Ann. Inst. Statist. Math.
Vol. 46, No. 1, 83–93 (1994)

ROBUST ESTIMATION OF k-COMPONENT UNIVARIATE NORMAL MIXTURES

B. R. CLARKE[1] AND C. R. HEATHCOTE[2]

[1]*School of Mathematical and Physical Sciences, Murdoch University, Western Australia 6150, Australia*
[2]*Department of Statistics, Australian National University, GPO Box 4, Canberra ACT 2601, Australia*

(Received June 25, 1992; revised March 25, 1993)

Abstract. The estimating equations derived from minimising a L_2 distance between the empirical distribution function and the parametric distribution representing a mixture of k normal distributions with possibly different means and/or different dispersion parameters are given explicitly. The equations are of the M estimator form in which the ψ function is smooth, bounded and has bounded partial derivatives. As a consequence it is shown that there is a solution of the equations which is robust. In particular there exists a weakly continuous, Fréchet differentiable root and hence there is a consistent root of the equations which is asymptotically normal. These estimating equations offer a robust alternative to the maximum likelihood equations, which are known to yield nonrobust estimators.

Key words and phrases: Influence function, weak continuity, mixtures of normals, Fréchet differentiability, consistency, asymptotic normality, selection functional, minimum distance estimator.

1. Introduction

There is a substantial literature concerning the estimation of parameters in a mixture of normal distributions

$$F(x;\boldsymbol{\theta}) = \sum_{j=1}^{k} \epsilon_j \Phi\{(x-\mu_j)/\sigma_j\}. \tag{1.1}$$

Here

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)dy, \qquad \phi(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2),$$

$\sum_{j=1}^{k} \epsilon_j = 1$, and $\boldsymbol{\theta} \in \Theta$ is the vector of the $3k-1$ parameters $\epsilon_1, \ldots, \epsilon_{k-1}$, $\mu_1, \ldots, \mu_k$, $\sigma_1, \ldots, \sigma_k$ which are to be estimated on the basis of the sample $X_1, X_2,$

$\ldots, X_n$. We assume that mixing proportions are positive and component distributions are distinct. Comprehensive accounts of mixtures can be found in Everitt and Hand (1981), and Titterington *et al.* (1985) whilst McLachlan and Basford (1988) pay particular attention to the fitting of finite mixture distributions.

Our concern is with the robust estimation of $\boldsymbol{\theta}$ using a combination of the methods of Clarke (1983, 1989) and Heathcote and Silvapulle (1981). In particular $\boldsymbol{\theta}$ will be estimated by minimising

$$J_n(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \{F_n(x) - F(x;\boldsymbol{\theta})\}^2 dx, \tag{1.2}$$

where $F_n(x)$ is the empirical distribution function of the sample. It will be shown that this yields a bounded influence function estimator that is weakly continuous and Fréchet differentiable. Consequently there is a robust estimator minimising (1.2), obtained by solving (2.7) below, that is consistent and asymptotically normally distributed.

Quadratic distances of the form (1.2) seem to have been first used generally for parametric estimation by Knüsel (1969) and in the case of mixture distributions by Choi and Bulgren (1968). A succinct discussion is given in Section 4.5 of Titterington *et al.* (1985). It is easily checked that maximising the likelihood in (1.1) leads to estimating equations that contain unbounded terms as well as being analytically complicated whereas minimum distance estimation has advantages in both respects.

2. The estimating equations

Minimising $J_n(\boldsymbol{\theta})$ of (1.2) leads to

$$(\partial/\partial\boldsymbol{\theta})J_n(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} -2(F_n(x) - F(x;\boldsymbol{\theta}))\{(\partial/\partial\boldsymbol{\theta})F(x;\boldsymbol{\theta})\}dx = 0, \tag{2.1}$$

which, on integrating by parts, is

$$n^{-1}\sum_{l=1}^{n}\int_{-\infty}^{X_l}\{(\partial/\partial\boldsymbol{\theta})F(y;\boldsymbol{\theta})\}dy - \int_{-\infty}^{\infty}\int_{-\infty}^{x}\{(\partial/\partial\boldsymbol{\theta})F(y;\boldsymbol{\theta})\}dydF(x;\boldsymbol{\theta}) = 0. \tag{2.2}$$

It is convenient to first consider the $k-1$ derivatives with respect to the mixing proportions

$$(\partial/\partial\epsilon_i)F(x;\boldsymbol{\theta}) = \Phi\{(x-\mu_i)/\sigma_i\} - \Phi\{(x-\mu_k)/\sigma_k\}, \qquad i = 1, \ldots, k-1.$$

Let

$$A_{ik}(x,\boldsymbol{\theta}) = \int_{-\infty}^{x}(\partial/\partial\epsilon_i)F(y;\boldsymbol{\theta})dy,$$

which is

$$\begin{aligned} A_{ik}(x,\boldsymbol{\theta}) = (x-\mu_k)\{\Phi\{(x-\mu_i)/\sigma_i\} - \Phi\{(x-\mu_k)/\sigma_k\}\} + \sigma_i\phi\{(x-\mu_i)/\sigma_i\} \\ - \sigma_k\phi\{(x-\mu_k)/\sigma_k\} + (\mu_k-\mu_i)\Phi\{(x-\mu_i)/\sigma_i\}. \end{aligned}$$

The identities

$$\int_{-\infty}^{\infty} \phi\{(x-\mu_i)/\sigma_i\}\phi\{(x-\mu_j)/\sigma_j\}dx = \left\{(\sigma_i\sigma_j)/\sqrt{\sigma_i^2+\sigma_j^2}\right\}\phi\left\{(\mu_i-\mu_j)/\sqrt{\sigma_i^2+\sigma_j^2}\right\}, \tag{2.3}$$

$$\int_{-\infty}^{\infty} \frac{1}{\sigma_j}\phi\{(x-\mu_j)/\sigma_j\}\Phi\{(x-\mu_i)/\sigma_i\}dx = \Phi\left\{(\mu_j-\mu_i)/\sqrt{\sigma_i^2+\sigma_j^2}\right\} \tag{2.4}$$

yield the following result concerning derivatives with respect to $\epsilon_1, \ldots, \epsilon_{k-1}$:

LEMMA 2.1.

$$\begin{aligned} B(\boldsymbol{\theta}; j, i, k) &= \int_{-\infty}^{\infty} \sigma_j^{-1}\phi\{(x-\mu_j)/\sigma_j\}A_{ik}(x,\boldsymbol{\theta})dx \\ &= (\mu_j-\mu_k)\left\{\Phi\left\{(\mu_j-\mu_i)/\sqrt{\sigma_j^2+\sigma_i^2}\right\}\right. \\ &\qquad \left.-\Phi\left\{(\mu_j-\mu_k)/\sqrt{\sigma_j^2+\sigma_k^2}\right\}\right\} \\ &\quad + \sqrt{\sigma_j^2+\sigma_i^2}\,\phi\left\{(\mu_j-\mu_i)/\sqrt{\sigma_j^2+\sigma_i^2}\right\} \\ &\quad - \sqrt{\sigma_j^2+\sigma_k^2}\,\phi\left\{(\mu_j-\mu_k)/\sqrt{\sigma_j^2+\sigma_k^2}\right\} \\ &\quad + (\mu_k-\mu_i)\Phi\left\{(\mu_j-\mu_i)/\sqrt{\sigma_j^2+\sigma_i^2}\right\}. \end{aligned}$$

Thus for instance (2.2) will involve terms

$$E_{\boldsymbol{\theta}}[A_{ik}(x,\boldsymbol{\theta})] = \sum_{j=1}^{k-1}\epsilon_j[B(\boldsymbol{\theta}; j, i, k) - B(\boldsymbol{\theta}; k, i, k)] + B(\boldsymbol{\theta}; k, i, k).$$

Terms involving partial derivatives with respect to μ_i use (2.4) to give

$$\begin{aligned} &E_{\boldsymbol{\theta}}\left[\int_{-\infty}^{x}(\partial/\partial\mu_i)F(y;\boldsymbol{\theta})dy\right] \\ &\quad = \epsilon_i\left[-\Phi\left\{(\mu_k-\mu_i)/\sqrt{\sigma_k^2+\sigma_i^2}\right\}\right. \\ &\qquad + \sum_{j=1}^{k-1}\epsilon_j\left[\Phi\left\{(\mu_k-\mu_i)/\sqrt{\sigma_k^2+\sigma_i^2}\right\}\right. \\ &\qquad \left.\left.-\Phi\left\{(\mu_j-\mu_i)/\sqrt{\sigma_j^2+\sigma_i^2}\right\}\right]\right] \\ &\quad = \epsilon_i C_{i,k}(\boldsymbol{\theta}) \quad \text{say.} \end{aligned} \tag{2.5}$$

Similarly it follows from (2.3) that terms involving partial derivatives with respect to σ_i lead to

$$
\begin{aligned}
(2.6) \qquad & E_{\boldsymbol{\theta}}\left[\int_{-\infty}^{x}(\partial/\partial\sigma_i)F(y;\boldsymbol{\theta})dy\right] \\
&= \epsilon_i\left[\left\{\sigma_i/\sqrt{\sigma_i^2+\sigma_k^2}\right\}\phi\left\{(\mu_i-\mu_k)/\sqrt{\sigma_i^2+\sigma_k^2}\right\}\right. \\
&\qquad + \sum_{j=1}^{k-1}\epsilon_j\left[\left\{\sigma_i/\sqrt{\sigma_i^2+\sigma_j^2}\right\}\phi\left\{(\mu_i-\mu_j)/\sqrt{\sigma_i^2+\sigma_j^2}\right\}\right. \\
&\qquad \left.\left. - \left\{\sigma_i/\sqrt{\sigma_i^2+\sigma_k^2}\right\}\phi\left\{(\mu_i-\mu_k)/\sqrt{\sigma_i^2+\sigma_k^2}\right\}\right]\right] \\
&= \epsilon_i D_{i,k}(\boldsymbol{\theta}) \qquad \text{say.}
\end{aligned}
$$

Combining these calculations gives the estimating equations

$$
(2.7) \qquad n^{-1}\sum_{i=1}^{n}\psi(X_i;\boldsymbol{\theta}) = \mathbf{0},
$$

details of which are as follows:

THEOREM 2.1. *Let* $\psi(x;\boldsymbol{\theta}) = (\psi_1(x;\boldsymbol{\theta}),\ldots,\psi_{3k-1}(x;\boldsymbol{\theta}))'$ *be given by*

$$
\begin{aligned}
&\psi_i(x;\boldsymbol{\theta}) = A_{ik}(x,\boldsymbol{\theta}) - B(\boldsymbol{\theta};k,i,k) - \sum_{j=1}^{k-1}\epsilon_j\{B(\boldsymbol{\theta};j,i,k) - B(\boldsymbol{\theta};k,i,k)\}; \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1,\ldots,k-1, \\
&\psi_{i'}(x;\boldsymbol{\theta}) = -\Phi\{(x-\mu_i)/\sigma_i\} - C_{i,k}(\boldsymbol{\theta}); \qquad i' = k,\ldots,2k-1; \quad i = i'-k+1, \\
&\psi_{i'}(x;\boldsymbol{\theta}) = \phi\{(x-\mu_i)/\sigma_i\} - D_{i,k}(\boldsymbol{\theta}); \qquad i' = 2k,\ldots,3k-1; \quad i = i'-2k+1.
\end{aligned}
$$

Then the estimator of $\boldsymbol{\theta}$ *obtained by minimising* $J_n(\boldsymbol{\theta})$ *of* (1.2) *satisfies* (2.7).

3. Boundedness of the influence function

As in the development of Hampel *et al.* (1986) let $T(F_n)$ be the functional estimating the parameter $T(F_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$. Dependence on the parameter $\boldsymbol{\theta}$ will be indicated by writing $F = F_{\boldsymbol{\theta}}$. Then in a standard notation the influence function is

$$
IF(x,F_{\boldsymbol{\theta}}) = \lim_{\epsilon\downarrow 0}[T\{(1-\epsilon)F_{\boldsymbol{\theta}} + \epsilon\delta_x\} - T(F_{\boldsymbol{\theta}})]/\epsilon,
$$

which (see p. 230 of Hampel *et al.*) in our case can be written as

$$
(3.1) \qquad IF(x,F_{\boldsymbol{\theta}}) = -M(\boldsymbol{\theta})^{-1}\psi(x;\boldsymbol{\theta}),
$$

where $M(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{(\partial/\partial\boldsymbol{\theta})\psi(X;\boldsymbol{\theta})\}$. Clearly the function $\psi(x;\boldsymbol{\theta})$ of Theorem 2.1 is bounded in the observation space variable for any $\boldsymbol{\theta}$ that is the parameter vector

of a non-degenerate mixture distribution. Consequently the influence function of (3.1) is bounded, provided the matrix $M(\boldsymbol{\theta})$ is nonsingular.

LEMMA 3.1. *Assume that there does not exist a nonzero vector* $\boldsymbol{b}' = (b_1, \ldots, b_{3k-1})$ *such that*

$$\boldsymbol{b}'(\partial/\partial\boldsymbol{\theta})F(x;\boldsymbol{\theta}) = \sum_{l=1}^{3k-1} b_l(\partial/\partial\theta_l)F(x;\boldsymbol{\theta}) = 0 \quad \text{for every} \quad x \in (-\infty, \infty). \tag{3.2}$$

Then for $\boldsymbol{\theta} \in \Theta$ *the matrix* $M(\boldsymbol{\theta})$ *is nonsingular when* ψ *is given by Theorem* 2.1.

PROOF. From the construction of $\psi(x;\boldsymbol{\theta})$ see that

$$\det\left[E_{\boldsymbol{\theta}}\left\{(\partial^2/\partial\boldsymbol{\theta}^2)\frac{1}{2}J_n(\boldsymbol{\theta})\right\}\right] = \epsilon_1^2 \cdots \epsilon_k^2 \det[M(\boldsymbol{\theta})],$$

where det denotes determinant. The matrix on the left has elements

$$\lambda_{lm}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \{(\partial/\partial\theta_l)F(y;\boldsymbol{\theta})\}\{(\partial/\partial\theta_m)F(y;\boldsymbol{\theta})\}dF(y;\boldsymbol{\theta}).$$

Given an arbitrary nonzero vector $\boldsymbol{b}$ there is an x for which

$$|\boldsymbol{b}'(\partial/\partial\boldsymbol{\theta})F(x;\boldsymbol{\theta})| = \eta(x;\boldsymbol{\theta}) > 0,$$

by condition (3.2). The function $\eta(x;\boldsymbol{\theta})$ is continuous and there exists a $\delta > 0$ such that $\eta(y;\boldsymbol{\theta}) \geq \eta(x;\boldsymbol{\theta})/2$ for $y \in [x-\delta, x+\delta]$. Then setting $\Lambda(\boldsymbol{\theta}) = (\lambda_{l,m}(\boldsymbol{\theta}))$ it follows that

$$\begin{aligned}\boldsymbol{b}'\Lambda(\boldsymbol{\theta})\boldsymbol{b} &= \int_{-\infty}^{\infty} |\boldsymbol{b}'(\partial/\partial\boldsymbol{\theta})F(y;\boldsymbol{\theta})|^2 dF(y;\boldsymbol{\theta}) \\ &\geq (\eta(x,\boldsymbol{\theta})/2)^2\{F(x+\delta;\boldsymbol{\theta}) - F(x-\delta;\boldsymbol{\theta})\} > 0.\end{aligned}$$

This implies the matrix $\Lambda(\boldsymbol{\theta})$ is positive definite, and consequently $M(\boldsymbol{\theta})$ is nonsingular.

Condition (3.2) requires in particular that no two component distributions $\Phi\{(x-\mu_i)/\sigma_i\}$ be the same. It is perhaps stronger than but linked to identifiability. Discussion of identifiability for finite mixtures can be found in Section 1.5 of McLachlan and Basford (1988) and Section 3.1 of Titterington *et al.* (1985). On a related issue Aitkin and Rubin (1985) invoke an ordering $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_k$ as a way of identifying the k components to avoid a lack of identifiability caused by permutations of the components.

4. Weak continuity and Fréchet differentiability

There may be more than one solution of equations (2.7). A more precise definition of the functional T is then achieved by invoking a selection functional $\rho_0(G,\tau)$ defined on $\mathcal{G}\times\Theta$. To do this, for any distribution function $G\in\mathcal{G}$, let $S(\psi,G)$ be the set of solutions of

$$\int_{-\infty}^{\infty}\psi(x,\tau)dG(x)=0. \tag{4.1}$$

The functional T, which depends on both ψ and ρ_0, is defined via

$$\inf_{\tau\in S(\psi,G)}\rho_0(G,\tau)=\rho_0(G,T[\psi,\rho_0,G]).$$

If $S(\psi,G)$ is empty then $T[\psi,\rho_0,G]=+\infty$. The selection functional discussed in Clarke (1991) may be the loss function from which equations (2.7) are derived or some auxiliary criteria. The L_2-distance estimator is achieved by choosing

$$\rho_0(G,\tau)=\int_{-\infty}^{\infty}\{G(x)-F(x;\tau)\}^2dx,$$

and ψ given by Theorem 2.1. This choice of selection functional is not easily computed, even in the case of $\rho_0(F_n,\tau)=J_n(\boldsymbol{\theta})$. But it is not necessary to choose ρ_0 such that $\psi(x,\tau)=(\partial/\partial\tau)\rho_0(G,\tau)$ as above. For example, one of the multiple roots (if this should be the case) of moment equations of the form (2.7) is selected on the basis of alternative sample based criteria. This is done by several authors listed on p. 76 of Titterington *et al.* (1985). Criteria for selection functionals include relative closeness of fitted higher moments to the sample versions, χ^2 goodness of fit statistics, and the likelihood.

Moreover local asymptotic theory for the root of (2.7) can be established using the auxiliary selection functional $\rho(G,\tau)=\|\tau-\boldsymbol{\theta}\|$ where $\|\cdot\|$ is the usual Euclidean norm. Using ψ defined by Theorem 2.1 it is shown that $T[\psi,\rho,F_n]$ is an estimator satisfying Hampel's (1971) definition of robustness. It is enough to show $T[\psi,\rho,\cdot]$ is continuous at $F_{\boldsymbol{\theta}}$ with respect to the Prokhorov metric, given by $d_p(F,G)=\inf\{\delta : F(A)\le G(A^\delta)+\delta \text{ for all events } A\}$, where A^δ is the set of all points whose distance from A is less than δ.

THEOREM 4.1. *Let $\psi(x;\boldsymbol{\theta})$ be given by Theorem 2.1 and assume $M(\boldsymbol{\theta})$ is nonsingular for a given $\boldsymbol{\theta}\in\Theta$. Given $\kappa>0$ there exists a $\delta>0$ such that $d_p(F_{\boldsymbol{\theta}},G)<\delta$ implies $T[\psi,\rho,G]$ exists and is such that $\|T[\psi,\rho,G]-\boldsymbol{\theta}\|<\kappa$. Further for this $\delta>0$ there is a $\kappa^*>0$ such that if $U_{\kappa^*}(\boldsymbol{\theta})\subset\Theta$ is the ball of radius κ^* about $\boldsymbol{\theta}$, then $S(\psi,G)\cap U_{\kappa^*}(\boldsymbol{\theta})=T[\psi,\rho,G]$. For any null sequence of positive numbers $\{\delta_k\}$ let G_k be an arbitrary sequence for which $d_p(F_{\boldsymbol{\theta}},G_k)\le\delta_k$. Then*

$$\lim_{k\to\infty}T[\psi,\rho,G_k]=T[\psi,\rho,F_{\boldsymbol{\theta}}]=\boldsymbol{\theta}.$$

Theorem 4.1 follows from Theorem 3.2 of Clarke (1983). We need to show Conditions A(1–4) of that paper are satisfied for the particular choice of ψ. Condition

A_1 requires $E_{\boldsymbol{\theta}}[\psi(X;\boldsymbol{\theta})] = 0$. This is seen by taking expectations of equations (2.1), and interchanging integration and expectation. Condition A_2 is satisfied since it is easily checked that both $\psi(x,\tau)$ and $(\partial/\partial\tau)\psi(x,\tau)$ are uniformly bounded functions on some compact set D contained in the interior of Θ and containing $\boldsymbol{\theta}$ in its interior. A_3 requires $M(\boldsymbol{\theta})$ to be nonsingular. Remark 6.2 of Clarke (1983) guarantees Condition A_4 holds with respect to the Prokhorov metric.

Theorem 4.1 gives weak continuity and consequently robustness of the functional $T[\psi,\rho,\cdot]$. In addition it specifies uniqueness of a solution of equations (4.1) in a region $U_{\kappa^*}(\boldsymbol{\theta})$ for small enough Prokhorov neighbourhoods of $F_{\boldsymbol{\theta}}$. An immediate corollary implies existence of a unique consistent root of the equations (2.7) since Varadarajan (1958) proves that F_n converges weakly to $F_{\boldsymbol{\theta}}$ almost surely, whereupon a result of Prokhorov (1956) gives that $d_p(F_n, F_{\boldsymbol{\theta}}) \to 0$ almost surely.

COROLLARY 4.1. *Let $X_1,\ldots,X_n$ be independent identically distributed according to the distribution $F_{\boldsymbol{\theta}}$ given by* (1.1). *Then there exists a $\kappa^* > 0$ such that $T[\psi,\rho,F_n]$ exists and is unique in $U_{\kappa^*}(\boldsymbol{\theta})$. Moreover*

$$\|T[\psi,\rho,F_n] - \boldsymbol{\theta}\| \to 0 \qquad \textit{almost surely.}$$

Kiefer (1978) proved a similar local convergence result for a solution of the likelihood equations though such a solution does not enjoy properties of robustness.

Theorem 1 of Hampel (1971) says that the weak continuity of the functional $T[\psi,\rho,\cdot]$ at a distribution $F_{\boldsymbol{\theta}}$ together with the continuity of each T_n (considered as a function of the sample) implies the robustness of $\{T_n\}$ at $F_{\boldsymbol{\theta}}$. Since $T_n = T[\psi,\rho,F_n]$ satisfies (2.7) it follows that it is a continuous function of the observations whenever

$$M(\tau,F_n)|_{\tau=T[\psi,\rho,F_n]} = \int (\partial/\partial\tau)\psi(x,\tau)dF_n(x)|_{\tau=T[\psi,\rho,F_n]} \tag{4.2}$$

is nonsingular. This follows even if F_n is generated from a G in a small Prokhorov neighbourhood of $F_{\boldsymbol{\theta}}$ (Prokhorov (1956)) since $d_p(F_n,G) \to 0$ almost surely. By the triangle inequality this implies that for all sufficiently large n, $d_p(F_n,F_{\boldsymbol{\theta}})$ is small and then the weak continuity of $T[\psi,\rho,\cdot]$ and Lemma 3.1 of Clarke (1983) imply that the matrix (4.2) is nonsingular. Note that in addition to the continuity of T with respect to the Prokhorov distance we have the differentiability of T with respect to the Kolmogorov distance.

THEOREM 4.2. *Let ψ be given by Theorem* 2.1 *and assume $M(\boldsymbol{\theta})$ is nonsingular at $\boldsymbol{\theta} \in \Theta$. Then $T[\psi,\rho,\cdot]$ is Fréchet differentiable at $F_{\boldsymbol{\theta}}$ with respect to the Kolmogorov distance metric on $\mathcal{G}$. That is*

$$\|T[G] - T[F_{\boldsymbol{\theta}}] - T'_F(G - F_{\boldsymbol{\theta}})\| = o(d_k(F_{\boldsymbol{\theta}},G))$$

as $d_k(F_{\boldsymbol{\theta}},G) = \sup_{-\infty<x<\infty} |F_{\boldsymbol{\theta}}(x) - G(x)| \to 0$. The derivative is given by

$$T'_{F_{\boldsymbol{\theta}}}(G - F) = -M(\boldsymbol{\theta})^{-1} \int_{-\infty}^{\infty} \psi(x;\boldsymbol{\theta}) d(G - F_{\boldsymbol{\theta}})(x).$$

Since Conditions A of Clarke (1983) are met with respect to the Kolmogorov metric and $\psi(x;\boldsymbol{\theta})$ is a function of total bounded variation in the observation space variable (it is a linear combination of exponential densities and normal distributions) this result follows from Theorem 5.1 of that paper. Note that the influence function (3.1) is obtained by evaluating the Fréchet derivative $T'_{F_{\boldsymbol{\theta}}}$ at the difference $\delta_x - F_{\boldsymbol{\theta}}$. From Section 2.5 of Huber (1981) and noting that $\sqrt{n}$ $o(d_k(F_n, F_{\boldsymbol{\theta}})) = o_p(1)$ as a consequence of the Dvoretzky-Kiefer-Wolfowitz inequality (Theorem 2.1.3A of Serfling (1980)), the following corollary to Theorem 4.2 results.

COROLLARY 4.2. *Assume the conditions of Corollary* 4.1. *Then* $\sqrt{n}(T[\psi, \rho, F_n] - \boldsymbol{\theta})$ *converges in distribution to a multivariate normal random variable with mean zero and variance covariance matrix* $\sigma^2(T, F_{\boldsymbol{\theta}})$ *where*

$$\sigma^2(T, F_{\boldsymbol{\theta}}) = M(\boldsymbol{\theta})^{-1} \int_{-\infty}^{\infty} \psi(x;\boldsymbol{\theta})\psi(x;\boldsymbol{\theta})' dF_{\boldsymbol{\theta}}(x) \{M(\boldsymbol{\theta})^{-1}\}'.$$

Here integration is carried out componentwise.

5. Simulation results

We report a small scale simulation that illustrates the performance of the L_2 estimator of the mixing parameter ϵ. The case considered was (1.1) with $k = 2$ and all five parameters to be estimated. One hundred samples each of size $n = 200$ were generated from seven parent distributions whose parameters are listed in Table 1. These are the same parameter configurations as those used by Clarke (1989) when ϵ was the only parameter to be estimated. Table 1 gives the mean squared errors over the 100 samples for the L_2 estimate ϵ_n^* and for the maximum likelihood estimate $\hat{\epsilon}_n$. In both cases, (2.7) with the appropriate choice of ψ was solved iteratively using a routine written by Powell and described in Rabinowitz (1970). Non convergent attempts at solution were discarded and it is believed that this has not favoured either of the two procedures. There were more non convergent cases when using maximum likelihood, even if the initial values were the true parameter values. This indicates the relative stability of the L_2 estimator. The parameter sets were chosen as in Clarke (1989) and the first six were characterised by having mixture densities whose components are close together in the sense that the difference in means $\mu_1 - \mu_2$ is small or moderate relative to the standard deviations. For such parameter configurations Table 1 illustrates the generally superior performance of the L_2 estimator of ϵ, dominating maximum likelihood in five of the sets and being comparable in the others. With wider separation maximum likelihood is expected to dominate and this is supported, though marginally, by parameter set (7).

The relatively good performance of the L_2 estimator was maintained when estimating the other components of the parameter vector. Briefly, the average over the seven parameter sets of the relative mean squared error of the L_2 estimator to the maximum likelihood estimator was found to be: 0.60 for μ_1, 0.50 for μ_2, 0.86 for σ_1 and 1.31 for σ_2.

Table 1. Mean squared errors of ϵ_n^* and $\hat{\epsilon}_n$ from 100 samples each of $n = 200$ from (1.1) with $k = 2$ and seven parameter sets as listed.

	Parameter set						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ϵ	0.75	0.50	0.75	0.50	0.50	0.50	0.50
μ_1	−1.00	−1.00	−1.00	−1.00	0.00	−1.50	−5.00
μ_2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
σ_1	1.00	1.00	2.00	1.00	1.00	1.00	1.00
σ_2	1.00	2.00	1.00	1.00	2.00	1.00	2.00
	Mean squared errors						
ϵ_n^*	0.0147	0.0266	0.0261	0.0189	0.0581	0.0220	0.0012
$\hat{\epsilon}_n$	0.0278	0.0325	0.0518	0.0639	0.0441	0.0504	0.0011

Table 2. Mean squared errors of ϵ_n^* and $\hat{\epsilon}_n$ from 100 samples each of $n = 200$ from mis-specified (1.1) with a mixture of $k = 2$ $t(5)$ distributions in place of the mixture of normals. Parameter sets are as in Table 1.

	Parameter set						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean squared errors						
ϵ_n^*	0.0113	0.0259	0.0179	0.0356	0.0379	0.0289	0.0005
$\hat{\epsilon}_n$	0.0423	0.0495	0.0396	0.1544	0.0497	0.0927	0.0013

One advantage of using ϵ_n^* instead of the maximum likelihood estimator is the absence of singularities, and attendant numerical problems, as may occur in the likelihood surface (Titterington *et al.* (1985)). Robustness considerations are also relevant. Suppose for example that data originated from a mixture of t distributions rather than the purported mixture of normals. Table 2 gives the mean squared errors of ϵ_n^* and $\hat{\epsilon}_n$ for a mixture of $k = 2$ t distributions with 5 degrees of freedom and means and variances in the seven cases as for Table 1. Clearly the L_2 estimator out performs the maximum likelihood estimator in this case.

6. Conclusion

It is observed in Clarke and Heathcote (1978) that many minimum distance estimators can be employed in estimating parameters in mixtures of normal distributions. The particular L_2 distance, $J_n(\boldsymbol{\theta})$, yields M-estimating equations with bounded ψ functions for the model (1.1). These estimators have reasonably high relative efficiency when estimating parameters from related distributions as is noted in Heathcote and Silvapulle (1981) and Clarke (1989). There is also

the advantage over characteristic function methods, as proposed by Bryant and Paulson (1983), in not requiring knowledge of parameters θ to choose optimal weighting functions. This would be cumbersome when attempting to estimate all components of θ simultaneously. The estimator proposed in this paper has obvious advantages over the moment generating function estimator of Quandt and Ramsey (1978), as discussed in Clarke (1989).

A quantity that is easily calculated is the Cramér-Von Mises distance

$$\rho_0(F_n, \tau) = \int_{-\infty}^{\infty} \{F_n(x) - F_\tau(x)\}^2 dF_\tau(x) \tag{6.1}$$
$$= n^{-1} \sum_{i=1}^{n} \left[F_\tau(X_{(i)}) - \left\{ \left(i - \frac{1}{2} \right) \Big/ n \right\} \right]^2 + (12n^2)^{-1},$$

where $X_{(i)}$ is the i-th order statistic. Woodward *et al.* (1984) minimised this distance to estimate the parameters of a mixture of two normal distributions and showed through simulations that the estimator performed well when component populations deviated from normality. This distance is well known for its role as a goodness of fit statistic, as in D'Agostino and Stephens (1986). For these reasons we advocate the combination of the ψ-function of Theorem 2.1 and the selection statistic of (6.1) in determining parameters. Theorem 4.1 and Corollary 4.1 indicate existence of a robust unique consistent root of equations (2.7) in a local neighbourhood of θ. Numerical nonlinear equation solving routines can be employed starting from suitable initial estimates from within that neighbourhood. If multiple roots are determined from several searches then the root which minimises (6.1) is selected as the estimator. This avoids the problem of calculating the distance $J_n(\theta)$ which may also be sensitive to outlying observations.

References

Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models, *J. Roy. Statist. Soc. Ser. B*, **47**, 67–75.

Bryant, J. L. and Paulson, A. S. (1983). Estimation of mixing proportions via distance between characteristic functions, *Comm. Statist. Theory Methods*, **12**, 1009–1029.

Choi, K. and Bulgren, W. B. (1968). An estimation procedure for mixtures of distributions, *J. Roy. Statist. Soc. Ser. B*, **30**, 444–460.

Clarke, B. R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations, *Ann. Statist.*, **11**, 1196–1205.

Clarke, B. R. (1989). An unbiased minimum distance estimator of the proportion parameter in a mixture of two normal distributions, *Statist. Probab. Lett.*, **7**, 275–281.

Clarke, B. R. (1991). The selection functional, *Probab. Math. Statist.*, **11**, *Fasc.*, **2**, 149–156.

Clarke, B. R. and Heathcote, C. R. (1978). Comment on "Estimating mixtures of normal distributions and switching regressions" by Quandt and Ramsey, *J. Amer. Statist. Assoc.*, **73**, 749–750.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-Fit Techniques*, Marcel Dekker, New York.

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

Hampel, F. R. (1971). A general qualitative definition of robustness, *Ann. Math. Statist.*, **42**, 1887–1896.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

Heathcote, C. R. and Silvapulle, M. J. (1981). Minimum mean squared estimation of location and scale parameters under misspecification of the model, *Biometrika*, **68**, 501–514.

Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.

Kiefer, N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model, *Econometrica*, **46**, 427–434.

Knüsel, L. F. (1969). Über Minimum-Distance-Schätzungen, Ph.D. Thesis, ETH, Zürich.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models, Inference and Applications to Clustering*, Marcel Dekker, New York.

Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory, *Theory Probab. Appl.*, **1**, 157–214.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions, *J. Amer. Statist. Assoc.*, **73**, 730–738.

Rabinowitz, P. (ed.) (1970). *Numerical Methods for Non-Linear Algebraic Equations*, Gordon and Breach, London.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

Varadarajan, V. S. (1958). On the convergence of probability distributions, *Sankhyā*, **19**, 23–26.

Woodward, W. A., Parr, W. C., Schucany, W. R. and Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion, *J. Amer. Statist. Assoc.*, **79**, 590–598.