ON THE ASYMPTOTIC PERFORMANCE OF THE LOG LIKELIHOOD RATIO STATISTIC FOR THE MIXTURE MODEL AND RELATED RESULTS

by

Jayanta Kumar Ghosh
Indian Statistical Institute, Calcutta

and

Pranab Kumar Sen
Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1467

September 1984

ON THE ASYMPTOTIC PERFORMANCE OF THE LOG LIKELIHOOD RATIO STATISTIC FOR THE MIXTURE MODEL AND RELATED RESULTS[1]

JAYANTA KUMAR GHOSH[2]
Indian Statistical Institute, Calcutta

PRANAB KUMAR SEN[3]
University of North Carolina, Chapel Hill

Summary. The classical distribution theory of the log likelihood ratio test statistic does not hold for testing homogeneity (i.e., no mixture) against mixture alternatives. Asymptotic theory for this problem is developed. For some special cases, asymptotically locally minimax tests are also found. It is pointed out that the main problem is lack of identifiability of the usual parameterization even when the mixtures are identifiable; if one chooses an identifiable parameterisation, then there is a problem of differentiability of the density.

AMS Subject Classification Numbers: 62E20, 62F05.

Key Words & Phrases: Asymptotic distribution; asymptotic local minimaxity; identifiability; likelihood ratio test statistic; mixture model.

1) This is one of the three examples presented by the first author at the Neyman-Kiefer Conference.

2) Work done partly at the University of California, Berkeley, supported by the ONR Grant N00014-80-C-0163.

3) Work partially supported by the National Heart, Lung and Blood Institute, Contract NIH-NHLBI-71-2243-L from the National Institutes of Health.

1. Introduction. Consider a family of probability desnities and mixtures $(1-\pi)g(x,\theta^{(1)}) + \pi g(x,\theta^{(2)})$, $0 \leq \pi \leq 1$. We assume the mixtures are identifiable in the sense that if $\pi \neq 0$, $\pi \neq 1$ and $\theta^{(1)} \neq \theta^{(2)}$, then the equality

$$(1-\pi)g(x,\theta^{(1)}) + \pi g(x,\theta^{(2)}) = (1-\pi')g(x,\theta^{(3)}) + \pi' g(x,\theta^{(4)}) \quad (1.1)$$

implies $\pi = \pi'$, $\theta^{(1)} = \theta^{(3)}$, $\theta^{(2)} = \theta^{(4)}$ or $\pi = 1 - \pi'$, $\theta^{(1)} = \theta^{(4)}$, $\theta^{(2)} = \theta^{(3)}$. Note that because of this, $g(x,\theta) = g(x,\theta')$ implies $\theta = \theta'$. (Both here and in (1.1) the relations between two densities hold almost every where with respect to the dominating σ-finite measure μ.)

Typically in cluster analysis one models data exhibiting two clusters by postulating a mixture of two densities. In this context it is important to test whether the observed clusters are real or merely a matter of appearance caused by random sampling from a homogeneous population. Formally, denoting the true density by f, one wishes to test

$$H_o; f = g(x,\theta) , \quad \theta \in \Theta \quad (1.2)$$

against the mixture alternatives H_1 considered above with $\pi \neq 0$, $\pi \neq 1$ and $\theta^{(1)} \neq \theta^{(2)}$.

The identifiability assumption (1.1) ensures that H_0 and H_1 have no overlap. But nonetheless the classical asymptotic theory for likelihood ratio tests is not applicable. Of course as pointed out in the literature, the null hypothesis is in some sense on the boundary of the parameter space of this problem, rather than its interior as assumed in classical theory. However, Chernoff (1954) has shown how to handle this kind of departure from standard assumptions; see also Feder (1968). The real

problem is that though the mixtures are identifiable, the parameters π, $\theta^{(1)}$, $\theta^{(2)}$ are not so. If the alternative hypothesis H_1 is true and the true density is written as $f(x,\pi,\theta^{(1)},\theta^{(2)})$ then there is exactly another set of parameter values, namely $(1-\pi,\theta^{(2)},\theta^{(1)})$, which will give exactly the same density; it will be seen in Section 5 that this kind of non-uniqueness is not hard to take care of. However, if H_0 is true and the true density f is $g(x,\theta^{(0)}),\theta^{(0)}$ fixed, then the same density is represented by three curves: $\pi = 0$ and $\theta^{(1)} = \theta^{(0)}$ or $\pi = 1$ and $\theta^{(2)} = \theta^{(0)}$ or $\theta^{(2)} = \theta^{(0)}$ and $\theta^{(1)} = \theta^{(0)}$. Another way of expressing this fact is to observe that we can pass to the one dimensional space of H_0 by specifying only one co-ordinate at a time -- and not two -- in the three dimensional space of H_1. Of course one can try a parametrisation which is identifiable, i.e., one which sets up Euclidean parameters in one to one correspondence with the mixing distribution. Then the problem becomes one of lack of differentiability of the density with respect to these parameters, at points in the space of H_0. For example, we may try $\lambda_1 = (1-\pi)\theta^{(0)} + \pi\theta^{(1)}$, λ_2 = the θ corresponding to $\text{Min}(\pi,1-\pi)$ (with a suitable convention for $\pi = \frac{1}{2}$), and $\lambda_3 = \{\text{Min}(\pi,1-\pi)\}\{2\lambda_2-\theta^{(0)}-\theta^{(1)}\}$.

We shall return to this problem after considering in detail a similar but simpler one which may be called the case of strongly identifiable mixtures. Here one considers two families of probability densities $g_1(x,\theta_1)$ and $g_2(x,\theta_2)$, $\theta_1 \in \Theta_1 \subset R^{p_1}$ and $\theta_2 \in \Theta_2 \subset R^{p_2}$. It is notationally convenient to replace π by θ_0 in the mixture $f(x,\theta_0,\theta_1,\theta_2) = (1-\theta_0)g_1(x,\theta_1) + \theta_0 g_2(x,\theta_2)$. If $\theta_0^{(1)} \neq 0$ or 1, we assume that $f(x,\theta_0^{(1)},\theta_1^{(1)},\theta_2^{(1)}) = f(x,\theta_0^{(2)},\theta_1^{(2)},\theta_2^{(2)})$ implies $\theta^{(1)} = \theta^{(2)}$ where $\theta = (\theta_0,\theta_1,\theta_2)$. Such mixtures may arise as models of partial slippage,

contamination or cluster analysis with some information about the direction of additional clustering. We wish to consider the null hypothesis

$$H_0: \quad f = g_1(x,\theta_1) \ , \quad \theta_1 \in \Theta_1 \tag{1.3}$$

against the strongly identifiable mixture alternatives. Note that if H_0 is true the parameters are still not identifiable. Here is an example of this sort which will be worked out in detail in Section 4.

Example 1. $\quad g_1 \equiv N(\theta_1,1) \ , \quad \theta_1 < a$

$$g_2 \equiv N(\theta_2,1) \ , \quad b \leq \theta_2 \leq c \ , \quad c > b > a \ .$$

Here $N(\theta,1)$ stands for a normal density with mean θ and variance one.

To motivate our main result in the strongly identifiable case, we must make a few general remarks about the asymptotic behaviour of the mle (maximum likelihood estimate) when the parameter is not identifiable. Suppose, in fact, all of Wald's (1949) conditions for the consistency of mle hold except for identifiability of θ. Suppose also, to fix ideas, that the parameter space is the three dimensional Euclidean space and all points non-identifiable from the true value θ^0 lie on a curve Γ. The best that one can hope for is that the maximum of the likelihood will eventually be attained in a neighbourhood of this curve. Actually, Redner (1981) has observed that essentially Wald's proof under Wald's conditions (sans identifiability) guarantees this; Redner calls it convergence of the mle in the topology of the quotient space obtained by collapsing Γ into a single point. This general fact has the following implication in the strongly identifiable case. When the true density is $g_1(x,\theta_1^o)$, the first two components of the mle $\hat{\theta}$, namely $\hat{\theta}_o$ and $\hat{\theta}_1$, will converge almost

surely to their true values $\theta_o = 0$ and $\theta_1 = \theta_1^o$. Of course there is no true value of θ_2 to which $\hat{\theta}_2$ can converge. (In fact under the assumptions made for Theorem 2.1 in Section 3, it can be shown that $\hat{\theta}_2$ cannot converge almost surely to a constant.)

The preceding facts will not be used explicitly in the sequel but they motivate what is done here. Among other things one sees that when H_0 is true, one cannot confine attention to a neighbourhood of a single point in order to maximise the likelihood. This means that the usual quadratic approximation to the likelihood is available only with respect to the first two components θ_0 and θ_1 for which the mle is consistent. However, under certain assumptions, we can still utilise these partial quadratic approximations to show (Theorem 2.1) that in the limit the likelihood ratio statistic is distributed as a certain functional $W^2 I_{\{W>0\}}$, where $W = \sup\{T(\eta_2)\}$ and $T(\cdot)$ is a Gaussian process with zero mean and covariance kernel depending on the true value θ_1^o under H_0. In Remark 2.1 we propose a family of other tests with simpler limiting distribution. Note that our treatment does not follow from Chernoff (1954) or Feder (1968), because they were able to exploit the existence of a consistent solution of the likelihood equation in the identifiable case. To follow their approach, one would have to develop first results in solutions of the likelihood equation in the non-identifiable case. This can be done but use of techniques similar to those of Redner (1981) seemed aesthetically more satisfactory.

The fact that the likelihood is not locally approximable by a quadratic has another repercussion. The proof of asymptotic local minimaxity of the likelihood ratio test via approximation by Bayes tests also breaks down.

In view of this it seems natural to introduce a prior G on θ_2 and then work with the integrated likelihood ratio statistic

$$\sup_{\theta_o,\theta_1} \int \prod_1^n f(x_i,\theta_o,\theta_1,\theta_2)G(d\theta_2) / \sup_{\theta_1} \prod_1^n g(y_i,\theta_1)$$

or some other functional on the integrated likelihood. One should probably choose G so that the associated test is asymptotically locally minimax. It is plausible that such a G and an associated test always exists under reasonable conditions. As a first step, it is proved in Section 4 that for Example 1 the prior degenerate at b and the corresponding likelihood ratio test assuming $\theta_2 = b$ works in this sense. A similar result is proved for general exponentials. On the other hand it is not hard to show (though we do not prove it here) that the likelihood ratio test is not asymptotically locally minimax for these examples -- these are thus new instances of the failure of the principle of maximising the likelihood.

For the case of (not strongly) identified mixtures, a result analogous to Theorem 2.1 is obtained for the likelihood ratio test of the hypothesis in (1.2). To do this we assume a separation condition $||\theta^{(1)}-\theta^{(2)}|| \geq \varepsilon$ where $\varepsilon > 0$ is a fixed quantity. In a subsequent communication we shall try to remove this condition.

A review of the literature on this topic is available in Bock (1981) and Gupta and Huang (1981) and, the final chapter of Everitt and Hand (1981). Even though there is no overlap with our results, a paper of Moran (1973) deserves to be mentioned. Moran derives the asymptotic distribution of the likelihood ratio test of homogeneity against special mixture alternatives in two cases, namely, Poissan and Gamma. For his alternative Moran considers

mixing distributions $G\{(\theta-\lambda_o)/\alpha^{\frac{1}{2}}\}$ where G is a fixed known distribution of which one needs only that the third moment about mean is zero. In our set-up of two point mixtures, this would correspond to assuming $\theta_0 = \Pi = \frac{1}{2}$, so that H_0 is equivalent to the scale parameter $|\theta^{(1)}-\theta^{(0)}| = 0$.

We conclude by making a few remarks about mixtures of $N(\mu,\sigma^2)$. The theorem in section 5 applies directly only the case of mixtures of means with known α and μ in a compact set. However, we feel a similar result would hold if σ is unknown but (μ,σ) lies in a compact set, $\alpha \neq 0$ and only mixtures of mean are allowed. If one also allows scale mixtures, substantial changes are needed in the treatment since the derivative of the likelihood with respect to π ceases to be square integrable. Note that a similar phenomenon occurs in Moran's case if G has non-zero third moment about mean. It should be possible to handle such cases by suitable truncation of the derivatives. Finally, it may be noted that though we have confined ourselves to the case of mixtures our main conclusions hold for other cases of non-identifiable parameters.

2. The Main Result for Strongly Identifiable Mixtures. Let $\theta = (\theta_o,\theta_1,\theta_2)$ and $f(x,\theta)$ be as in the introduction and

$$L_n(\theta) = \sum_1^n \log f(X_i,\theta)$$

be the log likelihood based on n i.i.d. observations. Suppose H_o of (1.3) is true and the true density is $g_1(x,\theta_1^o)$. All expectations and probabilities will be computed in this and the next section under this assumption but this will not be displayed in the notation.

We now sketch an argument leading to Theorem 2.1, introducing notations

as we go along. The details for handling several remainder terms as well as the necessary assumptions (A1 through A5) are collected in the next section. Here we only remark briefly on the nature of the assumptions. The assumptions are similar to those in the classical case but have to be strengthened suitably to ensure uniformity in θ_2 at various places. The latter as well as tightness of the Gaussian process $T_n(\cdot)$ introduced below makes it convenient to work with Θ_2 as a closed bounded interval. However all we really need is compactness of Θ_2 or its closure. The restriction to dimension one for Θ_2 is made so that the tightness Assumption A (vide Section 3) is easy to write down; it may be extended by making use of analogous conditions in Bickel and Wichura (1972). As usual we take Θ_1 as an open rectangle in R^p.

Among other things the assumptions of the next section guarantee that all quantities introduced below are well-defined.

We now begin by rescaling the parameters through

$$\theta_o = \theta_o^o + \eta_o/\sqrt{n} \quad \text{where} \quad \theta_o^o = 0$$

$$\theta_1 = \theta_1^o + \eta_1/\sqrt{n}$$

$$\theta_2 = \eta_2$$

Let $L_n(\theta)$ be denoted as $V_n(\eta)$ when regarded as a function of $\eta = (\eta_o, \eta_1, \eta_2)$. Note that $V_n(0,0,\eta_2)$ is free of η_2 and hence may be written simply as $V_n(0)$. Let

$$U_{no}(\eta_2) = n^{-\frac{1}{2}} \frac{\partial L_n}{\partial \theta_o}\Bigg|_{(0,\theta_1^o,\eta_2)}$$

$$= n^{-\frac{1}{2}} \sum_1^n \{g_2(X_i,\eta_2)/g_1(X_i,\theta_1^o) - 1\}$$

be the normalised derivative with respect to θ_o and let

$$U_{n1} = n^{-\frac{1}{2}} \frac{\partial L_n}{\partial \theta_1}\Bigg|_{(0,\theta_1^o,\eta_2)}$$

$$= n^{-\frac{1}{2}} \sum \frac{\partial \log g_1(X_i,\theta_1)}{\partial \theta_1}\Bigg|_{\theta_1^o}$$

be the $1\times p$ (row) vector of normalised derivatives with respect to the components of θ_1. The presence or absence of η_2 in $U_{no}^{(\eta_2)}$, U_{n1} indicates dependence on η_2 or lack of it. The same convention is followed below. Let

$$I_{oo}(\eta_2) = E(\{g_2(X_1,\eta_2)/g_1(X_1,\theta_1^o) - 1\}^2)$$

$$I_{01}^{1\times p}(\eta_2) = E(\{g_2(X_1,\eta_2)/g_1(X_1,\theta_1^o)\}\{\frac{\partial \log g_1(X_1,\theta_1)}{\partial \theta_1}\Bigg|_{\theta_1^o}\})$$

$$I_{11}^{p\times p} = E \frac{\partial \log g_1(X_1,\theta_1)}{\partial \theta_1}\Bigg|_{\theta_1^o}\}^T\{\frac{\partial \log g_1(X_1,\theta_1)}{\partial \theta_1}\Bigg|_{\theta_1^o}\}$$

and

$$I^{(p+1)\times(p+1)}(\eta_2) = \begin{bmatrix} I_{oo}(\eta_2) & I_{o1}(\eta_2) \\ I_{o1}(\eta_2)^T & I_{11} \end{bmatrix}, \quad I^{-1}(\eta_2) = \begin{bmatrix} I^{oo}(\eta_2) & I^{o1}(\eta_2) \\ I^{o1}(\eta_2)^T & I^{11}(\eta_2) \end{bmatrix}$$

By Assumption A1, I_{ij}'s are related to the second order derivatives of $\log f(X_1,\theta)$ in the usual way.

Expanding $V_n(\eta)$ with respect to the first two co-ordinates, by A1

$$V_n(\eta) = V_n(0) + A_n(\eta) + R_{n1} \tag{2.1}$$

where the remainder term R_{n1} is $o_p(1)$ on bounded sets of η_o, η_1 uniformly in η_2 and

$$A_n(\eta) = \eta_o U_{no}(\eta_2) + \eta_1 U_{n1}^T - \tfrac{1}{2}[\eta_o^2 I_{oo}(\eta_2) + 2\eta_1 I_{11}\eta_1^T + \eta_o I_{o1}(\eta_2)\eta_1^T]$$

We prove in the next section that uniformly in η_2,

$$L_n(\eta_2) \underset{\text{def}}{=} \sup_{\substack{0\le\theta_o\le 1 \\ \theta_1\in\Theta_1}} L_n(\theta) = V_n(0) + \sup_{\substack{\eta_o\ge 0 \\ \eta_1\in R^p}} A_n(\eta) + o_p(1) \tag{2.2}$$

(The proof is similar to the classical case but one has to ensure uniformity in η_2).

By the well-known Kuhn-Tucker-Lagrange theorem (viz., McCormick (1967)] the supremum of $A_n(\eta)$ is

$$\tfrac{1}{2}(U_{no}(\eta_2),\ U_{n1})I^{-1}(\eta_2)(U_{no}(\eta_2),\ U_{n1})^T \tag{2.3}$$

if

$$T_n(\eta_2) \underset{\text{def}}{=} \{I^{oo}(\eta_2)U_{no}(\eta_2) + I^{o1}(\eta_2)U_{n1}^T\}/\{I^{oo}(\eta_2)\}^{\frac{1}{2}} \ge 0 \tag{2.4}$$

and the supremum is

$$\tfrac{1}{2}U_{n1}I_{11}^{-1}U_{n1}^T \quad \text{if} \quad T_n(\eta_2) > 0 \tag{2.5}$$

Similarly

$$L_n(H_o) \underset{\text{def}}{=} \sup_{\substack{\theta_o=0 \\ \theta_1\in\Theta_1}} L_n(\theta) = V_n(0) + \tfrac{1}{2}U_{n1}I_{11}^{-1}U_{n1}^T + o_p(1) \qquad (2.6)$$

Hence,

$$\lambda_n(\eta_2) \underset{\text{def}}{=} 2\{L_n(\eta_2) - L_n(H_o)\} = o_p(1) \qquad (2.7)$$

if $T_n(\eta_2) < 0$. If $T_n(\eta_2) \geq 0$,

$$\lambda_n(\eta_2) = [U_{no}(\eta_2), U_{n1}]I^{-1}(\eta_2)[U_{no}(\eta_2), U_{n1}]^T - U_{n1}I_{11}^{-1}U_{n1}^T + o_p(1)$$

$$= T_n^2(\eta_2) + o_p(1) \ . \qquad (2.8)$$

So the likelihood ratio statistic is by definition

$$\lambda_n = \sup_{\eta_2} L_n(\eta_2) - L_n(H_o) = \sup \lambda_n(\eta_2) = W_n^2 I_{W_n\geq 0} + o_p(1) \qquad (2.9)$$

where $W_n = \sup_{\eta_2} T_n(\eta_2)$.

Assume $\Theta_2 = [b,c]$. By A4 and A5 of Section 3, the stochastic process $T_n(\cdot)$ taking values in $C[b,c]$ converges weakly to a Gaussian process $T(\cdot)$ on $C[b,c]$ whose mean is zero and the covariance kernel K is the same as that of $T_1(\cdot)$ and easy to write down. The covariance (under θ_1^o) of $T(\eta_{21})$ and $T(\eta_{22})$ is given below assuming U_{n1} is a scalar:

$$K(\eta_{21},\eta_{22},\theta_1^o) = I^{oo}(\eta_{21})I^{oo}(\eta_{22})J(\eta_{21},\eta_{22}) + I^{oo}(\eta_{21})I^{o1}(\eta_{22})I_{01}(\eta_{21})$$
$$+ I^{oo}(\eta_{22})I^{o1}(\eta_{21})I_{o1}(\eta_{22}) + I^{o1}(\eta_{21})I^{o1}(\eta_{22})I_{11}\}/\{I_{oo}(\eta_{21})I_{oo}(\eta_{22})\}^{\frac{1}{2}},$$

where $J(\eta_{21},\eta_{22})$ is the covariance of $U_{10}(\eta_{21})$ and $U_{10}(\eta_{22})$ under θ_1^o. Note that $\text{Var}(T_1(\eta_2)) = 1 \ \forall \ \eta_2$. Since $\lambda_n = \phi o Y_n(\cdot)$ where ϕ is a continuous functional, we have

Theorem 2.1. Under the assumptions A1 through A5 of Section 3, λ_n converges in distribution to $\phi o T(\cdot)$.

Remark 2.1 (a) The limiting distribution of λ_n simplifies a little when X_i assumes only k distinct values. Identifiability of the mixtures would require that $k-1 \geq \dim \Theta_1 + \dim \Theta_2 + 1$.

(b) The limiting distributions of $T_n(\theta_2)$ and $\lambda_n(\theta_2)$ are given explicitly and applied to Example 1 in Section 4.

(c) To get alternative test statistics whose limiting distributions are easier to compute, one may approximate Θ_2 by a finite set $\theta_2^1,\ldots,\theta_2^m$ and then consider the statistics $T_n(\theta_2^i)$, $i = 1,\ldots,m$ which are asymptotically multivariate normal with zero mean under H_o; the dispersion matrix can be consistently estimated since the mle is consistent for θ_1 when H_o is true. One can use as a test statistic any suitable function of $T_n(\theta_2^i)$'s; for example, choosing a suitably positive definite quadratic form in T_n's and then estimating the coefficients, one would get a limiting χ^2-distribution.

3. Assumptions and Details of Proof. As in Section 2 all expectations are computed under a fixed $g_1(x,\theta_1^o)$. The assumptions are marked A1 through A5. Instead of collecting them at one place we shall present them as the need arises, in course of supplying some of the details left out in Section 2.

Let $\theta_{o1} = (\theta_o,\theta_1)$, $\theta_{o1}^o = (0,\theta_1^o)$. A similar convention is followed for η_{o1}, η_{o1}^o. Let $D_o = \frac{\partial}{\partial\theta_o}$, $D_j = \frac{\partial}{\partial\theta_{1j}}$, $j = 1,\ldots,p$, unless otherwise stated the derivations are evaluated at (θ_{o1}^o,θ_2).

A1. (i) Θ_1 is an open set of R^p, and Θ_2 a closed bounded interval $[b,c]$ of R^1.

((ii) $f(x,\theta)$ is continuous in θ and twice continuously differentiable with respect to θ_{o1}.

(iii) $E(D_j \log f) = 0$,

$E(D_j D_{j'} \log f) = -E(D_j \log f \times D_{j'} \log f)$

(iv) $E\{ \sup_{||\theta_{o1}-\theta^o_{o1}||<\delta, \theta_2\in\Theta_2} |D_j D_{j'} \log f(X,\theta_{o1},\theta_2) - D_j D_{j'} \log f(X,\theta^o_{o1},\theta_2)| \} \to 0$

as $\delta \to 0$, $j,j' = 0,1,\ldots,p$.

To handle the remainder in (2.2) we proceed in three stages, Assumption A2 will allow us to restrict attention to a compact subset of Θ_1 while calculating the supremum of $L_n(\theta)$. Then with the help of A3 we work with arbitrary but fixed neighbourhoods of θ^o_{o1}, which is replaced at the third and final stage by neighbourhoods that shrink like $n^{-\frac{1}{2}}$, i.e., bounded neighbourhoods in the η_{o1}-plane. The fact that R_{n1} in (2.1) is $o_p(1)$ uniformly on bounded η_{o1}-sets now completes the proof.

A2. There exists a compact neighbourhood N of θ^{o1} such that

$E(\psi(X_1,\theta_2)) < 0$, where

$$\psi(X_1,\theta_2) = \sup_{\theta_{o1}\in[0,1]\times N^c} \log \{f(X_1,\theta)/g_1(X_1,\theta^o_1)\}$$

Moreover, $\psi(X_1,\cdot)$ is continuous on Θ_2, $|\psi(X_1,\theta_2)| \le H(X_1)$ $\forall$ θ_2 and $E(H(X_1)) < \infty$.

By the uniform strong law of large numbers (USLLN) applied to $n^{-1}\Sigma\psi(X_i,\cdot)$ and the fact that $n^{-1}\Sigma\psi(X_i,\theta_2) \ge \sup_{\theta_{o1}\in[0,1]\times N^c} L_n(\theta) - V_n(0)$, we get

$$\sup_{\theta_{o1}\in[0,1]\times H_1} L_n(\theta) - \sup_{\theta_{o1}\in[0,1]\times N} L_n(\theta) = o_p(1) \qquad (3.1)$$

uniformly in θ_2.

A3. For each $\theta_{o1} \neq \theta^o_{o1}$, there exists an open ball with centre θ_{o1} and radius δ_o such that if $U = U(\theta_{o1},\delta_o)$ is its intersection with $[0,1] \times \Theta_1$ and

$$\psi = \psi(X_1,U,\theta_2) = \sup_{\theta'_{o1}\in U} \log \{f(X_1,\theta'_{o1},\theta_2)/g_1(X_1,\theta^o_1)\}$$

then

$$|\psi(X_1,U,\theta_2)| \leq H(X_1) \quad \forall\ \theta_2 \quad \text{and} \quad E(H(X_1) < \infty \ .$$

By A3 and continuity of f

$$E(\psi(X_1,U(\theta_{o1},\delta),\theta_2) \to E(\log \{f(X_1,\theta_{o1},\theta_2)/g_1(X_1,\theta^o_1)\}) < 0$$

So we can choose δ_1 (depending on θ_{o1}) such that $E(\psi(X_1,U(\theta_{o1},\delta_1),\theta_2)) < 0$. Let $U_\delta = \{\theta_{o1};\ 0 < \theta_o < \delta,\ ||\theta_1-\theta^o_1|| < \delta\}$. Consider an open cover of $U^c_\delta \cap [0,1] \times N$ by sets $U(\theta_{o1},\delta_1)$ and choose a finite subcover $U_1,\ldots,U_m$. Now apply the USLLN to $n^{-1}\Sigma\psi(X_i,U_j,\theta_2)$, $j = 1,\ldots,m$, $\theta_2 \in \Theta_2$ to conclude that

$$\sup_{\theta_{o1}\in[0,1]\times N} L_n(\theta_{o1},\theta_2) - \sup_{\theta_{o1}\in U_\delta} L_n(\theta_{o1},\theta_2) = o_p(1)$$

uniformly in θ_2. This completes the second stage of the proof.

At the third and final stage, note that, by Taylor's theorem and A1,

$$L_n(\theta) = V_n(0) + (U_{no}(\eta_2),U_{n1})\eta^T_{o1} + \tfrac{1}{2}\Sigma\eta_i\eta_j J_{ij}(\eta)$$

where $|J_{ij}(\eta) + I_{ij}(\eta_2)| = \delta \cdot O_p(1)$ uniformly in $U_\delta \times \Theta_2$. We now use

A4. $I(\theta_2)^{(p+1)\times(p+1)}$ is continuous in θ_2 and its minimum eigen value is greater than $\varepsilon_o > 0$ $\forall\ \theta_2$; and

A5. $E|D_o \log f(X_1,\theta_{o1}^o,\theta_2) - D_o \log f(X_1,\theta_{o1}^o,\theta_2')|^\alpha \le K|\theta_2-\theta_2'|^{1+\gamma}$ for some $\alpha,\gamma > 0$.

A5 ensures tightness of $U_{no}(\cdot)$. (To see this one has to use the theorem of Dharmadhikari et al. (1968)). Hence, $\sup_{\eta_2} U_{no}(\eta_2)$ is $O_p(1)$. Also by A1, U_{n1} is also $O_p(1)$. Hence (by A4) for given ε, we can find $\varepsilon' < \varepsilon_o$, K and n_o such that for $n > n_o$,

$$P\left\{\begin{array}{l}\sup_{\eta_2} U_{no}(\eta_2) + |U_{n1}| < K\ , \\ \\ \text{the smallest eigen value of } [J_{ij}(\eta)]^{(p+1)\times(p+1)} > \varepsilon'\ ,\end{array} \quad \forall\ \eta \in U_\delta \times H_2\right\} \ge 1-\varepsilon\ .$$

Then by first making a suitable orthogonal transformation, one can find M such that for $n > n_o$,

$$P\{V_n(\eta) < V_n(0) \text{ if } \eta \in U_\delta \times H_2 \text{ and } ||\eta_{o1}|| > M \text{ and } A_n(\eta) < A_n(0) \text{ if } ||\eta_{o1}|| > M\} > 1-\varepsilon\ ,$$

where M depends only on K and ε'. Thus with probability $> 1-\varepsilon$ for $n > n_o$, the supremum of $L_n(\theta)$, i.e., $V_n(\eta)$ (over U_δ) and that of $A_n(\eta)$ (over $[0,1] \times R^p$) is attained in $||\eta_{o1}|| \le M$. Since on this set R_{n1} of (2.1) is $o_p(1)$ by A1, the proof of (2.2) is complete.

The proof of the similar result (2.6) follows along similar lines from A1 through A4 which are of course much stronger than what we need for (2.6).

Remark 3.1 (a) The tightness assumption A5 holds if

$$|D_o \log f(X,\theta^o_{o1},\theta_2) - D_o \log f(X,\theta^o_{o1},\theta_2') \leq \psi(X)|\theta_2-\theta_2'| \qquad (3.2)$$

and $E(\psi(X))^\beta < \infty$ for some $\beta > 1$.

(b) The limiting distribution of $T_n(\cdot)$ depends on the true value of θ_1. It is weakly continuous in θ_1 provided (i) the covariance kernel of $T(\cdot)$ is continuous in θ_1; and (ii) the Lipschitz condition (3.2) is suitably strengthened to be uniform over θ_1-neighbourhoods, (i) guarantees convergence of finite dimensional distributions of $T(\cdot)$ as $\theta_1 \to \theta^o_1$ and (ii) guarantees tightness. Under these conditions the limiting distribution of λ is also weakly continuous in θ_1. Suppose this is so and let $t(\theta_1,\alpha)$ be such that $\lim P_{\theta_1}\{\lambda \geq t(\theta_1,\alpha)\} = \alpha$. If for each θ_1, t exists, is unique and a point of continuity of the limiting distribution of λ under θ_1, then t is continuous in θ_1. By Redner's (1981) result, $\hat{\theta}_1$ is consistent for θ_1 under H_o. Hence, as pointed out to us by Peter Bickel, $\lim P_{\theta_1}\{\lambda \geq t(\hat{\theta}_1,\alpha)\} = \alpha$. Thus the test which rejects H_o if $\lambda \geq t(\hat{\theta}_1,\alpha)$ would be asymptotically similar provided the conditions assumed here hold. They are easy to check under A1 to A5 if Θ_2 is a finite set.

4. Asymptotically Locally Minimax Tests in some Examples. We shall calculate the asymptotic properties of tests based on $\lambda_n(\theta_2)$, where $\lambda_n(\theta_2)$ is defined in (2.2) and (2.7), and show that it is asymptotically locally minimax for problems like Example 1.

Fix θ^o_1 as in Section 2 and a sequence of alternatives K_n corresponding to a fixed $\eta = (\eta_o,\eta_1,\eta_2)$. We fix also a value b of θ_2 and consider the limiting distribution of $T_n(b)$ under θ^o_1 and K_n

where $T_n(b)$ is defined in (2.4).

Let $Z_n^* = V_n(\eta) - V_n(0)$. Then by (2.1), Z_n^* is asymptotically $N(-\frac{1}{2}\eta_{01}I(\eta_2)\eta_{01}^T, \eta_{01}I(\eta_2)\eta_{01}^T)$. By a well-known result of LeCam, namely his first lemma on contiguity [cf. Hájek and Sidák (1967, p. 204)], this shows K_n is contiguous to θ_1^o. Since $T_n(b)$ and Z_n^* are asymptotically bivariate normal, by another well-known result of LeCam, namely, his third lemma on contiguity (vide Hájek and Sidák (1967)], $T_n(b)$ is asymptotically normal under K_n with same asymptotic variance as under θ_1^o and mean under K_n equal to mean under θ_1^o plus the asymptotic covariance under θ_1^o. Moreover, $\lambda_n(b) = T_n^2(b)I_{\{T_n(b)\geq 0\}} + o_p(1)$ under K_n, since the same relation holds under θ_1^o and K_n is contiguous.

Under θ_1^o, $T_n(b)$ is asymptotically normal with mean zero and variance unity. Also the asymptotic covariance of Z_n^* and $T_n(b)$ under θ_1^o is

$$\rho = \{I^{oo}(b)\}^{-\frac{1}{2}}[\eta_o I^{oo}(\eta_2)\mathrm{Cov}(U_{no}(b), U_{no}(\eta_2)) + \eta_o \sum_{j=1}^{p} I_j^{o1}(\eta_2)\mathrm{Cov}(U_{n1j}, U_{no}(\eta_2))]$$

where I_j^{o1} and U_{n1j} are the j-th components of I^{o1} and U_{n1}. Note ρ depends only on η_o and η_2 and so may be written $\rho(\eta_o, \eta_2)$. By the remarks in the preceding paragraph, the following result is true.

Theorem 4.1. Assume the conditions of Section 3. Then

$$\lim P_{\theta_1^o}\{\lambda_n(b) \geq x\} = 1 - \phi(\sqrt{x}) \quad \text{if} \quad x > 0$$

$$= 1 \quad \text{if} \quad x = 0$$

$$\lim P_{K_n}\{\lambda_n(b) \geq x\} = 1 - \phi(\sqrt{x} - \rho(\eta_o, \eta_2)) \quad \text{if} \quad x > 0$$

$$= 1 \quad \text{if} \quad x = 0$$

where ϕ is the standard normal distribution function.

Consider now Example 1 given in the introduction and fix $\alpha < .5$. Let the limiting power of a sequence of tests ϕ_n of size α under K_n be denoted by $\beta(\{\phi_n\},\theta_1^o,\eta_o,\eta_1,\eta_2)$. Let us say $\{\phi_n^o\}$ is asymptotically locally minimax if for all sequences $\{\phi_n\}$ which have limiting power,

$$\inf_{\substack{\eta_1\in R^1 \\ \eta_2\in\Theta_2}} \beta(\{\phi_n\},\theta_1^o,\eta_o,\eta_1,\eta_2) \leq \inf_{\substack{\eta_1\in R^1 \\ \eta_2\in\Theta_2}} \beta(\{\phi_n^o\},\theta_1^o,\eta_o,\eta_1,\eta_2) \quad ,$$

for every $\eta_o > 0$ and $\theta_1^o \in \Theta_1$.

By the classical theory for the likelihood ratio test ϕ_n^o based on $\lambda_n(b)$, taking $\eta_2 = b$,

$$\inf_{\eta_1\in R^1} \beta(\{\phi_n\},\theta_1^o,\eta_o,\eta_1,b) \leq \inf_{\eta_1\in R^1} \beta(\{\phi_n^o\},\theta_1^o,\eta_1,\eta_2,b)$$

for every $\eta_o > 0$ and $\theta_1^o \in \Theta_1$.

Hence asymptotic local minimaxity of ϕ_n^o will follow from Theorem 4.1 if we show $\rho(\eta_o,\eta_2) \geq \rho(\eta_o,b)$ $\forall\ \eta_2 \geq b$. For Example 1 this follows easily by direct calculation. However the following lemma shows this property is true for general exponentials. Before stating the lemma we introduce some notations.

Let $g_\theta = g(x,\theta) = A(\theta)\exp\{\theta x\}h(x)$, $\theta \in$ some open interval J, be a family of probability densities. Let $a < b$ be fixed elements of J. Let

$$\psi(\theta) = I_{11}\cdot \mathrm{Cov}\left(\frac{g_\theta}{g_a}, \frac{g_b}{g_a}\right) - I_{o1}(b)\,\mathrm{Cov}\left(\frac{g_\theta}{g_a}, \frac{g_a'}{g_a}\right)$$

where the covariances are computed under $g_a, g_a' = \left.\frac{dg_\theta(x)}{d\theta}\right|_{\theta = a}$,

$$I_{11} = E_{g_a}\left(\frac{g_a'}{g_a}\right)^2 \quad , \quad I_{o1}(b) = \mathrm{Cov}_{g_a}\left(\frac{g_b}{g_a}, \frac{g_a'}{g_a}\right) \quad .$$

To relate to Example 1 (and similar problems) note that

$$\rho(\eta_o, \eta_2) = \eta_o\{I^{oo}(b)\}^{-\frac{1}{2}}[\det I(b)]^{-1}\cdot\psi(\eta_2)$$

with $\theta_1^o = a$.

We assume $\psi(\theta)$ is finite on J.

Lemma 4.1. $\psi(\theta) \geq \psi(b)$ if $\theta \geq b$.

Proof. Note that

$$\psi(a) = 0 < \psi(b) \tag{4.1}$$

Also $\psi(\theta)$ can be expressed as

$$\int\{I_{11}\frac{g_b}{g_a} - I_{o1}(b)\frac{g_a'}{g_a} - I_{11}\}g_\theta d\mu = \int\phi(x)g_\theta d\mu \quad \text{say}$$

Since ϕ is convex, for any constant K, $\phi(x) - K$ can have at most two sign changes and if there are two, they must be from positive to negative and negative to positive. Hence by Karlin's well-known result on sign diminishing properties of the exponential densities (see, e.g. Karlin (1968)), $\psi(\theta) - K$ has similar sign change properties. If there exists $\theta' > b$ such that $\psi(\theta') < \psi(b)$ then this sign change property would be contradicted at the points a, b, θ' provided we choose K such that $\mathrm{Max}\{0, \psi(\theta')\} < K < \psi(b)$. This proves the lemma.

5. The Case of Identifiable Mixtures. Let Θ be an open bounded real interval and $\overline{\Theta}$ its closure. Let $g(x,\theta)$, $\theta \in \Theta$ be a family of densities and consider the mixtures $(1-\theta_o)g(x,\theta_1) + \theta_o g(x,\theta_2)$, $\theta_i \in \Theta$, where $0 < \theta_o < 1$ and $|\theta_2-\theta_1| \geq \varepsilon$, ε being a fixed positive number. Without loss of generality we may take $0 < \theta_0 \leq \frac{1}{2}$. We wish to test homogeneity, i.e., H_o: $\theta_o = 0$ against the above mixture alternatives. In the sequel θ stands for the vector $(\theta_o,\theta_1,\theta_2)$.

Suppose H_o is true and the true density is $g(x,\theta_1^o)$. We make the following blanket assumption

B1. Let Θ_1 be any open set containing θ_1^o and Θ_2 any closed set such that $\Theta_1 \cap \Theta_2 = \emptyset$. Then A1, A3, A4, A5 hold with this Θ_1,Θ_2. (Since Θ_2 is compact by assumption, A2 is dropped.)

One can now imitate the arguments in Sections 2 and 3. As in Section 3, we can show that in order to maximise the log likelihood $L_n(\theta)$ we may restrict to $0 < \theta_o < \delta$ and $|\theta_1-\theta_1^o| < \delta$ $(\delta < \varepsilon)$. Hence θ_2 may be restricted to $\theta_2 \leq \theta_o - (\varepsilon-\delta)$ and $\theta_2 \geq \theta_o + (\varepsilon-\delta)$. Call this set $\Theta_2^{(1)}$. Define $\eta, V_n(\eta), A_n(\eta)$ etc. as in Section 2. Then one can prove as in Section 3, that with probability tending to one both $L_n(\theta)$ and $A_n(\eta)$ attain their maximum in a bounded neighbourhood of $\eta_{o1}^o = (0,0)$. Hence uniformly in θ_2,

$$\sup_{\theta_{o1}} L_n(\theta) = V_n(0) + \sup A_n(\eta) + o_p(1)$$

where the maximisation of A_n is over the set

$$V_{\theta_2} = \{\eta_{o1}^o\} \cup \{\eta_{o1};\ \eta_o > 0,\ |\theta_1^o + \eta_1 n^{-\frac{1}{2}} - \theta_2| \geq \varepsilon\}\ .$$

Because of the nature of V_{θ_2}, for given ε' one can find K and n_o such that with probability $\geq 1 - \varepsilon$, the maximum of A_n (over V_{θ_2}) is attained at η_o^o, η_{o1}^o if $\theta_1^o - \varepsilon + Kn^{-\frac{1}{2}} \leq \theta_2 \leq \theta_1^o - \varepsilon + \delta$ or $\theta_1^o + \varepsilon - \delta < \theta_2 \leq \theta_1 + \varepsilon - Kn^{-\frac{1}{2}}$.

An easy calculation now shows

$$\sup_{\theta} L_n(\theta) = V_n(0) + \sup_{\eta_2 \in \Theta_2} \sup_{\eta_{o1} \in R^+ \times R} A_n(\eta) + o_p(1)$$

where $\Theta_2^{(2)} = \{\theta_2 \ \bar{H}: \ \theta_2 \leq \theta_1^o - \varepsilon \text{ or } \geq \theta_1^o + \varepsilon\}$.
The supremum of $L_n(\theta)$ under H_o has therefore the same expression as in Section 2 with $\Theta_2 = \Theta_2^{(2)}$. Since the expression for the supremum of $L_n(\theta)$ under H_o remains unaltered, the conclusion of Theorem 2.1 is valid, i.e., the following is true.

Theorem 5.1. Assume B1. Then the limiting distribution of the likelihood ratio test under θ_1^o is the same as that in Theorem 2.1 with $\Theta_2 = \Theta_2^{(2)}$.

ACKNOWLEDGEMENT

Thanks are due to the referee whose comments clarified many issues and led to a better presentation.

REFERENCES

Bickel, P.J. and Wichura, M.J. (1971). Convergence criteria for multi-parameter stochastic processes and some applications. Ann. Math. Statist., 42, 1656-1670.

Bock, H.H. (1981). Statistical testing and evaluation methods in cluster analysis. Paper presented at the ISI Golden Jubilee Conference, to appear in the Proceedings.

Chernoff, H. (1954). On the distribution of the likelihood ratio. Ann. Math. Statist., 25, 573-578.

Dharmadhikari, S.W., Fabian, V. and Jogdeo, K. (1968). Bounds on moments of martingales. Ann. Math. Statist., 39, 1719-1723.

Everitt, B.S. and Hand, D.J. (1981). Finite Mixture Distributions. Chapman and Hall, London.

Feder, P.I. (1968). On the distribution of the log likelihood ratio test statistic when the true parameter is near the boundaries of the hypothesis region. Ann. Math. Statist., 39, 2044-2055.

Gupta, S.S. and Huang, Wen-Tao (1981). On mixtures of distributions: a survey and some new results on ranking and selection. Sankhyā B 43, 245-290.

Hájek, J. and Sidák, Z. (1967). Theory of Rank Tests. Academic Press, New York.

Karlin, S. (1968). Total Positivity, Vol. 1, Stanford University Press, Stanford, California.

McCormick, S.P. (1967). Nonlinear Programming. McGraw Hill, New York.

Moran, P.A.P. (1973). Asymptotic properties of homogeneity tests. Biometrika, 60, 79-85.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for non-identifiable distributions. Ann. Statist., 9, 224-227.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist., 20, 595-601.