3 we can show that $\{\partial \log p(Y_t \mid \phi_k^0)/\partial\phi_{kr}\}$ is a zero mean, geometrically α-mixing process, the zero mean arising because ϕ_k^0 is by definition a critical point of $\int p(y \mid \phi_{k_0}^0) \log p(y \mid \phi_k)\mu(dy)$ and hence, by C3', $E[\partial \log p(Y_t \mid \phi_k^0)/\partial\phi_{kr}] = 0$. It follows that $|\sum_{t=1}^T \partial \log p(Y_t \mid \phi_k^0)/\partial\phi_{kr}|$ obeys the law of the iterated logarithm. Applying this bound in (A.3) leads us to the conclusion that $(\lambda_{min}^+(H_k^0) + o(1))^2\|\hat{\phi}_{k,T} - \phi_k^0\|^2 = O(\log\log T/T)$, from which we can conclude that (A.5), and hence (A.2), is $O(\log\log T)$.

If ϕ_k^0 lies on the boundary $\Phi_k \backslash \mathrm{Int}\{\Phi_k\}$ then $\hat{\phi}_{k,T}$ and ϕ_k^0 need no longer correspond to stationary points of $\log L^m(Y^T \mid \phi_k)$ and $E[\log p(Y_t \mid \phi_k)]$, respectively, and the above derivation must be appropriately modified. In this case set $\mathcal{R}_\epsilon(\phi_k^0) = \mathcal{B}_\epsilon(\phi_k^0) \cap (\Phi_k \backslash \mathrm{Int}\{\Phi_k\})$, $\epsilon > 0$, and let $\hat{\phi}_{k,T}^{\mathcal{R}} = \mathrm{argmax}_{\phi_k \in \mathcal{R}_\epsilon(\phi_k^0)} \log L^m(Y^T \mid \phi_k)$. Then $\|\hat{\phi}_{k,T} - \phi_k^0\| < \epsilon$ for T sufficiently large and the difference $\log L^m(Y^T \mid \hat{\phi}_{k,T}) - \log L^m(Y^T \mid \phi_k^0)$ equals the sum of $\log L^m(Y^T \mid \hat{\phi}_{k,T}) - \log L^m(Y^T \mid \hat{\phi}_{k,T}^{\mathcal{R}}) \geq 0$ and $\log L^m(Y^T \mid \hat{\phi}_{k,T}^{\mathcal{R}}) - \log L^m(Y^T \mid \phi_k^0) \geq 0$. We will establish below that both of these terms are $O(\log\log T)$.

First consider $\log L^m(Y^T \mid \hat{\phi}_{k,T}^{\mathcal{R}}) - \log L^m(Y^T \mid \phi_k^0)$. By assumption C2', if $\phi_k \in \mathcal{R}_\epsilon(\phi_k^0)$ then $r(\phi_k) = (r_1(\phi_k), \ldots, r_m(\phi_k))' = 0$ where $m < \varphi_k$. Using the Implicit Function Theorem and following the argument used by Apostol (1974) pages 381-383 to justify Langrange's method of constrained optimization, we know that there exists a continuous reparameterisation of the form $\phi_k = h(\varrho_n)$, $\varrho_n \in \mathcal{R}^n$, $n = k - m$, such that $\log L^m(Y^T \mid \hat{\phi}_{k,T}^{\mathcal{R}}) = \log L^m(Y^T \mid h(\hat{\varrho}_{n,T})) \geq \log L^m(Y^T \mid h(\varrho_n))$ for all $\varrho \in \mathcal{B}_\eta(\varrho_n^0)$ where $\phi_k^0 = h(\varrho_n^0)$ and $\varrho \in \mathcal{B}_\eta(\varrho_n^0)$ implies that $\phi_k \in \mathcal{R}_\epsilon(\phi_k^0)$. Similarly, $\int p(y \mid \phi_{k_0}^0) \log p(y \mid h(\varrho_n))\mu(dy)$ has a local interior maximum at $\varrho_n = \varrho_n^0$. Using a repetition of the derivation employed previously when $\phi_k^0 \in \mathrm{Int}\{\Phi_k\}$ we can deduce that $\log L^m(Y^T \mid h(\hat{\varrho}_{n,T})) - \log L^m(Y^T \mid h(\varrho_n^0)) = O(\log\log T)$.

Aside: To illustrate the constructions used in the previous two paragraphs, suppose that a k component mixture is fitted to data from a process of order $k_0 < k$ where $p(y \mid \phi_{k_0}^0) = \sum_{i=1}^{k_0} \pi_i^0 f(y \mid \theta_i^0)$. Then $\hat{\phi}_{k,T}$ converges to the parameter set Φ_k^0 where $\Phi_k^0 = \{\phi_k \in \Phi_k : \phi_k = (\pi_1^0, \ldots, \pi_{k_0}^0, 0, \ldots, 0, \theta_1^0, \ldots, \theta_{k_0}^0, \theta_{k_0+1}^*, \ldots, \theta_k^*)\}$ where $\theta_{k_0+1}^*, \ldots, \theta_k^*$ denote arbitrary points in Θ. The restrictions that define $\mathcal{R}_\epsilon(\phi_k^0)$ are $r_i(\phi_k) = \pi_{k_0+i} = 0$, $i = 1, \ldots, m-1$, $r_m(\phi_k) = \sum_{i=1}^{k_0} \pi_i - 1 = 0$, $m = k - k_0$, and the restricted estimate $\hat{\phi}_{k,T}^{\mathcal{R}} = (\hat{\pi}_{1,T}, \ldots, \hat{\pi}_{k_0,T}, 0, \ldots, 0, \hat{\theta}_{1,T}, \ldots, \hat{\theta}_{k_0,T}, \theta_{k_0+1,T}^*, \ldots, \theta_{k,T}^*)$ where $\hat{\pi}_{1,T}, \ldots, \hat{\pi}_{k_0,T}$ and $\hat{\theta}_{1,T}, \ldots, \hat{\theta}_{k_0,T}$ are the unrestricted estimates obtained from the k_0 component model. The reparameterisation induced by the restrictions yields $\varrho_n = (\pi_1, \ldots, \pi_{k_0-1}, \theta_1, \ldots, \theta_k)$ where $n = \varphi_k - k + k_0$. The maximizing value $\hat{\varrho}_{n,T} = (\hat{\pi}_{1,T}, \ldots, \hat{\pi}_{k_0-1,T}, \hat{\theta}_{1,T}, \ldots, \hat{\theta}_{k_0,T}, \theta_{k_0+1,T}^*, \ldots, \theta_{k,T}^*)$, and the maximized log-likelihood equals $\sum_{t=1}^T \log\{\sum_{i=1}^{k_0} \hat{\pi}_{i,T} f(y_t \mid \hat{\theta}_{i,T})\}$. The Lagrangians associated with the constrained optimum are

$$
\begin{aligned}
\ell_i &= \sum_{t=1}^T \left(\frac{f(y_t \mid \theta_{k_o+i,T}^*) - f(y_t \mid \theta_{k,T}^*)}{\sum_{i=1}^{k_0} \hat{\pi}_{i,T} f(y_t \mid \hat{\theta}_{i,T})} \right), \text{ for } i = 1, \ldots, m-1, \text{ and} \\
\ell_m &= \sum_{t=1}^T \left(1 - \frac{f(y_t \mid \theta_{k,T}^*)}{\sum_{i=1}^{k_0} \hat{\pi}_{i,T} f(y_t \mid \hat{\theta}_{i,T})} \right) .
\end{aligned}
$$

in this example.

Now consider $\log L^m(Y^T \mid \hat{\phi}_{k,T}) - \log L^m(Y^T \mid \hat{\phi}_{k,T}^{\mathcal{R}})$. By Lemma 6 of this appendix this term is $O(T\|\nabla L_T^m(\phi_k^0)/T\|^2)$. To evaluate the order of magnitude of $\|\nabla L_T^m(\phi_k^0)/T\|$ it is sufficient to observe, from the Chain-Rule, that

$$
\frac{\partial \log L^m(Y^T \mid h(\varrho_n^0))}{\partial \varrho_n} = \frac{\partial h'(\varrho_n)}{\partial \varrho_n} \frac{\partial \log L^m(Y^T \mid h(\varrho_n))}{\partial \phi_k} \bigg|_{\varrho_n = \varrho_n^0}
$$

and therefore

$$
\begin{aligned}
\|\nabla L_T^m(\phi_k^0)/T\| &\leq \| \left(\frac{\partial h(\varrho_n^0)}{\partial \varrho_n'} \frac{\partial h'(\varrho_n^0)}{\partial \varrho_n} \right)^\dagger \frac{\partial h(\varrho_n^0)}{\partial \varrho_n'} \| \cdot \| \frac{\partial \log L^m(Y^T \mid h(\varrho_n^0))}{T \partial \varrho_n} \| \\
&= \left(\mathrm{tr} \left(\frac{\partial h(\varrho_n^0)}{\partial \varrho_n'} \frac{\partial h'(\varrho_n^0)}{\partial \varrho_n} \right)^\dagger \right)^{\frac{1}{2}} \cdot \| \frac{\partial \log L^m(Y^T \mid h(\varrho_n^0))}{T \partial \varrho_n} \| .
\end{aligned}
$$

Since we can show that $\{\partial \log p(Y_t \mid h(\varrho_n^0))/\partial \varrho_{nr}\}$ is a zero mean process, α-mixing at a geometric rate, it follows that $\|\nabla L_T^m(\phi_k^0)/T\| = O(\sqrt{\log\log T/T})$ and that $\log L^m(Y^T \mid \hat{\phi}_{k,T}) - \log L^m(Y^T \mid \hat{\phi}_{k,T}^{\mathcal{R}}) = O(\log\log T)$, as required.

Finally, if the sequence $\{\hat{\phi}_k\}$ converges, then $\{\hat{\phi}_k\} = \{\hat{\phi}_{k,T}\}$ and we are finished. Otherwise, Φ_k^0 must contain at least two cluster points of $\{\hat{\phi}_k\}$. Let $\{\phi_k^{0(1)}, \ldots, \phi_k^{0(N)}\}$ denote a finite collection of subsequential limit points chosen together with N and $\epsilon > 0$ such that $\bigcup_{n=1}^N \mathcal{B}_\epsilon(\phi_k^{0(n)})$ forms a uniform cover of Φ_k^0, which is possible by the Heine-Borel Covering Theorem. Let $\{\hat{\phi}_{k,T}^{(n)}\} = \{\hat{\phi}_k : T = T_1, T_2, \ldots\}$ denote a subsequence converging to $\phi_k^{0(n)}$. Then $\|\hat{\phi}_{k,T}^{(n)} - \phi_k^{0(n)}\| < \epsilon$ for $T > T_\epsilon$, T_ϵ sufficiently large, and since $\hat{\phi}_k \to \Phi_k^0$ almost surely $\hat{\phi}_k \in \mathcal{B}_{\epsilon_{T_\epsilon}}(\phi_k^{0(n)})$ where $\epsilon_{T_\epsilon} = \|\hat{\phi}_{k,T_\epsilon}^{(n)} - \phi_k^{0(n)}\|$ for at least one n and some $T > T' \geq T_\epsilon$. Hence $|\log L^m(Y^T \mid \hat{\phi}_k) - \log L^m(Y^T \mid \hat{\phi}_{k,T}^{(n)})| = O(\log\log T)$ by Lemma 6. But $\log L^m(Y^T \mid \hat{\phi}_{k,T}^{(n)}) - \log L^m(Y^T \mid \phi_k^{0(n)}) = O(\log\log T)$ for all n and the triangular inequality now yields $\log L^m(Y^T \mid \hat{\phi}_k) - \log L^m(Y^T \mid \phi_k^{0(n)}) = O(\log\log T)$. ∎

Proof of Lemma 5: For any $E \subseteq \mathcal{Y}^T$ let $\nu_{\phi_k}^m(E) = \int_E L^m(y^T \mid \phi_k)\mu^T(dy^T)$, $\nu_{\phi_k}(E) = \int_E p(y^T \mid \phi_k)\mu^T(dy^T)$ and consider the event $A_T = \{Y^T : \frac{1}{2}\log\{L^m(Y^T \mid \phi_k^0)/L^m(Y^T \mid \phi_{k_0}^0)\} \geq M \log\log T + \log T\}$. Clearly we have $A_T = \{Y^T : L^m(Y^T \mid \phi_k^0) \geq T^2(\log T)^{2M} L^m(Y^T \mid \phi_{k_0}^0)\}$ and therefore $\nu_{\phi_{k_0}^0}^m(A_T) \leq \nu_{\phi_k^0}^m(A_T)/T^2(\log T)^{2M} \leq (1/T^2(\log T)^{2M})$. Thus the series $\sum_T \nu_{\phi_{k_0}^0}^m(A_T)$ is convergent and so by the Borel-Cantelli lemma $\lim_{N\to\infty} \nu_{\phi_{k_0}^0}^m(\bigcup_{T=N}^\infty A_T) = 0$.

If $\mathcal{Y}^T$ is the sample space of a simple random sample of observations from a finite mixture with density $p(y \mid \phi_{k_0}^0)$ then we are finished because $p(y^T \mid \phi_{k_0}^0) = \prod_{t=1}^T p(y_t \mid \phi_{k_0}^0) = L^m(y^T \mid \phi_{k_0}^0)$ and $\nu_{\phi_{k_0}^0} = \nu_{\phi_{k_0}^0}^m$. Otherwise, observe that

$$\begin{aligned}
\nu_{\phi_{k_0}^0}(E) &= \int_E p(y^T \mid \phi_{k_0}^0)\mu^T(dy^T) \\
&= \int_E \psi(y^T \mid \phi_{k_0}^0) L^m(y^T \mid \phi_{k_0}^0)\mu^T(dy^T) \\
&= \int_E \psi_0(y^T)\nu_{\phi_{k_0}^0}^m(dy^T) \qquad \text{(A.6)}
\end{aligned}$$

where the likelihood ratio $\psi_0(y^T) = p(y^T \mid \phi_{k_0}^0)/L^m(y^T \mid \phi_{k_0}^0)$ and $\psi_0(\cdot) : \mathcal{Y}^T \to [0,\infty)\cup\{\infty\}$. Fix $b > 0$ and set $B_T = \{Y^T : \psi_0(Y^T) < bT(\log T)^M\}$. By virtue of equation (A.6) and the definitions of A_T and B_T we have the inequalities

$$\nu_{\phi_{k_0}^0}(A_T \cap B_T) < bT(\log T)^M \nu_{\phi_{k_0}^0}^m(A_T \cap B_T) \leq bT(\log T)^M \nu_{\phi_{k_0}^0}^m(A_T) \leq b/T \log T^M \ .$$

Applying Markov's inequality to $\psi_0(Y^T)$ we can also bound $\nu_{\phi_{k_0}^0}(\overline{B_T})$ by $\gamma/bT(\log T)^M$ where $\gamma = \int_{\mathcal{Y}^T} \psi_0(y^T)\nu_{\phi_{k_0}^0}(dy^T)$. It follows from the inequality $\nu_{\phi_{k_0}^0}(A_T \cap \overline{B_T}) \leq \nu_{\phi_{k_0}^0}(\overline{B_T})$ that

$$\begin{aligned}
\nu_{\phi_{k_0}^0}(A_T) &= \nu_{\phi_{k_0}^0}(A_T \cap B_T) + \nu_{\phi_{k_0}^0}(A_T \cap \overline{B_T}) \\
&\leq \frac{b^2 + \gamma}{bT(\log T)^M} \ .
\end{aligned}$$

Given $M > 1$ we can conclude that $\sum_T \nu_{\phi_{k_0}^0}(A_T) < \infty$ and hence, using the Borel-Cantelli lemma once again, $\nu_{\phi_{k_0}^0}(A_T\, i.o.) = 0$ and the lemma is proved. ∎

Proof of Theorem 2: By definition

$$\Delta_T(k) - \Delta_T(k_0) = \log \frac{L^m(Y^T \mid \hat{\phi}_k)}{L^m(Y^T \mid \hat{\phi}_{k_0})} - g(T)(h(k) - h(k_0))$$

and it suffices to show that for $k > k_0$ we will eventually have $\Delta_T(k) - \Delta_T(k_0) \leq 0$. Dividing by $\log T$ we find

$$\begin{aligned}\overline{\lim}_{T\to\infty}(\log T)^{-1}[\Delta_T(k) - \Delta_T(k_0)] \quad &\leq \quad \overline{\lim}_{T\to\infty}(\log T)^{-1} \log \frac{L^m(Y^T \mid \hat{\phi}_k)}{L^m(Y^T \mid \hat{\phi}_{k_0})} \\ &\qquad - \underline{\lim}_{T\to\infty}(\frac{g(T)}{\log T})(h(k) - h(k_0))\end{aligned}$$

and under the presumption that $\underline{\lim}_{T\to\infty} g(T)/\log T \geq 1$ the desired inequality translates into the condition that

$$(h(k) - h(k_0)) \geq \overline{\lim}_{T\to\infty}(\log T)^{-1} \log \frac{L^m(Y^T \mid \hat{\phi}_k)}{L^m(Y^T \mid \hat{\phi}_{k_0})}\ .$$

By Lemmas 4 and 5, however,

$$\begin{aligned}\overline{\lim}_{T\to\infty}(\log T)^{-1} \log \frac{L^m(Y^T \mid \hat{\phi}_k)}{L^m(Y^T \mid \hat{\phi}_{k_0})} \quad &= \quad \overline{\lim}_{T\to\infty}(\log T)^{-1} \Bigg[\log \frac{L^m(Y^T \mid \hat{\phi}_k)}{L^m(Y^T \mid \phi^0_k)} + \\ &\log \frac{L^m(Y^T \mid \phi^0_k)}{L^m(Y^T \mid \phi^0_{k_0})} - \log \frac{L^m(Y^T \mid \hat{\phi}_{k_0})}{L^m(Y^T \mid \phi^0_{k_0})}\Bigg] \qquad \text{(A.7)}\end{aligned}$$

is less than or equal to two with probability one. Equation (A.7) indicates that the requirement that $\Delta_T(k) - \Delta_T(k_0) \leq 0$ is assured whenever $(h(k) - h(k_0)) \geq 2$, which completes the proof. ∎

Proof of Theorem 3: Consider $\hat{k}_{II}$. Theorem 1 implies that for all $\epsilon > 0$ there exists a $T_\epsilon < \infty$ such that $\hat{k}_I \geq k_0$ for all $Y^T \in \mathcal{Y}^T \backslash A_T$ where the events $A_T \subset \mathcal{Y}^T$ have measure $\nu_{\phi^0_{k_0}}(\bigcup_{T=N}^{\infty} A_T) < \epsilon$, $N > T_\epsilon$. Similarly, from the expression for $R(Y^T, k)$ we see that Lemmas 2 and 3 imply that with probability one $\lim_{T\to\infty} R(Y^T, k) = 0$ if $k > k_0$ and $\lim_{T\to\infty} R(Y^T, k) = \{K(\phi^0_{k_0}, \phi^0_{k-1}) - K(\phi^0_{k_0}, \phi^0_k)\} + \{K(\phi^0_{k_0}, \phi^0_k) - K(\phi^0_{k_0}, \phi^0_{k+1})\} > 0$ if $k \leq k_0$. This means that if $\eta_T \to 0$ as $T \to \infty$ such that $R(Y^T, k) = o(\eta_T)$ for $k > k_0$ then there exists sets $B_T \subset \mathcal{Y}^T$ such that for every $Y^T \in \mathcal{Y}^T \backslash B_T$ $R(Y^T, k) > \eta_T$ if $1 \leq k \leq k_0$ and $R(Y^T, k)/\eta_T < \delta$, $\delta > 0$, if $k_0 < k < K$, where $\nu_{\phi^0_{k_0}}(\bigcup_{T=N}^{\infty} B_T) < \epsilon$ whenever $N > T'_\epsilon$, T'_ϵ sufficiently large. By the definition of $\mathcal{K}_T(\hat{k}_I)$ we therefore have that $k \in \mathcal{K}_T(\hat{k}_I)$ if and only if $k \leq k_0$ on the set $(\mathcal{Y}^T \backslash A_T \cap \mathcal{Y}^T \backslash B_T)$ and $\nu_{\phi^0_{k_0}}(\bigcup_{T=N}^{\infty}\{A_T \bigcup B_T\}) < 2\epsilon$ for $N > \max\{T_\epsilon, T'_\epsilon\}$. Since $\epsilon > 0$ is arbitrary we can conclude that $\hat{k}_{II}$ converges to k_0 almost surely. The proof that $\hat{k}'_{II}$ converges to k_0 is identical and is therefore omitted. ∎

Lemma 6 *Let* $\{\hat{\phi}_{k,T}\} = \{\hat{\phi}_k : T = T_1, T_2, \ldots\}$ *denote a subsequence converging to* $\phi^0_k \in \Phi^0_k$. *Then for* T *sufficiently large the bound* $\log L^m(Y^T \mid \hat{\phi}_{k,T}) - \log L^m(Y^T \mid \phi_k) = O(\|\nabla L^m_T(\phi^0_k)\|^2/T)$ *applies for all* $\phi_k \in \mathcal{B}_\epsilon(\phi^0_k)$, $\epsilon \leq \|\phi^0_k - \hat{\phi}_{k,T}\|$.

Proof: A first order Taylor series expansion of $\log L^m(Y^T \mid \hat{\phi}_{k,T})$ about $\log L^m(Y^T \mid \phi_k)$

and the Cauchy-Schwartz inequality yield the result that

$$\begin{aligned}\log\{L^m(Y^T \mid \hat{\phi}_{k,T})/L^m(Y^T \mid \phi_k)\} &= (\hat{\phi}_{k,T}-\phi_k)'\nabla L^m_T(\phi_k) + o(\|\hat{\phi}_{k,T}-\phi_k\|)\\ &\leq 2\|\hat{\phi}_{k,T}-\phi^0_k\|\cdot\|\nabla L^m_T(\phi_k)\| + o(\|\hat{\phi}_{k,T}-\phi^0_k\|)\end{aligned} \qquad (A.8)$$

since for T sufficiently large $\|\hat{\phi}_{k,T}-\phi^0_k\| < \delta$ for any $\delta > 0$ and $\|\hat{\phi}_{k,T}-\phi_k\| \leq 2\|\hat{\phi}_{k,T}-\phi^0_k\|$. But by Assumption C3 and Theorem 12.9 of Apostol (1974) there exists a constant $0 < C_1 < \infty$ such that $T^{-1}\|\nabla L^m_T(\phi^0_k) - \nabla L^m_T(\phi_k)\| \leq C_1\|\phi_k - \phi^0_k\| \leq 3C_1\|\hat{\phi}_{k,T}-\phi^0_k\|$, implying that

$$\|T^{-1}\nabla L^m_T(\phi_k)\| \leq \|T^{-1}\nabla L^m_T(\phi^0_k)\| + O(\|\hat{\phi}_{k,T}-\phi^0_k\|)\ . \qquad (A.9)$$

The inequalities (A.8) and (A.9) produce the upper bound

$$\log L^m(Y^T \mid \hat{\phi}_{k,T}) - \log L^m(Y^T \mid \phi_k) \leq O(\max\{\|\hat{\phi}_{k,T}-\phi^0_k\|\cdot\|\nabla L^m_T(\phi^0_k)\|, T\|\hat{\phi}_{k,T}-\phi^0_k\|^2\})\ .$$

As in (A.3), however,

$$\nabla L^m_T(\hat{\phi}_{k,T}) - \nabla L^m_T(\phi^0_k) = TH_T(\bar{\phi}^0_{k,T})(\hat{\phi}_{k,T}-\phi^0_k)\ . \qquad (A.10)$$

Let u_T denote a unit vector perpendicular to $\nabla L^m_T(\hat{\phi}_{k,T})$, that is, $\|u_T\| = 1$ and $u_T'\nabla L^m_T(\hat{\phi}_{k,T}) = 0$. Then $\hat{\phi}^0_{k,T} = \hat{\phi}_{k,T} + (u_T'\phi^0_k)\cdot\left|1 - |u_T'\hat{\phi}_{k,T}|/|u_T'\phi^0_k|\right|\cdot u_T$ is the projection of ϕ^0_k onto the plane that passes through $\hat{\phi}_{k,T}$ perpendicular to $\nabla L^m_T(\hat{\phi}_{k,T})$. Note that $\|\hat{\phi}_{k,T} - \hat{\phi}^0_{k,T}\| = \left||u_T'\phi^0_k| - |u_T'\hat{\phi}_{k,T}|\right| \leq \|\hat{\phi}_{k,T}-\phi^0_k\|$ and $\|\hat{\phi}^0_{k,T}-\phi^0_k\| \leq 2\|\hat{\phi}_{k,T}-\phi^0_k\|$. Multiplying (A.10) by $(\hat{\phi}_{k,T}-\hat{\phi}^0_{k,T})'/T$ gives

$$(\hat{\phi}^0_{k,T}-\hat{\phi}_{k,T})'(\nabla L^m_T(\phi^0_k)/T) = (\hat{\phi}_{k,T}-\hat{\phi}^0_{k,T})'H_T(\bar{\phi}^0_{k,T})(\hat{\phi}_{k,T}-\phi^0_k)\ . \qquad (A.11)$$

The right hand side of (A.11) equals

$$(\hat{\phi}_{k,T}-\phi^0_k)'H_T(\bar{\phi}^0_{k,T})(\hat{\phi}_{k,T}-\phi^0_k) - (\hat{\phi}^0_{k,T}-\phi^0_k)'H_T(\bar{\phi}^0_{k,T})(\hat{\phi}_{k,T}-\phi^0_k)\ . \qquad (A.12)$$

Applying the Rayleigh-Ritz theorem to the first term and using the Cauchy-Schwartz inequality together with the submultiplicative property of the Euclidean norm and the inequality $\|\hat{\phi}_{k,T}-\phi_k\| \leq 2\|\hat{\phi}_{k,T}-\phi^0_k\|$ on the second we can bounded (A.12) below by

$$\lambda_{min}(H_T(\bar{\phi}^0_{k,T}))\|\hat{\phi}_{k,T}-\phi^0_k\|^2 - \|H_T(\bar{\phi}^0_{k,T})\|\cdot\|\hat{\phi}_{k,T}-\phi^0_k\|^2\ .$$

Thus the right hand side of (A.11) is bounded below by a term $O(\|\hat{\phi}_{k,T}-\phi^0_k\|^2)$. The left hand side of (A.11), on the other hand, is bounded above by

$$\|\hat{\phi}^0_{k,T}-\hat{\phi}_{k,T}\|\cdot\|\nabla L^m_T(\phi^0_k)/T\| = O(\|\hat{\phi}_{k,T}-\phi^0_k\|\cdot\|\nabla L^m_T(\phi^0_k)/T\|)\ .$$

Thus we can conclude that $\|\hat{\phi}_{k,T}-\phi^0_k\| = O(\|\nabla L^m_T(\phi^0_k)/T\|)$ and that $\log L^m(Y^T \mid \hat{\phi}_{k,T}) - \log L^m(Y^T \mid \phi_k)$ is of order at most $T\|\nabla L^m_T(\phi^0_k)/T\|^2$. ∎

References

Apostol, T. M. (1974). *Mathematical Analysis, 2nd Edition.* Reading: Addison-Wesley.

Bickel, P. J., Ritov, Y. & Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *Annals of Statistics* **26**, 1614–1635.

Böhning, D., Schlattmann, P. & Lindsay, B. (1992). Computer assisted analysis of mixtures: Statistical algorithms. *Biometrics* **48**, 283–303.

Chen, J. & Kalbfleisch, J. (1996). Penalized minimum distance estimates in finite mixture models. *Canadian Journal of Statistics* **2**, 167–176.

Chung, S., Moore, J., Xia, L., Premkumar, L. & Gages, P. (1990). Characterization of single channel currents using digital signal processing techniques based on hidden markov models. *Philosophical Transactions Royal Society of London* **B 329**, 265–285.

Churchill, G. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin Mathematical Biology* **51**, 79–94.

Churchill, G. (1992). Hidden markov chains and the analysis of genome structure. *Computer Chemistry* **16**, 107–155.

Csiszar, I. & Shields, P. C. (2000). The consistency of the BIC markov order estimator. *Annals of Statistics* **28**, 1601–1619.

Dacunha-Castelle, D. & Gassiat, E. (1997). The estimation of the order of a finite mixture model. *Bernoulli* **3**, 279–299.

Davidson, J. (1994). *Stochastic Limit Theory.* Oxford: Oxford University Press.

Feng, Z. & McCulloch, C. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of Royal Statistical Society* **B 58**, 609–617.

Fredkin, D. & Rice, J. (1992a). Bayesian restoration of single channel patch clamp recording. *Biometrics* **48**, 427–448.

Fredkin, D. & Rice, J. (1992b). Maximum likelihood estimation and identification directly from single-channel recordings. *Proceedings of the Royal Society of London* **B 249**, 125–132.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycles. *Econometrica* **57**, 357–384.

Hamilton, J. (1996). Specification testing in markov switching time series models. *Journal of Econometrics* **70**, 127–157.

Hansen, B. (1992). The likelihood ratio test under nonstandard conditions: testing the markov switching model of gnp. *Journal of Applied Econometrics* **70**, 127–157.

Hansen, B. (1996). Erratum: The likelihood ratio test under nonstandard conditions: testing the markov switching model of gnp. *Journal of Applied Econometrics* **11**, 195–198.

Hartigan, J. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of Berkeley Conference in Honour of Jerzy Neyman and Jack Kieffer, Vol. 2.*, L.LeCam & R. Olshen, eds. Berkeley: Berkeley Press, pp. 807–810.

Heckman, J. J., Robb, R. & Walker, J. R. (1990). Testing the mixture of exponentials hypothesis and estimating the mixing distribution by the method of moments. *Journal of American Statistical Association* **85**, 582–589.

Henna, J. (1985). On estimating the number of the constituents of a finite mixture of continuous distributions. *Annals Institute of Statistical Mathematics* **37**, 235–240.

Juang, B. & Rabiner, L. (1990). Hidden markov models for speech recognition. *Technometrics* **33**, 251–272.

Leroux, B. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics* **20**, 1350–1360.

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

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal Statistics* **5**, 81–91.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. Berkeley: Institute Mathematical Statistics.

Marsden, J. E. (1974). *Elementary Classical Analysis*. San Francisco: W. H. Freeman.

McLachlan, G. & Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

Oodairia, H. & Yoshihara, K. I. (1971). The law of the iterated logarithm for stationary processes satisfying mixing conditions. *Kōdai Mathematical Seminar Reports* **23**, 311–334.

Poskitt, D. & Chung, S. (1996). Markov chain models, time series analysis and extreme value theory. *Advances In Applied Probability* **28**, 405–425.

Richardson, S. & Green, P. (1997). On bayesian analysis of mixtures with an unknown number of components, with discussion. *Journal of Royal Statistical Society* **B 59**, 731–792.

Rydén, T. (1995). Estimating the order of hidden markov models. *Statistics* **26**, 345–354.

Schelp, F., Vivatanasept, P., Sitaputra, P., Sormani, S., Pongpaew, P., Vudhivai, N., Egormaiphol, S. & Böhning, D. (1990). Relationship of the morbidity of under-fives to anthropometric measurements and community health intervention. *Tropical Medicine and Parasitology* **41**, 121–126.

Symons, M., Grimson, R. & Yuan, Y. (1983). Clustering of rate events. *Biometrics* **39**, 193–205.

Titterington, D., Smith, A. & Markov, V. (1985). *Statistical Analysis of Finite Mixture Distribution*. New York: Wiley.

Turner, C., Startz, R. & Nelson, C. (1990). A markov model of heteroscedasticity. *Journal of Financial Economics* **23**, 3–22.

Zhang, J. & Stine, R. (2001). Autocovariance structure of markov regime switching models and model selection. *Journal of Time Series Analysis* **22**, 107–124.