Statistics & Probability Letters 7 (1989) 275–281
North-Holland February 1989

AN UNBIASED MINIMUM DISTANCE ESTIMATOR OF THE PROPORTION PARAMETER IN A MIXTURE OF TWO NORMAL DISTRIBUTIONS

Brenton R. CLARKE

School of Mathematical and Physical Sciences, Murdoch University, Murdoch, Western Australia

Received March 1988
Revised May 1988

Abstract An estimator that minimizes an L_2 distance used in studies of estimation of the location parameter is shown here to give an explicit formulation for the estimator of proportion in a mixture of two normal distributions when other parameters are known. This can prove to be an advantage over other minimum distance methods and the maximum likelihood estimator. Monte Carlo simulation demonstrates this and highlights good small sample behaviour of the estimator. It is shown that the estimator is also qualitatively robust both empirically and asymptotically, the latter being evidenced by the existence of a Fréchet derivative.

AMS 1980 Subject Classifications: Primary 62F10, Secondary 62E25.

Keywords: mixtures of two normal distributions, unbiased estimator, Monte Carlo simulation, minimum distance estimation, Fréchet derivative.

1. Introduction

The well known problem of singularities in the likelihood surface, described in Example 4.3.2 of Titterington, Smith and Makov (1985), was one of the reasons that prompted Choi and Bulgren (1968) to consider estimating parameters in the mixture of two normal distributions by minimizing a Cramér von Mises type distance with an empirical weight function. MacDonald (1971) observed that a reduction in small sample bias of the estimate of proportion could be achieved by minimizing the proper Cramér von Mises distance with the parametric distribution as the weight function, so that the estimator minimizes

$$n^{-1}\sum_{i=1}^{n}\left[F_\theta(x_i)-\frac{\left(i-\frac{1}{2}\right)}{n}\right]^2+\left(12n^2\right)^{-1} \tag{1.1}$$

Here x_i is the i'th order statistic. The advantages of minimizing such a distance are further outlined in a paper by Woodward et al. (1984). In that paper the above minimum distance estimator is shown to be more robust when the actual underlying models are mixtures of heavytailed densities such as exponentials or student-t's.

On the other hand Clarke and Heathcote (1978) noted that several distance estimators are applicable to the same problem. In particular, it transpires in the discussion of the present paper, that a particularly attractive distance useful in the estimation of the proportion parameter is

$$J_n(\theta)=\int_{-\infty}^{+\infty}\left(F_n(x)-F_\theta(x)\right)^2\,\mathrm{d}x, \tag{1.2}$$

where $F_n(x)$ is the empirical distribution function. This distance has already shown to give robust and

0167-7152/89/$3.50 © 1989, Elsevier Science Publishers B.V (North-Holland)

efficient estimators of location and scale for the normal parametric family in studies by Heathcote and Silvapulle (1981) and Boos (1981). It has an added advantage in the estimation of the proportion parameter alone since it gives an estimator with an explicit formulation, as opposed to the implicit formulations of the maximum likelihood estimator and most other forms of minimum distance estimator that are currently in the literature. As a consequence it is easily programmable. Moreover, the estimator of proportion is unbiased and can have high efficiencies in comparison to both the maximum likelihood estimators and its competitors. In simulation studies, for one parameter configuration where the true proportion parameter is near the boundary the minimum distance estimator of this paper remains stable, whereas the maximum likelihood estimator has markedly increased mean squared error. Further numerical results comparing the estimators of James (1978) show this estimator in a favourable light. It is noted also that the estimating functional is Fréchet differentiable using the results of Clarke (1983) and hence qualitatively robust. This is illustrated further in Figure 1 of chapter 4 in a comparison between this estimator and the form of estimator given in Quandt and Ramsey (1978).

2. Unbiasedness, consistency and asymptotic normality

The distribution of a mixture of two normal distributions can be conveniently written

$$F_\theta(x)=\varepsilon\Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right)+(1-\varepsilon)\Phi\left(\frac{x-\mu}{\sigma_1}\right),$$

$$\Phi(x)=\int_{-\infty}^{x}\phi(y)\,\mathrm{d}y,\qquad \phi(y)=\frac{1}{\sqrt{2\pi}}\exp(-y^2/2). \tag{2.1}$$

Here Δ represents the separation between the two component normal populations, which have means $\mu_1=\mu-\Delta$ and $\mu_2=\mu$ and variances σ_1^2, σ_2^2 respectively. The parameter $\theta=(\varepsilon,\ \mu,\ \Delta,\ \sigma_1,\ \sigma_2)$.

Given an independent identically distributed sample $X_1,\dots,X_n$ taken from the distribution (2.1), the estimator which is a solution of the minimizing equation for the distance (1.2) can be written as a function of the empirical distribution function, $T[F_n]$. For brevity this will be written T_n^*. The explicit form of T_n^* is shown in the appendix to be a solution of the minimizing equation which takes the form

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\ \theta)=0, \tag{2.2}$$

where $\psi(x,\ \theta)=\xi(x;\ \mu,\ \Delta,\ \sigma_1,\ \sigma_2)-\varepsilon$ and ξ does not depend on ε. Consequently, T_n^* has the explicit formulation

$$T_n^*=\frac{1}{n}\sum_{i=1}^{n}\xi(X_i;\ \mu,\ \Delta,\ \sigma_1,\ \sigma_2). \tag{2.3}$$

The form for $\xi(\cdot)$ is

$$\xi(x;\ \mu,\ \Delta,\ \sigma_1,\ \sigma_2)$$
$$=\frac{A(x;\ \mu,\ \Delta,\ \sigma_1,\ \sigma_2)-\Delta\Phi\left(\Delta/\sqrt{\sigma_i^2+\sigma_2^2}\right)-\sqrt{\sigma_1^2+\sigma_2^2}\,\phi\left(\Delta/\sqrt{\sigma_1^2+\sigma_2^2}\right)+\sigma_2/\sqrt{\pi}}{\Delta\left(1-2\Phi\left(\Delta/\sqrt{\sigma_1^2+\sigma_2^2}\right)\right)+(\sigma_1+\sigma_2)/\sqrt{\pi}-2\sqrt{\sigma_1^2+\sigma_2^2}\,\phi\left(\Delta/\sqrt{\sigma_1^2+\sigma_2^2}\right)},$$

where

$$A(x;\ \mu,\ \Delta,\ \sigma_1,\ \sigma_2) = (x-\mu)\left\{\Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) - \Phi\left(\frac{x-\mu}{\sigma_2}\right)\right\} + \sigma_1\phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right)$$
$$-\sigma_2\phi\left(\frac{x-\mu}{\sigma^2}\right) + \Delta\Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right).$$

The formula (2.3) is calculated quickly and efficiently using Hasting's approximation to calculate the cumulative normal distribution (see 26.2.17 in Abramowitz and Stegun, 1970). The formula is easily programmed and can be put onto the programmable pocket calculator. Distributional properties of the estimator are described by the following theorem.

Theorem. *For a random sample of independent random variables with a nondegenerate distribution* (2.1), *so that* $0 < \varepsilon < 1$ *and either or both of* $\Delta \neq 0$ *or* $\sigma_1 \neq \sigma_2$ *hold, the estimator* T_n^* *is an unbiased consistent estimator of* ε. *Moreover,* $\sqrt{n}\,(T_n^* - \varepsilon)$ *converges in distribution to a normal random variable with mean zero and finite variance given by* $n\ \mathrm{var}[T_n^*] = \mathrm{var}[\psi(X,\ \theta)] < \infty$.

Proof. It is shown in the appendix that the minimizing equation for the distance (1.2) gives a solution

$$T_n^* = \frac{\displaystyle\int_{-\infty}^{+\infty}\left(F_n(x) - \Phi\left(\frac{x-\mu}{\sigma_2}\right)\right)a(x-\mu)\ \mathrm{d}x}{\displaystyle\int_{-\infty}^{+\infty} a^2(x-\mu)\ \mathrm{d}x} \tag{2.4}$$

Here $a(x-\mu)$ is an abbreviation for

$$a(x-\mu;\ \Delta,\ \sigma_1,\ \sigma_2) = \Phi\left(\frac{x-\mu+\Delta}{\sigma_1}\right) - \Phi\left(\frac{x-\mu}{\sigma_2}\right).$$

The denominator of the expression (2.4) is equal to that in the expression for ξ. For a nondegenerate mixuture distribution it is therefore nonzero. Consequently T_n^* is the average of independent identically distributed bounded random variables $\xi(X_i;\ \mu,\ \Delta,\ \sigma_1,\ \sigma_2)$. Consider the summands of equation (2.3) for example. The strong law of large numbers implies that T_n^* converges almost surely to the expected value $E_\theta[\xi(X_i;\ \mu,\ \Delta,\ \sigma_1,\ \sigma_2)] < \infty$. Because expectation is a linear operator, this also is equal to $E_\theta[T_n^*]$. But this limit is also realized by noting from the Glivenko–Cantelli lemma that the empirical distribution function $F_n(x)$ converges almost surely to $F_\theta(x) = \varepsilon a(x-\mu) + \Phi((x-\mu)/\sigma_2)$ uniformly in x. By dominated convergence in (2.4) this limit is ε. Hence T_n^* has expectation $E_\theta[T_n^*] = \varepsilon$, and it is strongly consistent. Since ξ is bounded, both the variance of ξ and that of $\psi = \xi - \varepsilon$ are finite. Since also $\sqrt{n}\,(T_n^* - \varepsilon) = (1/\sqrt{n})\Sigma_{i=1}^n \psi(X_i;\ \theta)$ the usual central limit theorem for a normed sum of independent random variables completes the proof.

The form of equations (2.2) is exactly that studied by Clarke (1983). The function ψ is continuously differentiable in both parameters x and ε, satisfies a Fisher consistency requirement $E_\theta[\psi(X;\ \theta)] = 0$, and is bounded. Conditions A of that paper are fulfilled and consequently the estimating functional T is Fréchet differentiable at the parametric distribution F_θ. It is a robust estimator which can be written as an M-functional.

3. Numerical results

To test the performance of the estimator empirically several parameter configurations were chosen so that the component distributions were reasonably separated and yet they overlapped (to be slightly more precise, given independent random variables X, Y with respective distributions corresponding to the two component normal distributions which make up the mixture distribution in (2.1), then X, Y are overlapped whenever $P(X < Y)$ is neither too near zero nor too near one and var$[X]$ is of comparable magnitude to var$[Y]$). Such parameter sets are common in practice and enough accuracy of the estimator can be obtained without necessarily a large number of observations. For these configurations the asymptotic efficiency of the estimator T_n^* relative to the maximum likelihood estimator $\hat{T}_n$ is given as an illustration of the high asymptotic performance of T_n^* that can be obtained.

The performance of T_n^* is equivalent for any distribution specified from a particular configuration. A Monte Carlo study was carried out using 500 replications of data generated from a distribution in each configuration and mean squared errors of T_n^* and $\hat{T}_n$ were recorded. If either of T_n^* or the solution of the maximum likelihood equation obtained numerically was outside the region (0, 1), the estimator was truncated to either zero or one, whichever was the closest. Results are given in Table 2 along with the variance of T_n^*. The close approximation for the latter and observed mean squared errors tends to support the results of the simulation and suggests that the asymptotic distribution of T_n^* can reasonably be used for inference.

Mean squared errors of T_n^* were comparable with those of the maximum likelihood estimator apart from values of ε near the boundary. In the particular example from configuration (8) the maximum likelihood estimator suffered difficulties. Even though a smaller proportion of values were truncated to zero, 21% as opposed to 25% of the values of T_n^*, the maximum likelihood estimator exhibited a large positive small sample bias. The difference between the average value of the estimate and the true value of ε was 0.2413 as opposed to 0.0295 for T_n^*. The mean squared error was also remarkably greater. For other configurations ε is well away from the boundary. For these values the proportion of truncation was below 1%.

For component populations with variances that are approximately equal and means that are close together, it transpires that a large number of observations can be required to obtain reasonable accuracy for any estimate of ε (Hill 1963). For this reason James (1978) proposed estimators dependent on auxillary parameters, which are chosen to obtain the best efficiency that is possible for the estimator of his formulation. Given the auxillary parameters the estimators have the advantage of simplifying data collection. Nevertheless, T_n^* can be easily computed if the whole sample is recorded on the computer and is unlikely to be influenced by a few gross errors that are likely to occur with large data collections. Moreover T_n^* is explicit, obviating any need to choose other parameters, the optimal choice of which may intrinsically depend on a knowledge of the unknown parameter ε. The asymptotic efficiency of T_n^* was examined for comparison with that of estimators $\hat{p}_r$ and $\hat{p}_{st}$ of James (1978) using the 60 parameter configurations in Table 3 of that paper. In 34 cases T_n^* performed better than $\hat{p}_r$, performances more frequently being better in instances of small separation. In 15 of the 30 cases corresponding to

Table 1
Percentage relative asymptotic efficiencies of T_n^*

Parameters									
	Δ/σ_1	1.0	1.0	0.5	1.0	0.0	1.5	5.0	1.5
	σ_2/σ_1	1.0	2.0	0.5	1.0	2 0	1.0	2.0	2.0
	ε	0.75	0.5	0.75	0.5	0.5	0 5	0.5	0.1
Configuration		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Efficiency		93	95	88	99	98	100	74	87

Table 2
Mean squared errors in a Monte Carlo simulation of T_n^* and $\hat{T}_n$

Configuration	n	$\hat{T}_n$	T_n^*	var[T_n^*]	n	$\hat{T}_n$	T_n^*	var[T_n^*]
(1)	200	0.0050	0.0053	0 0045 : (6)	50	0.0075	0.0073	0.0131
(2)	200	0.0038	0.0056	0.0061 : (7)	50	0.0008	0.0013	0.0062
(3)	200	0.0061	0.0085	0.0076 : (8)	50	0.1842	0.0158	0.0223
(4)	50	0.0217	0.0209	0.0247 . (3)	50	0.0233	0 0280	0.0310
(5)	50	0.0271	0.0263	0 0708 : (2)	20	0.0491	0.0587	0.0608

approximately equal scale parameters, $\sigma_2/\sigma_1 = 0.9$ or 1.0, T_n^* performed better than $\hat{p}_{st}$ with s, t assumed to be optimal, that is dependent on ε.

4. Other minimum distance methods

The above empirical and asymptotic comparisons highlight the good performance of T_n^* against the estimator which is asymptotically efficient, namely $\hat{T}_n$. On the other hand several minimum distance estimators have been put forward as alternative candidates for estimation of parameters in a mixture of normal distributions since the early papers of Choi (1969) and Choi and Bulgren (1968), as is noted in the paper of Clarke and Heathcote (1978), and the review work of Titterington et al. (1985).

Amongst these, the Cramér von Mises estimator used by Macdonald (1971) and also Woodward et al. (1984) to estimate ε proved to be closest to T_n^* in terms of mean squared errors for the parameter configurations listed in Table 2. However, only in one instance, where the mean squared errors were almost equal, was T_n^* beaten. Notably though, the Cramér von Mises estimator is subject to small sample bias. For example, for the configuration (5) with $\Delta = 0$ the bias incurred was -0.041 ± 0.0016, where 0.0016 represents the accuracy of the Monte Carlo estimation. For the unbiased estimator T_n^* a value close enough to zero was recorded in the Monte Carlo simulation of 0.0038 ± 0.0009. Unlike T_n^* the Cramér von Mises estimator requires a numerical minimization algorithm to minimize (2.1).

Another distance that can give an explicit form of estimator for ε when minimized is the integrated squared error of Heathcote (1977) and Thornton and Paulson (1977). Bryant and Paulson (1983) perform simulations using a weight function $e^{-\alpha^2 u^2}$ for specific choices of $\sigma_1 = 1$ and various choices of σ_2, showing competitive performance, when the scaling factor α is chosen optimally as a function of σ^2. In practice, some guidance must be given to the choice of α. With a fixed value of $\alpha = 1$ this estimator proved inferior, mean squared errors being sensitive to the configuration (further tables are available in Clarke 1980). This phenomenon also occurred for weighting functions with discrete weight functions. In the case of an exponential weight function, the integrated squared error estimator coincides with the estimator of Kumar et al. (1979), setting the parameters $\eta_1 = \eta_2$. For $\eta_1 \neq \eta_2$ the estimator of that paper must be evaluated numerically.

In a different context to the study of the current paper where interest is in the simultaneous estimation of all 5 parameters in θ, it is made apparent in the survey work of Titterington et al. (1985) that the performance of estimators is both dependent on the theoretical formulation of the estimator and its numerical implementation. Analogous to the problem of multiple local maxima of the likelihood surface is the problem of multiple local minima of the distance function. Though this is the case for the minimization of (2.2) the author has shown in his Ph.D thesis that there exists a Fréchet differentiable root of the minimizing equations. Robustness of the corresponding estimator θ_n^* is a consequence and this can be reflected in numerical studies similar to those of Woodward et al. (1984). This contrasts with nonrobust minimum distance proposals such as those of Quandt and Ramsey (1978). As a quick empirical indicator of the robustness of T_n^* versus the estimator T_n^{mgf} obtained from minimizing the moment generating

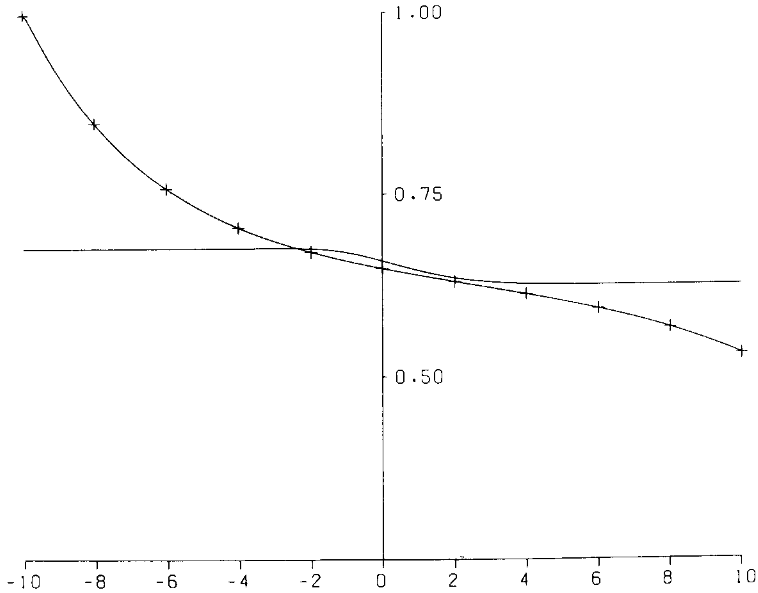

Fig. 1. Plots of influence on T_n^*; and T_n^{mgf}

function distance. Figure 1 exhibits plots of the estimators evaluated from a sample of $n = 51$ observations where 50 observations are selected randomly from the mixture distribution defined by $\theta = (0.5, 1.0, 2.0, 1.0, 2.0)$ and the 51st observation is given a value x. The plot of T_n^* compared with T_n^{mgf} shows how x has bounded influence on T_n^*, but can have disastrous consequences for the estimator T_n^{mgf}.

Acknowledgements

This paper was written while the author was on sabbatical leave and visiting the Department of Statistics, University of Edinburgh, and also the Fachgruppe für Statistik, Eidgenössische Technische Hochschule. Particular thanks are to Professor C.R. Heathcote for some early discussions.

References

Boos, D.D. (1981), Minimum distance estimators for location and goodness of fit. *J. Am Statist. Assoc* **76**, 663–670.

Bryant, J.L and A.S. Paulson (1983), Estimation of mixing proportions via distance between characteristic functions, *Commun. Statist Theor Math.* **12** (9), 1009–1029.

Choi, K. (1969), Estimators for the parameters of a finite mixture of distributions, *Ann Inst Statist. Math* **21**, 107–116.

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

Clarke, B.R. (1980), Robust estimation; limit theorems and their applications, Australian National University PhD thesis.

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

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

Heathcote, C.R. (1977), The integrated squared error estimation of parameters, *Biometrika* **64**, 255–264.

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

Hill, B.M. (1963), Information for estimating the proportions in mixtures of exponential and normal distributions, *J Am. Statist. Assoc* **58**, 918–932.

James, I.R. (1978), Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements, *Biometrics* **34**, 265–275.

Kumar, K.D., E.H. Nicklin and A.S. Paulson (1979), Comment on "Estimating mixtures of normal distributions and switching regressions" by Quandt and Ramsey, *J. Am. Statist. Assoc.* **74**, 52–55.

Macdonald, P.D.M. (1971), Comment on "An estimation procedure for mixtures of distributions" by Choi and Bulgren, *J.R. Statist. Soc. Ser. B* **33**, 326–329.

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

Thornton, J.C., and A.S. Paulson (1977), Asymptotic distribution of characteristic functions based estimators for the stable laws, *Sankhyā A* **39**, 341–354.

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

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