THE ESTIMATION OF FIVE-PARAMETER MIXED NORMAL DISTRIBUTIONS

by

J. G. Fryer and C. A. Robertson

Department of Biostatistics
University of North Carolina at Chapel Hill
and
Department of Statistics
University of California at Riverside

Institute of Statistics Mimeo Series No. 818

MAY 1972

The Estimation of Five-Parameter Mixed Normal Distributions

by

J. G. Fryer[1]
and
C. A. Robertson

Department of Biostatistics
University of North Carolina at Chapel Hill
and
Department of Statistics
University of California at Riverside

S U M M A R Y

Fisher's method of maximum likelihood breaks down when applied to the problem of estimating the five parameters of a mixture of two normal densities from a continuous random sample of size n. The two alternatives considered here are moment estimates on the one hand, and multinomial maximum likelihood and minimum χ^2 estimates obtained by grouping the underlying variable on the other. The methods are compared both for bias (to n^{-1}), and mean-squared-error (to n^{-2}) for a variety of mixed distributions. In terms of bias, there seems to be little to choose between them. As regards mean-squared-error, the grouped estimates are shown to be more accurate than the moment estimates for most distributions, though the moment estimates do seem to be preferable for distributions which are particularly difficult to estimate. It is also found that the accuracy levels of the grouped maximum likelihood and minimum χ^2 estimates do not differ greatly. For any of the estimates, we show in many cases that the n^{-2} term of the mean-squared-

---

[1]This author's contribution was supported in part by the National Institutes of Health, Institute of General Medical Sciences Grant No. GM-12868.

error cannot be neglected unless the sample size is very large indeed, and that only very large samples give an accuracy level which most experimenters would find acceptable.

1. Introduction

The moment solution to the problem of estimating the five parameters of an arbitrary mixture of two unspecified normal densities was first given by Karl Pearson, as long ago as 1894. Yet, despite the fact that many random phenomena have been subsequently shown to follow this distribution, it is only recently that the estimation problem has been seriously reconsidered. Since 1950 or so, a number of papers have appeared which are directly or indirectly concerned with some reduced form of the problem, such as when the variances of the two components are assumed to be equal or when certain of the parameters are assumed to be known. Some principal references here are Rao (1948), Preston (1953), Hill (1963), Bhattacharya (1967), Choi and Bulgren (1968), Kabir (1968) and Day (1969). The earliest recent reference for the full five-parameter problem seems to be Hasselblad's paper of 1966. Since then, the problem has also attracted the attention of Cohen (1967), Robertson and Fryer (1970) and Behboodian (1970). Cohen shows how the computation of Pearson's moment method, which requires the solution of a ninth degree equation, can be lightened to some extent. He also considers the estimates resulting from maximising the likelihood function of the sample for variation in the mixing proportion, given that the first four moment equations are satisfied, but does not go on to derive appropriate formulae for their biases and covariances. Our earlier paper gives formulae for the biases and accuracies of Pearson's moment estimate to order n^{-2}. The other two papers are concerned with maximum likelihood estimation of the five parameters, but before commenting on them we ought first to draw attention to an irregularity in the behaviour of the likelihood function itself.

Let

$$L = \prod_{i=1}^{n} \left[\frac{P}{\sigma_1\sqrt{2\pi}} \exp\left\{ -\tfrac{1}{2}\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2 \right\} + \frac{q}{\sigma_2\sqrt{2\pi}} \exp\left\{ -\tfrac{1}{2}\left(\frac{x_i-\mu_2}{\sigma_2}\right)^2 \right\} \right] \qquad (1)$$

denote the likelihood function of a sample from a two-component normal mixture. To simplify matters take the values of P, μ_2 and σ_2 to be fixed. Setting $\mu_1 = x_i$ and allowing $\sigma_1 \to 0$ we find immediately that L is unbounded for variation in μ_1 and σ_1. Day (1969) has also remarked on this. Consider now some other paths to the point $x_i = \mu_1$, $\sigma=0$. Denote $(x_i-\mu_1)$ by t and examine the effect on L of the limiting operation $t\to 0$ with $\sigma = ct^r$. Clearly if $r \leq 1$ then along any such path to the (t,σ) origin L is again unbounded. Evidently the maximum likelihood estimate does not exist for a continuous sample from a 5-parameter normal mixture. In practice we might be tempted to rule out this complicating part of the parameter space, but this can hardly be considered satisfactory. Alternatively, we might try to find the estimate corresponding to the largest stationary maximum of the likelihood function (assuming that one exists) and this is precisely what Behboodian proposes. It is also what Hasselblad proposes in effect, since although he starts by considering grouped maximum likelihood estimates, he then assumes the width of each group to be sufficiently small to allow us to replace each group probability divided by its length by the appropriate value of the density function.

But what are the properties of this method? Three results on the estimation of the overall mean of a mixture $\mu = (p\mu_1 + q\mu_2)$ suggest to us that the efficiencies of the individual parameter estimates are likely to vary considerably over the parameter space. Firstly Behboodian has proved that using his estimates we are led to estimate μ by $\bar{X}$, the sample mean. Secondly, Tukey (1960) has effectively shown that if $\mu_1 = \mu_2$ then the large-sample efficiency of $\bar{X}$ for μ against the sample median can be made arbitrarily small for any p by increasing or decreasing the ratio σ_1/σ_2. Of course this is what usually happens in long-tailed distributions. Thirdly, it is easily shown that $\bar{X}$ is asymptotically fully efficient for μ in mixtures with $\sigma_1 = \sigma_2$. Others may not find this argument convincing. Either way it seems to us that the efficiency question is well worth resolving.

In this paper, however, we aim to compare the performances of Pearson's moment estimates with those of two of the best known estimates obtained from grouping the underlying variable. The two grouped procedures that we examine are multinomial maximum likelihood and minimum χ^2, and we shall always assume the grouping to be coarse enough to counter the peculiar behaviour of the likelihood function for ungrouped data. The criteria we have used for comparison are simply bias (to n^{-1}) and mean-squared-error (to n^{-2}). Bowman and Shenton in an unpublished Oak Ridge National Laboratory Report (referred to as B-S from now on) have already given the biases and covariances of grouped maximum likelihood estimates to n^{-2} and we shall make use of their results. But in order to contrast them with minimum χ^2 estimates, we shall have to derive the analogous formulae ourselves. The set of distributions for which we calculate the biases and accuracies of the various estimates has been deliberatly chosen to cover the whole range of possibilities.

2. Notation and some other preliminaries

The notation that we shall use in the development of formulae for the biases and covariances of minimum χ^2 estimators is identical to that of B-S. Since their paper is not yet immediately accessible, we shall briefly outline their notation with a few necessary extensions here.

In this multinomial situation n_r will be used to denote the observed proportion of the sample values occurring in group r, and p_r to denote the corresponding group probability. These probabilities depend on the k unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$ which appear in the underlying distributional form. The deviation $(n_r - p_r)$ will be written as e_r. It will be convenient to write E_i for the coefficient of n^{-i} in the expansion of an expectation in powers of n^{-1}. Likewise, B_i, C_i and M_i will denote the coefficients of n^{-i} in the expansion of a bias, covariance and mean-squared-error, respectively.

To simplify the appearance of results, we make extensive use of a summation convention in which summation signs are dropped in the presence of a repeated suffix or suffices. There are two kinds of summation involved in the use of this convention; sums over the k parameters and sums over the classes or groups. We distinguish between these by adopting the convention that repeated Greek letters, α, β, γ, ... are used for summation over parameters, and Roman letters r, s, t, ... for summation over the groups. A quantity which is fundamental in this kind of problem is the derivative $\partial^m \log p_r / \partial\theta_{\alpha_1} \partial\theta_{\alpha_2} \ldots \partial\theta_{\alpha_m}$, and we shall always denote it here by $\Gamma^r_{\alpha_1\alpha_2\ldots\alpha_m}$. Similarly, the related derivative $\partial^m p_r^{-1} / \partial\theta_{\alpha_1} \partial\theta_{\alpha_2} \ldots \partial\theta_{\alpha_m}$ will be denoted by $\Delta^r_{\alpha_1\alpha_2\ldots\alpha_m}$. Terms like

$$\sum_r p_r \frac{\partial^2 \log p_r}{\partial\theta_\alpha \partial\theta_\beta} \cdot \frac{\partial^2 \log p_r}{\partial\theta_\gamma \partial\theta_\delta} \cdot \frac{\partial \log p_r}{\partial\theta_\varepsilon}$$

are also frequently met in this kind of work and for this sum, for example, we use the notation $(\alpha\beta, \gamma\delta, \varepsilon)$. Similarly,

$$\sum_r p_r \frac{\partial^3 \log p_r}{\partial\theta_\alpha \partial\theta_\beta \partial\theta_\gamma} \cdot \frac{\partial \log p_r}{\partial\theta_\delta}$$

is denoted by $(\alpha\beta\gamma, \delta)$ or $(\delta, \beta\gamma\alpha)$, because of the commutativity of differentiation and so on. By a slight extension of the B-S notation, we shall write

$\sum_r \Gamma^r_\alpha$ as (Γ_α), $\sum_r \Gamma^r_{\alpha\beta}$ as $(\Gamma_{\alpha\beta})$, $\sum_r \Gamma^r_\alpha \Gamma^r{}_\beta$ as $(\Gamma_\alpha \Gamma_\beta)$, and likewise for similar terms.

In order to write the formulae for minimum χ^2 estimators in a form similar to the corresponding formulae for grouped maximum likelihood estimators, we will need to make use of the following relationships between the Γ's and the Δ's.

$$p_r \Delta^r_\alpha = -\Gamma^r_\alpha \tag{2}$$

$$p_r \Delta^r_{\alpha\beta} = -\Gamma^r_{\alpha\beta} + \Gamma^r_\alpha \Gamma^r_\beta \tag{3}$$

$$p_r \Delta^r_{\alpha\beta\gamma} = -\Gamma^r_{\alpha\beta\gamma} + \sum^3 \Gamma^r_{\alpha\beta} \Gamma^r_\gamma - \Gamma^r_\alpha \Gamma^r_\beta \Gamma^r_\gamma \tag{4}$$

$$p_r \Delta^r_{\alpha\beta\gamma\delta} = -\Gamma^r_{\alpha\beta\gamma\delta} + \sum^4 \Gamma^r_{\alpha\beta\gamma}\Gamma^r_\delta + \sum^3 \Gamma^r_{\alpha\beta}\Gamma_{\gamma\delta} - \sum^6 \Gamma^r_{\alpha\beta}\Gamma^r_\gamma\Gamma^r_\delta + \Gamma^r_\alpha\Gamma^r_\beta\Gamma^r_\gamma\Gamma^r_\delta \tag{5}$$

In these equations, the summation over r is temporarily suspended. The sums $\sum^i$ describe sums over all the i permutations of the suffices inside producing terms of the same form, for example,

$$\sum^3 \Gamma^r_{\alpha\beta} \Gamma^r_\gamma = \Gamma^r_{\alpha\beta} \Gamma^r_\gamma + \Gamma^r_{\alpha\gamma} \Gamma^r_\beta + \Gamma^r_{\beta\gamma} \Gamma^r_\alpha.$$

To evaluate the final expectations for the minimum χ^2 estimators, we make use of some results of Bowman and Shenton (1963) on the expected values of products of linear combinations of the e_i's, such as $E_2\{(h_r e_r)(i_s e_s)(j_t e_t)(k_u e_u)\}$. In addition, we need the following non-linear relationships.

$$E_2\{(h_r e_r^2)(i_s e_s)\} = (p_r h_r i_r - 2p_r^2 h_r i_r) - (p_r h_r - 2p_r^2 h_r)(p_r i_r) \tag{6}$$

$$E_2\{(h_r e_r^2)(i_s e_s^2)\} = 2(p_r^2 h_r i_r - 2p_r^3 h_r i_r) + (p_r h_r - p_r^2 h_r)(p_r i_r - p_r^2 i_r) + 2(p_r^2 h_r)(p_r^2 i_r) \tag{7}$$

$$E_2\{(h_r e_r^2)(i_s e_s)(j_t e_t)\} = 2(p_r^2 h_r i_r j_r) - 2(p_r i_r)(p_r^2 h_r j_r) - 2(p_r j_r)(p_r^2 h_r i_r)$$
$$+ (p_r h_r - p_r^2 h_r)(p_r i_r j_r)$$
$$+ (3\, p_r^2 h_r - p_r h_r)(p_r i_r)(p_r j_r) \qquad (8)$$

3. Algebraic details for the minimum χ^2 estimators

In this section, we shall derive results for the variance-covariance matrix of minimum χ^2 estimators to the second order, and for their bias to the first order. We then compare them with the corresponding formulae for grouped maximum likelihood estimates. Following B-S, we first introduce a fundamental local stochastic Taylor expansion for the minimum χ^2 estimators $\underline{\hat{\theta}} = \{\hat{\theta}_\alpha\}$, namely

$$\hat{\theta}_\alpha = \theta_\alpha + \phi_1^\alpha + \frac{1}{2!}\phi_2^\alpha + \frac{1}{3!}\phi_3^\alpha + \ldots \qquad (9)$$

where

(i) $\phi_1^\alpha = e_r \dfrac{\bar{\partial}\theta_\alpha}{\partial n_r}$, $\qquad \phi_2^\alpha = e_r e_s \dfrac{\bar{\partial}^2\theta_\alpha}{\partial n_r \partial n_s}$, and so on.

(ii) $\dfrac{\bar{\partial}\theta_\alpha}{\partial n_r} = \dfrac{\partial\hat{\theta}_\alpha}{\partial n_r}$, evaluated at $n_r = p_r$, $\underline{\hat{\theta}}=\underline{\theta}$, and similarly for higher derivatives.

We can evaluate the derivatives occurring in (9), indirectly, by using the estimating equations,

$$\sum_r n_r^2\, \Delta_\alpha^r = 0, \qquad \alpha=1,2,\ldots,k. \qquad (10)$$

Differentiating (10) with respect to n_r gives

$$2n_r\, \Delta_\alpha^r + (n_r^2\, \Delta_{\alpha\beta}^r) \frac{\partial\hat{\theta}_\beta}{\partial n_r} = 0, \qquad (11)$$

where the term in brackets is summed over r. Hence, $(p_r\Gamma_\alpha^r\Gamma_\beta^r - p_r\Gamma_{\alpha\beta}^r)\dfrac{\bar{\partial}\theta_\beta}{\partial n_r} = 2\Gamma_\alpha^r$, using (2) and (3), and so

$$\frac{\bar{\partial}\theta_\alpha}{\partial n_r} = L^{\alpha\beta}\Gamma^r_\beta \ , \tag{12}$$

where $L^{\alpha\beta}$ is the inverse of $L_{\alpha\beta} = (\alpha,\beta)$. Using (12) in (9), we find immediately for the estimates $\hat{\theta}_a$ and $\hat{\theta}_b$, that

$$\begin{aligned} E_1\{(\hat{\theta}_a-\theta_a)(\hat{\theta}_b-\theta_b)\} &= E_1(\phi^a_1\phi^b_1) \\ &= E_1(L^{a\gamma}L^{b\delta}\Gamma^r_\gamma\Gamma^s_\delta e_r e_s) \\ &= L^{ab}, \end{aligned} \tag{13}$$

as it should be.

The process illustrated by (11), (12) and (13) can be extended to give results for $B_1(\hat{\theta}_a)$ and $C_2(\hat{\theta}_a,\hat{\theta}_b)$. If we differentiate (11) with respect to n_s, we find

$$2L_{\alpha\beta}\frac{\bar{\partial}^2\theta_\beta}{\partial n_r\partial n_s} = -2\delta_{rs}\Delta^r_\alpha - 2p_r\Delta^r_{\alpha\beta}\frac{\bar{\partial}\theta_\beta}{\partial n_s} - 2p_s\Delta^s_{\alpha\beta}\frac{\bar{\partial}\theta_\beta}{\partial n_r} - (p^2_r\Delta^r_{\alpha\beta\gamma})\frac{\bar{\partial}\theta_\beta}{\partial n_r}\cdot\frac{\bar{\partial}\theta_\gamma}{\partial n_s}, \tag{14}$$

where δ_{rs} is the usual Kronecker delta function. Further differentiation of (14) with respect to n_t, gives

$$\begin{aligned} 2L_{\alpha\beta}\frac{\bar{\partial}^3\theta_\beta}{\partial n_r\partial n_s\partial n_t} &= -2\sum^3 \delta_{rs}L^{\beta\gamma}\Delta^r_{\alpha\beta}\Gamma^t_\gamma - 2\sum^3 p_r L^{\beta\delta}L^{\gamma\varepsilon}\Delta^r_{\alpha\beta\gamma}\Gamma^s_\delta\Gamma^t_\varepsilon \\ &\quad - 2\sum^3 p_r\Delta^r_{\alpha\beta}\frac{\bar{\partial}^2\theta_\beta}{\partial n_s\partial n_t} - (p^2_r\Delta^r_{\alpha\beta\gamma})\sum^3\frac{\bar{\partial}^2\theta_\gamma}{\partial n_s\partial n_t}L^{\beta\delta}\Gamma^r_\delta \\ &\quad - (p^2_r\Delta^r_{\alpha\beta\gamma\delta})L^{\beta\varepsilon}L^{\gamma\zeta}L^{\delta\theta}\Gamma^r_\varepsilon\Gamma^s_\zeta\Gamma^t_\theta \end{aligned} \tag{15}$$

Now we find from the expansion (9) for $\hat{\theta}_a$ and $\hat{\theta}_b$, that

$$B_1(\hat{\theta}_a) = \frac{1}{2}E_1(\phi^a_2) \tag{16}$$

$$E_2\{(\hat{\theta}_a-\theta_a)(\hat{\theta}_b-\theta_b)\} = \frac{1}{2} E_2(\phi_1^a\phi_2^b + \phi_2^a\phi_1^b) + \frac{1}{6} E_2(\phi_1^a\phi_3^b + \phi_3^a\phi_1^b) + \frac{1}{4} E_2(\phi_2^a\phi_2^b) \tag{17}$$

Using the formulae (2) to (5), (12), (14), (15), the results of Bowman and Shenton (1963) on the expected values of linear combinations of the e's, and the non-linear results (6) to (8), we can evaluate all of the terms in (16) and (17). The algebra becomes very heavy; the final result for bias is

$$B_1(\hat{\theta}_a) = \frac{1}{2} L^{a\beta}(\Gamma_\beta) + \frac{1}{2} L^{a\beta}L^{\gamma\delta}\{(\beta\gamma\delta) + 2(\beta\gamma,\delta) - (\beta,\gamma,\delta)\} \tag{18}$$

and $E_2\{(\hat{\theta}_a-\theta_a)(\hat{\theta}_b-\theta_b)\}$ is given by

$$\begin{aligned}
E_2\{(\hat{\theta}_a-\theta_a)(\hat{\theta}_b-\theta_b)\} = {} & -L^{ab} + L^{a\beta}L^{b\alpha}\{\tfrac{1}{2}(\Gamma_\alpha\Gamma_\beta) + \tfrac{1}{4}(\Gamma_\alpha)(\Gamma_\beta) + (\Gamma_{\alpha\beta})\} \\
& + L^{a\beta}L^{b\alpha}L^{\gamma\delta}[-\tfrac{5}{2}(\alpha,\beta,\gamma,\delta) - (\alpha\gamma,\beta,\delta) - (\beta\gamma,\alpha,\delta) \\
& - \tfrac{3}{2}(\alpha\beta,\gamma,\delta) - \tfrac{1}{2}(\gamma\delta,\alpha,\beta) + 2(\alpha\gamma,\beta\delta) - \tfrac{1}{2}(\alpha\beta,\gamma\delta) \\
& + (\alpha\beta\gamma,\delta) + \tfrac{1}{2}(\alpha\beta\gamma\delta) + \tfrac{1}{2}(\Gamma_\beta)\{(\alpha\gamma,\delta) - \tfrac{1}{2}(\alpha,\gamma,\delta) + \tfrac{1}{2}(\alpha\gamma\delta)\} \\
& + \tfrac{1}{2}(\Gamma_\alpha)\{(\beta\gamma,\delta) - \tfrac{1}{2}(\beta,\gamma,\delta) + \tfrac{1}{2}(\beta\gamma\delta)\} \\
& + \tfrac{1}{2}(\Gamma_\delta)\{2(\alpha\beta\gamma) + (\beta\gamma,\alpha) + (\alpha\gamma,\beta)\}] \\
& + L^{a\beta}L^{b\alpha}L^{\gamma\varepsilon}L^{\delta\zeta}\{\tfrac{1}{4}(\beta\gamma\varepsilon)(\alpha\delta\zeta) + \tfrac{3}{2}(\beta\gamma\delta)(\alpha\varepsilon\zeta) \\
& + (\alpha\beta\zeta)(\gamma\delta\varepsilon) + 2(\alpha\beta\zeta)(\gamma\delta,\varepsilon) + \tfrac{1}{2}(\gamma\delta\varepsilon)(\beta\zeta,\alpha) \\
& + \tfrac{1}{2}(\gamma\delta\varepsilon)(\alpha\zeta,\beta) + \tfrac{1}{2}(\beta\delta\zeta)(\alpha\varepsilon,\gamma) + \tfrac{1}{2}(\beta\gamma\zeta)(\delta\varepsilon,\alpha) \\
& + \tfrac{1}{2}(\alpha\gamma\zeta)(\delta\varepsilon,\beta) + 2(\beta\gamma\delta)(\alpha\varepsilon,\zeta) + \tfrac{1}{2}(\alpha\delta\zeta)(\beta\varepsilon,\gamma) \\
& + 2(\alpha\gamma\delta)(\beta\varepsilon,\zeta) + (\beta\varepsilon,\gamma)(\alpha\zeta,\delta) + (\beta\varepsilon,\delta)(\alpha\zeta,\gamma) \\
& + (\beta\zeta,\varepsilon)(\gamma\delta,\alpha) + (\beta\zeta,\alpha)(\gamma\delta,\varepsilon) + (\alpha\zeta,\varepsilon)(\gamma\delta,\beta) \\
& + (\alpha\zeta,\beta)(\gamma\delta,\varepsilon) + \tfrac{1}{4}(\beta,\gamma,\varepsilon)(\alpha,\delta,\zeta) + \tfrac{1}{2}(\beta,\delta,\varepsilon)(\alpha,\gamma,\zeta)
\end{aligned}$$

$$- \frac{1}{4}(\alpha\delta\zeta)(\beta,\gamma,\varepsilon) - \frac{1}{4}(\beta\gamma\varepsilon)(\alpha,\delta,\zeta) - \frac{1}{2}(\beta\gamma\delta)(\alpha,\varepsilon,\zeta)$$

$$- \frac{1}{2}(\beta,\gamma,\delta)(\alpha\varepsilon\zeta) - (\alpha\beta\zeta)(\gamma,\delta,\varepsilon) - \frac{1}{2}(\gamma,\delta,\varepsilon)(\beta\zeta,\alpha)$$

$$- \frac{1}{2}(\gamma,\delta,\varepsilon)(\alpha\zeta,\beta) - \frac{1}{2}(\beta,\gamma,\varepsilon)(\alpha\zeta,\delta) - \frac{1}{2}(\alpha,\delta,\zeta)(\beta\varepsilon,\gamma)$$

$$- \frac{1}{2}(\beta,\varepsilon,\zeta)(\gamma\delta,\alpha) - \frac{1}{2}(\alpha,\varepsilon,\zeta)(\gamma\delta,\beta)\} \qquad (19)$$

The minimum χ^2 bias may be compared with that of the grouped maximum likelihood estimate (θ^*),

$$B_1(\theta^*_a) = \frac{1}{2} L^{a\beta}L^{\gamma\delta}\{2(\beta\gamma,\delta) + (\beta\gamma\delta)\} \qquad (20)$$

The M_2 result for grouped maximum likelihood estimates is somewhat simpler than that for the minimum χ^2 estimates, and is given in B-S as

$$E_2\{(\theta^*_a-\theta_a)(\theta^*_b-\theta_b)\} = - L^{ab} + L^{a\beta}L^{b\alpha}L^{\gamma\delta}\{(\alpha\delta,\beta,\gamma) + (\beta\delta,\alpha,\gamma) + (\alpha\beta\gamma\delta) + 3(\alpha\delta,\beta\gamma)$$

$$+ 2(\alpha\beta\gamma,\delta) + \frac{1}{2}(\beta\gamma\delta,\alpha) + \frac{1}{2}(\alpha\gamma\delta,\beta)\}$$

$$+ L^{a\beta}L^{b\alpha}L^{\gamma\varepsilon}L^{\delta\zeta}\{\frac{1}{2}(\alpha\gamma\zeta)(\beta,\delta,\varepsilon) + \frac{1}{2}(\beta\gamma\zeta)(\alpha,\delta,\varepsilon) + (\alpha\beta\gamma)(\delta\varepsilon\zeta)$$

$$+ \frac{5}{2}(\alpha\gamma\delta)(\beta\varepsilon\zeta) + \frac{1}{4}(\beta\gamma\varepsilon)(\alpha\delta\zeta) + (\delta\varepsilon,\alpha)(\beta\gamma\zeta) + (\delta\varepsilon,\beta)(\alpha\gamma\zeta)$$

$$+ 2(\alpha\beta\zeta)(\gamma\delta,\varepsilon) + 3(\alpha\zeta,\varepsilon)(\beta\gamma\delta) + 3(\beta\zeta,\varepsilon)(\alpha\gamma\delta) + \frac{1}{2}(\alpha\zeta,\beta)(\gamma\delta\varepsilon)$$

$$+ \frac{1}{2}(\beta\zeta,\alpha)(\gamma\delta\varepsilon) + \frac{1}{2}(\alpha\gamma\varepsilon)(\beta\delta,\zeta) + \frac{1}{2}(\beta\gamma\varepsilon)(\alpha\delta,\zeta) + (\alpha\delta,\varepsilon)(\gamma\zeta,\beta)$$

$$+ (\beta\delta,\varepsilon)(\gamma\zeta,\alpha) + (\alpha\delta,\beta)(\gamma\zeta,\varepsilon) + (\beta\delta,\alpha)(\gamma\zeta,\varepsilon)$$

$$+ (\alpha\delta,\gamma)(\beta\varepsilon,\zeta) + (\beta\gamma,\varepsilon)(\alpha\delta,\zeta)\} . \qquad (21)$$

In view of the fact that Bowman and Shenton have applied a number of checks to equation (21), we can take it to be correct. We have similarly applied checks to equation (19). For example, it may be shown that in the case where there is one parameter, θ, which is the probability parameter of a simple binomial distribution, then $E_2(\hat{\theta}-\theta)^2$ is zero when calculated from (19), even though a large number of non-zero terms are involved. Checks of this kind, taken together with the fact that (19) and (21) lead to values of comparable magnitude when computed for various grouped mixed normal distributions, strongly suggest that (19) has also been correctly derived.

Two general points should be made about the formula at (21) (and also about (13) and (19)). Firstly, it is easily seen that the grouped maximum likelihood estimates may not strictly exist if all of the observations fall into one group. For example, if in the case of a single two-parameter normal distribution, all of the observations fall in the j^{th} group, we can make the grouped likelihood function arbitrarily close to unity if we take the mean to be any point in the j^{th} group, and the variance arbitrarily small. The formula at (21), however, assumes that the grouped maximum likelihood estimate comes from the estimating equations in a regular way. Faced with this kind of difficulty, we should really use the maximum likelihood estimate conditional on not all of the observations falling in the same group. But, since with sensible grouping this conditioning event has very high probability, (21) can be used as a working approximation to the properly conditioned formula. Secondly, and probably more important, we should take note of Brillinger's (1964) general warning about the representation of moments by power series in n^{-1}.

In passing, it also seems well worth recording that under an obvious limit process, (21) simultaneously gives the required formula for the 'regular' maximum likelihood estimates in the continuous case, to order n^{-2}. Denoting the density function of a typical observation by $f(x; \underline{\theta})$, then for the continuous case all we need do is replace terms like $(\alpha\beta,\gamma,\delta)$ by

$$E\left\{\frac{\partial^2 \log f}{\partial\theta_\alpha \partial\theta_\beta} \cdot \frac{\partial \log f}{\partial\theta_\gamma} \cdot \frac{\partial \log f}{\partial\theta_\delta}\right\} ,$$

$L_{\alpha\beta}$ by

$$E\{\frac{\partial \log f}{\partial\theta_\alpha} \cdot \frac{\partial \log f}{\partial\theta_\beta}\} ,$$

and so on. It also seems worth mentioning that applying the same limiting process to (19), for example, we get the n^{-2} results for estimates, which are the continuous analogues of minimum χ^2 estimates, and which are obtained from minimising $\sum_i \{f(x_i; \underline{\theta})\}^{-1}$. These and similar estimates are seen to have the same asymptotic efficiency as the continuous maximum likelihood estimates to order n^{-1}, and in the multiparameter case, at least, might well prove to be effective competitors, especially for not too large sample sizes.

4. Some numerical comparisons

A mixture of two normal densities can have either two or four points of inflection. Distributions with two inflections have only one mode and so are termed unimodal. Those with four inflections may be either bimodal or unimodal. For the last kind, we use the term bitangential. A full description of the parameter space of mixed normal densities in terms of the three distinct types is given in Robertson and Fryer (1969). We have compared the performances of the moment and grouped estimates for a large number of mixtures of all three types. In this paper, we include details of three distributions of each kind (Tables 1 to 4), and these results can be considered fairly typical. Three of the chosen populations do not represent any known random phenomena. Population 4 is also artificial, but has been used by Cohen (1967). The remaining five populations all correspond to real data, and a description of them is given in Table 5.

The biases and mean-squared-errors of the grouped estimates were calculated for two different groupings of the underlying variable. First, the central part of the distribution (some 90%) was split into four equal-length groups, the two end-groups making six in all. Second, the distribution was divided into ten groups in a similar way. The first result in the tables for maximum likelihood refers to six groups and the second to ten. Since the minimum χ^2 calculation for six groups was almost identical to its maximum likelihood counterpart, only the ten group figures for minimum χ^2 are included. The method of grouping was dictated to a large extent by practical considerations, and it is not claimed that the groupings are in any sense optimal.

There seem to be a number of points worth making about the results in Tables 1 to 4. Firstly, as expected, the maximum likelihood results for ten groups are almost always substantial improvements on those for six groups. Population 6 shows that the change can be quite dramatic. We therefore need only compare moment estimators with the ten-group results. As regards bias, minimum χ^2 estimators seem to be slightly better than grouped maximum likelihood estimators, but the difference between the two is seldom large. Moment estimators are sometimes better and sometimes worse than the grouped estimators, and furthermore the differences are often considerable. For all three estimators, there is a definite tendency for bias to decrease as the number of inflections and modes in the parent distribution increases, indicating that we can indeed make some inferences about the quality of our estimates by looking at the shape of the distribution. Looking now at the first order terms in mean-squared-error, we find that the grouped estimators are usually markedly superior to the moment estimators. This superiority, however, is not completely uniform, since the performance of the moment estimates is often preferable to that of the grouped estimates for well-mixed distributions, with parameters that are very difficult to estimate. When we take the second-order terms into consideration, the general position of the moment estimators is slightly improved. Again, for all three estimates there is a strong tendency for mean-squared-error to decrease as the component distributions show themselves more clearly.

From the results, it is clear that for many parameter values a very large sample size is needed for the first order approximations to be adequate. (This is particularly noticeable for populations 2 and 5.) It is also frequently true that, for many parameter values, very large sample sizes are necessary to give sufficient accuracy to the estimators, in the sense of leading to estimators with reasonably small coefficient of variation. Martin's (1936) unsuccessful attempts to apply the moment solution using relatively small samples may be a reflection of this fact. In some cases, at least, low accuracy levels may be partly explained by the fact that mixed normal distributions can have very different parameter values, but virtually indistinguishable distribution functions. For example, denoting a normal density with mean μ and standard deviation σ by $N(\mu,\sigma)$, the two mixtures $\{0.0668.N(4.80, 1.90) + 0.9332.N(7.14, 1.08)\}$ and $\{0.1451.N(6.04, 2.10) + 0.8549.N(7.10, 1.02)\}$ are very similar over the whole range of the random variable. In such cases, a knowledge of the n^{-3} terms would also be desirable. It seems, however, that the calculations needed to obtain these terms for the grouped estimators would be prohibitive, both in terms of algebraic complexity and computational time; and the corresponding calculations for moment estimators, although not quite so bad, are heavy enough to make them impracticable at the present time.

Finally, there is an interesting comparison we can make, prompted by the results for population 2. Our calculations are based, of course, on the assumption that the two component population variances are not necessarily equal (although for a particular problem they may happen to be). Rao's (1948) first-order results for this distribution do involve the assumption that the variances are equal, so that there are only four parameters to estimate. It turns out that Rao's estimators are much more accurate. In terms of the estimator $\hat{p}$, for example, Rao's first-order variance is 11.8/n, whereas ours is of the order 500/n. This could imply that the assumption of equal variances leads to a great improvement in accuracy, when one can legitimately make it. The explanation for this might be that in such a case, the actual distribution is less likely to be confused with another of the same type. At the same time, this discrepancy in accuracy also suggests that making an unjustified assumption of equal variances may give very misleading results. However, the level of accuracy in this case is so poor that we would have to seriously question whether it was worth trying to identify all five parameters. In cases where a knowledge of the individual components is crucial, we can do little else. In other cases, however, we would be well advised to consider which functions of the parameters could be reliably estimated and whether or not we might be better off trying to estimate the percentage points of the mixture only, for example.

REFERENCES

Ashford, J. R., Brimblecombe, F. S. W. and Fryer, J. G. (1968) Birthweight and perinatal mortality in England and Wales, 1956-65. Chapter 1 of Problems and Progress in Medical Care, Third Series, ed. G. McLachlan, Oxford University Press.

Behboodian, J. (1970) On a mixture of normal distributions. Biometrika, 57, 215-7.

Bhattacharya, C. G. (1967) A simple method of resolution of a distribution into Gaussian components. Biometrics, 23, 115-37.

Brillinger, D. R. (1964) The asymptotic behaviour of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. Rev. Inst. Int. Statist., 32, 202-6.

Choi, K. and Bulgren, W. G. (1968) An estimation procedure for mixtures of distributions. J. R. Statist. Soc. B, 30, 444-60.

Cohen, A. C. (1967) Estimation in mixtures of two normal distributions. Technometrics, 9, 15-28.

Day, N. E. (1969) Estimating the components of a mixture of normal distributions. Biometrika, 56, 463-74.

Hald, A. (1952) Statistical Theory with Engineering Applications, New York, Wiley.

Hasselblad, V. (1966) Estimation of parameters for a mixture of normal distributions. Technometrics, 8, 431-44.

Hill, B. (1963) Information for estimating proportions in mixtures of exponential and normal distributions. J. Amer. Statist. Ass., 58, 918-32.

Kabir, A. B. M. L. (1968) Estimation of parameters of a finite mixture of distributions. J. R. Statist. Soc. B, 30, 472-82.

Martin, E. S. (1936) A study of an Egyptian series of mandibles with special reference to mathematical methods of sexing. Biometrika, 28, 149-78.

Pearson, K. (1894) Contributions to the mathematical theory of evolution. Phil. Trans. Roy. Soc., 185A, 71-110.

Pearson, K. (1914) A study of trypanosome strains. Biometrika, 10, 85-143.

Preston, E. J. (1953) A graphical method for the analysis of statistical distributions into two normal components. Biometrika, 40, 460-64.

Rao, C. R. (1948) The utilisation of multiple measurements in problems of biological classification. J. R. Statist. Soc. B, 10, 159-203.

Robertson, C. A. and Fryer, J. G. (1969) Some descriptive properties of normal mixtures. Skand. Aktuarietidskr, Parts 3-4, 137-46.

Robertson, C. A. and Fryer, J. G. (1970) The bias and accuracy of moment estimators. Biometrika, 57, 57-65.

Shenton, L. R. and Bowman, K. (1963) Higher moments of a maximum-likelihood estimate. J. R. Statist. Soc. B, 25, 305-17.

Tukey, J. W. (1960) A survey of sampling from contaminated distributions. Contributions to Probability and Statistics (essays in honour of Harold Hotelling), ed. Olkin et al., Stanford University Press, 448-85.

TABLE 1 - THE BIAS AND ACCURACY OF ESTIMATES FOR SOME UNIMODAL DISTRIBUTIONS

Population Number	Estimated Parameter and its value		Bias and Accuracy of Estimate							
			First order bias coefficient			Mean-squared-error, first order coefficient		Mean-squared-error, second order coefficient		
	Parameter	Value	M.M.	M.L.	M.χ^2	M.M.	M.L. and M.χ^2	M.M.	M.L.	M.χ^2
1	P	0.60	-2.93,1	-6.12,2 -5.52,1	-4.73,1	6.03,1	2.07,2 3.00,1	-2.08,3	2.46,6 2.19,4	2.27,4
	μ_1	0.00	-7.33,1	-5.74,2 -6.84,1	-5.95,1	3.12,1	9.79,1 1.91,1	5.06,4	2.34,6 3.25,4	3.23,4
	μ_2	0.60	3.97,1	-1.73,2 -2.07,0	-1.31,0	2.14,1	6.42,1 1.17,1	-2.76,3	2.33,5 2.20,3	2.24,3
	σ_1	1.00	-1.77,1	4.70,2 1.52,1	1.74,1	1.34,1	9.29,1 1.42,1	3.59,1	1.88,6 9.95,3	9.67,3
	σ_2	0.60	-6.20,1	2.10,2 8.14,0	4.98,0	4.66,1	9.12,1 1.66,1	3.38,3	3.18,5 5.82,3	6.87,3
2	P	0.57	-2.87,2	-1.08,3 -4.59,2	-4.50,2	4.72,2	2.38,3 8.08,2	-2.98,6	-1.19,8 -1.09,7	-1.10,7
	μ_1	0.00	-3.77,2	-1.33,3 -5.66,2	-5.57,2	7.76,2	4.01,3 1.34,3	-4.56,6	-1.92,8 -1.72,7	-1.74,7
	μ_2	1.44	-4.27,1	-6.73,1 -1.09,2	-8.74,1	1.31,3	6.27,3 2.19,3	-8.02,6	-2.98,8 -2.90,7	-2.92,7
	σ_1	1.00	-4.30,1	1.07,2 -1.13,1	-9.84,0	2.12,2	1.32,3 4.03,2	-1.11,6	-4.96,7 -4.41,6	-4.44,6
	σ_2	1.00	1.11,2	6.92,2 2.54,2	2.45,2	3.33,2	1.52,3 5.52,2	-1.09,6	-4.82,7 -4.29,6	-4.33,6
3	P	0.10	2.30,1	2.95,1 0.07,0	6.64,0	4.26,0	2.17,1 1.39,0	-2.67,3	-1.07,4 4.95,2	6.23,2
	μ_1	0.00	-9.90,1	-1.37,3 -7.77,1	-3.15,1	1.09,2	7.76,2 5.24,1	7.45,4	1.57,7 4.58,4	3.77,4
	μ_2	0.68	7.97,0	3.37,0 1.31,0	1.34,0	5.81,1	1.54,0 0.39,0	-3.23,2	-4.09,2 1.18,1	9.50,0
	σ_1	1.00	-1.12,2	2.08,3 -1.85,1	5.82,0	6.04,1	1.57,3 5.46,1	-2.94,4	4.08,7 2.07,4	1.72,4
	σ_2	0.45	-9.81,0	0.32,0 -0.17,0	-1.38,0	0.99,0	1.71,0 0.30,0	-8.58,2	-2.90,2 4.59,1	5.76,1

NOTE: In this and other tables, we use a decimal-exponent notation for the terms in n^{-1} and n^{-2}, e.g., 7.76,2 = 776. Also, M.M. denotes method of moments; M.L. maximum likelihood; and M.χ^2 minimum χ^2.

TABLE 2 - THE BIAS AND ACCURACY OF ESTIMATES FOR SOME BITANGENTIAL DISTRIBUTIONS

Population Number	Estimated Parameter and its value		Bias and Accuracy of Estimate							
			First order bias coefficient			Mean-squared-error, first order coefficient		Mean-squared-error, second order coefficient		
	Parameter	Value	M.M.	M.L.	$M.\chi^2$	M.M.	M.L. and $M.\chi^2$	M.M.	M.L.	$M.\chi^2$
4	P	0.30	5.00,1	-1.76,3 3.05,1	2.83,1	9.31,0	5.36,1 5.28,0	7.07,4	3.88,7 1.02,4	1.14,4
	μ_1	0.00	1.41,2	-8.15,3 1.12,2	1.04,2	1.77,2	1.09,3 1.07,2	9.40,5	8.19,8 1.56,5	1.75,5
	μ_2	1.75	1.59,1	-9.08,2 -1.60,0	-1.76,0	4.21,0	9.87,0 1.85,0	1.85,4	9.00,6 -7.82,2	-6.40,2
	σ_1	1.00	9.28,1	-1.19,4 1.24,2	1.14,2	1.42,2	2.17,3 1.61,2	5.62,5	1.71,9 1.94,5	2.22,5
	σ_2	0.53	-3.47,1	4.61,2 -1.37,1	-1.15,1	2.54,0	5.51,0 1.23,0	2.58,4	3.02,6 2.62,3	2.81,3
5	P	0.41	6.68,1	4.09,2 1.54,2	1.52,2	3.13,1	1.16,2 4.67,1	1.44,5	3.36,6 3.34,5	3.55,5
	μ_1	0.00	9.15,1	8.71,2 3.34,2	3.28,0	2.42,2	9.19,2 3.83,2	7.26,5	2.16,7 2.05,6	2.15,6
	μ_2	1.62	1.74,1	1.11,2 1.98,1	1.93,1	2.28,1	6.95,1 2.59,1	5.33,4	9.41,5 2.59,4	2.59,4
	σ_1	1.00	5.87,1	7.65,2 3.13,2	3.06,2	1.44,2	7.38,2 3.54,2	3.50,5	1.68,7 1.78,6	1.86,6
	σ_2	0.70	-4.29,1	-1.91,2 -6.41,1	-6.21,1	1.01,1	2.84,1 1.12,1	5.76,4	9.39,6 9.32,4	9.65,4
6	P	0.80	-9.54,1	-2.94,4 -1.63,1	-1.67,1	3.23,1	2.10,3 3.19,0	1.24,5	1.26,10 2.10,3	2.11,3
	μ_1	0.00	-1.03,2	-4.01,4 -2.14,1	-2.44,1	2.77,1	3.46,3 5.25,0	1.41,5	2.22,10 3.62,3	3.61,3
	μ_2	1.00	1.88,1	-3.64,3 -4.50,0	-5.01,0	7.59,1	1.56,1 5.64,0	5.18,4	9.54,7 6.40,2	5.77,2
	σ_1	1.00	-3.76,1	-2.72,4 -1.11,1	-8.96,0	4.76,0	1.44,3 4.01,0	1.14,4	9.81,9 1.38,3	1.37,3
	σ_2	0.40	-7.23,1	2.66,4 9.70,0	1.04,1	1.76,2	2.73,3 4.97,0	1.11,5	1.30,10 2.74,3	2.71,3

TABLE 3 - THE BIAS AND ACCURACY OF ESTIMATES FOR SOME BIMODAL DISTRIBUTIONS

Population Number	Estimated Parameter and its value		Bias and Accuracy of Estimate							
			First order bias coefficient			Mean-squared-error, first order coefficient		Mean-squared-error, second order coefficient		
	Parameter	Value	M.M.	M.L.	M.χ^2	M.M.	M.L. and M.χ^2	M.M.	M.L.	M.χ^2
7	P	0.35	3.26,0	1.63,1 4.53,0	5.03,0	3.20,0	6.97,0 2.12,0	4.10,2	2.37,3 3.07,2	3.34,2
	μ_1	0.00	0.63,0	5.64,1 1.32,1	1.40,1	5.19,1	1.04,2 3.38,1	1.66,3	3.01,4 3.91,3	3.91,3
	μ_2	2.42	0.45,0	4.92,0 2.35,0	1.84,0	8.30,0	1.58,1 5.41,0	-3.11,2	-3.34,3 8.43,1	8.41,1
	σ_1	1.00	-1.62,0	8.71,1 1.76,1	1.69,1	5.80,1	2.40,2 5.05,1	2.08,3	1.10,5 7.03,3	6.89,3
	σ_2	0.78	0.41,0	4.79,0 0.22,0	0.39,1	6.69,0	1.31,1 5.84,0	-1.81,2	-2.11,2 1.88,2	1.72,2
8	P	0.50	0.00,0	3.92,0 -2.96,0	-3 15,0	6.70,0	1.67,1 6.65,0	-2.21,2	-1.83,3 3.56,1	5.24,1
	μ_1	0.00	-3.73,0	2.05,1 -4.95,0	-5.25,0	4.30,1	9.75,1 4.05,1	-1.99,3	-1.37,4 -1.10,3	-1.20,3
	μ_2	2.50	-3.73,0	-1.30,0 -8.81,0	-1.11,1	4.30,1	9.69,1 4.17,1	-1.99,3	-2.16,4 -1.16,1	2.70,1
	σ_1	1.00	-0.65,0	5.54,1 5.96,0	4 82,0	4.11,1	1.75,2 4.57,1	-1.30,3	2.37,4 9.42,1	-1.77,2
	σ_2	1.00	-0.65,0	3.47,1 2.09,1	2.22,1	4.11,1	1.13,2 5.90,1	-1.30,3	-9.54,3 4.81,3	4.80,3
9	P	0.79	-4.29,0	-0.40,0 -3.04,0	-3.22,0	1.65,0	0.86,0 0.72,0	-1.64,2	1.07,2 1.02,2	1.02,2
	μ_1	0.00	-7.97,0	0.25,0 -5.61,0	-5.45,0	8.78,0	4.42,0 3.82,0	-7.35,2	4.13,2 3.46,2	3.27,2
	μ_2	2.80	-8.36,0	-1.59,0 -1.21,1	-1.18,1	3.85,1	1.51,1 1.35,1	-7.54,2	2.80,3 1.99,3	1.88,3
	σ_1	1.00	-7.41,0	7.19,0 -5.85,0	-7.21,0	1.57,1	2.34,1 1.18,1	-8,76,2	2.07,3 6.02,2	5.05,2
	σ_2	0.68	2.62,0	5.82,0 1.19,1	1.04,1	2.29,1	1.30,1 1.40,1	2.23,2	3.18,3 2.56,3	2.41,3

TABLE 4

First Order Coefficients of Generalised Variances

Population Number	M.M.	M.L, M.χ^2 (6 groups)	M.L, M.χ^2 (10 groups)
1	1.320,4	9.800,4	5.658,3
2	1.362,6	2.848,7	4.149,6
3	1.417,2	3.435,3	4.958,1
4	2.310,2	5.838,3	1.891,2
5	4.432,3	4.634,4	1.104,4
6	1.014,4	2.060,5	1.189,2
7	6.917,2	5.120,3	5.762,2
8	5.811,3	5.964,4	8.963,3
9	2,689,2	4.335,2	1.807,2

TABLE 5

Details of the Five Real Populations

Population Number	Description of Data	Source
2	Heights of 454 plants.	Rao (1948)
3	Birthweights of 8,533 Devon infants, 1965.	Ashford, Brimblecombe and Fryer (1968)
5	Breadth of forehead of 1,000 Naples crabs	Pearson (1894)
7	Length of 1,000 individuals of Trypanosoma Gambiense	Pearson (1914)
9	Ash content of 430 samples of peat.	Hald (1952, p.156)