BIOMETRICS 56, 742–747
September 2000

Likelihood-Ratio Tests for Hidden Markov Models

Paolo Giudici,[1,*] Tobias Rydén,[2] and Pierre Vandekerkhove[1]

[1]Department of Economics and Quantitative Methods, University of Pavia,
Via San Felice 5, 27100 Pavia, Italy
[2]Centre for Mathematical Sciences, Lund University, Box 118, 221 00 Lund, Sweden
[*]*email:* giudici@unipv.it

SUMMARY. We consider hidden Markov models as a versatile class of models for weakly dependent random phenomena. The topic of the present paper is likelihood-ratio testing for hidden Markov models, and we show that, under appropriate conditions, the standard asymptotic theory of likelihood-ratio tests is valid. Such tests are crucial in the specification of multivariate Gaussian hidden Markov models, which we use to illustrate the applicability of our general results. Finally, the methodology is illustrated by means of a real data set.

KEY WORDS: Gaussian hidden Markov model; Likelihood-ratio test; Multivariate hidden Markov model; Temporal graphical model.

1. Introduction

Hidden Markov models (HMMs) are a versatile class of models for weakly dependent random phenomena. An HMM consists of two parts, a nonobservable finite-state Markov chain $\{X_k\}$ and an observable stochastic process $\{Y_k\}$. Given $\{X_k\}$, the Y's are conditionally independent with the conditional distribution of Y_k depending on X_k only. Hence, X_k governs the distribution of Y_k, and for this reason, $\{X_k\}$ is sometimes called the regime. The word 'hidden' is motivated by the nonobservability of $\{X_k\}$; inferences, predictions, etc., must be carried out solely in terms of $\{Y_k\}$.

HMMs have, during the last decade, become wide spread for modeling sequences of weakly dependent random variables, with applications in areas like speech processing (Rabiner, 1989), neurophysiology (Fredkin and Rice, 1992), biology (Leroux and Puterman, 1992), and finance (Rydén, Teräsvirta, and Åsbrink, 1998). (See also the monograph by MacDonald and Zucchini (1997).) Commonly, the conditional distributions of Y_k given X_k all belong to a single parametric family, such as the normal or Poisson families, so that X_k selects the parameter used to generate Y_k. The distribution of Y_k, i.e., the marginal distribution of $\{Y_k\}$, will then be a finite mixture from the parametric family. Mixtures are frequently used in i.i.d. settings to increase the dispersion governed by a specific parametric family, and this effect is obviously found in the marginal distribution of an HMM as well. In addition, $\{Y_k\}$ is dependent. HMMs can thus be viewed as an extension of Markov chains but also as an extension of mixture models.

The topic of the present paper is likelihood-ratio (LR) testing for HMMs. Drawing on results of Bickel, Ritov, and Rydén (1998), we show that, under certain conditions, the standard asymptotic theory for such tests is valid, i.e., we arrive at a χ^2 distributional limit. This problem is particularly crucial in the detection of multivariate Gaussian HMMs, which will be considered in order to illustrate the wide applicability of our general results. In these multivariate models, a natural problem is, when comparing different models, testing for zeros in the precision matrices of the mixture densities.

We shall study a trivariate data set on London mortality, taken from Harrison, West, and Pole (1994). The observed variables are daily deaths (adjusted to remove the effect of a flu epidemic), average temperature, and sulfur dioxide level (on the log scale, as a measure of pollution) for 102 days in London during the winter of 1958. The main substantive problem is to understand how these three variables are interrelated. In particular, a very important problem of interest is to assess whether or not daily deaths are independent of the pollution once the effect of the average temperature is taken into account. The main issue underlying our approach is that the degree of dependence between these two variables may change over time; e.g., other unmeasured variables may sometimes be effective, and, therefore, conditional independence may hold sometimes but not for the whole batch of data. In order to properly take into account the above temporal aspect, we need a model that allows switching between states representing presence and absence of conditional independence as above. To achieve this aim, we shall propose an HMM with a multivariate response variable and two latent states corresponding to the two alternative conditional independence patterns.

More generally, we would like to select a multivariate HMM whose latent states correspond to all association structures that receive support from the data—not always, but at least for considerable periods of time. The methodology is suited to performing model comparisons between alternative condi-

tional independence representations. In the present paper, we provide the LR test theory suited for this purpose.

We remark that the applicability of multivariate HMMs is quite wide: It applies to any multivariate time series whose dependency structure is thought to change considerably over time. The LR tests are useful when the objective of the research is to perform model comparison between alternative temporal association models. Further important examples include, among others, environmental data, typically multivariate and never measured exhaustively, and financial times series, where the state of a national economy, e.g., is a powerful qualitative mechanism that determines changes in the correlation structure among the considered variables.

The paper is organized as follows. In the next section, we discuss multivariate Gaussian HMMs in some detail, and in Section 3, we set the notation and present some background material. In Section 4, we give our results on the analysis of LR tests for HMMs. Section 5 is dedicated to the illustration of our methodology to a real data set. Finally, the Appendix contains formal assumptions, theorems, and proofs.

2. Multivariate Gaussian Hidden Markov Models

Consider an HMM with $\{Y_k\}$ being multidimensional and with the conditional distribution of Y_k given $X_k = i$ being $\mathrm{N}(0, \Sigma^i)$, i.e., multivariate Gaussian with zero mean and covariance matrix Σ^i. The unconditional distribution of Y_k is thus a mixture of multivariate Gaussian distributions. Such a multivariate time series model may be of interest in several areas, as mentioned in Section 1.

In the present paper, our primary interest lies in the precision matrices $K^i = (\Sigma^i)^{-1}$ rather than in the covariance matrices themselves. More precisely, we are interested in testing for two or more zero elements in one or several of the K's. This is because such zeros correspond to conditional independence structures within the components of the Y's. Indeed, for a multivariate Gaussian random variable Y with precision matrix $K = (k_{ij})$, $k_{ij} = 0$ is equivalent to the ith and jth coordinates being conditionally independent given the remaining ones. For example, these coordinates may be mortality and pollution in the example of Section 1. In our HMM, the precision matrix K is governed by X_k, and hence X_k governs the dependence structure within Y_k. As $\{X_k\}$ is a random process, this structure may change over time. Moreover, obviously, the state X_k also carries information about the numerical values of variances and covariances of Y_k. It may well be the case that different values of X_k correspond to the same dependence structure within Y_k, although with different variances and/or covariances. We note that the model formulated here may be described in terms of Gaussian graphical models (cf., Lauritzen, 1996), and we refer to an extended technical report available from the second author for further reading on this.

As noted in the Appendix, identifiability of the model is crucial for the validity of the LR tests. Identifiability of finite mixtures of multivariate Gaussian distributions has been established by Yakowitz and Spragins (1968), whence the results of Section 4 may be applied to multivariate Gaussian HMMs. However, as discussed in Section 4, we may not compare two models with a different number of states. In our computations in Section 5, we assume that we have a fixed number of states (two states), and we can then test for zeros in the corresponding precision matrices. An assumption we need to make, however, is that no two states coincide in the sense of having identical covariance matrices since then we would effectively have one state less than specified by the model.

3. Notation

Before proceeding, we need to introduce some notation. We let $\{X_k\}_{k=1}^{\infty}$ be a stationary Markov chain on $\{1, \ldots, m\}$ with transition probabilities $\alpha(i,j) = P(X_{k+1} = j \mid X_k = i)$. We also let $\{Y_k\}$ be a $\mathcal{Y}$-valued sequence such that, given $\{X_k\}$, $\{Y_k\}$ is a sequence of conditionally independent random variables, Y_k having (conditional) density $g(y \mid X_k)$ with respect to some σ-finite measure ν on $\mathcal{Y}$. Usually $\mathcal{Y}$ is a subset of $\mathbb{R}^q$ for some q, but it may also be a higher dimensional space. Moreover, both $\{\alpha(i,j)\}$ and $\{g(\cdot \mid i)\}$ depend on a parameter ϑ, i.e., $\alpha(i,j) = \alpha_\vartheta(i,j)$ and $g(\cdot \mid i) = g_\vartheta(\cdot \mid i)$, where ϑ is to be estimated from a realization of $\{Y_k\}$. The set to which ϑ belongs is denoted by Θ, and we assume that $\Theta \subseteq \mathbb{R}^d$ is d dimensional. Note that the stationary distribution of $\{X_k\}$, denoted by $\{\pi(i)\}_{i=1}^m$, also depends on ϑ.

The most common set-up is one where ϑ contains the transition probabilities themselves, together with some parameters characterizing the g's. In particular, it is often the case that $g_\vartheta(y \mid i) = f(y; \phi(i))$ for some parametric family $f(y;\phi)$. We refer to this case as the usual parameterization.

The joint density of $(Y_1, \ldots, Y_\mathrm{n})$ may be compactly written as

$$p_\vartheta(y_1, \ldots, y_\mathrm{n}) = \pi_\vartheta \left\{ \prod_{k=1}^{n} G_\vartheta(y_k) A_\vartheta \right\} \mathbf{1}, \tag{1}$$

where $A_\vartheta = \{\alpha_\vartheta(i,j)\}$, $G_\vartheta(y) = \mathrm{diag}\{g_\vartheta(y \mid i)\}$, and $\mathbf{1}$ is an $m \times 1$ vector of ones. The computational complexity of (1) is linear in n. The maximum likelihood estimator (MLE), denoted by $\hat{\vartheta}_\mathrm{n}$, maximizes $p_\vartheta(Y_1, \ldots, Y_\mathrm{n})$ over the parameter set Θ. In many cases, we may renumber the state space of $\{X_k\}$ and the g's, leaving the likelihood unchanged, and the MLE is then not unique. In particular, we may do so if the usual parameterization is employed. This ambiguity is obviously not a big concern, though. Finally, the log likelihood will be denoted by $L_\mathrm{n}(\vartheta) = \log p_\vartheta(Y_1, \ldots, Y_\mathrm{n})$.

4. Likelihood-Ratio Testing

As usual, an LR test may be employed to test whether the parameter ϑ equals some specific value, i.e., the null hypothesis H_0 is a single point, H_0: $\vartheta = \vartheta_0$, and the alternative H_1 is H_1: $\vartheta \neq \vartheta_0$. The LR test statistic for this null hypothesis is $\lambda_\mathrm{n} = 2\{L_\mathrm{n}(\hat{\vartheta}_\mathrm{n}) - L_\mathrm{n}(\vartheta_0)\}$, where $\hat{\vartheta}_\mathrm{n}$ as above is the MLE over Θ. Under some regularity assumptions, stated in the Appendix, under the null hypothesis and for large n, λ_n has approximately a χ^2 distribution with d d.f. (recall that d is the dimension of Θ). Hence, we obtain a test with size approximately equal to α if we reject H_0 if $\lambda_\mathrm{n} > \chi^2_{d,1-\alpha}$, where $\chi^2_{d,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2 distribution with d d.f. This result is stated formally as Theorem 1 in the Appendix.

We now proceed to composite null hypotheses. Assume that we want to test whether the parameter ϑ belongs to a certain $(d-r)$-dimensional subset Θ_0, determined by a set of $r \leq d$ restrictions given by the equations $R_i(\vartheta) = 0$, $1 \leq i \leq r$, of the parameter space Θ.

The LR test statistic for testing the null hypothesis H_0: $\vartheta \in \Theta_0$ versus the alternative H_1: $\vartheta \notin \Theta_0$ is denoted by λ_n and given by $\lambda_n = 2\{\sup_{\vartheta\in\Theta} L_n(\vartheta) - \sup_{\vartheta\in\Theta_0} L_n(\vartheta)\}$. Under some regularity assumptions, stated in the Appendix, under the null hypothesis and for large n, λ_n has approximately a χ^2 distribution with r d.f. Hence, we obtain a test with size approximately equal to α if we reject H_0 if $\lambda_n > \chi^2_{r,1-\alpha}$. This result is stated formally as Theorem 2 in the Appendix.

It is important to notice this result does not hold for the case when testing for the number of states of an HMM. For example, assume that our model is $Y_k \mid X_k = i \sim N(\mu_i, 1)$, that our null hypothesis Θ_0 is that there is only one state ($m = 1$), and that the full model Θ is that there are two states ($m = 2$). In this situation, the asymptotic theory described above is not valid. The very same problem occurs when testing for the number of components in mixture models (cf., Titterington, Smith, and Makov, 1985, p. 153).

5. An Application

The application we consider to illustrate the methodology concerns the London mortality data described in Section 1. The sample variances of the mortality, temperature, and pollution are 650.38, 9.15, and 0.27, respectively. Our main objectives are to model this trivariate time series as a multivariate Gaussian HMM, to estimate the transition probabilities $\alpha(i, j)$ and the conditional covariance matrices Σ^i, and to carry out LR tests discriminating between different models. This task can clearly become quite formidable, both in terms of computational complexity and in sample-size requirements. Therefore, in order to illustrate the methodology, we considered, without loss of theoretical generality, situations where the dimension m of the state space is small.

More precisely, we took $m = 2$, so that $\{X_k\}$ is a Markov chain on $\{1, 2\}$. The observable process $\{Y_k\}$ is trivariate, and conditional on $X_k = i$, Y_k has distribution $N(0, \Sigma^i)$. Write $K^i = (\Sigma^i)^{-1}$. As we were mainly interested in the covariance structure, we subtracted the sample mean from each component of the series. In other words, some parameters of the model (the population means) were not estimated by maximum likelihood but by a moment method. We nevertheless applied our results to the adjusted data, expecting the effect caused by adjusting the means to be negligible. Finally, we assume that the component variances in the two states agree, i.e., $\sigma^1_{ii} = \sigma^2_{ii}$ for $i = 1, 2, 3$. Hence, the particular outcome of X_k only affect the dependencies between the components of the data.

Our first objective was to understand whether and how often excess mortality and pollution are dependent, conditionally on the temperature. In order to do this, we considered two models, C3 and D (this notation is explained below). Under model C3, in state 1, mortality and pollution are conditionally independent given the temperature. We can write this as $k^1_{13} = 0$, which is the restriction defining Θ_0 in Section 4. In state 2, there are no restrictions, i.e., Σ^2 may be chosen freely. In model D, no restrictions were put on any of the covariance matrices.

The MLEs of the transition probability matrix A and the covariance matrices Σ^1 and Σ^2 were as follows:

Model C3:

$$A = \begin{pmatrix} 0.469 & 0.531 \\ 0.107 & 0.893 \end{pmatrix},$$

$$\Sigma^1 = \begin{pmatrix} 664.44 & 74.67 & -0.57 \\ 74.67 & 9.32 & -0.07 \\ -0.57 & -0.07 & 0.26 \end{pmatrix},$$

$$\Sigma^2 = \begin{pmatrix} 664.44 & -11.14 & 8.61 \\ -11.14 & 9.32 & -0.39 \\ 8.61 & -0.39 & 0.26 \end{pmatrix}.$$

Model D:

$$A = \begin{pmatrix} 0.315 & 0.685 \\ 0.136 & 0.864 \end{pmatrix},$$

$$\Sigma^1 = \begin{pmatrix} 643.19 & 75.99 & -2.82 \\ 75.99 & 9.72 & -0.05 \\ -2.82 & -0.05 & 0.26 = \end{pmatrix},$$

$$\Sigma^2 = \begin{pmatrix} 643.19 & -11.70 & 8.39 \\ -11.70 & 9.72 & -0.45 \\ 8.39 & -0.45 & 0.26 \end{pmatrix}.$$

We note that the diagonal elements of the Σ's are close to the sample variances reported above. The Σ^2 matrices are very similar, and the largest difference in the Σ^1 matrices is in the element σ^1_{13}, which indeed is not a free parameter in model C3. The elements σ^1_{23} are also a bit different but not to the same extent.

In order to perform an LR test, we also evaluated the maximal log likelihood under each model. They were -782.68 (model C3) and -779.05 (model D). Therefore, the LR statistic is 7.26. This should be compared to the $\chi^2(1)$ distribution, yielding a p-value of 0.007. The test hence clearly indicates that model C3 is to be rejected. Thus, conditionally on the temperature, mortality and pollution are found to be always dependent.

Our second objective was to extend the above procedure to a complete stepwise LR testing scheme, selecting a final candidate among eight different models that we now describe. Here M, T, and P stand for mortality, temperature, and pollution, respectively, and, e.g., $M \perp P \mid T$ means that mortality and pollution are conditionally independent given temperature. The eight models differ by their dependence structure in state 1; none of the models have any restrictions in state 2. The eight models are as follows:

Model A: In state 1, M, T, and P are independent.
Model B1: In state 1, $M \perp P \mid T$, $T \perp P \mid M$.
Model B2: In state 1, $M \perp T \mid P$, $T \perp P \mid M$.
Model B3: In state 1, $M \perp T \mid P$, $M \perp P \mid T$.
Model C1: In state 1, $M \perp T \mid P$.
Model C2: In state 1, $T \perp P \mid M$.
Model C3: In state 1, $M \perp P \mid T$.
Model D: In state 1, no restrictions.

We briefly illustrate the results obtained with a forward approach that starts from the simplest model, model A, and in which we step-by-step test for additional dependencies in state 1. We also carried out a backward selection procedure,

starting from model D and using step-by-step testing for removal of dependencies in state 1. The two approaches led to the same final model.

Step 1: Model A was compared to models B1, B2, and B3. The maximal log likelihoods were −785.32 (A), −782.69 (B1), −782.20 (B2), and −784.60 (B3), giving the LR statistics 5.27 (A vs. B1), 6.25 (A vs. B2), and 1.45 (A vs. B3). These should be compared to the $\chi^2(1)$ distribution, whose 95% quantile is 3.84. Based on these tests, we choose model B2, i.e., we extended model A by removing the conditional independence between mortality and pollution in state 1.

Step 2: Model B2 was compared to models C1 and C2. The maximal log likelihoods for the new models were −779.14 (C1) and −781.43 (C2), giving the LR statistics 6.10 (B2 vs. C1) and 1.54 (B2 vs. C2). Again, these statistics should be compared to the $\chi^2(1)$ distribution. Hence, we choose model C1, i.e., we extend model B2 by removing the conditional independence between temperature and pollution in state 1.

Step 3: Model C1 was compared to model D. The maximal log likelihood for the new model was −779.05, giving the LR statistic 0.20.

Thus, our final model is C1.

A common problem in the literature on likelihood-ratio model selection is that the sampling theory properties of the model selection procedure are not known. It is often suggested to accompany the stepwise selection results with a penalized log-likelihood score for each model. Therefore, Table 1 reports, for each entertained model, the maximal loglikelihood, its dimension (number of free parameters), and both the Akaike information criterion (AIC) and Bayesian information criterion (BIC) scores.

From Table 1, we note that the best (largest) AIC and BIC scores are both obtained for model C1, which is consistent with the stepwise selection procedure.

In the final model C1, in state 1, mortality and temperature are conditionally independent given the pollution. The MLEs of the parameters of the final model were as follows:

$$A = \begin{pmatrix} 0.860 & 0.140 \\ 0.684 & 0.316 \end{pmatrix},$$

$$\Sigma^1 = \begin{pmatrix} 641.36 & -14.60 & 8.37 \\ -14.60 & 9.73 & -0.45 \\ 8.37 & -0.45 & 0.26 \end{pmatrix},$$

$$\Sigma^2 = \begin{pmatrix} 641.36 & 75.85 & -2.72 \\ 75.85 & 9.73 & -0.04 \\ -2.72 & -0.04 & 0.26 \end{pmatrix}.$$

The stationary probabilities π for this transition probability matrix are (0.83, 0.17), so the restricted model (state 1) accounts for the better part of the observations.

It should be noted that, for mixtures of multivariate Gaussian distributions with zero mean, the likelihood is unbounded if no further restrictions are put on the parameter space. This is because we can pick a particular observation y_k and then find positive definite matrices Σ such that $y_k^{\mathrm{T}}\Sigma^{-1}y_k$ stays bounded while $\det\Sigma$ tends to zero. Hence, the likelihood of y_k for the state of the mixture having covariance matrix Σ will tend to infinity. Of course, this state will, in the limit, yield a zero likelihood for other observations, but the likelihoods given by other states do not vanish and so the overall likelihood grows indefinitely. In the univariate case ($q = 1$), this problem does not occur. In our numerical computations, we carried out 200 likelihood optimizations for each model, each time starting from a randomly chosen point. The overall best parameters were chosen, excluding those for which any of the Σ^i had condition number larger than 10^{12}.

Table 1
Maximal loglikelihood, dimensionality, and AIC and BIC for the eight models for the London mortality data

Model	Max L_n	Dimensionality	AIC	BIC
A	−785.32	8	−793.32	−803.82
B1	−782.69	9	−791.69	−803.50
B2	−782.20	9	−791.20	−803.01
B3	−784.60	9	−793.60	−805.41
C1	−779.14	10	−789.14	−802.27
C2	−781.43	10	−791.42	−804.55
C3	−782.68	10	−792.68	−805.80
D	−779.05	11	−790.04	−804.48

Acknowledgements

This work has been supported by EU TMR network ERB-FMRX-CT96-0096 on computational and statistical methods for the analysis of spatial data. The second author was supported by the Swedish Natural Sciences Council grant M-AA/MA 10538-308 and by a grant from the Royal Swedish Academy of Sciences' exchange program with the United Kingdom.

Résumé

Nous considérons des modèles de Markov cachés comme une classe versatile de modèles pour des phénomènes aléatoires faiblement dépendants. Le sujet de ce présent papier est de tester le rapport de vraisemblance pour les modèles de Markov cachés, et nous montrons que sous des conditions appropriées, la théorie asymptotique standard des tests du rapport de vraisemblance est valide. De tels tests sont cruciaux dans la spécification des modèles de Markov cachés gaussiens multivariés, que nous utilisons pour illustrer l'applicabilité de nos résultats généraux. Enfin, la méthodologie est illustrée sur des données réelles.

References

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

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

Harrison, J., West, M., and Pole, A. (1994). *Applied Bayesian Forecasting and Time Series Analysis.* London: Chapman and Hall.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford: Oxford University Press.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and Their Applications* **40,** 127–143.

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

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series.* London: Chapman and Hall.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77,** 257–284.

Rydén, T., Teräsvirta, T., and Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model of absolute returns. *Journal of Applied Econometrics* **13,** 217–244.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Chichester, U.K.: Wiley.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics* **39,** 209–214.

Received March 1999. *Revised November* 1999. *Accepted January* 2000.

Appendix

Formal Assumptions, Theorems, and Proofs

In this Appendix, the true parameter is denoted by ϑ_0. We deliberately replace the subindex ϑ_0 by 0 in notation like P_{ϑ_0} (becoming P_0), etc. Differentiation with respect to ϑ is denoted by dots, with one dot forming the gradient and two dots forming the Hessian. Also, the Fisher information matrix for $\{Y_k\}$, denoted by $\mathcal{J}_0$, will be used. Intuitively, $\mathcal{J}_0$ can be thought of either as the limiting covariance matrix of $n^{-1/2}\dot{L}_n(\vartheta_0)$ or as the limit (in some suitable sense) of $-n^{-1}\ddot{L}_n(\vartheta_0)$; both of these definitions are valid.

The following assumptions will be referred to in the sequel:

(A1) The transition probability matrix $\{\alpha_0(i,j)\}$ is ergodic, i.e., irreducible and aperiodic.

(A2) For all i and j, the map $\vartheta \mapsto \alpha_\vartheta(i,j)$ has two continuous derivatives in some neighborhood $G = \{\vartheta : |\vartheta - \vartheta_0| < \delta\}$ of ϑ_0. Similarly for $\vartheta \mapsto \pi_\vartheta(i)$. For all i and $y \in \mathcal{Y}$, the map $\vartheta \mapsto g_\vartheta(y \mid i)$ has two continuous derivatives in the same neighborhood.

(A3) Write $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$. There exists a $\delta > 0$ such that (i) for all $1 \le i \le d$ and all a, $E_0\{\sup_{|\vartheta-\vartheta_0|<\delta} |(\partial/\partial\vartheta_i) \log g_\vartheta(Y_1 \mid a)|^2\} < \infty$; (ii) for all $1 \le i,j \le d$ and all a, $E_0\{\sup_{|\vartheta-\vartheta_0|<\delta} |(\partial^2/\partial\vartheta_i\partial\vartheta_j) \log g_\vartheta(Y_1 \mid a)|\} < \infty$; (iii) for $j = 1, 2$, all $1 \le i_\ell \le d$, $\ell = 1, \ldots, j$, and all a, $\int \sup_{|\vartheta-\vartheta_0|<\delta} |(\partial^j/\partial\vartheta_{i_1}\cdots\partial\vartheta_{i_j}) g_\vartheta(y \mid a)|\, \nu(dy) < \infty$.

(A4) There exists a $\delta > 0$ such that, with $\rho_0(y) = \sup_{|\vartheta-\vartheta_0|<\delta} \max_{1\le i,j\le m}\{g_\vartheta(y \mid i)/g_\vartheta(y \mid j)\}$, $P_0\{\rho_0(Y_1) = \infty \mid X_1 = i\} < 1$ for all i.

(A5) ϑ_0 is an interior point of Θ.

(A6) The maximum likelihood estimator is strongly consistent.

In A6, we assume that $\hat{\vartheta}_n \to \vartheta_0 P_0$ a.s. as $n \to \infty$ (up to a possible permutation of states). Consistency of the MLE is discussed by Leroux (1992). A necessary condition for consistency is identifiability, i.e., for $\vartheta \ne \vartheta_0$, the laws of $\{Y_k\}$ are not the same (P_ϑ and P_0 are singular). In particular, if the usual parameterization is employed, identifiability is ensured if finite mixtures of the parametric family $\{f(y;\cdot)\}$ are identifiable. We note that A6 cannot hold (in a simple sense) if the model is overparameterized in the way of specifying more states than there actually are (m is too large). This is because ϑ_0 is then not unique, and there is no unique point around which to expand the log likelihood when analyzing the LR tests.

THEOREM 1. *Assume that A1–A6 hold and that $\mathcal{J}_0$ is nonsingular. Then $2\{L_n(\hat{\vartheta}_n) - L_n(\vartheta_0)\} \to \chi^2(d) P_0$ weakly as $n \to \infty$.*

Proof. Notice that, by A2, A5, and A6, for large n, we can make a Taylor expansion of $L_n(\vartheta)$ about $\hat{\vartheta}_n$, yielding

$$\begin{aligned} L_n(\vartheta_0) - L_n(\hat{\vartheta}_n) &= 0 + \frac{1}{2}(\vartheta_0 - \hat{\vartheta}_n)^T \ddot{L}_n(\bar{\vartheta}_n)(\vartheta_0 - \hat{\vartheta}_n) \\ &= \frac{1}{2}\left\{n^{1/2}(\vartheta_0 - \hat{\vartheta}_n)\right\}^T \\ &\quad \times \left\{n^{-1}\ddot{L}_n(\bar{\vartheta}_n)\right\}\left\{n^{1/2}(\vartheta_0 - \hat{\vartheta}_n)\right\}, \end{aligned}$$

where $\bar{\vartheta}_n$ is a point on the line segment between $\hat{\vartheta}_n$ and ϑ_0. Since $\hat{\vartheta}_n \to \vartheta_0$ P_0 a.s., so does $\bar{\vartheta}_n$, and using Lemma 2 and Theorem 1 in Bickel et al. (1998), the proof is complete.

Before we proceed to the more general result, we assume that the specification $R_i(\vartheta) = 0$, $1 \le i \le r$, of Θ_0 may equivalently be given as a transformation $\vartheta_1 = h_1(\nu_1, \ldots, \nu_{d-r}), \ldots, \vartheta_d = h_d(\nu_1, \ldots, \nu_{d-r})$, where $\nu = (\nu_1, \ldots, \nu_{d-r})$ belongs to a subset of $\mathbb{R}^{d-r}$. We denote by ν_0 the point such that $\vartheta_0 = h(\nu_0)$ and assume that ϑ_0 is an interior point of Θ_0. Each R_i and h_i is assumed to be continuously differentiable in a neighborhood of ϑ_0 and ν_0, respectively, and the matrices

$$C_\vartheta = \left(\frac{\partial R_i}{\partial \vartheta_j}\right)_{r\times d} \quad \text{and} \quad D_\vartheta = \left(\frac{\partial h_i}{\partial \nu_j}\right)_{d\times(d-r)}$$

are assumed to have ranks r and $d - r$, respectively, in the same neighborhoods.

THEOREM 2. *Assume that* A1–A6 *and the above additional assumptions hold and that $\mathcal{J}_0$ is nonsingular. Then $\lambda_n \to \chi^2(r)$ P_0 weakly as $n \to \infty$.*

Proof. The proof essentially follows Serfling (1980, Section 4.4.4). Let $b_\vartheta = (R_1(\vartheta), \ldots, R_r(\vartheta))$, let $\hat{\vartheta}_n$ be the estimate that maximizes L_n over Θ, and let ϑ_n^* be the estimate that maximizes L_n over Θ_0, i.e., under the constraints $R_i(\vartheta) = 0$, $1 \le i \le r$. Equivalently, ϑ_n^* can be written as $\vartheta_n^* = h(\hat{\nu}_n) = (h_1(\hat{\nu}_n), \ldots, h_d(\hat{\nu}_n))$, where $\hat{\nu}_n$ is the MLE of ν in the reparameterization specified by the null hypothesis. The Fisher information matrix for $\hat{\nu}_n$ is $D_0^T \mathcal{J}_0 D_0$, which is nonsingular since D_0 is assumed to be of full rank.

Define the Wald-type statistic

$$W_n = n b_{\hat{\vartheta}_n} \left(C_0 \mathcal{J}_0^{-1} C_0^T \right)^{-1} b_{\hat{\vartheta}_n}.$$

The key idea of the proof is an asymptotic comparison between the W_n and λ_n. Expanding b_ϑ about ϑ_0, noting that $b_0 = 0$ and using Lemma 1 in Bickel et al. (1998), it follows that $n^{1/2} b_{\hat{\vartheta}_n} \to N(0, C_0 \mathcal{J}_0^{-1} C_0^T)$ P_0 weakly, and hence $W_n \to \chi^2(r)$ P_0 weakly. Furthermore, we can expand b_ϑ about ϑ_n^*, yielding

$$b_{\hat{\vartheta}_n} = b_{\hat{\vartheta}_n} - b_{\vartheta_n^*} = (\hat{\vartheta}_n - \vartheta_n^*) C_{\bar{\vartheta}_n}^T,$$

where $\bar{\vartheta}_n$ is a point on the line segment between $\hat{\vartheta}_n$ and ϑ_n^*. Since $\hat{\vartheta}_n$ and ϑ_n^* converge to ϑ_0 P_0 a.s., so does $\bar{\vartheta}_n$. Using continuity of C_ϑ and the fact that both $\hat{\vartheta}_n - \vartheta_0$ and $\vartheta_n^* - \vartheta_0$ are $O_{P_0}(n^{-1/2})$, we obtain

$$b_{\hat{\vartheta}_n} = (\hat{\vartheta}_n - \vartheta_n^*) C_0^T + o_{P_0}\left(n^{-1/2}\right)$$

and hence

$$W_n = n(\hat{\vartheta}_n - \vartheta_n^*) C_0^T \left(C_0 \mathcal{J}_0^{-1} C_0^T \right)^{-1} C_0 (\hat{\vartheta}_n - \vartheta_n^*)^T + o_{P_0}(1). \tag{2}$$

We now turn to the LR statistic. For n large, we can make a Taylor expansion of $L_n(\vartheta)$ about $\hat{\vartheta}_n$, yielding

$$L_n(\vartheta_n^*) - L_n(\hat{\vartheta}_n) = \frac{1}{2} \left\{ n^{1/2} (\vartheta_n^* - \hat{\vartheta}_n) \right\}^T \times \left\{ n^{-1} \ddot{L}_n(\tilde{\vartheta}_n) \right\} \left\{ n^{1/2} (\vartheta_n^* - \hat{\vartheta}_n) \right\},$$

where $\tilde{\vartheta}_n$ is a point on the line segment between $\hat{\vartheta}_n$ and ϑ_n^*. Thus, by Lemma 2 in Bickel et al. (1998),

$$\lambda_n = n(\hat{\vartheta}_n - \vartheta_n^*)^T \mathcal{J}_0 (\hat{\vartheta}_n - \vartheta_n^*) + o_{P_0}(1). \tag{3}$$

Now using the arguments of linear algebra in Serfling (1980, pp. 159–160), it can be proven that the asymptotic representations (2) and (3) are equivalent, and this completes the proof.

References

Asymptotic Normality of the Maximum-Likelihood Estimator for General Hidden Markov Models

Peter J. Bickel; Ya'acov Ritov; Tobias Ryden

The Annals of Statistics, Vol. 26, No. 4. (Aug., 1998), pp. 1614-1635.

Stable URL:

http://links.jstor.org/sici?sici=0090-5364%28199808%2926%3A4%3C1614%3AANOTME%3E2.0.CO%3B2-I

Maximum Likelihood Estimation and Identification Directly from Single-Channel Recordings

Donald R. Fredkin; John A. Rice

Proceedings: Biological Sciences, Vol. 249, No. 1325. (Aug. 22, 1992), pp. 125-132.

Stable URL:

http://links.jstor.org/sici?sici=0962-8452%2819920822%29249%3A1325%3C125%3AMLEAID%3E2.0.CO%3B2-N

Maximum-Penalized-Likelihood Estimation for Independent and Markov- Dependent Mixture Models

Brian G. Leroux; Martin L. Puterman

Biometrics, Vol. 48, No. 2. (Jun., 1992), pp. 545-558.

Stable URL:

http://links.jstor.org/sici?sici=0006-341X%28199206%2948%3A2%3C545%3AMEFIAM%3E2.0.CO%3B2-I