CLASSIFICATION MODELS AND PROPORTION ESTIMATION

Wayne A. Woodward and H. L. Gray
Center for Applied Mathematical and Statistical R
Southern Methodist University

1. Introduction

A common objective in remote sensing is the estimation of the proportions $p_1, p_2, \ldots, p_m$ in the mixture density

$$f(x)=p_1f_1(x)+p_2f_2(x)+\ldots+p_mf_m(x) \quad (1.1)$$

where m is the number of components(crops) in the mixture and for component i, $f_i(x)$ is a (possibly multivariate) density. In practice this density has been assumed to be (multivariate) normal with X being the reflected energy in four bands of the light spectrum, certain linear combinations of these readings, or other derived "feature" variables. Generally the parameter estimation is accomplished using maximum likelihood techniques. In this paper we examine the use of minimum distance estimation as an alternative to maximum likelihood and we will compare the performance of the two estimation techniques when dealing with mixtures of normal and of non-normal densities with varying amounts of separation. We will focus on the mixture of two univariate distributions given by

$$f(x)=pf_1(x)+(1-p)f_2(x) \quad (1.2)$$

We are also assuming 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

In this section we will assume that $f_1(x)$ and $f_2(x)$ in (1.2) are normal densities with mean and variance μ_1, σ_1^2 and μ_2, σ_2^2 respectively where it is assumed that all five parameters $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$, and p are unknown. Techniques for estimating these parameters will be discussed.

(a) Maximum Likelihood

Several recent articles have dealt with the problem of obtaining the maximum likelihood estimates of μ_1, σ_1^2, μ_2, σ_2^2, and p (Hasselblad(1966), Day(1969), Wolfe(1970), Hosmer(1975), Fowlkes(1979), Lennington and Rassbach(1979), and Redner(1980).) The MLEs are those parameter estimates which maximize the likelihood function

$$L = f(x_1)f(x_2)\ldots f(x_n) \quad (2.1)$$

or log(L), where n in (2.1) is the sample size. This maximum is usually found by setting the partial derivatives of log(L) with respect to each of the 5 parameters equal to zero and solving the resulting set of equations, called the likelihood equations. Since closed form solutions of these equations do not exist, they must be solved 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) have pointed out that this fixed point technique is essentially an application of the EM algorithm (see Dempster, Laird and Rubin(1977)) with the only difference being that using the EM algorithm, the estimates of σ_1^2 and σ_2^2 at step k involve the updated k step estimates of μ_1 and μ_2.

Fowlkes(1979), on the other hand, maximized the likelihood function directly by utilizing a quasi-Newton method for minimizing -log(L) and found that good starting values were crucial for acceptable performance. Hosmer(1975) stated that using the likelihood equations, starting values were not a serious problem in his experience. In order to determine which of the two techniques seemed preferable in our simulation studies we replicated simulations performed by Fowlkes in which various sets of poor starting values were used to initiate the minimization procedure. We simulated realizations from the mixture utilized by Fowlkes and estimated the parameters using both direct maximization and the EM algorithm. The results of our simulations indicate that the EM algorithm approach is preferable and hence we have used this technique for obtaining MLEs in our simulations.

(b) Minimum Distance

Although ML estimation procedures are known to have certain optimality properties, their sensitivity to violations of the underlying assumptions is well known. The development of estimation procedures which perform well even under moderate deviations from assumptions has been a topic of major interest in recent literature. One of these robust procedures which has received recent attention is that of minimum

distance(MD) estimation introduced by Wolfowitz(1957). Parr and Schucany(1980), for example, have shown that MD techniques provide robust estimators of the location parameter of a symmetric distribution. Minimum distance estimation has been used for parameter estimation in the mixture model by Choi and Bulgren(1968) and MacDonald(1971) with some success although, to our knowledge, the question of sensitivity to assumptions in this setting has not been addressed. These authors assume that the parameters of the component distributions are known and that only the mixing proportion(s) is to be estimated.

In order to briefly describe minimum distance estimation, we let $X_1, X_2, \ldots X_n$ denote a random sample from a population with distribution function F and let F_n denote the empirical distribution function, ie $F_n(x)=k/n$ where k is the number of observations less than or equal to x. Further, let $\mathcal{H}=\{H_\theta : \theta \in \Omega\}$ denote a family of distributions depending on the possibly vector valued parameter θ. The MD estimate of θ is that value of θ for which the distance between F_n and H_θ is minimized. Of course, when a mixture of two normals is assumed, H_θ becomes

$$H_\theta(x) = p \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + (1-p) \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2} .$$

Certain considerations become obvious at this point. First, we must define what we mean by the "distance" between two distributions. Several such distance measures have appeared in the literature. The reader is referred to the article by Parr and Schucany(1980) for a discussion of these measures. For our purposes we have chosen the Cramer-von Mises distance, W^2, between distribution functions G_1 and G_2 which is given by

$$W^2 = \int_{-\infty}^{\infty} (G_1(x)-G_2(x))^2 dG_2(x) .$$

In our setting a computing formula for the Cramer-von Mises distance between F_n and H_θ is given by

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

where Y_i is the ith order statistic. The similarity between W_n^2 and the sum of squared differences between the empirical distribution function F_n and H_θ used by Choi and Bulgren(1968) should be noted.

Another consideration involves the minimization procedure to be employed in minimzing W_n^2. Parr and Schucany used the IMSL quasi-Newton algorithm ZXMIN. Our comparisons have shown ,however, that the IMSL routine ZXSSQ which uses Marquardt's(1963) method for minimizing a sum of squares was significantly faster, usually taking no more than half the time required by ZXMIN. In the simulation studies reported in the next section we have used the Marquardt minimization procedure when calculating the MDE. It should be noted that minimization is subject to the constraints $\sigma_1^2>0$, $\sigma_2^2>0$, and $0<p<1$. Another finding which deserves mention before proceeding is that similar to the technique we have chosen for calculating the MLE, the MDE has the desirable property that it is relatively insensitive to starting values.

3. Starting Values

In order for the estimators discussed in the previous chapter to be used in practice, starting values for the iterative procedures must be provided. We have chosen to obtain starting values in this two component univariate setting using a partitioning technique which is very easy to implement. In the discussion to follow we will assume, without loss of generality, that $\mu_1 < \mu_2$. This technique involves first obtaining the initial estimate of p, denoted by p_0, and then estimating the remaining four parameters given p_0. Under the current implementation, only the 9 values .1,.2,...,.9 are allowed as possible values for p_0. For each allowable value of p_0, the sample is divided into two subsamples :

$$Y_1, Y_2, \ldots, Y_{n_1}$$

$$Y_{n_1+1}, Y_{n_1+2}, \ldots, Y_n$$

where Y_i is the ith order statistic and n_1 is np_0 rounded to the nearest integer. The value for p is that value of p for 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 for which the within cluster sum-of-squares is minimized. It is easy to show, however, that the within cluster sum-of-squares is minimized in the two cluster case when $p(1-p)(\bar{x}_1-\bar{x}_2)$ is maximized where $\bar{x}_j$ is the sample mean of cluster j and and $p=n_1/n$ with n_1 the number of sample values placed in cluster 1. Such a clustering is based upon a cut-point, c , for which all sample values below c are assigned to the cluster associated with population 1. It must be observed, however, that due

to 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 the sample mean from cluster 1 is likely to underestimate μ_1 while μ_2 is likely to be overestimated. In addition σ_1^2 and σ_2^2 are likely to be underestimated by s_1^2 and s_2^2. If we assume that the overlap between the two populations is not too severe, then the sample values in cluster 1 to the left of m_1 are relatively pure observations from population 1 in which case m_1 is a "good" estimate of the population mean in the case of symmetric distributions. This reasoning also indicates that m_1 and m_2 should provide better estimates of μ_1 and μ_2 than would $\bar{x}_1$ and $\bar{x}_2$. In order to estimate the variances of the component distributions we again will depend upon the fact that the values to the left of m_1 and to the right of m_2 are "pure" samples from populations 1 and 2 respectively. Thus, we will use only this data for estimation of the sample variances. We have used the fact that the semi-interquartile range of a standard normal distribution is .6745, to estimate σ_1^2 by

$$\sigma^2_{1(0)} = \left(\frac{m_1 - r_1^{(.25)}}{.6745}\right)^2$$

where $r_j^{(q)}$ is the qth percentile from the jth cluster, j=1,2. Similarly, $\sigma^2_{2(0)} = ((r_2^{(.75)} - m_2)/.6745)^2$.

In the next section we will discuss the results of a major simulation investigation comparing ML and MD estimation. In these simulations the iterative techniques were initiated by the starting values as discussed in the previous paragraph. A preliminary simulation to the one discussed in the next section investigated the performance of the starting values described here. In this study we compared the convergence starting at these starting values with that starting at the true parameter values. Convergence was almost always to the same parameter estimates, a result which held for both the MLE and MDE. For this reason and results to be shown in section 4, we believe this starting value procedure to be adequate.

4. Simulation Results

In the previous two sections we have discussed ML and MD estimators 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 are normal and when they are non-normal. In addition we have made our comparisons under varying degrees of separation between the two distributions. All computations were performed on the CDC 6600 at Southern Methodist University.

In our comparison of the MDE and MLE we have begun by comparing their performance when the normality assumption is valid, i.e., when the mixture distributions actually are normal. We should mention that because of the optimality properties of the MLE we would expect that the MLE would be superior in this situation. Since in practice the validity of the normality assumption is subject to question, we are also very interested in the performance of the MDE and MLE when the component distributions are not normal. To this end we have simulated mixtures in which the component distributions are distributed as a t with 4 degrees of freedom. We simulated 500 samples of size n=100 from mixtures of normal and of t(4) components for each of the following parameter configurations:

Mixing proportion
- .25
- .50
- .75

Variances

$\sigma_1^2 = \sigma_2^2$ $\quad$ $\sigma_1^2 = 2\sigma_2^2$

Since we are interested in the performance of the MDE and MLE under various levels of separation between the two component distributions it is necessary to define a measure of "overlap". Without loss of generality we assume that population 1 is centered to the left of population 2. We define "overlap" to be the probability of misclassification using the rule:

Classify an observation x as:
population 1 if $x < x_c$
population 2 if $x \geq x_c$

where x_c is the unique point between μ_1 and μ_2 such that

$$pf_1(x_c) = (1-p)f_2(x_c).$$

We have based our current study on "overlaps" of .03 and .10.

In Table 1 we display the simulation results. Although both estmation procedures provided estimates of all 5 of the parameters, only the results for the estimation of p are given here since the mixing proportion is the parameter of interest. In addition, when dealing

with the non-normal mixtures, the remaining parameter estimates often do not have a meaningful interpretation. In these simulations we have used the procedure discussed in the previous section to obtain starting values. It should be noted that although we refer to mixtures of t(4) distributions here, they are actually mixtures of disributions associated with the random variable T'=aT+b, where T has a t(4) distribution. These modifications are made in order to obtain the desired separation and variance ratios.

In Table 1 we show the estimates of the bias and MSE obtained from the simulations for purposes of comparing the performance of the MLE , MDE, and the starting value procedure. In particular, let $\hat{p}_i$ denote the estimate of p for the ith sample. Then based upon the simulations, estimates of the bias and MSE are given by:

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

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

where n_s is the number of samples. It should be noted that nMSE is the quantity actually given in the table.

An examination of the table indicates that all three techniques yielded estimates with small bias. Upon comparison of the MSEs, it can be seen, as expected, that the MLE was superior to the MDE when the components were normally distributed. This relationship between the estimators held for both overlaps. The MLE and MDE were in fact quite similar at p=.5 while for p=.25 and p=.75 the superiority of the MLE is more pronounced. A very unexpected result, however, is that the starting value routine produced estimates with lower MSE than either the MDE or MLE at .10 overlap. At .03 overlap, the starting values were generally the poorest with the exception being that their MSE was lower than that for the MDE at p=.25.

For the t(4) mixtures the relationship between MDE and MLE is reversed in that the MDE shows to perform better. In addition, this superiority of the MDE is greater at p=.5 than at the other values of p. As in the mixtures of normals, the starting values were generally the best estimates at .10 overlap and the poorest at .03 overlap. The superiority of the MDE in this case is due in part to the heavy tails in the t(4) mixture. 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. Due to the well known singularities associated with a zero variance estimate of a component distribution, Day(1969), we were concerned that the observed behavior of the MLE was due to the fact that we did not constrain the variances away from zero. However,simulation results in which equal variances were assumed (which removes the singularity) and also those which used a penalized MLE suggested by Redner(1980) were very similar to those quoted here.

5. Mixtures of Asymmetric Distributions

The simulation results of the previous chapter focus on the performance of the MLE and MDE under deviations from the assumption of normality. However, the t(4) distribution is symmetric, and recent studies have indicated that there is often a substantial asymmetry in the component distributions for variables of interest in agricultural remote sensing. A Monte Carlo examination of the performance of the MDE and MLE, assuming normal components, when in fact the component distributions were asymmetric, was performed, and the results of this examination will be discussed in this section.

For purposes of our examination, we simulated mixtures of χ^2 (9) distributions with p=.5. In these simulations the two distributions differed from each other only by a location shift. Actually the component distribution to the left is χ^2 (9) while that to the right is that of a "shifted" χ^2 (9) with origin no longer at 0. This shift was varied to provide overlaps of .01, .05, and .10. Since our estimation procedures involve a normality assumption, we used the means and variances of the two component χ^2 (9) distributions and the true mixing proportions as our starting values. The problem of obtaining starting values from the data in this case is being examined. In Table 2 we display the results of this simulation. Only when the two component distributions were widely separated (overlap=.01) do the two procedures provide reasonable results. However, when the two chi-square distributions are not widely separated, both estimators tend to seriously underestimate p. In Figure 1 we display the three mixture distributions on which these simulations were based. We see there that it is no surprise that the estimate of p is less than .5, especially for p=.10. Both estimation procedures view this as a mixture of normals, and therefore make the reasonable interpretation that the density to the left has a smaller variance and a mixing proportion less than .5. These results point out the

impact which skewed distributions can have on the proportion estimation in the mixture model when normal mixtures are assumed.

Current investigation into this area centers around modifying the estimation procedures by assuming that the underlying component distributions belong to some family of distributions whose members can be either symmetric or asymmetric depending on parameter configurations. At the present time, the Weibull distribution is being examined concerning its usefulness.

6. 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 areas for future investigation have already been identified in addition to the asymmetric components problem discussed in section 5.

First, simulations similar to the ones presented here should be performed without the assumption of only two populations in the mixture. The performance of the MDE and MLE should be compared when the number of populations is known and larger than two. In addition the applicability of the MDE to the problem of estimating the number of populations also warrants investigation. We plan to examine these possibilities.

Second, the problem of applying the MDE to the multivariate setting is of interest. Preliminary indications are that such an extension will be possible.

Third, the choice of distance measure in the MDE is a topic of interest. Our results are not meant to imply that W_2 is optimal.

Finally, the MDE and MLE must ultimately be compared on real data. Several related practical considerations have not yet been investigated. For example, when applying these estimators to LANDSAT data, the number of iterations allowed must be small due to time constraints. In the simulations described here, these constraints were not imposed and iteration was allowed to continue until convergence was obtained. The performance of the MDE and MLE under convergence restrictions should be examined.

REFERENCES

1. Choi,K. and Bulgren,W. G.(1968). "An Estimation Procedure for Mixtures of Distributions," JRSS-B 30, 444-460.

2. Cohen, A.C.(1967)."Estimation in Mixtures of Two Normal Distributions," Technometrics 9, 15-28.

3. Day, N. E.(1969). "Estimating the Components of a Mixture of Normal Distributions," Biometrika 56, 463-474.

4. Dempster, A.P., Laird, N.M., and Rubin, D.B.(1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm," JRSS-B 39, 1-38.

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

6. Fowlkes, E. B.(1979)."Some Methods for Studying the Mixture of two Normal (Lognormal) Distributions," JASA 74, 561-575.

7. Hall,P. (1981). "On the Non-parametric estimation of Mixture Proportions," JRSS-B 43 ,147-156.

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

9. 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.

10. Lennington, R. K. and Rassbach, M. E.(1979). "CLASSY-An Adaptive Maximum Likelihood Clustering Algorithm," Proc. Tech. Sessions, The LACIE Symposium, vol. II, 671-690.

11. MacDonald, P.D.M.(1971). "Comment on 'An Estimation Procedure for Mixtures of Distributions' by Choi and Bulgren," JRSS-B 33, 326-329.

12. Parr, W. C. and Schucany W. R.(1980)."Minimum Distance and Robust Estimation,"JASA 75, 616-624.

13. Pearson,K.(1894)."Contribution to the Mathematical Theory of Evolution," Phil trans A 185, 71-110.

14. Redner, R.A.(1980)."Maximum Likelihood Estimation for Mixture Models," JSC-16832, October 1980.

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

16. Wolfe, J.H.(1980). "Pattern Clustering by Multivariate Mixture Analysis," Multivariate Behavioral Research 5. 320-350.

17. Wolfowitz, J.(1957)."The Minimum Distance Method," Annals of Math. Statist. 28, 75-88.

TABLE 1

SIMULATION RESULTS

Sample size n=100, 500 Replications

	Normal						T(4)					
	Overlap = .10			Overlap = .03			Overlap = .10			Overlap = .03		
	nMSE			nMSE			nMSE			nMSE		
$\sigma_1^2 = \sigma_2^2$	MDE	MLE	Start	MDE	MLE	Start	MDE	MLE	Start	MDE	MLE	Start
P = .25	7.80	4.26	2.06	1.09	0.54	0.78	6.18	7.35	1.59	0.47	0.88	1.00
P = .50	3.86	3.21	1.22	.42	.38	0.63	1.82	5.59	1.21	0.27	0.44	0.50
$\sigma_1^2 = 2\sigma_2^2$												
P = .25	5.30	2.25	0.89	0.96	0.49	0.51	5.20	4.63	0.81	0.61	0.98	0.65
P = .50	2.79	2.41	1.85	0.44	0.42	0.87	1.80	4.49	1.52	0.30	0.55	0.78
P = .75	8.36	4.87	3.97	1.08	0.47	1.56	3.68	7.84	3.07	0.36	0.57	1.75

TABLE 2

Mixtures of $\chi^2(9)$ Components

Simulation Results:

200 repititions

n=100

p=.5

	Overlap		
	.01	.05	.10
Bias (MLE)	-.03	-.15	-.22
Bias (MDE)	-.05	-.13	-.22
nMSE(MLE)	.39	2.73	6.81
nMSE(MDE)	.47	2.27	6.59

Figure 1

Mixtures of $\chi^2(9)$ Components