Testing Homogeneity in a Mixture Distribution via the $L^2$ Distance Between Competing Models

Richard CHARNIGO and Jiayang SUN

Ascertaining the number of components in a mixture distribution is an interesting and challenging problem for statisticians. Chen, Chen, and Kalbfleisch recently proposed a modified likelihood ratio test (MLRT), which is distribution-free and locally most powerful, asymptotically. In this article we present a new method for testing whether a finite mixture distribution is homogeneous. Our method, the D test, is based on the $L^2$ distance between a fitted homogeneous model and a fitted heterogeneous model. For mixture components from standard parametric families, the D-test statistic has a closed-form expression in terms of parameter estimators, whereas likelihood ratio-type test statistics do not; the latter test statistics are nontrivial functions of both the parameter estimators and the full dataset. The convergence rates of the D-test statistic under a null hypothesis of homogeneity and an alternative hypothesis of heterogeneity are established. The D test is shown to be competitive with the MLRT when the mixture components come from a normal location family. However, in the exponential scale and normal location/scale cases, the relative performances of the D test and the MLRT are mixed. In cases such as these two, we propose to use a weighted D test, in which the measure underlying the $L^2$ distance is changed to accentuate the disparities between the homogeneous and heterogeneous models. Changing the measure is equivalent to computing the D-test statistic using a weighting function or to transforming the data before conducting the D test. Appropriately weighted D tests are competitive in both the exponential scale and normal location/scale cases. After applying the D test to a dataset in which the observations are measurements of firms’ financial performances, we conclude with discussion and remarks.

KEY WORDS: ?????

1. INTRODUCTION AND MOTIVATION

Mixture models are useful in describing complex populations with observed or unobserved heterogeneity. For example, the distribution of human heights may be modeled as a two-component mixture, one component for men and one for women. Substructures in a galaxy may be modeled as contaminations of a big initial galaxy; the evidence of substructures is important in the modern galaxy formation theory (Sun, Morrison, Harding, and Woodroffe 2002). There are also applications in econometrics, biology, genetics, medicine, agriculture, zoology, and population studies; examples are provided by Everitt and Hand (1985), Ott (1999), and Lemdani and Pons (1997). However, determining the number of components in a mixture has long been a challenging problem for statisticians—see, for example, Titterington, Smith, and Makov (1993) and Lindsay (1995)—because the problem is undistinguishable under the null hypothesis of homogeneity.

So far, influential procedures for testing homogeneity have included the $C(\alpha)$ test by Neyman and Scott (1966), moment-based testing procedures by Lindsay (1989), and likelihood ratio techniques, discussed later. The $C(\alpha)$ test is a dispersion test for which the alternative is general (i.e., there are at least two distinct components). The other two approaches can be used to test against a specific alternative (e.g., there are exactly two distinct components).

The likelihood ratio test (LRT) is locally most powerful, but its limiting distribution under the null hypothesis is not the standard $\chi^2$ distribution; see, for example, Bickel and Chernoff (1993), Chernoff and Lander (1995), Ghosh and Sen (1985), Hartigan (1985), and Titterington et al. (1985). This “mystery” has inspired continuous research, from the early use of mixture models by Rao (1948), to many studies cited by McLachlan and Basford (1988) and by Lindsay (1995), to more recent works by Chen and Chen (2001), Chen, Chen, and Kalbfleisch (2001), Lo, Mendell, and Rubin (2001), Garel (2001), and Lemdani and Pons (1999). Chen and Chen (2001) and Ghosh and Sen (1985) showed that, under some regularity conditions, the limiting distribution under the null is $\sup_{\theta \in \Theta_0} \left( W^{+}(\theta) \right)^2$, where $W^{+}$ is the positive part of a Gaussian process $W$ and $\Theta$ is the compact parameter space. Thus, the $p$ value of the LRT can be estimated, in principle, using Sun’s (1993) approximation to the tail probability of $\sup_{\theta} \left( W(\theta) \right)$, which does not require $W$ to have a finite Karhunen–Loève expansion. Because of complications in obtaining the approximation, Chen and Chen (2001) resorted to a computationally intensive bootstrap procedure. The modified likelihood ratio test (MLRT), proposed by Chen et al. (2001) for a mixture of a one-dimensional parametric family, is a clever penalized LRT procedure. The MLRT is distribution-free and also locally most powerful, asymptotically. Chen et al. (2001) reported that the MLRT is competitive with the $C(\alpha)$ test, the bootstrap test by McLachlan (1987), and the method of Davies (1987). We will use the MLRT as a benchmark in the comparisons that follow.

The LRT and MLRT are approximately equivalent to using Kullback–Leibler (K–L) information as a measure of the distance between the null and an alternative model. Indeed, the MLRT beats the $C(\alpha)$ test when the K–L information increases (Chen et al. 2001). In this article we propose a new test procedure that uses the $L^2$ distance between a fitted homogeneous model and a fitted heterogeneous model. This is related to Scott’s (1999) $L^2 E$ method for model selection but differs in that we propose a testing procedure rather than a technique for model selection; moreover, we compare two parametric density estimates to each other rather than compare each one to a nonparametric density estimate as Scott did. The $L^2$ distance is
more closely related to the visual separation between the competing models than the K–L distance is, in the following sense:

The area between the homogeneous and heterogeneous density curves is the $L^1$ distance; the $L^2$ distance is like the $L^1$ distance but places more emphasis on the larger separations between the curves. Moreover, using the $L^2$ distance leads to simple closed-form expressions for the test statistic in terms of parameter estimators, a feature absent from likelihood ratio-type tests.

We refer to our procedure as the D test. In Section 2 we describe the D test and illustrate the closed-form expressions of the test statistic for mixture components from standard parametric families. Asymptotic behavior is discussed in Section 3; in particular, we present convergence rates under the null and alternative hypotheses, and the shape of the distribution under the null hypothesis is investigated. The performance of the D test is assessed in Section 4. The D test is competitive with the MLRT when the mixture components come from a normal location family. The MLRT performs a little better for small sample sizes when the mixture components come from an exponential scale family, whereas the D test performs a little better for larger sample sizes. In this exponential scale case, there is little visual separation and, hence, little $L^2$ separation between the homogeneous and heterogeneous models, making heterogeneity difficult to detect via $L^2$ distances when the sample sizes are small. The MLRT of Chen et al. (2001) can be generalized straightforwardly to higher dimensional problems and its null distribution can be simulated. The generalization of the MLRT to two-dimensional problems tends to perform better than the D test when the mixture components come from a normal location/scale family. Therefore, with the exponential scale and normal location/scale cases in mind, in Section 5 we propose a weighted D test. The measure underlying the $L^2$ distance is changed to place greater weight on the disparities between the homogeneous and heterogeneous models. In the exponential scale case, for instance, the disparities are manifested more in the tail behavior than in the shape. Changing the measure underlying the $L^2$ distance is equivalent to computing the D-test statistic using a weighting function, which, in turn, is equivalent to transforming the data before conducting the D test. Appropriately weighted D tests are competitive in both the exponential scale and the normal location/scale cases. Our findings reinforce the principle of using a test statistic based on a suitable distance metric, so that the test is sensitive to the separation between the null and the specific alternative under consideration; this is quite a different idea from merely adapting or modifying the likelihood ratio test. In Section 6 we apply the D test to a dataset in which the observations are measurements of firms’ financial performances. In Section 7 we describe the extent to which the D test enjoys a computational advantage over the MLRT and present some conclusions. Proofs are given in the Appendix.

2. DESCRIPTION OF THE D TEST

Let $X_1, \ldots, X_n$ be a simple random sample from the mixture distribution $\sum_{i=1}^{k} p_i f(x|\theta_i)$, where $p_i \geq 0$, $\sum_{i=1}^{k} p_i = 1$, and $\{f(x|\theta): \theta \in \Theta\} \subseteq L^2$ is a family of probability density functions associated with a scalar or vector parameter $\theta$.

We are interested in testing

$$H_0: X_1 \sim f(x|\theta_0) \quad \text{vs.} \quad H_1: X_1 \sim \sum_{i=1}^{k} p_i f(x|\theta_i),$$

where $\theta_i \in \Theta$ for $i = 0, \ldots, k$, the $\theta_i$ are different for $i \geq 1$, and the $p_i$ are nonzero. Let $\hat{\theta}_0$ denote a consistent estimator of $\theta_0$ under $H_0$, the null hypothesis of homogeneity. Let $\hat{\theta}_i$ and $\hat{p}_i$ denote consistent estimators of $\theta_i$ and $p_i$ (for $i = 1, \ldots, k$) under the alternative hypothesis $H_1$. Our notation suppresses the estimators’ dependence on the sample size $n$; whenever we refer to a sequence of estimators, it is to be understood that this sequence is in $n$.

Until Section 7 we assume that the $\hat{\theta}_i$ and the $\hat{p}_i$ are maximum likelihood estimators obtained via an expectation maximization (EM) algorithm. We comment about the EM algorithm in Remark 2, and we note the possibility of using other estimators in our D test in Section 7.

The D-test statistic is defined as follows:

$$d(k,n) := \int \left[ \sum_{i=1}^{k} \hat{p}_i f(x|\hat{\theta}_i) - f(x|\hat{\theta}_0) \right]^2 dx$$

$$- \int \left[ \sum_{i=0}^{k} \hat{p}_i f(x|\hat{\theta}_i) \right]^2 dx,$$

(1)

where $\hat{\theta}_0 := -1$. In words, $d(k,n)$ is the square of the $L^2$ distance between the heterogeneous and homogeneous densities. It is apparent that $d(k,n)$ should be “small” when the mixture distribution is homogeneous, whereas $d(k,n)$ should be “large” when the mixture distribution is not homogeneous. Of course, “small” and “large” may depend on $k, n$, the family $\{f(x|\theta): \theta \in \Theta\}$, and the extent of the departure from homogeneity.

Remark 1. From (1) it is clear that the D-test statistic depends only on the $\hat{\theta}_i$ and the $\hat{p}_i$. In fact, any test statistic based on an integration in $dx$ will depend only on these parameter estimators. Other distance metrics besides $L^2$ that one might consider are $L^1$ distance and Hellinger distance. For standard parametric families, the $L^2$ distance metric leads to simpler closed-form expressions than the $L^1$ distance and Hellinger distance do. Likelihood ratio-type test statistics, on the other hand, do not involve an integration in $dx$; they depend on the full dataset, not just the estimators.

2.1 Examples

As illustrations, we provide the closed-form expressions for $d(k,n)$ in terms of parameter estimators when the components of the mixture are univariate normal, univariate gamma, and multivariate normal.

For the univariate normal case, the density is $1/(\sqrt{2\pi}\sigma) \cdot \exp(-(x-\mu)^2/(2\sigma^2))$, so that $\theta = (\mu, \sigma^2)$. Using the fact that $\exp(-(x-b)^2/(2a))$ integrates to $\sqrt{\pi a}$, we readily obtain the following closed-form expression:

$$d(k,n) = \sum_{i=0}^{k} \sum_{j=0}^{k} \frac{\hat{p}_i \hat{p}_j}{\sqrt{2\pi(\sigma_i^2 + \sigma_j^2)}} \exp\left[ -\frac{1}{2} \frac{(\hat{\mu}_i - \hat{\mu}_j)^2}{\sigma_i^2 + \sigma_j^2} \right].$$

(2)
If we want to treat $\sigma^2$ as known, we may do so. For instance, if $\sigma^2 = 1$,

$$d(k, n) = \sum_{i=0}^{k} \sum_{j=0}^{k} \frac{\hat{p}_i \hat{p}_j}{2\sqrt{\pi}} \exp \left[-\frac{1}{4}(\hat{\mu}_i - \hat{\mu}_j)^2 \right].$$  

(3)

For the univariate gamma case, the density is $\beta^\alpha / \Gamma(\alpha) \cdot e^{-\beta x} x^{\alpha-1}$, so that $\theta = (\alpha, \beta)^t$. If $\alpha = 1$, then the density is exponential. The respective expressions are

$$d(k, n) = \sum_{i=0}^{k} \sum_{j=0}^{k} \frac{\hat{p}_i \hat{p}_j}{\Gamma(\hat{\alpha}_i + \hat{\alpha}_j - 1)} \Gamma(\hat{\alpha}_i) \Gamma(\hat{\alpha}_j)(\hat{\beta}_i + \hat{\beta}_j)^{-\hat{\alpha}_i - \hat{\alpha}_j - 1}$$  

(4)

for the gamma case and

$$d(k, n) = \sum_{i=0}^{k} \sum_{j=0}^{k} \frac{\hat{p}_i \hat{p}_j}{\hat{\beta}_i + \hat{\beta}_j}$$  

(5)

for the exponential case.

For the $d$-dimensional multivariate normal case with known identity covariance matrix,

$$d(k, n) = \sum_{i=0}^{k} \sum_{j=0}^{k} \frac{\hat{p}_i \hat{p}_j}{2\pi^{d/2}} \exp \left[-\frac{1}{4}(\hat{\mu}_i - \hat{\mu}_j)^2 \right].$$  

(6)

### 3. ASYMPTOTIC BEHAVIOR

In this section we study the asymptotic behavior of the D test under the null and the alternative hypotheses. For simplicity, as in Chen et al. (2001), we consider two-component mixtures (i.e., $k = 2$) involving univariate X and scalar $\theta$. We also assume without loss of generality that $p_1 \geq p_2$, and where convenient we take $p_2 = \alpha$ and hence $p_1 = 1 - \alpha$.

The study of the asymptotic behavior of $d(2, n)$ is complicated by the lack of simple distributional results about $\hat{p}_1$ and $\hat{\theta}_2$; such results are unavailable because the respective parameters are not identifiable under the null hypothesis of homogeneity.

Our approach to the study of asymptotic behavior will, therefore, be twofold: First, we will show that $d(2, n) \sim O_p(1)$ under the alternative hypothesis and that $d(2, n) = O_p(n^{-1/2})$ under the null hypothesis. Second, for the normal location and exponential scale models indicated by expressions (3) and (5), we will examine the shape of the distribution of $d(2, n)$ under the null hypothesis for various choices of $n$.

#### 3.1 Convergence Rates for $d(2, n)$

The convergence rates for $d(2, n)$ will be established based on conditions A1–A5, given in the Appendix. Conditions A1–A5 are sufficient for the "generalized" Cramér’s regularity conditions on $\{f(x|\theta) : \theta \in \Theta\}$, which ensure that $(\hat{\theta}_0 - \theta_0) = O_p(n^{-1/2})$ (see thm. 7.3.1 of Lehmann 1999), and the conditions imposed in Section 2 of Chen and Chen (2001), which ensure strong identifiability and further smoothness. Many standard parametric families satisfy these conditions.

**Proposition 1.** Let $X_1, \ldots, X_n$ be iid with mixture components from a minimal exponential family (cf. Brown 1986) for which conditions A1–A3 are satisfied, with $f(x|\theta) = a(x) \exp[-b(\theta) + t(x)\theta]$. Suppose that $X_1 \sim (1-\alpha) f(x|\theta_1) + \alpha f(x|\theta_2)$, where $0 < \alpha \leq 1/2$ and $\theta_1 \neq \theta_2$. If $(\hat{\theta}_1, \hat{\theta}_2, \hat{\alpha})$ is a sequence of consistent roots of the likelihood equation, then $d(2, n) = O_p(1)$.

The proof, given in the Appendix, shows that $d(2, n)$ is not only $O_p(1)$ but also convergent in probability to a nonzero quantity; in addition, the requirement of exponential family membership may be dropped if we can assume that $\hat{\theta}_0$ converges to some value $\theta_0$ under $H_1$.

**Remark 2.** The existence of a sequence of consistent roots of the likelihood equation under $H_1$ was shown by Kiefer (1978) for model (2). In the EM algorithm, one may choose starting values either from moment estimates of the parameters or from examining an undersmoothed histogram. Both methods lead to comparable results based on our experiences. Clearly, the moment estimators are consistent and, hence, so are the estimators obtained from the EM algorithm. Lindsay (1995, p. 66) reported that the EM algorithm starting with moment estimates generally yields solutions close to the truth.

**Lemma 1.** Let $X_1, \ldots, X_n$ be iid and consider the null hypothesis $H_0: X_i \sim f(x|\theta_0)$, where the true $\theta_0$ is an interior point in the compact parameter space $\Theta$. Under conditions A1–A5,

$$\hat{p}_1(\hat{\theta}_1 - \theta_0) + \hat{p}_2(\hat{\theta}_2 - \theta_0) = O_p(n^{-1/2}),$$

$$\hat{\theta}_1 - \theta_0 \sim O_p(n^{-1/2}),$$

$$\hat{\theta}_2 - \theta_0 \sim O_p(n^{-1/2}).$$

These results can be obtained as by-products of Chen and Chen’s (2001) derivation of the limiting distribution for the LRT statistic. For details, see the Appendix.

**Proposition 2.** Under the conditions in Lemma 1, $d(2, n) = O_p(n^{-1})$.

In the Appendix we present the proof, which is based on Lemma 1 and Taylor expansions of $f(x|\theta)$.

**3.2 Shape of the Distribution Under the Null**

Given the convergence rate of the D-test statistic under the null hypothesis, it may be possible to obtain a limiting distribution for properly rescaled $d(2, n)$. The usefulness of such a limiting distribution, however, would depend on how quickly the rescaled test statistic converged in distribution.

Figures 1 and 2 pertain to models (3) and (5)—recall that the underlying parametric families are normal location (with known $\sigma^2 = 1$) and exponential scale—with $\mu_0$ and $\beta_0$ taken to equal 0 and 1, respectively. The associated parameter spaces are taken to be $[-M, M]$ and $[M^{-1}, M]$ for large $M$. The his-
Figure 1. Histograms of log(d(2, 40)) and log(d(2, 400)).

Figure 2. Histograms of log(d(2, 400)).

Figure 3. Histograms of log(d(2, 4000)).

tograms display values of log(d(2, n) + δ), for various n and an infinitesimal δ = 10^{-14}, based on simulations of size 10,000.

In Figure 1 are histograms corresponding to n = 40 and n = 400. When n = 40, there are two clear modes in both the normal location and the exponential scale cases. The left mode represents an accumulation of d(2, n) values that are essentially equal to 0. By contrast, when n = 400, the left mode has shrunk considerably; note the scale on the vertical axis. In Figure 2 are histograms for n = 4,000; the left mode is now quite small. Noting the scale on the horizontal axis, one also sees that the right mode is moving leftward with increasing n; in light of Proposition 2, this is to be expected.

From the histograms, we can see that the size of the left mode decreases with n, though the basic shape of the distribution does not seem to depend on the family \{f(x|θ): θ ∈ Θ\}. However, because the shape of the distribution changes so dramatically with n, direct estimation of critical values would seem necessary for small and modest n even if the limiting distribution were known.

Accordingly, we have tabulated critical values for models (3) and (5) when µ₀ = 0 and β₀ = 1; these tables are based on simulations of size 10,000 for each n. The tables and our software are available upon request. At first it may seem that we need to produce many more tables, corresponding to different choices of µ₀ (or the known σ²) and β₀. Proposition 3 shows that, in fact, only a few basic tables are required.

**Proposition 3.** For any fixed n, let \(d_{α}(\mu_0,σ^2)\) and \(d_{α}\exp(β_0)\) denote the level α critical values of \(d(k,n)\) based on null distributions of \(N(μ_0,σ^2)\) and \(\exp(β_0)\). Then \(d_{α}(\mu_0,σ^2) \approx d_{α;N(0,1)/σ} \) and \(d_{α}\exp(β_0) \approx β₀d_{α;N(1)}\).

The proof is given in the Appendix. Note that Proposition 3 is valid for an arbitrary fixed number of mixture components k. In practice, the approximation symbols may be treated as equalities. Thus, for example, if we have a two-component mixture from a normal location family but with \(σ^2 = 4\) instead of \(σ^2 = 1\), the appropriate critical value is one-half the tabulated value for model (3); similarly, if we have a two-component mixture from an exponential scale family and \(β₀ = 2\), the appropriate critical value is twice the tabulated value for model (5).

Finally, the asymptotic analysis in the preceding section provides a mechanism to obtain critical values for sample sizes n not included in our tables. Explicitly, let \(d_{α;n}\) denote the level α critical value of \(d(2, n)\); the \(O(n^{-1})\) rate of Proposition 2 suggests that, for \(n_1 < n_2 < n_3\),

\[
d_{α;n_2} \approx \frac{n_2^{-1} - n_3^{-1}}{n_1^{-1} - n_3^{-1}}d_{α;n_1} + \frac{n_1^{-1} - n_2^{-1}}{n_1^{-1} - n_3^{-1}}d_{α;n_3}. \tag{7}
\]

4. EVALUATION OF THE D TEST

To assess how well the D test performs, we generate data from mixture distributions that have various degrees of departure from homogeneity, then we see how frequently the critical values are exceeded; we set the Type I error probability α to .05. We first consider models (3) and (5), which are comparable to the cases examined by Chen et al. (2001). In particular, model (3) represents a basic one-dimensional situation...
on which any homogeneity test should be evaluated. We then proceed to model (2), in which the mixture components come from a normal location/scale family; this is a more complex (and often more realistic) situation than the one represented by model (3).

4.1 Normal Location and Exponential Scale Families

For model (3), we generate data from an equal-probability mixture of normals, $0.5N(-\mu, 1) + 0.5N(\mu, 1)$, for $\mu = 0.5, 1.0, 1.5, 2.0$. For each $\mu$ and $n$, we obtain 10,000 values of $d(2, n)$, which are compared to the critical values. Moreover, we simultaneously apply the MLRT of Chen et al. (2001) with their penalty parameter $C$ set to $\log(10)$, obtaining 10,000 values for this test statistic as well. The D test performs better than the MLRT for the larger sample sizes, but the D test does not perform as well for the smaller sample sizes, as illustrated for $a = 1$ and $a = 2$ in Figure 3.

This outcome may not be surprising: The D test is based on the $L^2$ distance between competing models, and the $L^2$ distance is geared more toward detecting disparities in shape than disparities in scale. With little difference in shape between homogeneous and heterogeneous exponential mixture densities (see the left panel of Fig. 4), small-sample variability makes it difficult for the D test to detect heterogeneity. In Section 5 we show how the D test can be adapted to better handle model (5) when $n$ is small.
4.2 Normal Location/Scale Family

For model (2), we again generate data from an equal-probability mixture of normals, .5N(μ, 1) + .5N(μ, 1), for
μ = .5, 1.0, 1.5, 2.0. This time, however, both the location and
scale parameters are treated as unknown and are estimated.
For each μ and n, we obtain 10,000 values of d(2, n), which
are compared to appropriate critical values. The MLRT was
proposed by Chen et al. (2001) for testing homogeneity in
a mixture whose components come from a one-dimensional
parametric family. However, the MLRT can be generalized to
higher-dimensional problems. Because no asymptotic distribu-
tion has been established for such problems, we obtain critical
values by simulation. Finally, we calculate 10,000 values of the
generalized MLRT statistic under the indicated heterogeneity,
comparing them to the critical values.

The generalized MLRT tends to be more powerful than the
D test. The solid and dotted curves in Figure 6 show the rel-
ative performances of the generalized MLRT and the D test
for μ = 1.0 and μ = 1.5. The approaches corresponding to the
other curves in Figure 6 will be explained in the next section.

5. WEIGHTING FUNCTIONS

We defined d(k, n) as an integral with respect to Lebesgue
measure. If we wish to emphasize disparities in right-tail be-
behavior instead of disparities in shape, as is appropriate for the
exponential scale case, we can choose some other underlying
measure, say, dv(x) := w(x)dx, where w(x) is a positive,
increasing function of x.

Another way to motivate the weighting function w(x) is
through transformation of the data. Suppose X1, ..., Xn ∼
∑k
i=1
pif(x, θi) and define Y := g(X). Then

Y1, ..., Yn ∼ ∑k
i=1
pif(g−1(y), θi)|dg−1(y)/dy|.

So, when Y is viewed as the response variable,

\[ d(k, n) = \int \left( \sum_{i=1}^{k} \hat{p}_i f \left( g^{-1}(y), \hat{\theta}_i \right) - f \left( g^{-1}(y), \theta_0 \right) \right)^2 \times \left| \frac{dg^{-1}(y)}{dy} \right|^2 dy \]

\[ = \int \left( \sum_{i=1}^{k} \hat{p}_i f \left( x, \hat{\theta}_i \right) - f \left( x, \theta_0 \right) \right)^2 \frac{dx}{dy} \left| \frac{dx}{dy} \right| \]

\[ = \int \left( \sum_{i=1}^{k} \hat{p}_i f \left( x, \hat{\theta}_i \right) - f \left( x, \theta_0 \right) \right)^2 w(x) dx, \quad (8) \]

where w(x) := |dx/dy|.

For example, the transformation Y := log(X) gives rise to
the weighting function w1(x) = x, whereas the transformation
Y := 1/X gives rise to the weighting function w2(x) = x².
Indeed, these transformations make the homogeneous and het-
erogeneous exponential mixtures much more separable in the
density space (see the right panel of Fig. 4) and in the L² met-
ric space (see Fig. 8, to be described later).

In the exponential scale case, this weighted D test with
w1(x) = x is competitive with the MLRT for both the smaller
and the larger sample sizes, as shown in the left panel of
Figure 5. The weighting function w2(x) = x² yields even better
results for small sample sizes and small departures from homo-
genility, as shown in the right panel of Figure 5; however, the
weighting function w1(x) = x seems better in general.

In the normal location/scale case, we do not want to choose
a weighting function that is increasing in x. It makes more
sense to consider weighting functions that depend on x through
(x − µ0)²/σ0². A family of such functions is given by

\[ w_c(x) := \exp\left[ -c(x - \hat{\mu}_0)^2/\hat{\sigma}_0^2 \right]. \quad (9) \]

If c > 0, w_c(x) places more weight in the central region
around µ0; if c < 0, more weight is placed in the tails. When
the heterogeneity is in location, placing more weight in the central region works well, as the disparities between the homogeneous and heterogeneous models are pronounced near the mode of the homogeneous model. Results for $c = 8, 16, \text{and } 32$ ($\mu = 1.0 \text{ and } \mu = 1.5$) are displayed in Figure 6; they compare favorably to the results shown for the generalized MLRT and are clearly superior to the results displayed for the unweighted D test. Results for $\mu = .5$ are worth mentioning: The power figures for both tests are close to the nominal significance level $\alpha = .05$ although the generalized D tests improved the performance slightly for large $n$. This is in contrast to the results for the normal location model (3) in Figure 3 when $\mu = .5$.

Interestingly, the weighting function suggested by formula (6) of Roeder (1994) is the one with $c = -1/2$, but this weighting function actually yields a much less powerful test. Roeder (1994), however, was proceeding under the assumption that $\sigma_1^2 = \sigma_2^2$; we have made no such assumption for model (2).

6. APPLICATION TO THE BANKRUPTCY DATA

To see how the D test might work in practice, we apply it to the bankruptcy data from Johnson and Wichern (2002). This dataset contains four measurements of financial performance for each of 46 firms as well as an indicator of whether each firm remained solvent over the next few years. Johnson and Wichern presented this dataset in the context of a classification problem: Can the measurements of financial performance be used to separate the firms that went bankrupt from those that remained solvent?

Similar error rates are obtained whether one classifies by tree or by quadratic discriminants. An advantage of the former approach is that it reveals which financial performance measurement is most relevant in separating the two kinds of firms (bankrupt and solvent). When the tree is optimally pruned according to a cross-validation criterion, only the first and third measurements of financial performance come into play; the initial split is based on the third. Moreover, when three rather unusual observations are omitted, the optimally pruned tree involves only the third measurement of financial performance.

Therefore, we confine our attention here to the third measurement of financial performance, a ratio of “current assets” to “current liabilities” that accounting professionals refer to as the “current ratio”; the current ratio reflects a firm’s ability to meet its short-term financial obligations, so it is reasonable that this particular measurement turned out to be most relevant in separating the two kinds of firms. Figure 7 shows a density estimate for the distribution of log current ratios based on Scott’s ASH plot (1985; code available online at ftp://ftp.stat.rice.edu/pub/scottdw/ASH.code) as implemented in X-Gobi (May 2000 release; available online at...
The smoothing parameter has been set at .130; bankrupt firms are marked with dots, whereas solvent firms are marked with crosses. For the moment, we ignore the distinction between the two kinds of firms and simply take all 46 values of the log current ratio as observations to which the D test will be applied.

For the purpose at hand, model (2) is more appropriate than model (3); we will estimate the $\sigma^2$ rather than fix them at a particular value. The realized value of $d(2, n)$ for the bankruptcy data is .02126, and the associated $p$ value is estimated to be .115; the $p$ value is fairly low, though it is above the usual significance demarcation of .05.

Is this a reasonable finding? To get an idea, we compute the Akaike information criterion (AIC) and Bayesian information criterion (BIC) for the fitted homogeneous and heterogeneous models. We find that the AIC is smaller for the heterogeneous model (72.99, compared to 75.10), whereas the BIC is smaller for the homogeneous model (78.76, compared to 82.13). So, based on the AIC and BIC, neither the homogeneous model nor the heterogeneous model is clearly preferred over the other, which reinforces the result of the D test and gives credibility to our method.

7. DISCUSSION AND CONCLUSIONS

We have provided a new procedure for testing homogeneity in a finite mixture distribution, one in which the test statistic has a closed-form expression in terms of parameter estimators. Because calculation of the LRT or MLRT statistic also requires that the full dataset be in hand (not just the parameter estimates), it is apparent that the D test enjoys an advantage relative to the LRT and MLRT. As shown in Table 1, this advantage is small if the full dataset is readily accessible for the calculation of the LRT or MLRT statistic, though the advantage increases as $n$ becomes larger. However, if the full dataset is not readily accessible—sometimes datasets are “processed” and then discarded after summary statistics and parameter estimates are calculated—one can only apply the LRT or MLRT after time has been expended to reacquire the dataset. In such a situation, the D test has a greater advantage.

In Table 1, we used Version 6.1.2 of S-PLUS for Linux on an AMD Athlon processor (858 MHz) workstation, invoking g77 FORTRAN subroutines to calculate maximum likelihood parameter estimates and the LRT statistic. The third column provides the time to compute the LRT statistic, given the maximum likelihood estimates and the full dataset; the time to compute the MLRT statistic, given the penalized maximum likelihood estimates and the full dataset, would be comparable. The fourth column provides the time to compute the D-test statistic, given the parameter estimates, in S-PLUS. If the LRT were computed in S-PLUS, it would be slower than what is reported in the third column. All times are reported to the nearest hundredth of a second.

In addition, although we have adopted a maximum likelihood estimation framework in this article, the D test could be used in conjunction with any reasonable framework for parameter estimation; for example, method-of-moments estimators are reasonable alternatives. Using a different estimation framework could provide a substantial savings in time, making the D test particularly attractive for data mining applications. Indeed, Table 1 indicates that the bulk of the time involved in using the D test, LRT, or MLRT lies with computing the maximum likelihood estimates rather than calculating the actual test statistics.

It is worth emphasizing that the D test can be applied in situations where $\theta$ is a vector or the data are multivariate. We considered the former situation in Sections 4.2 and 6, and we indicated the latter possibility in expression (6). There may be issues of computational feasibility in some such situations, although these issues may be mitigated by using a different framework for parameter estimation. Moreover, our method can be extended to test homogeneity in discrete mixture distributions, which will be the subject of a forthcoming paper.

Our procedure is designed for testing homogeneity versus a specific kind of heterogeneity, namely, that in which the density is a finite mixture from a particular parametric family. The motivation for our procedure is that a test based on a specific alternative is often more powerful under that alternative than a test based on a general alternative. However, if a general alternative were desired, the D test could be modified to accommodate a nonparametric density estimate.

The success of the weighted D test in handling the exponential scale case can be further clarified with the aid of Figure 8. For $a = 0, 1, 2$, let $f_0(x) := (1 + a/2)^{-1}\exp(-(1 + a/2)^{-1}x)$ and $f_1(x) := .5\exp(-x) + .5(1 + a)^{-1}\exp(-(1 + a)^{-1}x)$. The first panel shows the function $f_0(x) \cdot \log(f_1(x)/f_0(x))$. Up to a scaling factor, the LRT statistic is an empirical version of the integral of this function. The next three panels show $(f_1(x) - f_0(x))^2 w(x)$ for $w(x) = 1, x,$ and $x^2$, respectively. The corresponding (weighted) D-test statistics are empirical versions of the integrals of these functions. We see a clearer separation between $f_0$ and $f_1$ when a nonconstant weighting function is applied, which suggests that employing one of these nonconstant weighting functions will make departures from homogeneity easier to detect. Plots such as Figure 8 can also be adapted to provide graphical diagnostic procedures; these would be similar in spirit to Roeder’s (1994) procedures although somewhat different technically, as she uses a kernel estimate of the mixture density rather than a parametric estimate.

The performance of the D test for model (2), in which the family $\{f(x; \theta) : \theta \in \Theta\}$ is normal but $\theta$ is taken to be a vector, is different from its performance for model (3), in which the variance is known and $\theta$ is simply the mean. In the former situation, overdispersion alone does not provide evidence against $H_0$, which makes it more difficult to reject $H_0$ using the D test (and also using the MLRT). However, as we have seen, it is possible to improve upon the performance of the D test for model (2) by selecting a different measure on which to base the $L^2$ distance.

In summary, our article calls for using more powerful tests based on suitably chosen distance metrics, quite a different idea from merely adapting or modifying the likelihood ratio test.

---

### Table 1. Computation Times for Massive Datasets—Normal Location

<table>
<thead>
<tr>
<th>$n$</th>
<th>Parameter estimation</th>
<th>Computing LRT</th>
<th>Computing D test</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,000</td>
<td>.87</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>10,000</td>
<td>2.18</td>
<td>.02</td>
<td>.01</td>
</tr>
<tr>
<td>40,000</td>
<td>8.21</td>
<td>.08</td>
<td>.01</td>
</tr>
<tr>
<td>100,000</td>
<td>24.07</td>
<td>.21</td>
<td>.01</td>
</tr>
<tr>
<td>400,000</td>
<td>96.34</td>
<td>.82</td>
<td>.01</td>
</tr>
</tbody>
</table>
APPENDIX: REGULARITY CONDITIONS AND PROOFS

Regularity Conditions A1–A5

A1 (Support). The probability density function \( f(x|\theta) \) has a common support in \( x \) for all \( \theta \) in the parameter space \( \Theta \); that is, \( A := \{ x : f(x|\theta) > 0 \} \) is independent of \( \theta \). The parameter space \( \Theta \) is compact and the true \( \theta_0 \) (under \( H_0 \)) is an interior point of \( \Theta \).

A2 (Smoothness and integrability). The probability density function \( f(x|\theta) \) has three continuous partial derivatives with respect to \( \theta \), for which
\[
\left| \frac{\partial^j}{\partial \theta^j} \log f(x|\theta) \right| \leq M_0(x) \quad \forall x \in A \quad \text{and} \quad |\theta - \theta_0| < \epsilon(x_0),
\]
where \( E M_0(X) < \infty \) and \( K(x) \in L^1 \cap L^2 \). Moreover, \( f(x|\theta) \) and its first two partial derivatives with respect to \( \theta \) are jointly continuous in \( x \) and \( \theta \).

A3 (Identifiability). Let \( G(\theta) := (1-p)I_{[\theta \geq \theta_1]} + pI_{[\theta \geq \theta_2]} \). The mixture density
\[
\int f(x|\theta) \, dG(\theta) = (1-p) f(x|\theta_1) + p f(x|\theta_2)
\]
is identifiable in the sense that \( \int f(x|\theta) \, dG_1(\theta) = \int f(x|\theta) \, dG_2(\theta) \) for all \( x \) implies \( G_1 = G_2 \). The function \( f(x|\theta) \) and its two derivatives are also identifiable. For any distinct \( \theta_1, \theta_2 \in \Theta \),
\[
\sum_{j=1}^2 \left\{ a_j f(x|\theta_j) + b_j \frac{\partial}{\partial \theta} f(x|\theta_j) + c_j \frac{\partial^2}{\partial \theta^2} f(x|\theta_j) \right\} = 0 \quad \text{for all} \quad x
\]
implies that \( a_j = b_j = c_j = 0 \) for \( j = 1, 2 \).

A4 (Tightness). The processes \( n^{-1/2} \sum Y_i(\theta_1), \, n^{-1/2} \sum Y_i(\theta_2), \, n^{-1/2} \sum Z_i(\theta_1), \, n^{-1/2} \sum Z_i(\theta_2) \) are tight, where
\[
Y_i(\theta) := \frac{f(X_i|\theta) - f(X_i|\theta_0)}{\theta - \theta_0} f(X_i|\theta_0),
\]
\[
Z_i(\theta) := \frac{Y_i(\theta) - Y_i(\theta_0)}{\theta - \theta_0} \quad \text{for} \quad \theta \neq \theta_0 \quad \text{and} \quad Y_i(\theta_0) := \frac{\partial}{\partial \theta} f(X_i|\theta_0),
\]
and \( Y_i(\theta_0) :\) is independent of \( X_i \) and its two derivatives
\[
\int Y_i(\theta_0) \, dG_1(\theta_0) = \int Y_i(\theta_0) \, dG_2(\theta_0) \quad \text{for} \quad \theta \neq \theta_0 \quad \text{and} \quad Z_i(\theta_0) := Y_i(\theta_0).
\]

A5 (Uniform strong law condition of large numbers). There exist a function \( g \), for which \( Eg(X_i) < \infty \), and a number \( \delta > 0 \) such that
\[
|Y_i(\theta)|^{4+\delta} \leq g(X_i) \quad \text{and} \quad |Y_i(\theta)|^3 \leq g(X_i) \quad \text{for all} \quad \theta \in \Theta.
\]
For a fixed \( \theta_0 \), we can consider an open set in \( \Theta \) that contains \( \theta_0 \). So, conditions A1–A5 ensure conditions C1–C5, C6, and C7 required by theorem 7.3.1 of Lehmann (1999) for any consistent sequence of roots \( \hat{\theta} \) of the likelihood equation to satisfy \( \hat{\theta}_0 = \theta_0 = O_p(n^{-1/2}) \). Conditions A1–A5 are also sufficient for the conditions imposed in section 2 of Chen and Chen (2001) to establish the limiting distribution of the likelihood ratio test statistic.

Proof of Proposition 1

Let \( \theta_0 \) be defined by \( h'(\theta_0) = (1-\alpha)h'(\theta_1) + \alpha h'(\theta_2) \) and note that \( \hat{\theta}_0 \) converges in probability to \( \theta_0 \). From condition A2, it now follows that \( d(2, n) \) converges in probability to \( f(1-\alpha)f(x|\theta_1) + \alpha f(x|\theta_2) - f(x|\theta_0)]^2, \) so that \( d(2, n) = O_p(1) \).

Proof of Lemma 1

Let \( \varepsilon := \sup \Theta - \inf \Theta, \, m_1 := (1-\alpha)(\theta_1 - \theta_0) + \alpha(\theta_2 - \theta_0), \, m_2 := (1-\alpha)(\theta_1 - \theta_0)^2 + \alpha(\theta_2 - \theta_0)^2, \) and \( \varepsilon := m_1 + m_2 h(\theta_0), \) where \( h(\theta) := E(Y_i(\theta_0)Z_i(\theta))/E(Y_i^2(\theta_0)) \); refer to condition A4 for the definitions of \( Y_i(\theta) \) and \( Z_i(\theta) \).
Also, let $I_n(\alpha, \theta_1, \theta_2) := \sum_{i=1}^{n} \log[(1 - \alpha) f(X_i|\theta_1) + \alpha f(X_i|\theta_2)]$ be the log likelihood and let $W_i(\theta) := Z_i(\theta) - Y_i(\theta)h(\theta)$. From the computations in Chen and Chen (2001), we have the following string of inequalities:

$$2\ln(\hat{\alpha}, \hat{\theta}_1, \hat{\theta}_2) - 2\ln(1/2, \theta, \theta_0)$$
$$\leq 2\ln(1/2, \theta, \theta_0) + 2\hat{\alpha} \sqrt{n^{1/2}eO_P(1) + \sum W_i(\theta)}$$
$$- \left( \frac{\hat{\alpha}}{\sqrt{n}} + \hat{\gamma} \ln(1 + \alpha f(\theta)) \right) \left[ \sum W_i(\theta) \right]$$
$$\leq \left[ (1 + o_P(1))^{-1} \left( \sum W_i(\theta) \right) / \left( \sum W_i(\theta) \right) \right]$$
$$+ 2\hat{\alpha} \sqrt{n^{1/2}eO_P(1) + \sum W_i(\theta)}$$
$$- \left( \frac{\hat{\alpha}}{\sqrt{n}} + \hat{\gamma} \ln(1 + \alpha f(\theta)) \right) \left[ \sum W_i(\theta) \right]$$
$$\leq \left( \frac{\hat{\alpha}}{\sqrt{n}} + \hat{\gamma} \ln(1 + \alpha f(\theta)) \right) \left[ \sum W_i(\theta) \right]$$

The last term is the constant in the exponential scale case similar. If the parameter space is unbounded, we would have $\hat{\alpha}_k = \mu_0 + \hat{\alpha}_k$ and $\hat{\theta}_k = \hat{\theta}_k$ for $0 < i < k$, where $\hat{\theta}_k$ and $\hat{\alpha}_k$ are estimators corresponding to the standardized data $Z_i := (X_i - \mu_0)/\sigma$. Because the parameter space is bounded, the preceding equalities hold with probability $1 - \gamma$, where $\gamma$ will be essentially $0$ for modest $\mu_0$ and $\sigma^2$. When these equalities do hold, the value of $d(k, n)$ based on $X_1, \ldots, X_n$ is

$$\int \left\{ \frac{\hat{\mu}_i f(x|\hat{\mu}_i, \sigma^2) - f(x|\mu_0, \sigma^2)}{\sigma} \right\}^2 dx$$
$$= \int \left\{ \frac{\hat{\mu}_i f(x|\hat{\mu}_i, \sigma^2) - f(x|\mu_0, \sigma^2)}{\sigma} \right\}^2 dx$$
$$= \int \left\{ \frac{\hat{\mu}_i f(x|\hat{\mu}_i, \sigma^2) - f(x|\mu_0, \sigma^2)}{\sigma} \right\}^2 dx$$
$$= \int \left\{ \frac{\hat{\mu}_i f(x|\hat{\mu}_i, \sigma^2) - f(x|\mu_0, \sigma^2)}{\sigma} \right\}^2 dx$$

which we recognize as $1/\sigma$ times the value of $d(k, n)$ based on $Z_1, \ldots, Z_n$.

Now, let $A$ be the event that $\hat{\mu}_i = \mu_0 + \hat{\alpha}_i$ and $\hat{\theta}_k = \hat{\theta}_k$ for $0 < i < k$, let $B$ be the event that $d(k, n)$ based on $Z_1, \ldots, Z_n$ is less than or equal to $d_{\alpha,N}(0,1)$, and let $C$ be the event that $d(k, n)$ based on $X_1, \ldots, X_n$ is less than or equal to $d_{\alpha,N}(0,1)/\sigma$. The following equalities establish bounds for $P(C) = 1 - \alpha + \gamma = 1 - \{P(\hat{\alpha}) + P(\hat{\theta}) + P(d(k, n)) \}$.

Hence, $(1/\sigma)d_{\alpha,N}(0,1) = d_{\alpha,N}(\mu_0, \sigma^2)$ for some $\alpha$ with $|\alpha - \hat{\alpha}| < \gamma$.

[Received July 2002, Revised October 2003.]

REFERENCES


Charnigo and Sun: Testing Homogeneity in a Mixture Distribution


