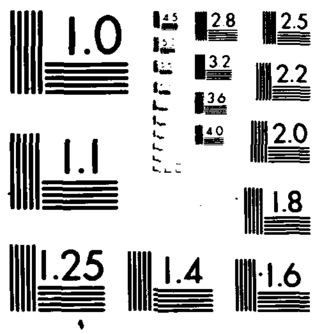
AD-A097 030 STANFORD UNIV CA DEPT OF STATISTICS F/G 12/1
UPDATING A DISCRIMINANT FUNCTION ON THE BASIS OF UNCLASSIFIED D--ETC(U)
NOV 80 G J MCLACHLAN, S GANESALINGAM N00014-75-C-0442
UNCLASSIFIED TR-47 NL

END
DATE
FILMED
4-81
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

10 **LEVEL II**

AD A097030

6

UPDATING A DISCRIMINANT FUNCTION ON THE BASIS OF UNCLASSIFIED DATA .

BY

10 G. J. McLACHLAN [redacted] S. GANESALINGAM

9 TECHNICAL REPORT NO. 47

11 NOVE[redacted]80

15

12 27 PREPARED UNDER CONTRACT N00014-75-C-0442

(NR-042-034)
OFFICE OF NAVAL RESEARCH
THEODORE W. ANDERSON, PROJECT DIRECTOR

14 TR-47

DTIC FILE COPY

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

81 3 27 163

330580

UPDATING A DISCRIMINANT FUNCTION ON THE BASIS OF UNCLASSIFIED DATA

by

G. J. McLACHLAN

and

S. GANESALINGAM

TECHNICAL REPORT NO. 47
NOVEMBER 1980

14

PREPARED UNDER CONTRACT N000-75-C-0442
(NR-042-034)
OFFICE OF NAVAL RESEARCH

Theodore W. Anderson, Project Director

Also prepared under Public Health Service Grant 5 R01 GM21215-06 and issued as Technical Report No. 62, Stanford University, Division of Biostatistics.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

UPDATING A DISCRIMINANT FUNCTION
ON THE BASIS OF UNCLASSIFIED DATA

G. J. McLachlan and S. Ganesalingam

ABSTRACT

The problem of updating a discriminant function on the basis of data of unknown origin is studied. There are observations of known origin from each of the underlying populations, and subsequently there is available a limited number of unclassified observations assumed to have been drawn from a mixture of the underlying populations. A sample discriminant function can be formed initially from the classified data. The question of whether the subsequent updating of this discriminant function on the basis of the unclassified data produces a reduction in the error rate of sufficient magnitude to warrant the computational effort is considered by carrying out a series of Monte Carlo experiments. The simulation results are contrasted with available asymptotic results.

1. Introduction

The problem of updating a discriminant function on the basis of unclassified data is considered. For simplicity it is assumed that each object belongs to one of two possible populations, say H_1 and H_2; the procedures to be discussed can be extended in a straightforward manner to cover an arbitrary number of populations. A discriminant function is to be formed for allocating an unclassified object to H_1 or H_2 on the basis of a p-dimensional feature vector, y, which can be observed on each object. The density function of y in H_i is denoted by $f_i(y)$, and π_{1y} and $\pi_{2y} = 1-\pi_{1y}$ denote the prior probabilities of y belonging to H_1 and H_2, respectively.

The optimal or Bayes rule of allocation assigns an unclassified object with observation y so as to maximize $\theta_i(y)$ over $i=1$ and 2, where

$$\theta_i(y) = \pi_{iy}\, f_i(y)/\{\pi_{1y}\, f_1(y) + \pi_{2y}\, f_2(y)\} \qquad (1.1)$$

is the posterior probability that the object belongs to H_i given y. In practice the densities are either unknown or, if their forms are known, their parameters are unknown. The estimation is usually carried out on the basis of m_i classified observations $x_{i1}, \ldots, x_{im_i}$, sampled from H_i $(i=1, 2)$. One way of proceeding is to assume some parametric form for the $f_i(y)$, such as the normal or the logistic families (Anderson, 1951 or Anderson, 1972), and estimate the unknown parameters of the particular family adopted. If this is not possible, some nonparametric approach must be used, such as kernel estimation (Remme et al., 1980, Aitchinson and Aitken, 1976, and Titterington, 1980). For the case of two populations

Greer (1979) has presented a solution to the problem of consistent nonparametric estimation of allocation rules that are best in a given class of linear rules.

2. Model

We consider the model where in addition to the $m = m_1+m_2$ classified observations there are subsequently available n unclassified observations $y_1, \ldots, y_n$. It is supposed here that they have been drawn from a mixture of H_1 and H_2 in some unknown proportions, say π_1 and $\pi_2 = 1-\pi_1$; that is, each y_i has the mixture density

$$f(y_i) = \pi_1 f_1(y_i) + \pi_2 f_2(y_i) , \qquad (i=1, \ldots, n) . \qquad (2.1)$$

This model is usually associated with two problems of somewhat different aims. With one problem the aim is to estimate the mixing proportion π_1; the classified data are assumed to have been obtained by sampling separately from H_1 and H_2, and so provide no information about π_1. This situation corresponds to a number of important problems in practice; see, for example, Hosmer (1973), Odell (1976), Odell and Basu (1976), Switzer (1979), and McLachlan (1980). The standard discriminant analysis approach is to use the classified data to form a discriminant function which can be applied to the unclassified data to obtain an estimate of π_1 given by the proportion assigned to H_1. Alternatively, if the form of the densities are known, we can apply the EM algorithm of Dempster et al. (1977) to obtain the maximum likelihood (ML) estimate of π_1 based on all the data. The latter involves more computation but is asymptotically more efficient providing regularity conditions hold. The efficiency of the former estimator of π_1 corrected for bias relative to the ML estimator has been derived asymptotically by Ganesalingam and McLachlan (1981) for two multivariate normal populations in which

$$y \sim N(\underset{\sim}{\mu}_i, \underset{\sim}{\Sigma}) \text{ in } H_i \quad (i=1, 2) . \tag{2.2}$$

They concluded that if the discriminant analysis approach gives disparate estimates of the mixing proportions, then one should proceed further and compute the ML estimates, particularly if n is large relative to m. Otherwise there may be a considerable loss in efficiency.

The other problem associated with the model (2.1) concerns the updating of the discriminant function formed initially from the classified data. Here the primary aim is not to estimate the mixing proportions, although they will have to be estimated along the way, but rather to use the unclassified data to improve the initial estimate of the densities $f_1(y)$ and $f_2(y)$ and hence the performance of the discriminant function as assessed by its overall error rate in allocating a subsequent unclassified observation. If the form of the densities is known, then the discriminant function formed initially from only the classified data can be updated using the ML estimates of the population parameters based on the combined data. Providing regularity conditions hold, there should be a reduction in the error rate, at least asymptotically, since the updated discriminant function is based on asymptotically more efficient estimates of the population parameters.

In the context of the first problem where interest is focused on the estimation of the mixing proportions, there is generally only a limited number of classified observations available, but there may be quite a large number of unclassified data. In the updating context there are also only limited classified data available, but the unclassified data may be limited too. For example, at any one time in a continuing discriminant situation, say in medical diagnosis, the n unclassified observations may consist of the data collected up to date on those objects whose true populations of origin are not known with certainty. Therefore, n may not be large, at least initially. Hence there is the question of how large n must be in order for updating to produce a reduction in the overall error rate which warrants the computational effort involved.

There would appear to be few small sample results on the possible gains from updating on the basis of n unclassified observations under the model (2.1), in particular as n varies for a given number of classified observations, m. O'Neill (1978) has studied asymptotically the performance of a discriminant function formed from classified and unclassified data combined. However, it follows from the work of Ganesalingam and McLachlan (1978, 1979a) for the cluster analysis problem ($m=0$) that the asymptotics do not always provide a reliable guide as to what happens with small sample sizes. Hence the updating problem is still essentially unresolved. Little (1978) has commented that there may be no discernible gain from updating.

In order to provide more information on the question of whether updating on the basis of a limited number of unclassified data is a worthwhile exercise, a series of simulations was performed over various combinations of the population parameters, the mixing proportions π_1 and π_2, and the sample sizes n and m. Attention is concentrated on the normality case (2.2). This is a straightforward situation to handle and, if updating does not produce any worthwhile gains in this instance then it is unlikely it will in more difficult situations where normality does not apply. Updating procedures appropriate for non-normal situations have been suggested by Murray and Titterington (1978) who expounded various approaches using distribution-free kernel methods and Anderson (1979) who gave a method for the logistic discriminant function. A Bayesian approach to the problem was considered by Titterington (1976) who also considered sequential updating.

3. Updating Procedure

Under (2.2) the rule based on (1.1) with parameters replaced by their usual estimates computed from the classified data reduces to allocating y to H_2 or H_1 according as

$$W(y) = a'y + b$$

is greater or less than the cut-off point $C = \log(\pi_{1y}/\pi_{2y})$, where

$$a = S^{-1}(\bar{x}_2-\bar{x}_1) \ ,$$

$$b = \frac{1}{2}\ (\bar{x}_1+\bar{x}_2)'\ S^{-1}(\bar{x}_1-\bar{x}_2) \ ,$$

and $\bar{x}_1$, $\bar{x}_2$, and S denote the sample means and pooled sample covariance matrix formed from the m_i classified observations from H_i $(i=1, 2)$. The vector a of discriminant coefficients is that originally obtained by Fisher (1936).

For the model (2.2) the estimates a and b can be updated on the basis of the n unclassified observations $y_1, \ldots, y_n$ by maximizing the combined likelihood

$$L = \prod_{i=1}^{2} \prod_{j=1}^{m_i} f_i(x_{ij}) \prod_{k=1}^{n} \{\pi_1 f_1(y_k) + \pi_2 f_2(y_k)\} .$$

The updated estimates, a_U and b_U, are given iteratively by

$$a_U = V^{-1}(\hat{\mu}_2-\hat{\mu}_1)/\{1-\pi_1^*\pi_2^*(\hat{\mu}_2-\hat{\mu}_1)' V^{-1}(\hat{\mu}_2-\hat{\mu}_1)\}$$

and

$$b_U = -\frac{1}{2} a_U'(\hat{\mu}_1+\hat{\mu}_2) ,$$

where

$$\hat{\pi}_i = \sum_{k=1}^{n} \hat{w}_{ik}/n , \qquad (i=1, 2) ,$$

$$\hat{w}_{1k} = 1 - \hat{w}_{2k} = \hat{\theta}_1(y_k) = 1/\{1 + \exp(a_U'y_k + b_U + \log(\hat{\pi}_2/\hat{\pi}_1)\} ,$$

$$\hat{\mu}_i = (m_i \bar{x}_i + \sum_{k=1}^{n} \hat{w}_{ik} y_k)/(m_i + n\hat{\pi}_i) , \qquad (i=1, 2) ,$$

$$\pi_i^* = (m_i + n\hat{\pi}_i)/(m+n) , \qquad (i=1, 2) ,$$

and V denotes the sample covariance matrix of the combined sample. The EM algorithm of Dempster et al. (1977) ensures the convergence of these estimates to a local maximum; see also Day (1969), O'Neill (1978), and Ganesalingam and McLachlan (1979b).

An obvious choice of starting values for a_U and b_U are the estimates based solely on the classified data, a and b. Ideally, one should try several starting points in an attempt to locate the global maximum. However, if starting the iterations with a and b does not lead to a solution which is near to the one corresponding to the global maximum, then the selection of more appropriate starting values would be a difficult exercise, particularly with high dimensional data. Therefore, if the updating procedure is to be implemented in a straightforward manner in practice, the use of a and b as starting values should lead to satisfactory estimates for the updated discriminant function coefficients. Hence in our simulations updating was performed starting with a and b always.

Frequently when no suitable estimate for π_{1y} is available, the convention $\pi_{1y} = \pi_{2y} = 0.5$ is adopted, which yields the minimax rule for $m_1=m_2$. In the updating example given in the previous section where y can be regarded as the (n+1)th unclassified observation to be recorded, $\pi_{1y} = \pi_1$ under the model (2.1), and so it can be estimated by the ML estimate of π_1 obtained during the updating process. In our simulations π_{1y} was not taken to be data dependent, but was set at a predetermined value. At least two levels of π_{1y}, including $\pi_{1y} = \pi_1$, were taken with each combination of the other parameters.

4. Relative Efficiency

Let $r(m,n)$ denote the overall unconditional error rate that the updated discriminant function, $W(y; a_U, b_U)$, misallocates the observation y with prior probabilities π_{1y} and $\pi_{2y} = 1-\pi_{1y}$ of belonging to H_1 and H_2 respectively; $r(m,0)$ and $r(m+n, 0)$ refer to the corresponding error rates for the initial discriminant function based solely on the classified data and for the discriminant function obtained if updating were performed knowing the true origin of each of the unclassified observations. For a given π_{1y}

$$\varepsilon(\pi_{1y}) = \{r(m,0) - r(m,n)\}/\{r(m,0) - r(m+n, 0)\} \qquad (4.1)$$

can be used as a measure of how efficient the updating is relat ·e to the standard procedure where the origin of each unclassified observation is known. The various unconditional error rates on the right-hand side of (4.1) can be investigated through simulation by using the sample means of their simulated conditional values which can be calculated exactly from the normal distribution.

A series of 30 trials was performed for each of the 32 different combinations of Δ, p, m, n, and π_1 considered, where $\Delta = \{(\mu_1-\mu_2)' \Sigma^{-1}(\mu_1-\mu_2)\}^{1/2}$ is the Mahalanobis distance between H_1 and H_2. On a given trial the same simulated data were used to compute the conditional error rates for different levels of π_{1y} in the cut-off point. The convenient canonical form

$$\mu_1 = -\mu_2 = (\tfrac{1}{2}\Delta, 0, \ldots, 0)' \quad \text{and} \quad \Sigma = ((\delta_{ij})) ,$$

was adopted without loss of generality. The method of Box and Muller (1958) was used to generate normal variables from uniformly distributed deviates which were produced by a multiplicative congruential generator of the form $x_{i+1} \equiv cx_i$ (modulo d), where $c = 14^{29}$ and $d = 2^{31} - 1$.

The updating problem is only of interest in those instances where the performance of the discriminant function based on the classified data is well below that of the optimal version; that is, in situations where m is not large relative to the number of dimensions p. Consequently in the simulations m was taken to be small relative to p. Various levels of n were taken for a given level of m. On all the trials the m classified observations were obtained by sampling $\frac{1}{2}$ m observations from H_1 and from H_2.

5. Simulation Results

The simulated values obtained for the relative efficiency measure (4.1) for the updating procedure are displayed in Table 1 for the various combinations of the parameters considered. All entries are expressed as percentages, and an entry for (π_1, π_{1y}) corresponds also to the case $(1-\pi_1, 1-\pi_{1y})$.

For widely separated populations such as with $\Delta=3$ the discriminant function formed initially from the classified data should be able to provide a fair degree of separation between the populations, and so the unclassified data should be able to be used quite effectively in the updating process to reduce the overall error rate. This is clearly supported by the simulation results in Table 1 which show that the reduction in error rate from updating is generally an appreciable proportion of the reduction possible where updating is performed knowing the true classification of the data. The relative efficiency is for most combinations well above 50%.

For populations which are not widely separated a discriminant function based on only a small number of classified observations is unable to provide good discrimination, and so it is of central interest to see to what extent updating on the basis of unclassified data is able to reduce its error rate. The simulation results for $\Delta=2$ in Table 1 demonstrate that in such situations some worthwhile reduction in the error rate can be achieved by updating if the unclassified data have been sampled in disparate proportions from each population. Otherwise the results suggest that if p is not very small updating would have to be performed on the basis of an extremely large number n of unclassified observations relative to p to produce any practical gain in the error rate. Indeed, for four combinations with $\Delta=2$ and $\pi_1 = 0.5$ the change in the error rate is simulated as an increase. In these instances n/p is at its lowest level (12.5) which apparently represents a situation where there are insufficient unclassified data relative to p. For higher levels of n relative to p at the same levels of the other parameters a reduction in the error rate was obtained as a result of updating.

Regarding the effect of increasing n on the results in Table 1, it can be seen that for most combinations the simulated relative efficiency of the updating procedure increases with n. On the effect of different π_{1y} for a given π_1, there is generally not an appreciable change in the relative efficiency as π_{1y} varies over 0.25 and 0.5, and also 0.75 for $\pi_1 = 0.25$ (for $\pi_1 = 0.5$ the relative efficiencies are the same at $\pi_{1y} = 0.25$ and 0.75). For most combinations the relative efficiency decreases as π_{1y} increases from 0.25 to 0.5, and increases as π_{1y} increases further to 0.75 for $\pi_1 = 0.25$.

As the aim of updating a discriminant function is to reduce its error rate, it is worth examining further those combinations in Table 1 for which an increase in the overall unconditional error rate was reported as a consequence of updating. In these cases for which $\pi_1 = 0.5$ and p is either equal to 4 or 8, the decrease in error due to updating is either so small that it is simulated as an increase or the error rate has actually increased. In order to investigate this somewhat further another 30 trials were generated for each of the relevant combinations. On this occasion positive values were obtained for the simulated relative efficincies, namely 21%, and 4% at $\pi_{1y} = 0.25$ and 0.50 respectively with $\pi_1 = 0.5$, $p = 4$, $m = 40$, $n = 50$, and 4%, and 1% at $\pi_{1y} = 0.25$ and 0.50 respectively with $\pi_1 = 0.5$, $p = 8$, $m = 40$, $n = 100$. On the basis of the combined 60 trials per combination, the change in error rate due to updating was simulated still as an increase in all but one of the four cases. However, as the differences between the expectations of the error rates are apparently not large relative to the standard errors of their simulated values, it would require an extremely large number of simulation trials in order to demonstrate with a high degree of confidence that the error rate has been increased after updating in these instances.

For the cluster analysis problem where there are no classified data, Ganesalingam and McLachlan (1979a) have reported some very encouraging results in the univariate and bivariate cases for forming a linear discriminant function which provides adequate separation even in small samples from populations close together. They noted, however, as did Day (1969), that there are problems with multiple maxima for $p \geq 3$. The results in Table 1 for $p = 4$ and 8 show that even when we have some classified data available to provide what would hopefully be reasonable starting values in the search for the global estimates, updating does not necessarily improve the performance of a linear discriminant function if the unclassified data are limited and drawn in approximately equal proportions from the respective underlying populations.

6. Asymptotic Results

It is of interest to compare the simulations of the previous section with available asymptotic results in order to assess how applicable the latter are to small sample sizes. O'Neill (1978) has considered asymptotically the relative efficiency measure,

$$\{r(m+n),\ 0) - r(\infty,0)\}/\{r(m,n) - r(\infty,0)\}\ ,$$

where $r(\infty,0)$ refers to the overall error rate of the optimal discriminant function. His underlying model also differed from the present one in that the classified data were obtained by mixture sampling in the proportions π_1 and π_2 and that π_{1y} was set equal to the updated estimate of π_1, π_1^*. These last two conditions are important from an analytical point of view as the problem can be then reparametrized in terms of a_U and $b_U^* = b_U + \log(\pi_2^*/\pi_1^*)$ without difficulty, which subsequently enables the information matrix for a_U and b_U^*, and hence the asymptotic error rates, to be derived. In a similar manner we can derive the asymptotic relative efficiency based on our measure (4.1), providing of course these two conditions are retained. The asymptotic relative efficiency so obtained should be fairly similar to that in the case of known π_{1y} equal to π_1, and in Table 2 it is contrasted with our simulated efficiencies for these combinations with $\pi_{1y} = \pi_1$.

It can be seen that there is good agreement for $p=1$; the simulated relative efficiency always exceeds the corresponding asymptotic value. However, for higher levels of p, the simulated relative efficiencies are always less than the asymptotic predictions. There is still reasonable agreement except for combinations with $\Delta=2$ and $\pi_1 = 0.5$ where the simulated relative efficiencies are appreciably below the asymptotic values.

7. Conclusions

The simulations conducted for the updating of a discriminant function by maximum likelihood on the basis of unclassified p-dimensional data drawn from a mixture of the underlying populations suggest that the error rate can be reduced by a substantial percentage for widely separated populations. In situations where the number of classified observations is small relative to p and the populations are not far apart, and so where an efficient updating of the discriminant function is most needed, the results are not so encouraging. Indeed, if the n unclassified observations have been sampled in approximately the same proportions from the populations, then there appears to be little if any gain from updating in cases with $p > 2$, say, unless n is quite large relative to p. A comparison of the simulations with available asymptotic results appropriate for a similar model suggests that the asymptotics give a reasonable guide as to what happens with finite sample sizes for univariate populations and in those instances where the multivariate populations are widely separated or are represented in the unclassified data in disparate proportions.

8. Discussion

If it is not appropriate to adopt the mixture sampling scheme (2.1) for the unclassified data, then one might consider iterati ely updating a discriminant function by applying it to the unclassified data and then recomputing the estimates of the population parameters on the basis of the combined observations with the unclassified data partitioned accordingly, and so on (McLachlan, 1975). This process may be viewed as applying the so-called classification maximum likelihood approach with starting values equal to the estimates based solely on the classified data. With this approach there is an identifying label associated with each unclassified observation, and the labels are treated as unknown parameters to be estimated; see Hartley and Rao (1968), Scott and Symons (1971), John (1970), and Sclove (1977). It is well known (Marriott, 1975 and Bryant and Williamson, 1978) that this approach does not yield consistent estimates of the population parameters. The results of McLachlan (1975, 1977) suggest that it should not be used unless one can be sure that the unclassified observations are present in approximately the same proportions from each population. Some recent Monte Carlo experiments undertaken by Ganesalingam and McLachlan (1980) in a cluster analysis context suggest that, even if the unclassified observations are obtained by sampling separately from the individual populations, maximum likelihood estimation performed on the basis of mixture sampling leads to reasonable results.

Note. This manuscript was prepared while the first author was on leave with the Department of Statistics at Stanford University.

REFERENCES

Aitchison, J. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. Biometrika 63, 413-420.

Anderson, J. A. (1972). Separate sample logistic discrimination. Biometrika 59, 19-35.

Anderson, J. A. (1979). Multivarate logistic compounds. Biometrika 66, 7-16.

Anderson, T. W. (1951). Classification by multivariate analysis. Psychometrika 16, 13-50.

Box, G. E. P. and Muller, M. E. (1958). A note on the generation of random normal deviates. Ann. Math. Statist. 29, 610-611.

Bryant, P. and Williamson, J. A. (1978). Asymptotic behaviour of classification maximum likelihood estimates. Biometrika 65, 273-281.

Day, N. E. (1969). Estimating the components of a mixture of two 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. J. Roy. Statist. Soc. B 39, 1-38.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eug. 7, 179-188.

Ganesalingam, S. and McLachlan, G. J. (1978). The efficiency of a linear discriminant function based on unclassified initial samples. Biometrika 65, 658-662.

Ganesalingam, S. and McLachlan, G. J. (1979a). Small sample results for a linear discriminant function estimated from a mixture of normal populations. J. Statist. Comput. Simul. 9, 151-158.

Ganesalingam, S. and McLachlan, G. J. (1979b). A case study of two clustering methods based on maximum likelihood. Statistica Neerlandica 33, 81-90.

Ganesalingam, S. and McLachlan, G. J. (1980). A comparison of the mixture and classification approaches to cluster analysis. Commun. Statist. - Theor. Meth. A9, 923-933.

Ganeslaingam, S. and McLachlan, G. J. (1981). Some efficiency results for the estimation of the mixing proportion in a mixture of two normal distributions. Biometrics 37 (to appear).

Greer, R. L. (1979). Consistent nonparametric estimation of best linear classification rules/solving inconsistent systems of linear inequalities. Technical Report No. 129, Department of Statistics, Stanford University.

Hartley, H. O. and Rao, J. N. K. (1968). Classification and estimation in analysis of variance problems. Rev. Inter. Statist. Inst. 36, 141-147.

Hosmer, D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics 29, 761-770.

John, S. (1970). On identifying the population of origin of each observation in a mixture of observations from two normal populations. Technometrics 12, 553-563.

Little, R. J. A. (1978). Consistent regression methods for discriminant analysis with incomplete data. J. Amer. Statist. Assoc. 73, 319-322.

Marriott, F. H. C. (1975). Separating mixtures of normal distributions. Biometrics 31, 767-769.

McLachlan, G. J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation. J. Amer. Statist. Assoc. 70, 365-369.

McLachlan, G. J. (1977). Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. J. Amer. Statist. Assoc. 72, 403-406.

McLachlan, G. J. (1980). Estimation of mixing proportions by the EM algorithm. Proceedings of the Statistical Computing Section of the Annual Meeting of the American Statistical Association, Houston.

Murray, G. D. and Titterington, D. M. (1978). Estimation problem with data from a mixture. Appl. Statist. 27, 325-334.

Odell, P. L. (1976). Current and past roles of the statistician in space applications. Commun. Statist. - Theor. Meth. A5, 1077-1089.

Odell, P. L. and Basu, J. P. (1976). Concerning several methods for estimating crop averages using remote sensing data. Commun. Statist. - Theor. Meth. A5, 1091-1114.

O'Neill, T. J. (1978). Normal discrimination with unclassified observations. J. Amer. Statist. Assoc. 73, 821-826.

Remme, J., Habbema, J. D. F., and Hermans, J. (1980). A simulative comparison of linear, quadratic and kernel discrimination. J. Statist. Comput. Simul. 11 (to appear).

Sclove, S. L. (1977). Population mixture models and clustering algorithms. Commun. Statist. - Theor. Meth. A6, 417-434.

Scott, A. J. and Symons, M. L. (1971). Clustering methods based on likelihood ratio criterion. Biometrics 27, 387-397.

Switzer, P. (1979). Extensions of linear discriminant analysis for statistical classification of remotely sensed imagery. Technical Report No. 30, Department of Statistics, Stanford University.

Titterington, D. M. (1976). Updating a diagnostic system using unconfirmed cases. Appl. Statist. 25, 238-247.

Titterington, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. Technometrics 22, 259-268.

TABLE 1

Simulated Relative Efficiency as a Percentage of Updating Procedure for Unclassified Data in Unknown Proportions π_1 and $1-\pi_1$ and Defined in Terms of Overall Error Rate for a Subsequent Observation with Known Priors π_{1y} and $1-\pi_{1y}$

			π_1									
			0.25						0.5			
			π_{1y}									
			0.25		0.5		0.75		0.25		0.5	
			Δ									
p	m	n	2	3	2	3	2	3	2	3	2	3
1	20	50	17	73	31	68	33	71	37	85	34	74
		100	37	89	38	84	53	86	64	90	59	78
		200	74	93	60	87	49	88	86	86	77	87
4	40	50	29	41	30	38	38	40	-3	58	-25	57
		100	61	60	54	58	52	63	13	67	11	66
		200	55	76	49	72	48	73	24	70	15	70
8	40	100	43	54	35	52	41	55	-12	54	-16	51
		200	52	78	42	76	47	80	22	76	9	73

TABLE 2

Simulated Relative Efficiency of Updating Procedure for $\pi_{1y} = \pi_1$ Versus Asymptotic Relative Efficiency for $\pi_{1y} = \pi_1^*$ (in Parentheses)

			π_1			
			0.25		0.5	
			Δ			
p	m	n	2	3	2	3
1	20	50	31	68	34	74
			(23)	(62)	(28)	(67)
		100	38	84	59	78
			(33)	(74)	(40)	(77)
		200	60	87	77	87
			(48)	(84)	(55)	(86)
4	40	50	30	38	-25	57
			(33)	(68)	(28)	(66)
		100	54	58	11	66
			(45)	(78)	(40)	(77)
		200	49	72	15	70
			(59)	(87)	(55)	(86)
8	40	100	35	52	-16	51
			(48)	(79)	(40)	(77)
		200	42	76	9	73
			(62)	(87)	(55)	(86)

TECHNICAL REPORTS

OFFICE OF NAVAL RESEARCH CONTRACT N00014-67-A-0112-0030 (NR-042-034)

1. "Confidence Limits for the Expected Value of an Arbitrary Bounded Random Variable with a Continuous Distribution Function," T. W. Anderson, October 1, 1969.

2. "Efficient Estimation of Regression Coefficients in Time Series," T. W. Anderson, October 1, 1970.

3. "Determining the Appropriate Sample Size for Confidence Limits for a Proportion," T. W. Anderson and H. Burstein, October 15, 1970.

4. "Some General Results on Time-Ordered Classification," D. V. Hinkley, July 30, 1971.

5. "Tests for Randomness of Directions against Equatorial and Bimodal Alternatives," T. W. Anderson and M. A. Stephens, August 30, 1971.

6. "Estimation of Covariance Matrices with Linear Structure and Moving Average Processes of Finite Order," T. W. Anderson, October 29, 1971.

7. "The Stationarity of an Estimated Autoregressive Process," T. W. Anderson, November 15, 1971.

8. "On the Inverse of Some Covariance Matrices of Toeplitz Type," Raul Pedro Mentz, July 12, 1972.

9. "An Asymptotic Expansion of the Distribution of "Studentized" Classification Statistics," T. W. Anderson, September 10, 1972.

10. "Asymptotic Evaluation of the Probabilities of Misclassification by Linear Discriminant Functions," T. W. Anderson, September 28, 1972.

11. "Population Mixing Models and Clustering Algorithms," Stanley L. Sclove, February 1, 1973.

12. "Asymptotic Properties and Computation of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance," John James Miller, November 21, 1973.

13. "Maximum Likelihood Estimation in the Birth-and-Death Process," Niels Keiding, November 28, 1973.

14. "Random Orthogonal Set Functions and Stochastic Models for the Gravity Potential of the Earth," Steffen L. Lauritzen, December 27, 1973.

15. "Maximum Likelihood Estimation of Parameters of an Autoregressive Process with Moving Average Residuals and Other Covariance Matrices with Linear Structure," T. W. Anderson, December, 1973.

16. "Note on a Case-Study in Box-Jenkins Seasonal Forecasting of Time series," Steffen L. Lauritzen, April, 1974.

17. "General Exponential Models for Discrete Observations," Steffen L. Lauritzen, May, 1974.

18. "On the Interrelationships among Sufficiency, Total Sufficiency and Some Related Concepts," Steffen L. Lauritzen, June, 1974.

19. "Statistical Inference for Multiply Truncated Power Series Distributions," T. Cacoullos, September 30, 1974.

Office of Naval Research Contract N00014-75-C-0442 (NR-042-034)

20. "Estimation by Maximum Likelihood in Autoregressive Moving Average Models in the Time and Frequency Domains," T. W. Anderson, June 1975.

21. "Asymptotic Properties of Some Estimators in Moving Average Models," Raul Pedro Mentz, September 8, 1975.

22. "On a Spectral Estimate Obtained by an Autoregressive Model Fitting," Mituaki Huzii, February 1976.

23. "Estimating Means when Some Observations are Classified by Linear Discriminant Function," Chien-Pai Han, April 1976.

24. "Panels and Time Series Analysis: Markov Chains and Autoregressive Processes," T. W. Anderson, July 1976.

25. "Repeated Measurements on Autoregressive Processes," T. W. Anderson, September 1976.

26. "The Recurrence Classification of Risk and Storage Processes," J. Michael Harrison and Sidney I. Resnick, September 1976.

27. "The Generalized Variance of a Stationary Autoregressive Process," T. W. Anderson and Raul P.Mentz, October 1976.

28. "Estimation of the Parameters of Finite Location and Scale Mixtures," Javad Behboodian, October 1976.

29. "Identification of Parameters by the Distribution of a Maximum Random Variable," T. W. Anderson and S.G. Ghurye, November 1976.

30. "Discrimination Between Stationary Guassian Processes, Large Sample Results," Will Gersch, January 1977.

31. "Principal Components in the Nonnormal Case: The Test for Sphericity," Christine M. Waternaux, October 1977.

32. "Nonnegative Definiteness of the Estimated Dispersion Matrix in a Multivariate Linear Model," F. Pukelsheim and George P.H. Styan, May 1978.

TECHNICAL REPORTS (continued)

33. "Canonical Correlations with Respect to a Complex Structure," Steen A. Andersson, July 1978.

34. "An Extremal Problem for Positive Definite Matrices," T.W. Anderson and I. Olkin, July 1978.

35. " Maximum likelihood Estimation for Vector Autoregressive Moving Average Models," T. W. Anderson, July 1978.

36. "Maximum likelihood Estimation of the Covariances of the Vector Moving Average Models in the Time and Frequency Domains," F. Ahrabi, August 1978.

37. "Efficient Estimation of a Model with an Autoregressive Signal with White Noise," Y. Hosoya, March 1979.

38. "Maximum Likelihood Estimation of the Parameters of a Multivariate Normal Distribution,"T.W. Anderson and I. Olkin, July 1979.

39. "Maximum Likelihood Estimation of the Autoregressive Coefficients and Moving Average Covariances of Vector Autoregressive Moving Average Models," Fereydoon Ahrabi, August 1979.

40. "Smoothness Priors and the Distributed Lag Estimator," Hirotugu Akaike, August, 1979.

41. "Approximating Conditional Moments of the Multivariate Normal Distribution," Joseph G. Deken, December 1979.

42. "Methods and Applications of Time Series Analysis - Part I: Regression, Trends, Smoothing, and Differencing," T.W. Anderson and N.D. Singpurwalla, July 1980.

43. "Cochran's Theorem, Rank Additivity, and Tripotent Matrices." T.W. Anderson and George P.H. Styan, August, 1980.

44. "On Generalizations of Cochran's Theorem and Projection Matrices," Akimichi Takemura, August, 1980.

45. "Existence of Maximum Likelihood Estimators in Autoregressive and Moving Average Models," T.W. Anderson and Raúl P. Mentz, Oct. 1980.

46. "Generalized Correlations in the Singular Case," Ashis Sen Gupta, November 1980.

47 "Updating a Discriminant Function on the Basis of Unclassified Data," G.J. McLachlan and S. Ganesalingam, November 1980.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 47	2. GOVT ACCESSION NO. AD-A097030	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) UPDATING A DISCRIMINANT FUNCTION ON THE BASIS OF UNCLASSIFIED DATA		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) G. J. McLACHLAN and S. GANESALINGAM		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0442
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, California		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (NR-042-034)
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics and Probability Program Code 436 Arlington, Virginia 22217		12. REPORT DATE NOVEMBER 1980
		13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Also prepared under Public Health Service Grant 5 R01 GM21215-06 and issued as Technical Report no. 62, Stanford University, Division of Biostatistics.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Sample linear discriminant function; data from a mixture; updating; EM algorithm; simulated error rates.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

SEE REVERSE SIDE.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20.

ABSTRACT

The problem of updating a discriminant function on the basis of data of unknown origin is studied. There are observations of known origin from each of the underlying populations, and subsequently there is available a limited number of unclassified observations assumed to have been drawn from a mixture of the underlying populations. A sample discriminant function can be formed initially from the classified data. The question of whether the subsequent updating of this discriminant function on the basis of the unclassified data produces a reduction in the error rate of sufficient magnitude to warrant the computational effort is considered by carrying out a series of Monte Carlo experiments. The simulation results are contrasted with available asymptotic results.

47/62

LMED

8