A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion

Author(s): Wayne A. Woodward, William C. Parr, William R. Schucany and Hildegard Lindsey

Source: *Journal of the American Statistical Association*, Vol. 79, No. 387 (Sep., 1984), pp. 590-598
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2288405
Accessed: 12-02-2017 01:10 UTC

A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion

WAYNE A. WOODWARD, WILLIAM C. PARR, WILLIAM R. SCHUCANY, and HILDEGARD LINDSEY*

The estimation of mixing proportions in the mixture model is discussed, with emphasis on the mixture of two normal components with all five parameters unknown. Simulations are presented that compare minimum distance (MD) and maximum likelihood (ML) estimation of the parameters of this mixture-of-normals model. Some practical issues of implementation of these results are also discussed. Simulation results indicate that ML techniques are superior to MD when component distributions actually are normal, but MD techniques provide better estimates than ML under symmetric departures from component normality. Interestingly, an ad hoc starting value for the iterative procedures occasionally outperformed both the ML and MD techniques. Results are presented that establish strong consistency and asymptotic normality of the MD estimator under conditions that include the mixture-of-normals model. Asymptotic variances and relative efficiencies are obtained for further comparison of the MD and ML estimators.

KEY WORDS: Robustness; Relative efficiency; Simulation; Iterative routines; EM algorithm.

1. INTRODUCTION

The problem of estimating the mixing proportions p_1, $p_2, \ldots, p_m$ in the mixture density

$$f(x) = p_1 f_1(x) + p_2 f_2(x) + \cdots + p_m f_m(x),$$

where the number of components m is known and the component densities are only specified as belonging to some parametric family, such as the normal, is often encountered in statistical applications. For example, in aerospace remote sensing, the components might be different crops, with the variable of interest being the reflected energy in four bands of the light spectrum as measured by the LANDSAT satellite, certain linear combinations of these readings, or other derived feature variables.

Substantial attention has been given in the literature to the case in which the component densities are assumed to be normal and the parameters are estimated using maximum likelihood (ML) techniques. In this article we examine the use of a form of minimum distance (MD) estimation as an alternative to ML. Except for some general theoretical results in Section 5, we focus on the mixture of two univariate normal distributions. In Section 2 we provide background material concerning the estimators to be studied, and in Section 3 we discuss the practical issue of starting values for these iterative estimation techniques. In Section 4 we compare the robustness of the estimators to departures from component normality through simulation results. Section 5 presents asymptotic normality and strong consistency results for the Cramér–von Mises minimum distance estimator under conditions that include the mixture-of-normals model. We use these results in Section 6 to produce the asymptotic relative efficiencies of these ML and MD estimators (MLE and MDE).

In this article we assume that only data from the mixture distribution are available. Other sampling schemes in which training samples from the component distributions are also available have been discussed by Hosmer (1973), Redner (1980), and Hall (1981), among others.

2. ESTIMATION IN THE MIXTURE-OF-NORMALS MODEL

The results in Sections 2–4 are based on a mixture $f(x) = pf_1(x) + (1 - p)f_2(x)$, where f_1 and f_2 are normal densities with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. All five parameters μ_1, σ_1^2, μ_2, σ_2^2, and p are unknown, and various techniques for estimating them are described and compared. In our setting, which was motivated by the LANDSAT crop example, p is the parameter of paramount importance, and the unknown location and scale parameters of the components are viewed as nuisance parameters.

2.1 Maximum Likelihood

Several articles have dealt with the problem of obtaining the ML estimates of μ_1, σ_1^2, μ_2, σ_2^2, and p (Has-

* Wayne A. Woodward is Associate Professor of Statistics and Associate Director of the Center for Applied Mathematical and Statistical Research, and William R. Schucany is Professor of Statistics, both at Southern Methodist University, Dallas, TX 75275. William C. Parr is Associate Professor of Statistics, University of Florida, Gainesville, FL 32611. Hildegard Lindsey participated in this research as a graduate student in the Statistics Department at Southern Methodist University. She is employed by Marketing and Research Counselors, Inc., Dallas, TX 75234. This research was sponsored by National Aeronautics and Space Administration Contract NAS 9-16438 and Office of Naval Research Contract N00014-82-K-0207.

© Journal of the American Statistical Association
September 1984, Volume 79, Number 387
Theory and Methods Section

selblad 1966, Day 1969, Wolfe 1970, Hosmer 1973, Fowlkes 1979, Lennington and Rassbach 1979, and Redner 1980). Since the likelihood function L is not a bounded function in this case (see Day 1969), the objective in the ML approach is to find an appropriate local maximum. Since closed-form solutions of the likelihood equations do not exist, they must be solved by using iterative techniques. Hasselblad (1966) and Wolfe (1969) suggested that these equations be solved by taking advantage of their fixed-point form. Redner (1980) and Redner and Walker (1982) pointed out that this fixed-point technique is essentially an application of the EM algorithm (see Dempster, Laird, and Rubin 1977), the only difference being that in using the EM algorithm, the estimates of σ_1^2 and σ_2^2 at step k involve the updated kth step estimates of μ_1 and μ_2. Our implementation will use this updating. Fowlkes (1979), on the other hand, proceeded by utilizing a quasi-Newton method of maximizing $\log(L)$ and found that good starting values were crucial for acceptable performance. In Section 3 we compare these two techniques with regard to their sensitivity to poor starting values. Based on those results, we chose to use the EM algorithm to obtain the MLE in the simulations reported in Section 4.

2.2 Minimum Distance

Minimum distance estimation, introduced by Wolfowitz (1957), has often been shown to be more robust to departures from underlying assumptions than is maximum likelihood. Parr and Schucany (1980), for example, showed that MD techniques can provide robust estimators of the location parameter of a symmetric distribution. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a population with distribution function G, and let G_n denote the empirical distribution function; that is, $G_n(t) = (1/n) \sum_{i=1}^{n} I(X_i \leq t)$. Furthermore, let $\mathcal{F} = \{F_\theta: \theta \in \Theta\}$ denote a family of distributions, called the projection model, indexed by the vector-valued parameter θ. An MD estimate of θ is that value of θ_n for which the distance between G_n and F_θ is minimized. For the moment we ignore questions of the existence and uniqueness of such a value. It is not necessary that $G \in \mathcal{F}$. With a mixture of two normal components as the projection model,

$$F_\theta(x) = p\int_{-\infty}^{x} (2\pi\sigma_1^2)^{-1/2}\exp\left[-\frac{1}{2}\left(\frac{y-\mu_1}{\sigma_1}\right)^2\right] dy + (1-p)\int_{-\infty}^{x} (2\pi\sigma_2^2)^{-1/2}\exp\left[-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right] dy,$$

where $\theta' = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p)$.

Several distance measures have appeared in the literature (see Parr and Schucany 1980 for a discussion of these measures). In the mixture model setting, Choi and Bulgren (1968) and MacDonald (1971) estimated the mixture proportions (assuming the component distributions were known) by minimizing the sum-of-squares distance between the empirical and theoretical distribution functions. Quandt and Ramsey (1978) estimated the parameters in the mixture model by minimizing the sum-of-squares distance between the empirical and theoretical moment generating functions. Kumar, Nicklin, and Paulson (1979), however, showed that this technique is highly sensitive to starting values. Bryant and Paulson (1983) recently examined the empirical characteristic function in this setting. We chose to base the MDE on the Cramér–von Mises distance because Parr and Schucany (1980) found this estimator to perform well in a variety of settings. The Cramér–von Mises distance W^2 between distribution functions H_1 and H_2 is given by

$$W^2(H_1, H_2) = \int_{-\infty}^{\infty} [H_1(x) - H_2(x)]^2 dH_2(x).$$

We thus have the following definition.

Definition. A minimum Cramér–von Mises distance estimator of θ is any value $\theta_n \in \Theta$ such that

$$W^2(G_n, F_{\theta_n}) \leq \inf_{\theta \in \Theta} W^2(G_n, F_\theta) + 1/n. \qquad (2.1)$$

The additive $1/n$ term is a simple device to handle nonattainable infima. The computing formula for the Cramér–von Mises distance between G_n and F_θ in (2.1) is given by

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F_\theta(Y_i) - \frac{i - .5}{n}\right]^2,$$

where Y_i is the ith order statistic. The similarity between W_n^2 and the distance used by Choi and Bulgren (1968) should be noted.

Our implementation of the MDE employed the International Mathematical and Statistical Libraries (IMSL) subroutine ZXSSQ that uses Marquardt's (1963) method for minimizing a sum of squares. This routine was significantly faster than the quasi-Newton algorithm used by Parr and Schucany (1980). It should be noted that minimization is subject to the constraints $\sigma_1^2 \geq 0$, $\sigma_2^2 \geq 0$, and $0 \leq p \leq 1$.

3. STARTING VALUES

For either of the estimators discussed in the previous section to be used in practice, one must have starting values for the iterative procedures. Fowlkes (1979) indicated that the selection of starting values is crucial in ML estimation of the parameters of a mixture-of-normals model, and provided simulation results demonstrating the sensitivity to starting values. This contrasts with the results of Hosmer (1973) and Hasselblad (1969), which indicated that starting values are not that critical. We repeated the simulation analysis performed by Fowlkes and found that the susceptibility to starting values that he observed was evidently entirely a product of the direct maximization approach he used to calculate the MLE. The MLE (using the EM algorithm) and the MDE did not

exhibit this sensitivity for the starting value configurations used by Fowlkes.

Fowlkes (1979) suggested procedures that depend on the examination of Q–Q plots and the truncated normal. Using the truncated normal distribution was also suggested in this context by Hasselblad (1966), and cluster analysis methods have been used with some success (see Wolfe 1969). Pearson (1894) suggested method-of-moments procedures for estimating the parameters of the mixture model. We found, however, that the method-of-moments technique often produced unreasonable estimates, such as negative variances, in as many as 25% of realizations from certain mixture-of-normal models and in more than 60% of realizations from certain mixtures with nonnormal components.

We chose to obtain starting values by using an ad hoc quasi-clustering technique that is easy to implement. In the following we assume, without loss of generality, that $\mu_1 \leq \mu_2$. Under the implementation used, only the nine values .1, .2, . . . , .9 are allowed as possible values for the initial estimate of p, denoted by $\tilde{p}_0$. For each of these values of p, the sample is divided into two subsamples, $Y_1, Y_2, \ldots, Y_{n_1}$ and $Y_{n_1+1}, Y_{n_1+2}, \ldots, Y_n$, where Y_i is the ith order statistic and n_1 is np rounded to the nearest integer. $\tilde{p}_0$ is that value at which $p(1 - p)(m_1 - m_2)^2$ is maximized, where m_j is the sample median of the jth subsample. The criterion used here is a robust counterpart to the classical cluster analysis procedure of selecting the clusters so that the within-cluster sum of squares is minimized. It is easy to show that this sum of squares is minimized in the two-cluster case in which $p(1 - p)(\bar{x}_1 - \bar{x}_2)^2$ is maximized, where $\bar{x}_j$ is the sample mean of cluster j and $p = n_1/n$ with n_1 the number of sample values placed in cluster 1. Such a clustering is based on a cut-point c, below which all sample values are assigned to the cluster associated with population 1. It must be observed, however, that because of the overlap between the two mixture distributions, some sample points assigned to cluster 1 may be from population 2 and some observations from population 1 may be in cluster 2. The effect of this truncation of the right tail in population 1 is that (if σ_1^2 and σ_2^2 are approximately equal) the ML estimates based on cluster 1 are likely to underestimate both μ_1 and σ_1^2. If the overlap between the two populations is not severe, then the sample values in cluster 1 to the left of m_1 are relatively pure observations from population 1. Thus m_1 is an approximately unbiased estimate of the population mean in the case of symmetric distributions, and it is our choice for an initial estimator of μ_1. Our preliminary estimate of σ_1^2 also makes use of the pure observations to the left of m_1 and is given by

$$\tilde{\sigma}_{1(0)}^2 = ((m_1 - r_1^{(.25)})/.6745)^2,$$

where $r_j^{(q)}$ is the qth percentile from the jth cluster, $j = 1, 2$. Similarly, m_2 is used as our initial estimate of μ_2, and hence $\tilde{\sigma}_{2(0)}^2 = [(r_2^{(.75)} - m_2)/.6745]^2$.

In the next section we discuss the results of a large-scale simulation experiment comparing ML and MD estimation. In these simulations the iterative techniques were initiated by the starting values discussed in the previous paragraph. In a preliminary simulation study, we investigated the performance of these starting values by comparing the convergence initiated from the starting values with the convergence when the iterative procedures were started at the true parameter values for 100 realizations from each of the 10 parameter configurations used to generate the data in Table 1. The convergence from these two starts was almost always to the same parameter estimates, a result that held for both the MLE and MDE. This agreement held in at least 97 out of 100 realizations in 7 of the 10 configurations for both the MLE and MDE. Considering all 10, the MDE showed convergence agreement a minimum of 85 times out of 100, and the MLE had a minimum agreement of 94 out of 100. Hence this starting value procedure appears to be adequate for our purposes. Further support for this judgment appears in the next section.

4. SIMULATION RESULTS

In the previous sections we discussed MLE and MDE for the parameters of the mixture of two distributions. In this section we report the results of simulations designed to compare these two estimators when the component distributions in the simulated samples are normal and when they are not. In addition, we made our comparisons under varying degrees of separation between the two distributions. All computations were performed on the CDC 6600 computer at Southern Methodist University.

In these simulations, the mixing proportion p takes on the values .25, .50, and .75. For a given mixture, the component distributions differ from each other only by location and scale differences. In particular, $f_1(x)$ is taken to be the density associated with a random variable $X = aY$ and $f_2(x)$ is the density for $X = Y + b$, where $a > 0$, $b > 0$, and the distribution of Y is symmetric about 0. Thus a is the ratio of scale parameters for the densities f_1 and f_2, and similarly, b is the difference in location parameters. The random variable Y in our simulations was either normal, Student's t with two or four degrees of freedom, or double exponential. In our simulations we used $a = 1$ and $a = \sqrt{2}$, and b was selected to provide the desired separation between the component distributions. The number of modes of the mixture density depends to a large extent on this separation between the two component distributions. Although the mixture model has a characteristic bimodal shape for sufficient separation, Behboodian (1970) showed, for example, that a sufficient condition for the mixture density (of two normal components) to be unimodal is that $|\mu_1 - \mu_2| \leq 2\min(\sigma_1, \sigma_2)$. Of course, in this situation parameter estimation is more difficult than in the bimodal cases.

To quantify this separation between the components, we defined a measure of overlap between two distributions. Without loss of generality, we assumed that population 1 is centered to the left of population 2. We defined *overlap* as the probability of misclassification using this

rule: Classify an observation x as being from population 1 if $x < x_c$ and from population 2 if $x \geq x_c$, where x_c is the unique point between μ_1 and μ_2 such that $pf_1(x_c) = (1 - p)f_2(x_c)$. The overlaps examined in the current study are .03 and .10. Figure 1 displays the mixture densities associated with normal components and equal scale parameters. For each mixture, the scaled components $pf_1(x)$ and $(1 - p)f_2(x)$ are also shown. Note that the densities for $p = .75$ are not displayed here because when the scale parameters are equal, it follows that $f^p(x) = f^{1-p}(\mu_1 + \mu_2 - x)$, where $f^h(x)$ denotes the mixture density associated with a mixing proportion of h. Thus the shapes of the densities at $p = .75$ can be inferred from those at $p = .25$. Likewise, parameter estimation for $p = .75$ is not included in the results of the simulations when the scale parameters are equal.

For each set of parameter configurations, 500 samples of size $n = 100$ were generated from the corresponding mixture distribution. Simulations were based on the IMSL multiplicative congruential uniform random number generator GGUBS. Normal component observations were generated using the IMSL subroutine GGNPM, which uses the polar method, and $t(n)$ observations were based on the ratio of independent chi-squared and normal deviates, each obtained using IMSL routines. Double exponential components were based on $\ln(U)$, where U is uniform (0, 1), and were randomly assigned either a positive or negative sign. In all cases observations from the basic component distribution under investigation were simulated and then assigned to either component 1 or component 2, depending on whether an independent uniform (0, 1) was less than or greater than p. The observations were then scaled and shifted (with a and b) to provide observations from the appropriate component.

For each sample simulated, both the MDE and MLE were obtained. The iterative procedures discussed in Section 2 were implemented in such a way that acceptable parameter estimates were obtained for each sample. For example, if the iterative procedure failed to converge in the specified number of iterations, the last value obtained in the iteration was taken to be the estimate if this value was reasonable according to preset criteria. In general, if any of the conditions $\hat{\sigma}_1 > Y_n - Y_1$ (= sample range), $\hat{\sigma}_2 > Y_n - Y_1$, $\hat{\mu}_1 < Y_1 - (Y_n - Y_1)/10$, or $\hat{\mu}_2 > Y_n + (Y_n - Y_1)/10$ existed for an MD or ML estimate, the corresponding estimate was taken to be the starting value.

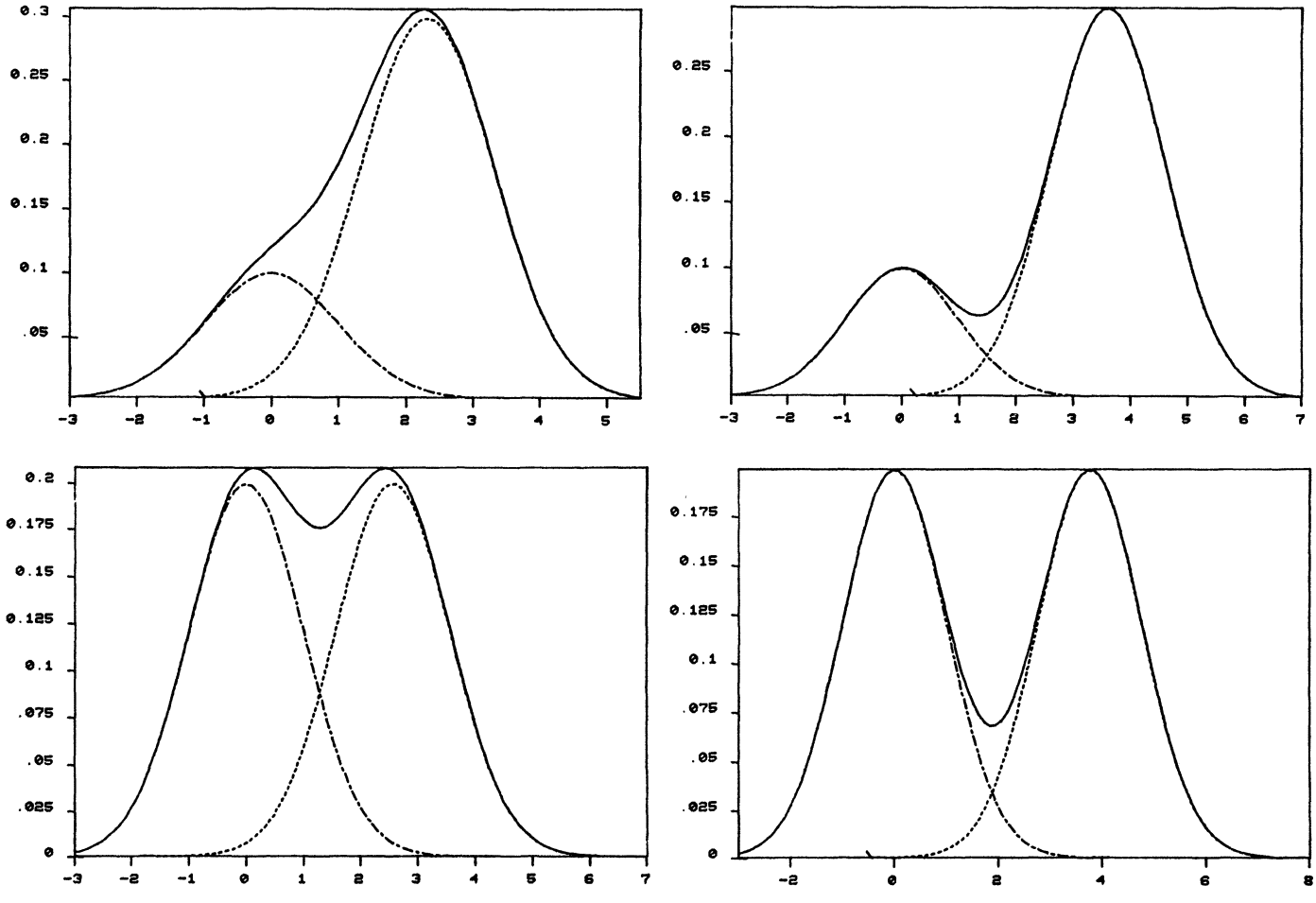

Figure 1. Mixture Densities. Graphs of densities associated with normal components and equal scale parameters are shown. For each configuration the mixture density is displayed using the solid curve, and the underlying components pf_1 and $(1 - p)f_2$ are shown using dot–dash and dotted curves, respectively. Upper left: $p = .25$; $\mu_1 = 0$; $\mu_2 = 2.32$; overlap = .10. Upper right: $p = .25$; $\mu_1 = 0$; $\mu_2 = 3.6$; overlap = .03. Lower left: $p = .5$; $\mu_1 = 0$; $\mu_2 = 2.56$; overlap = .10. Lower right: $p = .5$; $\mu_1 = 0$; $\mu_2 = 3.76$; overlap = .03.

This did not occur for most configurations in any of the 500 repetitions, but it did occur a maximum of 7 times out of 500 for MD estimates of the parameters of a mixture of $t(2)$ components. The extreme observations, which occasionally appear in samples from $t(2)$ mixtures, also forced a modification in the first step of the MLE iteration to avoid a division by zero. Although both estimation procedures provide estimates of all five of the parameters, we were most interested in the estimation of p. Results for estimating p are given in Tables 1 and 2. (In Table 3, at the end of the this section, we provide a few results for the other parameters.)

In Table 1 we present summary results of the simulations comparing the performance of the MLE and the MDE for mixtures of normal components; in Table 2 we display the results for the nonnormal components. Estimates of the bias and mean squared error (MSE) based on the simulations are given by

$$\hat{\text{bias}} = \frac{1}{n_s}\sum_{i=1}^{n_s}(\hat{p}_i - p)$$

and

$$\hat{\text{MSE}} = \frac{1}{n_s}\sum_{i=1}^{n_s}(\hat{p}_i - p)^2,$$

where n_s is the number of samples and $\hat{p}_i$ denotes an estimate of p for the ith sample. It should be noted that $n\hat{\text{MSE}}$ is the quantity actually given in the tables to correspond with the asymptotic variances σ^2 in Section 6. Since the MLE and MDE are both asymptotically normal and unbiased (which will be discussed for the MDE in Section 5), $n_s\hat{\text{MSE}}/(\sigma^2/n)$ is approximately $\chi^2(500)$. It is easy to show, then, that the approximate standard error of a tabled $n\hat{\text{MSE}}$ is (.0632) ($n\hat{\text{MSE}}$). In addition, we also provide the ratio

$$\hat{E} = \hat{\text{MSE}}\ (\text{MLE})/\hat{\text{MSE}}\ (\text{MDE}) \qquad (4.1)$$

as an empirical relative efficiency measure.

Table 1. Simulation Results for Mixtures of Normal Components

p	Ratio of Scale Factors (a)	Estimator	$\hat{\text{Bias}}$	$n\hat{\text{MSE}}$	$\hat{E}$	MDE Closer
				.10 Overlap		
.25	1	MDE	.125	7.80	.55	.38
		MLE	.052	4.26		
		Start	.084	2.06		
.50	1	MDE	.010	3.86	.83	.41
		MLE	.000	3.21		
		Start	−.005	1.22		
.25	$\sqrt{2}$	MDE	.084	5.30	.42	.32
		MLE	.002	2.25		
		Start	−.004	.894		
.50	$\sqrt{2}$	MDE	.005	2.79	.86	.43
		MLE	−.009	2.41		
		Start	−.089	1.85		
.75	$\sqrt{2}$	MDE	−.137	8.36	.58	.36
		MLE	−.086	4.87		
		Start	−.158	3.97		
				.03 Overlap		
.25	1	MDE	.026	1.09	.49	.39
		MLE	.008	.539		
		Start	.048	.782		
.50	1	MDE	.001	.420	.91	.46
		MLE	.000	.382		
		Start	.001	.634		
.25	$\sqrt{2}$	MDE	.027	.956	.51	.38
		MLE	.006	.489		
		Start	.014	.510		
.50	$\sqrt{2}$	MDE	.008	.441	.94	.45
		MLE	.009	.416		
		Start	−.048	.866		
.75	$\sqrt{2}$	MDE	−.024	1.08	.44	.42
		MLE	−.002	.470		
		Start	−.093	1.56		

NOTE: Values displayed include estimated bias ($\hat{\text{Bias}}$) and nMSE ($n\hat{\text{MSE}}$) of the ML and MD estimators and of the estimator defined as the starting point for iterations (Start). These estimates are based on 500 samples of size $n = 100$ from each mixture distribution considered. The quantity $\hat{E}$ is defined in (4.1); MDE Closer refers to the proportion of times that the MD estimator was closer to p than was the MLE.

To take advantage of the paired nature of our ML and MD estimates, we counted the proportion of samples for which $\hat{p}_D$ is closer to p than is $\hat{p}_L$, where $\hat{p}_D$ and $\hat{p}_L$ denote the MD and ML estimates, respectively. We present this proportion in the tables under the heading *MDE Closer*. It provides an estimate of $\Pr\{|\hat{p}_D - p| < |\hat{p}_L - p|\}$. The standard error of the binomial proportions shown in the tables is no greater than $[(.5)(.5)/500]^{1/2} = .022$.

Analyzing the results (and as can be seen by inspection), we find that the estimated bias and MSE associated with the MLE were generally smaller than those for the MDE when the components were actually normally distributed. This relationship between the estimators held for both overlaps. The MLE and MDE were similar at $p = .5$, but at $p = .25$ and $p = .75$ the superiority of the MLE was more pronounced.

For mixtures of nonnormal components, the relationship between MDE and MLE is reversed—the MDE generally has the smaller estimated bias and MSE, especially for $t(2)$ mixtures. The superiority of the MDE is due in part to the heavy tails in these components. The MLE often interpreted an extreme observation as being the only sample value from one of the populations with all remaining observations belonging to the other. Because of the well-known singularities associated with a zero variance estimate for a component distribution (Day 1969), we were concerned that the observed behavior of the MLE was a result of the variances' not being constrained away from zero. Simulation results in which equal variances were assumed (which removes the singularity), however, and those that used a penalized MLE (suggested by Redner 1980) were similar to those quoted here.

A surprising result here is that the starting values obtained using the procedure outlined in Section 3 produced estimators that were competitive with both the MLE and MDE. For both the normal and nonnormal mixtures, the $\hat{\text{MSE}}$'s associated with the starting values were generally

Table 2. Simulation Results for Mixtures of Nonnormal Components

p	Ratio of Scale Factors (a)	Estimator	.10 Overlap				.03 Overlap			
			$\widehat{Bias}$	$n\widehat{MSE}$	$\hat{E}$	MDE Closer	$\widehat{Bias}$	$n\widehat{MSE}$	$\hat{E}$	MDE Closer
					Double Exponential Components					
.25	1	MDE	.054	2.96	2.13	.66	.030	.545	1.18	.50
		MLE	.091	6.31			.026	.645		
		Start	.065	1.40			.078	1.04		
.50	1	MDE	.007	1.03	4.04	.69	−.001	.286	1.29	.54
		MLE	.007	4.16			−.001	.368		
		Start	−.004	1.17			.000	.414		
.25	$\sqrt{2}$	MDE	.102	4.42	1.40	.60	.035	.775	1.07	.48
		MLE	.034	6.17			.037	.832		
		Start	.011	.926			.050	.678		
.50	$\sqrt{2}$	MDE	.032	1.50	2.71	.68	.003	.259	1.44	.58
		MLE	.073	4.06			.009	.372		
		Start	−.088	1.86			−.035	.570		
.75	$\sqrt{2}$	MDE	−.037	2.20	2.94	.73	−.026	.344	.94	.44
		MLE	−.067	6.47			−.014	.323		
		Start	−.151	3.31			−.107	1.63		
					t(4) Components					
.25	1	MDE	.104	6.18	1.19	.61	.020	.466	1.89	.49
		MLE	.096	7.35			.029	.883		
		Start	.068	1.59			.072	.998		
.50	1	MDE	.004	1.82	3.07	.69	.000	.266	1.64	.53
		MLE	.015	5.59			−.005	.436		
		Start	.006	1.21			−.001	.496		
.25	$\sqrt{2}$	MDE	.098	5.20	.89	.53	.029	.605	1.61	.49
		MLE	.061	4.63			.044	.976		
		Start	−.010	.810			.036	.654		
.50	$\sqrt{2}$	MDE	.022	1.80	2.77	.67	.001	.300	1.85	.55
		MLE	.028	4.99			.010	.554		
		Start	−.072	1.52			−.046	.778		
.75	$\sqrt{2}$	MDE	−.058	3.68	2.13	.65	−.016	.361	1.57	.50
		MLE	−.076	7.84			−.012	.567		
		Start	−.137	3.07			−.108	1.75		
					t(2) Components					
.25	1	MDE	.076	3.42	4.30	.80	.024	.308	10.32	.65
		MLE	.199	14.7			.083	3.18		
		Start	.067	1.85			.096	1.37		
.50	1	MDE	−.001	1.34	9.03	.92	−.005	.264	9.24	.63
		MLE	.024	12.1			−.009	2.44		
		Start	−.004	1.39			−.002	.364		
.25	$\sqrt{2}$	MDE	.118	4.92	2.26	.69	.031	.452	7.70	.69
		MLE	.169	11.1			.106	3.48		
		Start	.006	1.18			.071	.962		
.50	$\sqrt{2}$	MDE	.016	1.52	7.76	.89	−.001	2.43	8.27	.68
		MLE	.028	11.8			.029	2.01		
		Start	−.078	2.08			−.032	.508		
.75	$\sqrt{2}$	MDE	−.059	2.99	5.79	.85	−.022	.300	11.40	.63
		MLE	−.186	17.3			−.045	3.42		
		Start	−.137	3.37			−.122	1.96		

NOTE: Values displayed include estimated bias ($\widehat{Bias}$) and nMSE ($n\widehat{MSE}$) of the ML and MD estimators and of the estimator defined as the starting point for iterations (Start). These estimates are based on 500 samples of size $n = 100$ from each mixture distribution considered. The quantity $\hat{E}$ is defined in (4.1); MDE Closer refers to the proportion of times that the MD estimator was closer to p than was the MLE.

lower than those for the MDE and MLE when overlap = .10. However, when overlap = .03, the starting value estimates were generally poorer than those for the MDE and MLE, except for the $t(2)$ mixtures, for which the MLE's were the poorest.

At this point the question arises of whether the results shown here for p are typical of those for the nuisance parameters μ_1, σ_1^2, μ_2, and σ_2^2. In Table 3 we display empirical relative efficiencies for these parameters for normal and $t(4)$ mixtures. It should be noted that only the

Table 3. Estimated Relative Efficiencies of the MDE Relative to the MLE for Mixture Model Parameters

		Normal										t(4)					
	Ratio of Scale Factors	.10					.03					10			.03		
p	(a)	μ_1	σ_1^2	μ_2	σ_2^2	p	μ_1	σ_1^2	μ_2	σ_2^2	p	μ_1	μ_2	p	μ_1	μ_2	p
.25	1	.69	.35	1.01	.65	.55	.47	.31	.83	.66	.49	3.34	48.71	1.19	2.20	1.73	1.89
.50	1	.90	.53	.94	.59	.83	.89	.64	.95	.79	.91	21.17	20.35	3.07	1.52	2.31	1.64
.25	$\sqrt{2}$	.78	.75	.90	.55	.42	.50	.33	.94	.70	.51	3.67	14.37	.89	1.77	1.13	1.61
.50	$\sqrt{2}$	.90	.63	1.12	.72	.86	.96	.70	.89	.74	.94	17.38	25.91	2.77	3.28	2.01	1.85
.75	$\sqrt{2}$	1.01	.59	.61	.44	.58	.76	.69	.35	.19	.44	26.51	17.94	2.13	1.40	13.33	1.57

NOTE: Tabled values are $\hat{E}$ (as defined in (4.1) for p) for parameters in the mixture model. Results are based on the simulation for Tables 1 and 2.

results for p, μ_1, and μ_2 are comparable when a mixture-of-normals model is used to fit data from a mixture of $t(4)$ components. From the table we see that as in Table 1, the MLE is superior for all parameters for mixtures of normal components, whereas as in Table 2, the MDE provides better results for $t(4)$ mixtures. More specifically, the efficiencies in estimating the component means are generally higher than those obtained for p, whereas the efficiencies for component variances are generally lower.

One additional point concerns the impact of sample size on the relationship between MDE and MLE. Results of similar simulations based on $n = 50$ and $n = 200$ from the 10 normal mixtures are given in Table 4 of Section 6, along with the asymptotic relative efficiencies. The empirical efficiencies at $n = 100$ are not substantially different from those at $n = 50$ or $n = 200$.

5. ASYMPTOTIC DISTRIBUTION THEORY FOR MINIMUM CRAMÉR–VON MISES DISTANCE ESTIMATION

Asymptotic theory for minimum Cramér–von Mises distance estimators for location parameters can be found in Parr and Schucany (1980), and for the general one-parameter case, in Parr and de Wet (1981). Bolthausen (1977) gave results for the mutiparameter case, but with conditions so strict as to rule out scale parameters for unbounded random variables (see his condition III). The purpose of the results in this section is to extend this previous work to cover multiparameter situations including, among others, the problem of normal mixtures.

Assume that at stage n we observe real-valued $X_1, X_2, \ldots, X_n$ iid from a distribution with cdf G and let G_n denote the usual empirical distribution function. Let $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq R^k\}$, the projection model, be a family of continuous distribution functions and assume that $G \in \mathcal{F}$; that is, $G = F_{\theta_0}$ for some $\theta_0 \in \Theta$. Furthermore, assume that there exists an open set $A \subset \Theta$ with $\theta_0 \in A$. Also consider the following continuity (C) and differentiability (D) conditions:

C. If $\theta_n \in \Theta$, $n = 1, 2, \ldots$, then

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} (F_{\theta_n}(x) - F_{\theta_0}(x))^2 dF_{\theta_0}(x) = 0$$

implies $\lim_{n\to\infty} \theta_n = \theta_0$.

D. There exists a function η: $(0, 1) \to R^k$ such that

$$\sup_{-\infty<x<\infty} | F_\theta(x) - F_{\theta_0}(x) - (\theta - \theta_0)'\eta(F_{\theta_0}(x)) | = o(\| \theta - \theta_0 \|)$$

as $\| \theta - \theta_0 \| \to 0$, where $\| \cdot \|$ is the usual Euclidean norm on R^k and $\int_0^1 \eta_i^2(u)du < \infty$ for $i = 1, 2, \ldots, k$, where $\eta'(u) = (\eta_1(u), \eta_2(u), \ldots, \eta_k(u))$.

Note 1. Condition C is satisfied if, for instance, $F_\theta(x)$ is continuous in θ at θ_0, pointwise in x (use dominated convergence). It can be interpreted as requiring that θ "continuously parameterize $\mathcal{F}$."

Note 2. If condition C is not satisfied, then $\sup_{-\infty<x<\infty} | F_\theta(x) - F_{\theta_0}(x) |$ can be arbitrarily small without having θ approach θ_0. In such a case, the search for any consistent estimator seems hopeless. In particular, in such a situation any consistent estimating functional must be discontinuous with respect to the sup-norm and hence highly nonrobust.

Note 3. Condition D is weaker than (implied by) quadratic mean differentiability of $f_\theta^{1/2}$—the canonical regularity condition for asymptotic normality of the maximum likelihood estimator (see LeCam 1970 and Pollard 1980).

Note 4. Usually,

$$\eta_i(u) = \left.\frac{\partial F_\theta(x)}{\partial \theta_i}\right|_{x = F_\theta^{-1}(u)}$$

and condition D simply states the uniform validity of the first-order Taylor approximation to $F_\theta(x)$. If $k = 1$ and θ is a location parameter, a sufficient condition to imply D is that F_θ possess a uniformly continuous density.

Before continuing, define the $k \times k$ symmetric matrices A and B with elements

$$a_{ij} = \int_0^1 \eta_i(u)\eta_j(u)du$$

$$b_{ij} = \int_0^1 \int_0^1 \{\min(u, v) - uv\}\, \eta_i(u)\eta_j(v)dudv$$

and assume A to be of full rank. We can now state and outline the proof of the following strong consistency and asymptotic normality results.

Theorem 1. Let θ_n be a minimum distance estimator of θ for all $n = 1, 2, \ldots$. Then, if condition C holds, $\theta_n \to \theta_0$ with probability one.

Proof. Clearly $\int (G_n - F_{\theta_0})^2 dF_{\theta_0} \to 0$ with probability one, and hence, also, $\inf_{\theta \in \Theta} \int (G_n - F_\theta)^2 dF_\theta \to 0$ with probability one. Now

$$\sup_\theta \left| \int (G_n - F_\theta)^2 dF_\theta - \int (F - F_\theta)^2 dF_\theta \right| \leq 4 \sup_{-\infty < t < \infty} | G_n(t) - F_{\theta_0}(t) | \to 0$$

with probability one. Hence

$$\int (F_{\theta_0} - F_{\theta_n})^2 dF_{\theta_n} = \int (F_{\theta_n} - F_{\theta_0})^2 dF_{\theta_0} \to 0$$

with probability one, and strong consistency of θ_n follows from the assumption.

Theorem 2. Assume conditions C and D and that A is of full rank. Then if $f_\theta(x)$ is continuous in θ at θ_0 for every x,

$$\sqrt{n}\,(\theta_n - \theta_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, A^{-1}BA^{-1}).$$

Proof. Set

$$K_n(\xi) = n \int (G_n - F_{\theta_0 + \xi/\sqrt{n}})^2 dF_{\theta_0 + \xi/\sqrt{n}} \quad \text{for } \xi \in R^k.$$

Then

$$\begin{aligned} K_n(\xi) &= n \int (G_n - F_{\theta_0} - (F_{\theta_0 + \xi/\sqrt{n}} - F_{\theta_0}))^2 dF_{\theta_0} \\ &\quad + n \int (G_n - F_{\theta_0} - (F_{\theta_0 + \xi/\sqrt{n}} - F_{\theta_0}))^2 \\ &\quad \times d[F_{\theta_0 + \xi/\sqrt{n}} - F_{\theta_0}] \\ &= o_p(1) + \int_0^1 (U_n(t) - \xi'\eta(t) - R_n(t))^2 dt \end{aligned}$$

uniformly in ξ for $\xi'\xi \leq C$, for any $C < \infty$, where $\sup_{0<t<1} | R_n(t) | \to 0$ with probability one, also uniformly in ξ for $\xi'\xi \leq C$. Here $U_n(t) = \sqrt{n}(G_n(F_{\theta_0}^{-1}(t)) - t)$, $0 < t < 1$. By an extension of the argument of Pyke (1970, pp. 29–30) to the present context, we obtain that the limiting law of the random variable minimizing $K_n(\xi)$ over ξ is also that of the value minimizing

$$\int_0^1 (B(t) - \xi'\eta(t))^2 dt,$$

where B is a Brownian bridge. The result then follows immediately.

It can be shown that the mixture-of-normals model satisfies the conditions of both Theorem 1 and Theorem 2.

6. ASYMPTOTIC RELATIVE EFFICIENCIES

Theorem 2 of the previous section indicates that for the mixture-of-normals model,

$$\sqrt{n}\,(\theta_n - \theta_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, A^{-1}BA^{-1}),$$

where $\theta_0' = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p)$ and θ_n is the vector of corresponding MD estimators using Cramér–von Mises distance. Likewise, it is well known that

$$\sqrt{n}\,(\hat{\theta}_L - \theta_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, I^{-1}(\theta_0)),$$

where $\hat{\theta}_L$ is the MLE of θ_0 and $I(\theta_0)$ is Fisher's information matrix. We will employ the usual terminology and refer to $A^{-1}BA^{-1}$ and $I(\theta_0)$ as asymptotic variance–covariance matrices and to their diagonal elements as asymptotic variances of the corresponding estimators. In this section we will present computed asymptotic variances for the MDE of p, which is denoted by $\hat{p}_D$, and compare these with the asymptotic variances associated with the MLE, denoted by $\hat{p}_L$.

The components of the matrix A were evaluated using the expression $\int_{-\infty}^{\infty} \xi_i(x)\xi_j(x)f_\theta(x)dx$, where $F_\theta(x)$ and $f_\theta(x)$ denote the distribution function and density function, respectively, for the mixture, θ_i is the ith component of θ, and $\xi_i(x) = \partial F_\theta(x)/\partial\theta_i$. This integral was evaluated using the IMSL subroutine DCADRE, which employs Romberg extrapolation to perform numerical integration of an integral over a finite interval. In our implementation, we used DCADRE to evaluate the integral $\int_L^U \xi_i(x)\xi_j(x)f_\theta(x)dx$, where $L = \min(-10\sigma_1 + \mu_1, -10\sigma_2 + \mu_2)$ and $U = \max(10\sigma_1 + \mu_1, 10\sigma_2 + \mu_2)$ with maximum allowable absolute error specified as 1.0×10^{-15} and relative error of 1.0×10^{-12}. The double integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{F_\theta(\min(x, y)) - F_\theta(x)F_\theta(y)\}\xi_i(x)\xi_j(y)f_\theta(x)f_\theta(y)dxdy$$

involved in calculating the elements of the matrix B is approximated by using the IMSL subroutine DBLIN to perform a Romberg integration of the integral

$$\int_L^U\int_L^U \{F_\theta(\min(x, y)) - F_\theta(x)F_\theta(y)\}\xi_i(x)\xi_j(y)f_\theta(x)f_\theta(y)dxdy$$

with maximum allowable absolute error specified as 1.0×10^{-9}.

The calculation of the information matrix for the mixture-of-normals model is discussed by Behboodian (1972). We have followed Behboodian's procedure and used Gauss–Hermite quadrature to approximate the integrals involved. Using 48-point quadrature, we obtain good agreement with Behboodian's tabled results.

In Table 4 we display asymptotic relative efficiencies

Table 4. Estimated Efficiencies for Various Sample Sizes and Asymptotic Relative Efficiencies for Estimating p in Normal Mixtures

		.10 Overlap				.03 Overlap			
	Ratio of Scale	$\hat{E}$				$\hat{E}$			
p	Factors (a)	$n = 50$	$n = 100$	$n = 200$	ARE	$n = 50$	$n = 100$	$n = 200$	ARE
.25	1	.68	.55	.43	.42	.58	.49	.46	.69
.50	1	.82	.83	.67	.65	1.03	.91	.90	.89
.25	$\sqrt{2}$	.50	.42	.38	.32	.47	.51	.43	.65
.50	$\sqrt{2}$	.96	.86	.69	.68	.93	.94	.83	.89
.75	$\sqrt{2}$	.72	.58	.45	.58	.50	.44	.43	.73

NOTE: Simulation results for $\hat{E}$ are based on 500 replications for each sample size. For each configuration and sample size, we display the value of $\hat{E}$ as defined in (4.1); under the ARE heading we display ARE = asyvar($\hat{p}_L$)/asyvar($\hat{p}_D$).

(ARE) when estimating p, calculated as ARE = asyvar $(\hat{p}_L)$/asyvar $(\hat{p}_D)$. These values are calculated for each of the parameter configurations used in Table 1 for the normal mixtures. As in Table 1, the asymptotic results indicate that the MDE compares more favorably with the MLE when $p = .5$, whereas its relative performance is not as good for $p = .25$ or $p = .75$. We have also provided simulation estimates for E obtained using $n = 50$, 100, and 200 to provide an indication of the relationship between asymptotic and finite sample properties. For most configurations, convergence to the asymptotic values is evident.

7. CONCLUDING REMARKS

We believe that the results of the preceding sections are of sufficient substance to motivate further research in the area of MD estimation in the mixture model. Our results indicate that the MDE is indeed more robust than the MLE in the sense that it is less sensitive to symmetric departures from the underlying assumption of normality of component distributions. Several topics for further research include the effects of multiple (>2) components of the mixture, skewness of the component distributions, and other possible choices of distances for minimization such as the Hellinger metric.

[*Received June 1983. Revised March 1984.*]

REFERENCES

BEHBOODIAN, J. (1970), "On the Modes of a Mixture of Two Normal Distributions," *Technometrics*, 12, 131–139.

——— (1972), "Information Matrix for a Mixture of Two Normal Distributions," *Journal of Statistical Computation and Simulation*, 1, 295–314.

BOLTHAUSEN, E. (1977), "Convergence in Distribution of Minimum-Distance Estimators," *Metrika*, 24, 215–228.

BRYANT, J.L, and PAULSON, A.S. (1983), "Estimation of Mixing Proportions via Distance Between Characteristic Functions," *Communications in Statistics*, A12, 1009–1029.

CHOI, K., and BULGREN, W.G. (1968), "An Estimation Procedure for Mixtures of Distributions," *Journal of the Royal Statistical Society*, Ser. B, 30, 444–460.

DAY, N.E. (1969), "Estimating the Components of a Mixture of Normal Distributions," *Biometrika*, 56, 463–474.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

FOWLKES, E.B. (1979), "Some Methods for Studying the Mixture of Two Normal (Lognormal) Distributions," *Journal of the American Statistical Association*, 74, 561–575.

HALL, P. (1981), "On the Nonparametric Estimation of Mixture Proportions," *Journal of the Royal Statistical Society*, Ser. B, 43, 147–156.

HASSELBLAD, V.A. (1966), "Estimation of Parameters for a Mixture of Normal Distributions," *Technometrics*, 8, 431–446.

——— (1969), "Estimation of Finite Mixtures of Distributions From the Exponential Family," *Journal of the American Statistical Association*, 64, 1459–1471.

HOSMER, D.W. (1973), "A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of Two Normal Distributions Under Three Different Types of Samples," *Biometrika*, 29, 761–770.

KUMAR, K.D., NICKLIN, E.H., and PAULSON, A.S. (1979), "Comment" on "Estimating Mixtures of Normal Distributions and Switching Distributions" by Quandt and Ramsey, *Journal of the American Statistical Association*, 74, 52–55.

LECAM, L. (1970), "On the Assumptions Used to Prove Asymptotic Normality of Maximum Likelihood Estimates," *Annals of Mathematical Satistics*, 41, 802–828.

LENNINGTON, R.K., and RASSBACH, M.E. (1979), "CLASSY—An Adaptive Maximum Likelihood Clustering Algorithm," *Proceedings of Technical Sessions, The LACIE Symposium* (Vol. 2), 671–690.

MACDONALD, P.D.M. (1971), "Comment on 'An Estimation Procedure for Mixtures of Distributions' by Choi and Bulgren," *Journal of the Royal Statistical Society*, Ser. B, 33, 326–329.

MARQUARDT, D.W. (1963), "An Algorithm for Least-Squares Estimation of Nonlinear Parameters" *Journal of the Society for Industrial Engineers*, 11, 431–441.

PARR, W.C., and DEWET, T. (1981), "On Minimum Cramér–von Mises–Norm Parameter Estimation," *Communication in Statistics*, A10, 1149–1166.

PARR, W.C., and SCHUCANY, W.R. (1980), "Minimum Distance and Robust Estimation," *Journal of the American Statistical Association*, 75, 616–624.

PEARSON, K. (1894), "Contribution to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society of London*, Ser. A, 71–110.

POLLARD, D. (1980), "The Minimum Distance Method of Testing," *Metrika*, 27, 43–70.

PYKE, R. (1970), "Asymptotic Results on Rank Statistics," in *Nonparametric Techniques in Statistical Inference*, ed. M.L. Puri, London: Cambridge University Press, 21–37.

QUANDT, R.E., and RAMSEY, J.B. (1978), "Estimating Mixtures of Normal Distributions and Switching Distributions," *Journal of the American Statistical Association*, 73, 730–738.

REDNER, R.A. (1980), "Maximum Likelihood Estimation for Mixture Models," Technical Report No. JSC-16832, National Aeronautics and Space Administration.

REDNER, R.A., and WALKER, H.F. (1982), "Mixture Densities, Maximum Likelihood, and the EM Algorithm," unpublished manuscript.

WOLFE, J.H. (1969), "Pattern Clustering by Multivariate Mixture Analysis," Research Memorandum SRM 69-17, San Diego: U.S. Naval Personnel Research Activity.

——— (1970), "Pattern Clustering by Multivariate Mixture Analysis,' *Multivariate Behavioral Research*, 5, 320–350.

WOLFOWITZ, J. (1957), "The Minimum Distance Method," *Annals of Mathematical Statistics*, 28, 75–88.