Biometrika (1979), 66, 1, *pp.* 17–26
Printed in Great Britain

Multivariate logistic compounds

BY J. A. ANDERSON
Department of Statistics, University of Newcastle upon Tyne

SUMMARY

A general method for estimating the proportions of a finite compound distribution is suggested, based on direct estimation of the likelihood ratio. The only distributional assumption is that the log likelihood ratio is linear in the observations and hence the method is robust and applicable to a number of families of distributions. Sample points are required from each component distribution and from the compound. In particular, the method handles continuous and discrete data with equal facility. Only a moderate number of parameters need estimation so that problems in quite high dimensions can be solved. The method can be used to update logistic discriminant functions using data points whose parent populations are unknown.

Some key words: Compound distribution; Discriminant update; Likelihood ratio; Logistic function; Maximum likelihood estimation; Quasi-Newton optimization.

1. INTRODUCTION

The resolution of a mixture of distributions into its separate components on the basis of sample data is a classical statistical problem, an early contribution being Pearson (1894).

Thus suppose that sample points $x^{\mathrm{T}} = (x_1, \ldots, x_p)$ are available from some or all of the distributions with likelihoods $f_s(x)$ $(s = 1, \ldots, k)$ and

$$g(x) = \sum_{s=1}^{k} \theta_s f_s(x),$$

where $\theta_s \geqslant 0$ $(s = 1, \ldots, k)$ and $\theta_1 + \ldots + \theta_k = 1$. In what follows $g(x)$ will be called a 'compound' distribution and the term 'mixture' will be reserved to distinguish between mixture and separate sampling in later sections.

In the above inference problem Hosmer (1973) identifies three distinct cases in which sample points are available respectively (i) only from the compound distribution $g(x)$, (ii) from $g(x)$ and some but not all the $\{f_s(x)\}$, and (iii) from $g(x)$ and all the $\{f_s(x)\}$.

Case (i) may be regarded as a special form of cluster analysis; solutions are possible only if the distributions $\{f_s(x)\}$ are specified rather closely and even then maximum likelihood methods for estimation lead to numerical optimization over many parameters. Day (1969) and Wolfe (1970) considered problems of this type with underlying multivariate normal distributions. In case (ii) structural assumptions about the $\{f_s(x)\}$ are also necessary; a medical example is given by Emery & Carpenter (1974) who estimate the proportion of cot deaths due to a specific cause. The present paper is concerned with case (iii) in which interest again focuses on the estimation of proportions from various sources. For example, Hosmer (1973) estimated the proportion of male and female fish in a population of halibut from univariate data. It should be noted also that case (iii) is equivalent to the situation in discriminant analysis using information from sample points not yet identified definitively. This will be discussed briefly further in § 6.

Most approaches to case (iii) for multivariate data make stringent assumptions about the component distributions. This means that maximum likelihood estimation can be used, giving asymptotic efficiency but at the cost of heavy computing. For example, compounds of multivariate normal distributions have been postulated (Odell & Basu, 1976; Tubbs & Coberley, 1976; Peters & Coberley, 1976). Maximum likelihood estimation here requires an iterative procedure for the very large number, $(k+1)p-1+\frac{1}{2}p(p+1)$, of parameters for the simplest case, that of equal dispersion matrices. Simpler, less efficient methods of estimation are available for these problems. One general method is to estimate $f_s(x)$ or its parameters from the sample from H_s and to take these as given when estimating the $\{\theta_s\}$ from the sample points from H_{k+1}. The information in this last sample about the $\{f_s(x)\}$ is thus not used, reducing the efficiency. All the above techniques lack robustness; the estimates are rather sensitive to the distributional forms postulated.

If the data are univariate, the above methods may be used or the distribution-free approach of Bartlett & Macdonald (1968) is available. These authors estimated the $\{\theta_s\}$ from the compound sample empirical density function, using empirical estimates of the $f_s(x)$ from the sample points just from H_s. Clearly this approach lacks efficiency and is probably best regarded as a method for very large samples. In principle the extension to multivariate data is obvious but the loss of efficiency is probably even more serious.

The logistic compound technique introduced here is based on a direct parameterization and estimation of the likelihood ratio. In this paper the only distributional assumption is that $\log\{f_s(x)/f_t(x)\}$ is linear in x so that the method is applicable to a wide class of continuous and discrete random variables. Less restrictive assumptions are discussed in § 7. In § 3 the parameters of the likelihood ratios are estimated by maximizing the likelihood so that there is no loss of asymptotic efficiency unless the functional forms of the distributions $\{f_s(x)\}$ are known *a priori*. In this case the full likelihood approach should be used, provided that this is computationally feasible. The logistic compound approach involves an iterative procedure to maximize the likelihood with respect to $(k-1)(p+1)+p-1$ parameters. By comparison, the iteration in the multivariate normal formulation above involves $O(p^2)$ parameters. Hence the logistic approach may be indicated on grounds of computational feasibility. An example of the use of the method is given in § 4. The technique introduced here is a natural extension of the method of logistic discrimination (Anderson, 1972; Cox, 1966; Day & Kerridge, 1967) and may be used to update the estimates of the discriminant functions using data whose origin is unknown, as described in § 6.

The method of this paper can be applied with equal facility to compounds of contingency tables, compounds of multivariate continuous data and compounds where some of the data is discrete and some continuous.

2. LOGISTIC COMPOUNDS OF TWO DISTRIBUTIONS

Attention is directed here to the case in which sample points are available from the compound distribution $g(x)$ and each of the separate component distributions $\{f_s(x)\}$ for $s = 1, \ldots, k$. For simplicity it will be assumed temporarily that $k = 2$. The ideas carry over to general k, as outlined in § 5.

Suppose then that sample points are available separately from populations H_1, H_2 and H_3 with likelihoods $f_1(x)$, $f_2(x)$ and $\theta_1 f_1(x)+\theta_2 f_2(x)$ respectively at the point x, $\theta_1+\theta_2 = 1$. Further, suppose that at x a total of $n(x)$ points are observed, with $n_s(x)$ from H_s $(s = 1, 2, 3)$. The sample size from H_s is $n_s = \Sigma n_s(x)$, giving total sample size $n = n_1+n_2+n_3$. The

likelihood is

$$L = \Pi\{f_1(x)\}^{n_1(x)}\{f_2(x)\}^{n_2(x)}\{\theta_1 f_1(x) + \theta_2 f_2(x)\}^{n_3(x)}. \tag{1}$$

Suppose for the moment that the functional form $f_s(.)$ is known and contains parameters w_s $(s = 1, 2)$. Maximum likelihood estimators of w_1, w_2 and θ_1 may be obtained in the obvious way by maximizing (1). There are two reasons for not proceeding in this way. The first, as noted in the previous section, is computational. Secondly, in many practical situations the functional forms $f_s(.)$ may not be known sufficiently well to give any confidence in the procedure. A thesis of this paper is that there are many situations where there is more confidence in the functional form of likelihood ratio $r(x) = f_1(x)/f_2(x)$. For example, the likelihood ratio is invariant under various forms of selection bias, including truncation, whereas the $\{f_s(.)\}$ may be modified quite severely. This has been noted by Anderson (1972), Mantel (1973) and Dawid (1976), the latter in the context of medical diagnostic distributions. It therefore seems sensible to consider statistical inference based on $r(.)$ rather than the $\{f_s(.)\}$. An added advantage of $r(.)$ is that it has a particularly simple form in some commonly occurring situations, which reduces the computational load. Thus $\log r(x)$ is linear in x for many families of distributions and perturbations of them, including the multivariate normal with equal dispersion matrices and the binary log linear model with equal second- and higher-order interactions. Anderson (1972) discusses these families in more detail.

We therefore suppose that

$$r(x) = \exp(\beta^{\mathrm{T}} x),$$

where $\beta^{\mathrm{T}} = (\beta_0, \beta_1, \ldots, \beta_p)$ and $x^{\mathrm{T}} = (1, x_1, \ldots, x_p)$. Then in (1) we may substitute

$$f_1(x) = \exp(\beta^{\mathrm{T}} x) f_2(x), \quad f_3(x) = \{\theta_1 \exp(\beta^{\mathrm{T}} x) + \theta_2\} f_2(x)$$

to obtain the likelihood of θ_1, β and $f_2(.)$ as

$$L = \Pi\{\exp(\beta^{\mathrm{T}} x)^{n_1(x)}\{\theta_1 \exp(\beta^{\mathrm{T}} x) + \theta_2\}^{n_3(x)}\{f_2(x)\}^{n(x)}. \tag{2}$$

Note that there are two constraints on L since $f_1(x)$ and $f_2(x)$ must sum or integrate to unity. The only constraint needed for $f_3(x)$ is then $\theta_1 + \theta_2 = 1$. The discrete sample case will be considered first.

3. Maximum likelihood estimation

3·1. *Discrete case*

The constraints may be written

$$c_1 = \Sigma_x f_1(x) - 1 = \Sigma_x \exp(\beta^{\mathrm{T}} x) f_2(x) - 1 = 0, \tag{3}$$

$$c_2 = \Sigma_x f_2(x) - 1 = 0. \tag{4}$$

No assumptions will be made about the functional form of $f_2(x)$, whose values as probabilities will be estimated at each point x of the sample space.

The likelihood (2) is now maximized using Lagrange's method of undetermined multipliers, which gives, with $\theta_2 = 1 - \theta_1$, the equations

$$\partial \log L/\partial \beta_j + \lambda\, \partial c_1/\partial \beta_j + \mu\, \partial c_2/\partial \beta_j = 0 \quad (j = 0, \ldots, p),$$

$$\partial \log L/\partial \theta_1 + \lambda\, \partial c_1/\partial \theta_1 + \mu\, \partial c_2/\partial \theta_1 = 0,$$

$$\partial \log L/\partial f_2(x) + \lambda\, \partial c_1/\partial f_2(x) + \mu\, \partial c_2/\partial f_2(x) = 0,$$

the last for all x. Applying these with (2), (3) and (4), we have that

$$\Sigma\,[n_1(x)+n_3(x)\,\theta_1\exp(\beta^{\mathrm{T}}x)/\{\theta_1\exp(\beta^{\mathrm{T}}x)+\theta_2\}+\lambda\exp(\beta^{\mathrm{T}}x)\,f_2(x)]\,x_j=0 \quad (j=0,\ldots,p), \tag{5}$$

$$\Sigma\,n_3(x)\,\{\exp(\beta^{\mathrm{T}}x)-1\}/\{\theta_1\exp(\beta^{\mathrm{T}}x)+\theta_2\}=0, \tag{6}$$

$$n(x)/f_2(x)+\lambda\exp(\beta^{\mathrm{T}}x)+\mu=0, \tag{7}$$

the last for all x. To solve these equations, note first that (7) multiplied by $f_2(x)$ and summed over x gives

$$n+\lambda+\mu=0. \tag{8}$$

Next, if we write

$$S(x)=\theta_1\exp(\beta^{\mathrm{T}}x)/\{\theta_1\exp(\beta^{\mathrm{T}}x)+\theta_2\},$$

then some rearrangement of (6) leads to $\Sigma n_3(x)\,S(x)=\theta_1 n_3$. Substitution into (5) with $j=0$ then gives, since $x_0=1$,

$$\lambda=-(n_1+\theta_1 n_3)=-n_1^*,$$

and (8) gives

$$\mu=-(n_2+\theta_2 n_3)=-n_2^*.$$

Substitution of λ and μ into (7) gives

$$f_2(x)=\frac{n(x)}{n_1^*\exp(\beta^{\mathrm{T}}x)+n_2^*}. \tag{9}$$

Finally, substitution for $f_2(x)$ and λ into (5) gives for $j=0,\ldots,p$

$$\Sigma\left\{n_1(x)+n_3(x)\frac{\theta_1\exp(\beta^{\mathrm{T}}x)}{\theta_1\exp(\beta^{\mathrm{T}}x)+\theta_2}-n(x)\frac{n_1^*\exp(\beta^{\mathrm{T}}x)}{n_1^*\exp(\beta^{\mathrm{T}}x)+n_2^*}\right\}x_j=0. \tag{10}$$

Equations (6) and (10) contain the $p+2$ parameters θ_1 and β only, and hence could be solved by standard iterative methods, for example, the Newton–Raphson technique.

It is worth noting that (10) is directly analogous to that for estimating β using data only from H_1 and H_2 (Anderson, 1972, equation (25)) in the following sense. In (10) the proportion of $n_3(x)$ 'from' H_1 is added to $n_1(x)$ and n_1^* and n_2^* represent the 'true number' of points from H_1 and H_2, respectively.

Now let

$$P(x)=\frac{n_1^*\exp(\beta^{\mathrm{T}}x)}{n_1^*\exp(\beta^{\mathrm{T}}x)+n_2^*}, \quad Q(x)=1-P(x). \tag{11}$$

Then substituting for $f_2(x)$ from (9) into the likelihood (2), we have that

$$L_{\max}=L^*\,\Pi\{n(x)\}^{n(x)},$$

where

$$L^*=\Pi\left\{\frac{1}{n_1^*}P(x)\right\}^{n_1(x)}\left\{\frac{1}{n_2^*}Q(x)\right\}^{n_2(x)}\left\{\frac{\theta_1}{n_1^*}P(x)+\frac{\theta_2}{n_2^*}Q(x)\right\}^{n_3(x)} \tag{12}$$

The maximum likelihood estimates of θ_1 and β may be obtained by maximizing L^* in (12), rather than directly solving the equations (6) and (10), for example using quasi-Newton methods (Fletcher & Powell, 1963) based on direct search and gradients. The gradients, or derivatives, of $\log L^*$ with respect to β are given by the left-hand side of (10).

The asymptotic dispersion matrix of the estimates may in principle be obtained using the method of constrained maximum likelihood estimation (Aitchison & Silvey, 1958). Anderson (1972) used this technique for a similar problem. Some technical difficulties have been found but a solution is conjectured in § 4·2.

The functional form of the likelihood L^* is interesting because it clearly demonstrates the dependence of the solution to the estimation problem on logistic functions and a compound logistic function.

3·2. *Continuous and polychotomous observations*

Although it was assumed above that x was discrete, the method does extend to the continuous case with some loss of efficiency. The simplest approach would be to subdivide the range of each continuous variate to make it discrete. Subject to the proviso about polychotomous variates later in this section, the method of § 3·1 would then be applicable with some loss of information due to the range subdivision.

Alternatively, a full maximum likelihood approach for continuous x would involve the simultaneous estimation of β, θ_1 and a functional form for $f_2(.)$. Good & Gaskins (1971) discuss the estimation of one likelihood function without the complicating interrelations between $f_1(.)$, $f_2(.)$ and their compound considered here. They found that a straightforward approach to maximizing the likelihood function was unbounded and involved Dirac delta functions. They showed how to avoid this by constraining the smoothness of the likelihood function. Unfortunately the problem of maximizing the likelihood L of (2) with respect to β, θ_1 and $f_2(.)$ with smoothness constraints on the latter has not yet been solved. When it is, the estimates for β and θ_1 may be different from those given in § 3·1.

In the meantime, a less efficient estimator of $f_2(.)$ is available, that given in (9) where now x ranges over a continuous space. This is neither the smoothness constrained nor the unconstrained maximum likelihood estimator, nor indeed is it smooth. However, these disadvantages are of decreasing importance as the sample size increases and this estimator of $f_2(.)$ does lead to the same estimation system for θ_1 and β, maximizing L^*, for continuous variates as for discrete variates. In this sense the estimation method given in § 3·1 is appropriate for continuous or discrete variates, with a slight loss of efficiency for the former. We recommend this approach until better methods are available.

It should be remembered that, although the results in § 3·1 are good for any discrete observations, the requirement in § 3 that the log likelihood ratios are linear in x is quite severe for all but binary variates. Thus consider the variable x_1 which can take the values 0, 1 or 2, possibly coded. In some cases it will be reasonable to assume that the log likelihood ratio is linear in x_1; in other cases, this will not be so and x_1 must be transformed into two binary variables to apply the results of this paper. Anderson (1972) considers these points in greater detail.

4. Maximum likelihood estimation

4·1. *Practical considerations*

From the results of the previous section, attention may now be concentrated on the unconstrained problem of maximizing $\log L^*$ in (12). There are many numerical algorithms available for maximizing functions over several variables. A selection of these are given in the N.A.G. Library Manual (1975). Since the gradients are available here in (6) and (10) we have chosen a quasi-Newton method, combining some of the advantages of direct search and the speed of convergence of the Newton–Raphson method near the optimum. Specifically the N.A.G. subroutine E04DDF (Gill & Murray, 1972) was used. This routine does not permit the range of the variables to be bounded, so that $\theta_1 \in [0, 1]$ was transformed such that

$$\theta_1 = e^{\phi}/(e^{\phi}+1). \tag{13}$$

The optimization of $\log L^*$ was then carried out with respect to the $\{\beta_j\}$ and ψ.

The only element of the gradient vector of $\log L^*$ to be changed is

$$\frac{\partial \log L^*}{\partial \psi} = \frac{\partial \log L^*}{\partial \theta_1}\frac{d\theta_1}{d\psi} = \frac{\partial \log L^*}{\partial \theta_1}\theta_1(1-\theta_1).$$

A Fortran program has been written for the Newcastle University IBM Computer Installation which carries out this optimization procedure using the N.A.G. routine E04DDF. Copies of the program are available from the author.

Starting values of zero have been found to be satisfactory for the $\{\beta_j\}$ and ψ in all of more than 30 runs of the program so far tried. However, it would be possible to derive starting values for the $\{\beta_j\}$ by using the standard logistic discriminant program (Anderson, 1972) on the sample points from H_1 and H_2. A starting value for θ_1 is obtained, by the method of moments, from the one-way marginal distributions by solving

$$\sum_{j=1}^{p}\bar{x}_{3j} = \theta_1\sum_{j=1}^{p}\bar{x}_{1j} + (1-\theta_1)\sum_{j=1}^{p}\bar{x}_{2j},$$

where $\bar{x}_{ij}$ is the sample mean of x_j in H_j ($i = 1, 2, 3$; $j = 1, \ldots, p$). A gain of between 3 and 5 iterations was found using these starting values in three trials. This gain is cancelled by the extra computing required to find the starting values. Hence zero starting values are recommended.

Note that the method of § 3 is computationally feasible with quite large data bases. For example, it has been used successfully when $n_1 = n_2 = 40$, $n_3 = 140$ and $p = 10$.

4·2. *Example*

The random variable x_1 has the $N(2, 1)$ and $N(0, 1)$ distributions in populations H_1 and H_2, respectively. Sets of sample points S_1, S_2 and S_3 were randomly generated from H_1, H_2 and H_3, the latter being the compound of H_1 and H_2 with $\theta_1 = 0{\cdot}25$. The values obtained were:

S_1:	1·15	0·25	2·31	2·44	3·28	3·34;
S_2:	0·74	−0·50	1·08	1·34	−0·74	0·15;
S_3:	−0·23	0·71	0·92	−0·53	−0·68	1·04
	0·61	−0·88	−0·61	0·59*	2·96*	2·59*.

The three starred values in S_3 were actually from H_1.

In the notation of §§ 2 and 3, the objective is to estimate the parameters, β_0 and β_1, of the likelihood ratio, and the compound proportion θ_1, from the sample data. The true values of β_0 and β_1, which are -2 and 2, respectively, are calculated using the normal distribution property. Recall that the estimation procedure does not use this normal property, only the assumption that the log likelihood ratio is linear.

The parameters β_0 and β_1 were estimated three ways. Estimates E_1 were found using ordinary logistic discrimination (Anderson, 1972) on samples S_1^* and S_2^*; these are S_1 and S_2 augmented by the points in S_3 allocated to their correct source population. Estimates E_2 were given by the logistic compound technique of § 3 on samples S_1, S_2 and S_3. Finally, estimates E_3 were given by ordinary logistic discrimination on the samples S_1 and S_2. All these estimates are given in Table 1 together with estimates of their standard errors. The latter were calculated from the second derivatives of $\log L^*$ evaluated at the maximum likelihood point. That is, regarding L^* as the likelihood, the usual estimates of the asymptotic standard errors were taken. By analogy with ordinary logistic discrimination, we conjecture

that these estimates of the standard errors of θ_1 and β will be close to the estimates of their asymptotic standard errors calculated from the likelihood L given in (2). Work is in progress to confirm this.

Table 1. *Estimates of the parameters of the likelihood ratio and the compound proportion*

	β_0	β_1	θ_1
True value	−2·00	2·00	0·25
E_1: standard logistic estimates, S_1^* and S_2^*	−2·05 ± 1·03	1·99 ± 0·83	—
E_2: compound logistic estimates, S_1, S_2 and S_3	−1·95 ± 1·29	1·76 ± 1·02	0·19 ± 0·18
E_3: standard logistic estimates, S_1 and S_2	−2·00 ± 1·31	1·72 ± 0·97	—
Method of moments	—	—	0·11

The compound fraction θ_1 was estimated using the compound logistic technique, E_2, and also by the method of moments given above. These results are reported in Table 1, together with a standard error for the E_2 estimate, calculated as above. A standard error for the method of moments is not given because this would be based on the actual form of the distributions $f_1(.)$ and $f_2(.)$. It would therefore not be comparable with the E_2 standard error which does not involve these assumptions.

It would be expected that, on average, the estimates would increase in accuracy from E_3 through E_2 to E_1 on the grounds of the amount of information available. In the results of the single simulation in Table 1 this happens. However, this may be a chance finding; the standard errors are relatively large.

In Table 1, all the estimates are reasonably close to their true values, perhaps closer by chance than would be expected, bearing in mind the size of the standard errors. However, these are the results from a single simulation.

Recall that the sample sizes are small, minimal distributional assumptions are being made and that compound distribution estimation problems are notoriously difficult (Day, 1969; Hosmer, 1973). Viewed in this light the compound logistic estimation system has worked well and can be recommended.

5. Logistic compounds of k distributions

The procedure introduced in §§ 2 and 3 for dealing with compounds of two distributions can be extended to deal with compounds of k distributions. An outline is presented here.

Suppose that sample points are available from populations H_s $(s = 1, \ldots, k)$ and from the compound distribution, H_{k+1}. The corresponding likelihoods are $f_s(x)$ $(s = 1, \ldots, k)$ and $g(x) = \Sigma\theta_s f_s(x)$. There are $n_s(x)$ observations at the point x from H_s $(s = 1, \ldots, k+1)$. Suppose $n(x) = \Sigma n_s(x)$ summed over s, $n_s = \Sigma n_s(x)$ summed over x and the total sample size, $n = \Sigma n_s$. Following the analysis in § 3, we estimate the parameters of the likelihood ratios $r_s(x) = f_s(x)/f_k(x)$ $(s = 1, \ldots, k-1)$, and the compound proportions $\theta^{\mathrm{T}} = (\theta_1, \ldots, \theta_k)$ by maximum likelihood under the model $r_s(x) = \exp(\beta_s^{\mathrm{T}} x)$, where $\beta_s^{\mathrm{T}} = (\beta_{s0}, \beta_{s1}, \ldots, \beta_{sk})$ $(s = 1, \ldots, k-1)$. The function $f_k(.)$ is a nuisance parameter.

The likelihood of the observations is, generalizing (2),

$$L_k = \Pi_x \left[\prod_{s=1}^{k-1} \{\exp(\beta_s^{\mathrm{T}} x)\}^{n_s(x)} \right] \left\{ 1 + \sum_{s=1}^{k-1} \exp(\beta_s^{\mathrm{T}} x) \right\}^{n_{k+1}(x)} \{f_k(x)\}^{n(x)}.$$

Restricting attention to the discrete case first, we may now maximize L_k with respect to the $\{\beta_s\}$, $\{\theta_s\}$ and $\{f_k(x)\}$, regarding the latter as probabilities to be estimated at each point of the sample space, as before. In this maximization θ_k is regarded as the function $1-\theta_1-\ldots-\theta_{k-1}$. Clearly there are constraints on the likelihoods $\{f_s(x)\}$:

$$c_s = \Sigma \exp(\beta_s^{\mathrm{T}} x) f_k(x) - 1 = 0 \quad (s = 1, \ldots, k-1), \quad c_k = \Sigma f_k(x) - 1 = 0.$$

When L_k is maximized using Lagrange's method of undetermined multipliers, it is found that for all x

$$f_k(x) = n(x) \Big/ \Big\{ n_k^* + \sum_{s=1}^{k-1} n_s^* \exp(\beta_s^{\mathrm{T}} x) \Big\},$$

where $n_s^* = n_s + \theta_s n_{k+1}$ $(s = 1, \ldots, k)$.

Generalizing (12), we let

$$L_k^* = \Pi_x \prod_{s=1}^{k} \Big\{ \frac{1}{n_s^*} P_s(x) \Big\}^{n_s(x)} \Big\{ \sum_{s=1}^{k} \frac{\theta_s}{n_s^*} P_s(x) \Big\}^{n_{k+1}(x)},$$

$$P_s(x) = n_s^* \exp(\beta_s^{\mathrm{T}} x) \Big/ \Big\{ n_k^* + \sum_{s=1}^{k-1} n_s^* \exp(\beta_s^{\mathrm{T}} x) \Big\} \quad (s = 1, \ldots, k-1),$$

$$P_k(x) = n_k^* \Big/ \Big\{ n_k^* + \sum_{s=1}^{k-1} n_s^* \exp(\beta_s^{\mathrm{T}} x) \Big\}.$$

It follows that the maximum likelihood estimates of the $\{\beta_s\}$ and $\{\theta_s\}$ from L_k may be obtained by maximizing the simpler function L_k^*. While this is not trivial, it involves only $(k-1)(p+1)+(p-1)$ parameters and so is computationally feasible using standard optimization routines for many values of k and p.

Extension of these results to a continuous space for x follow along the same lines as § 3·2.

6. Updating discriminant functions

In the notation of § 5 suppose that the objective is to discriminate between populations $H_1, \ldots, H_k$. Samples size n_s have been drawn from H_s $(s = 1, \ldots, k)$ and the discriminant functions have been estimated. In practice the rules will be used on further sample points and when their parent populations are eventually known, the discriminant functions can be updated by reestimating them with the extra sample points. However, at any one time in a continuing discriminant situation, say in medical diagnosis, there may be a large sample of size n_{k+1} of points whose true populations are not known with certainty. This sample has been taken from H_{k+1}, the compound of $H_1, \ldots, H_k$ in some proportions $\theta_1, \ldots, \theta_k$ and hence contains information about the discriminant functions. Unfortunately these sample points cannot be fed into standard procedures for estimating discriminant functions.

This problem of utilizing the information in H_{k+1} is exactly the same as that considered in previous sections. The notation for this section has been chosen to emphasize the similarities. Thus the techniques developed in this paper may be used with minor modifications to update estimates of logistic discriminant functions as derived by Anderson (1972), using sample points from H_{k+1}. The constants $\{\beta_{sj}\}$ as defined in § 5 give the likelihood ratios. If it is desired to use discriminant functions the $\{\beta_{s0}\}$ must be modified to give posterior probabilities (Anderson, 1972). Handling the discriminant updating problem in this way introduces no approximation, no unwanted order effects, while keeping the number of parameters to be estimated at a moderate level. Another advantage is that the proportion of sample points from H_s in H_{k+1} $(s = 1, \ldots, k)$ is not assumed to be the same as in the original

samples from $H_1, \ldots, H_k$. That is, it is not assumed that $\theta_s = n_s/(n-n_{k+1})$. Indeed, it may be of interest to test this equality if it is thought that the proportions from the various populations may change. It is in fact necessary to postulate that $\theta_s \neq n_s/(n-n_{k+1})$ for at least one value of $s = 1, \ldots, k$. Otherwise $n_s^* = n_s + \theta_s n_{k+1} = n\theta_s$. This implies that terms in L_k^*

$$\sum_{s=1}^{k} \frac{\theta_s}{n_s^*} P_s(x) = \frac{1}{n} \sum_{s=1}^{k} P_s(x) = \frac{1}{n}.$$

Hence if θ_s is constrained to be proportional to n_s $(s = 1, \ldots, k)$, there is no information in the sample from H_{k+1} about the $\{\beta_s\}$; it is informative only about $f_{k+1}(x)$. In practice the procedure is simple: the $\{\theta_s\}$ are not so constrained.

The ability to update the discriminant function estimates with information from sample points of uncertain provenance gives logistic discrimination a considerable advantage over Fisher's linear discriminant function which also has distribution-free properties. Titterington (1976) gives a Bayesian approach to the updating problem which uses distributional assumptions and introduces an unwanted order effect. Murray & Titterington (1978) have recently extended the kernel approach to discrimination to take account of sample points whose origins are uncertain. Thus their method provides an alternative to the logistic in some circumstances.

7. Discussion

For simplicity we again consider compounds of two distributions. Similar remarks apply to the k-compound case.

Although the methods presented in this paper are asymptotically efficient, there are occasionally problems with small samples. For example, if there is complete separation (Anderson, 1972) between the sample points from H_1 and H_2, the extra information from the H_3 sample may not always be utilizable. The behaviour of the likelihood L^* is complicated and requires further investigation. Similarly, the behaviour of the likelihood L^* when a binary or qualitative variable has a class empty in the sample from one of H_1 or H_2, but not both, is not clear. Anderson (1974) discusses this problem when there are no sample points available from the compound distribution H_3 and introduces an approximation to avoid the singularity in the likelihood. Again further investigation is required. These problems do not occur if the functional form $f_s(.)$ is specified.

In this paper, only one type of sampling has been considered: where sample points are available from each population H_1 and H_2 and the compound H_3, separately. However, in practice, the sample points may be from a mixture of H_1, H_2 and possibly H_3, in which case all the $\{n_s\}$ would be random variables. The relevant procedure in this case is to condition on the $\{n_s\}$, which leads straight back to the formulations of the likelihood given earlier. Hence, the method of estimation we have described is applicable, irrespective of the mode of sampling. Although it was assumed in § 3 that $\log r(x)$ is linear in x, the same approach may be used with other functional forms for $r(x)$. The results of § 3 are applicable if $\log r(x)$ is linear in the coefficients β; with this proviso, arbitrary functions of the observations x are permissible. A similar observation holds in ordinary linear regression. Thus if $\log r(x)$ is a quadratic function in x (Anderson, 1975), in principle the approach of § 3 may be used for estimation. In practice, the number of parameters to be estimated is $O(p^2)$ which may give a computational procedure too lengthy to be feasible. Anderson's (1975) device of taking a rank one approximation to the quadratic form to reduce the number of parameters to $O(p)$ can be used here.

I am very grateful to Dr Peter Phillips of the Medical Care Research Unit, University of Newcastle upon Tyne, for assistance with computing. I would also like to thank a referee for suggesting changes which have considerably improved the presentation.

REFERENCES

AITCHISON, J. & SILVEY, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29**, 813–28

ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

ANDERSON, J. A. (1974). Diagnosis by logistic discriminant function: further practical problems and results. *Appl. Statist.* **23**, 397–404.

ANDERSON, J. A. (1975). Quadratic logistic discrimination. *Biometrika* **62**, 149–54.

BARTLETT, M. S. & MACDONALD, P. D. M. (1968). Least squares estimation of distribution mixtures. *Nature* **217**, 195–6.

COX, D. R. (1966). Some procedures associated with the logistic qualitative response curve. In *Research Papers in Statistics: Festschrift for J. Neyman*, Ed. F. N. David, pp. 55–71. London: Wiley.

DAWID, A. P. (1976). Properties of diagnostic data distributions. *Biometrics* **32**, 647–58.

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463–74.

DAY, N. E. & KERRIDGE, D. F. (1967). A general maximum likelihood discriminant. *Biometrics* **23**, 313–23.

EMERY, J. L. & CARPENTER, R. G. (1974). Pulmonary mast cells in infants and their relation to unexpected death in infancy. *Proceedings of the Francis E. Camps International Symposium on Sudden and Unexpected Deaths in Infancy*. Toronto: The Canadian Foundation for the study of infant deaths.

FLETCHER, R. & POWELL, M. J. D. (1963). A rapidly convergent descent method for minimisation. *Computer J.* **6**, 163–8.

GILL, P. E. & MURRAY, W. (1972). Quasi-Newton methods for unconstrained optimization. *J. Inst. Math. & Applic.* **9**, 91–108.

GOOD, I. J. & GASKINS, R. A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58**, 255–77.

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**, 761–70.

MANTEL, N. (1973). Synthetic retrospective studies and related topics. *Biometrics* **29**, 479–86.

MURRAY, G. D. & TITTERINGTON, D. M. (1978). Estimation problems with data from a mixture. *Appl. Statist.* **27**, 325–34.

N.A.G. *Fortran Library Manual* (1975). Oxford: N.A.G. Executive.

ODELL, P. L. & BASU, J. P. (1976). Concerning several methods for estimating crop acreages using remote sensing data. *Comm. Statist.* A **5**, 1091–114.

PEARSON, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans.* A **188**, 71–110.

PETERS, C. & COBERLY, W. A. (1976). The numerical evaluation of the maximum-likelihood estimate of mixture proportions. *Comm. Statist.* A **5**, 1127–35.

TITTERINGTON, D. M. (1976). Updating a diagnostic system using unconfirmed cases. *Appl. Statist.* **25**, 238–47.

TUBBS, J. D. & COBERLY, W. A. (1976). An empirical sensitivity study of mixture proportion estimators. *Comm. Statist.* A **5**, 1115–25.

WOLFE, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Mult. Behav. Res.* **5**, 329–50.

[*Received July* 1977. *Revised October* 1978]