Noname manuscript No.
(will be inserted by the editor)

A Comparison of the L_2 Minimum Distance Estimator and the EM-Algorithm when Fitting *k*-Component Univariate Normal Mixtures

Brenton R. Clarke · Thomas Davidson · Robert Hammarstrand

the date of receipt and acceptance should be inserted later

Abstract The method of maximum likelihood using the EM-algorithm for fitting finite mixtures of normal distributions is the accepted method of estimation ever since it has been shown to be superior to the method of moments. Recent books testify to this. There has however been criticism of the method of maximum likelihood for this problem, the main criticism being when the variances of component distributions are unequal the likelihood is in fact unbounded and there can be multiple local maxima. Another major criticism is that the maximum likelihood estimator is not robust. Several alternative minimum distance estimators have since been proposed as a way of dealing with the first problem. This paper deals with one of these estimators which is not only superior due to its robustness, but in fact can have an advantage in numerical studies even at the model distribution. Importantly, robust alternatives of the EM-algorithm, ostensibly fitting t distributions when in fact the data are mixtures of normals, are also not competitive at the normal mixture model when compared to the chosen minimum distance estimator. It is argued for instance that natural processes should lead to mixtures whose component distributions are normal as a result of the Central Limit Theorem. On the other hand data can be contaminated because of extraneous sources as are typically assumed in robustness studies. This calls for a robust estimator.

Keywords EM algorithm; Minimum distance estimation; Robust estimation; Monte Carlo simulation.

Mathematics Subject Classification (2000) 62F10 · 62NO2

B.R. Clarke
Mathematics and Statistics, School of Eng. and I.T., Murdoch University, Murdoch, Western Australia, 6150
Tel.: +61-8-93602578 E-mail: B.Clarke@murdoch.edu.au

T. Davidson
Australian Bureau of Statistics, Perth, Western Australia, 6000

R. Hammarstrand
Mathematics and Statistics, School of Eng. and I.T., Murdoch University, Murdoch, Western Australia, 6150

1 Introduction

The method of maximum likelihood has been considered superior to the method of moments proposed by Pearson (1894) for the fitting of finite mixtures of normal distributions, particularly since the appearance of the EM algorithm in Dempster et al. (1977). Prior to that numerical studies by Tan and Chang (1972) and Fryer and Robertson (1972) presented the maximum likelihood estimator(MLE) as being superior to the method of moments in the case of equal variances. We consider in this paper the more general estimation of k component univariate normal mixture distributions of the form

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

Here

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

$\sum_{j=1}^{k} \epsilon_j = 1$ and $\boldsymbol{\theta} \in \Theta$ is the vector of $3k-1$ 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 univariate sample $X_1, X_2, \ldots, X_n$. An introduction to parameterization of mixtures is afforded in the two texts Titterington et al. (1985) and McLachlan and Peel (2000). We consider non-degenerate mixtures, where mixture parameters are assumed greater than zero and component distributions distinct. It is emphasized that we consider the number of components to be known prior to estimation. Questions of the order or number of components are not discussed. A recent review article on this is Depraetere and Vandebroek (2014). Also while identifiability of mixture distributions is important, we address only local solutions of the estimating equations, obviating the need to distinguish between models caused by swapping components and mixture proportions.

The parametric model with unequal component standard deviations or variances leads to singularities in the likelihood surface. Letting a component mean equal to one of the observations and taking the limit as the corresponding standard deviation to zero leads to unboundedness of the likelihood function (see Titterington et al. (1985), page 83, Example 4.3.2). Clearly then there are problems re consistency of the MLE, due to the likelihood being unbounded, albeit there exist works by Redner and Walker (1984) and Amemiya (1985, chapter 4) that claim consistency of a solution under certain conditions. See the summary discussion in sections 2.5 and 2.6 of McLachlan and Peel (2000). The well documented problem of singularities in the likelihood was the principal reason for motivating minimum distance methods in Choi and Bulgren (1968), MacDonald (1971), Quandt and Ramsey (1978) and Woodward et al. (1984). See also in particular Clarke and Heathcote (1978) and Cutler and Cordiero-Braña (1996) for discussions re robustness.

In a robustness perspective Clarke and Heathcote (1994), note the unbounded nature of the efficient score function to do with the MLE and instead give M-

estimating equations arrived at from minimizing the L_2 distance

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

That paper offers only a limited numerical study of comparing estimators of proportion for several parameter sets involving a mixture of two component distributions. For that study the MLE and the L_2 estimator were arrived at by solving their respective nonlinear equations using Newton's method, when all parameters are estimated. The underlying component distributions generating the data were either normal or t distributions with 5 degrees of freedom. A further comparison involving seismic data is offered in Clarke (2000) again when the estimators are arrived at by solving nonlinear equations. In this paper we give a more extensive reporting of the comparison involving all the parameters including multivariate measures of performance when the L_2 is compared to the MLE, where the latter estimator is arrived at by the EM algorithm. The EM algorithm has become the prevailing method of estimation for mixtures and is widely discussed in books such as Titterington et al. (1985) and McLachlan and Peel (2000). Comparisons given here involve the seven parameter sets in Clarke and Heathcote (1994) involving a mixture of two normal distributions and for purposes of illustration new examples with three component normal mixtures are included. In practice any finite number of component normals can be fitted.

The L_2 minimum distance estimator is implemented via solving a set of $3k-1$ nonlinear equations of the form

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(X_i, \boldsymbol{\theta}) = 0$$

where the form for the vector $\boldsymbol{\Psi}$ is given in Theorem 2.1 of Clarke and Heathcote (1994) and involves cumulative normal distributions and normal densities, whereby the smooth function $\boldsymbol{\Psi}$ is bounded in the observation space variable. Robustness qualities of the subsequent estimator are also described therein. Perhaps the relatively slow uptake of this estimator in the literature to date may be explained by the complicated nature of these equations, however the authors of the current paper make available an R function which can fit any finite mixture of univariate normal distribution functions and uses the nonlinear equation solving package "nleqslv" of Hasselman (2013). See R Development Core Team (2014). See also the two R files at the web site http://researchrepository.murdoch.edu.au/23755/ that are both used to analyse the data in our Example 2 below.

The EM algorithm is an iterative method for finding maximum likelihood estimates. It is widely discussed in the work McLachlan and Peel (2000) and McLachlan and Krishnan (2008) and is implemented in the R package in Benaglia et al. (2009). We use this package with the default tolerance of 10^{-5} and a maximum number of iterations equal to 1000 for fitting normal mixtures. The classical MLE arrived at assuming finite normal mixtures and implemented using the EM algorithm is thus denoted in tables as EM_N.

A prime motivation for the article Clarke and Heathcote (1994) was to demonstrate through theory and a limited computing simulation the robustness of the L_2 estimator when compared to the MLE when fitting a finite mixture of normal distributions. In fact, there has since arisen a body of argument that if one fits a mixture of t distributions then one can inherit robustness properties using the EM algorithm. See for example Peel and McLachlan (2000) and McLachlan and Peel (2000). We argue, however, that nature would rather present mixtures of normal distributions on which the L_2 estimator is modelled rather than mixtures of t distributions. If there is contamination of the samples or even outliers present the L_2 naturally caters for them, given its qualitative robustness qualities. The question is then asked "What are we estimating if indeed we have the model normal mixture and we fit t distributions?"

It is recognized by these authors that if the data are known to be generated by t distributions with known degrees of freedom then it would naturally be sensible to fit those distributions by maximum likelihood algorithms, such as the EM algorithm, for then the estimator would be efficient. The algorithm related in section 2.2 fits t distributions with input parameters of 4 degrees of freedom; we denote it EM_T4, and then the degrees of freedom are subsequently estimated along with other parameters. Theoretically, an L_2 distance estimator could be derived for t distributions, but since there are few if any practical situations known where t distributions are known to be the underlying distribution, other than through simulation, we do not pursue that here.

By performing a Monte Carlo simulation we show in section 2 that the L_2 estimator can have superior performance to the EM algorithm for a number of parameter sets~~, see Tables 1 and 6,~~ when the model normal mixture is used to generate the data. It also appears the L_2 algorithm thus fitted has more generally superior performance compared with the EM algorithm that implements t distributions. In calibrating the L_2 R function, a feature of the estimates is that they were roughly unbiased, meaning that the parameter estimates formed distributions that centered around the true component parameters for sample sizes 200 and 500 respectively (~~exceptions to are noted for parameter set (5) for a mixture of two normal distributions and parameter set (8) for a mixture of three normal distributions, where a very few number of estimates out of the 100 calculated tended to diverge; this raised the overall mean squared error for the estimates of individual parameters for one of the components say~~There were exceptions for a couple of parameter sets). In fact in estimating proportion alone, say in a mixture of two normal distributions, it is shown in Clarke (1989) that the L_2 estimate is unbiased. For this reason, we prefer to compare variances of the estimates via calculating sample covariance matrices of the successful iterations in addition to supplying in an appendix means and mean squared errors for individual component parameters. Subsequently we report multivariate measures of the generalized variance and total variation based on these estimated covariance matrices. These are the determinant and trace of the resultant covariance matrix respectively. An estimator with a lower determinant can be considered more efficient. Similar arguments follow for the trace. ~~In only one parameter set (7) for sample size $n = 500$ do we find determinant and trace indicating different estimators as being more efficient.~~It transpired that in estimating parameters from seven parameter sets, representing

mixtures of two normal distributions, the determinant and trace indicated the L_2 estimator was efficient in all but two cases; one where the component populations had the same location but different standard deviations, the other where the component populations were well separated. Superior performance for the L_2 estimator was also found in a number of parameter sets in a mixture of three normal distributions.

In section 3 we offer a graphic illustration using an extra single point contamination of a sample of size 200 of the non-robustness of the EM algorithm for fitting mixtures of normals when say compared to the L_2 estimator. In the final section we give two examples of robust fitting. In the first example we consider fitting daily returns for the company Bethlehem Steel from 3rd October 1984 to the 22nd of October 1990 and note the significant influence on the MLE and the relative stability of the L_2 estimator by the observations corresponding to the October "crash" period (September 23 to November 3, 1987). The second example is based on the heart weight measurements of cats given in Fisher (1947) where the EM and L_2 estimates are comparable and demonstrates that a measurement of standard errors of estimates for the L_2 estimate is afforded by the multivariate jackknife.

2 Fitting Mixtures of k-Component Normal Univariate Distributions

2.1 Comparison Simulations when Fitting Mixtures of Two Normal Distributions

To do a numerical comparison it should be recognized that the algorithms are completely different. The problems involve solutions which achieve stationary points of the surface. For the L_2 distance it is possible to converge to a root which is outside the parameter boundary or even a local minima. The EM algorithm on the other hand converges to a local maxima, but as noted by Seidel et al. (2000) the algorithm can converge to different local maxima and is heavily starting point dependent. See also Biernacki and Chretien (2003), Biernacki et al. (2003) and Seidel and Ševčíková (2004). In fact according to McLachlan and Krishnan (2008) on page 88 they write "The fixed points of the EM algorithm include all local maxima of the likelihood, and sometimes saddle points and local minima." Nevertheless Wu (1983) has given asymptotic results on convergence properties to a consistent root which is a local maximizer of the likelihood.

The root n asymptotically unique consistent estimator of the L_2 minimizing equations is the focus of this study. According to the theory of uniform convergence discussed in Clarke and Futshik (2007) the Newton algorithm will converge to this root for all sufficiently large n from within a small ball centered at the true underlying parameter. For this reason, for a correctly specified mixture of k normal distributions defined by the parameters in Clarke and Heathcote (1994) we carry out our simulations by starting both the considered algorithms at the true parameter used to define the distribution that generates the data. This is done for several parameter sets for $k = 2$ and $k = 3$ respectively. In principle any value of

Table 1 Parameter sets in a mixture of two normal distributions.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ϵ_1	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

Table 2 Number of samples that failed.

		Mixtures of 2 Normals						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
$n = 200$	L_2	26	6	9	19	9	12	1
$n = 200$	EM_N	16	1	4	13	2	13	0
$n = 200$	EM_T4	1	0	0	2	0	2	0
$n = 500$	L_2	11	1	3	11	0	8	0
$n = 500$	EM_N	40	2	7	70	1	29	0
$n = 500$	EM_T4	2	0	1	1	0	2	0

Table 3 Times (in seconds) ignoring failures.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
L_2	$n = 200$	9.30	5.38	3.82	9.48	5.41	5.29	2.25
EM_N	$n = 200$	6.52	1.92	3.34	5.91	3.60	5.92	0.41
EM_T4	$n = 200$	4.41	3.66	3.60	4.15	3.97	3.98	3.11
L_2	$n = 500$	12.69	4.44	5.30	14.10	7.13	4.28	2.97
EM_N	$n = 500$	13.77	2.89	5.23	12.48	4.04	14.76	0.66
EM_T4	$n = 500$	8.54	7.84	8.17	9.08	8.11	8.45	7.09

Table 4 Breakdown of L_2 failures.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
$n = 200$	Outside Parameter Space	22	4	3	15	6	3	0
$n = 200$	Maximum 150 Iterations Reached	4	2	6	4	3	8	1
$n = 200$	Algorithm Stalled	0	0	0	0	0	1	0
$n = 500$	Outside Parameter Space	10	0	0	7	0	6	0
$n = 500$	Maximum 150 Iterations Reached	1	1	3	4	0	2	0

$k \geq 2$ can be investigated.

For each parameter set with two component normals (Table 1), 100 successful iterations of sample estimates of θ for both the L_2 and EM algorithms are generated from samples of size 200. With the L_2 estimator equations, samples where there was either a failure of the nonlinear equation solver to converge within 150 iterations using a function value tolerance of $ftol = 10^{-3}$ or where the obtained solution was outside the parameter space were discarded. See Table 2. The aforesaid tolerance was chosen instead of a smaller tolerance of 10^{-5} say, in order for

the algorithm to converge within the specified number of 150 iterations. The point of Table 3 is to give researchers an idea of relative computing time taken for these methods; which of course differ in software packages that are programmed differently and have differing amounts of overhead processing expenses. These values are the elapsed times in seconds from a computer with an Intel(R) i7-4500U processor running at 1.80GHz, with 8GB RAM on a 64-bit Windows 8.1 operating system. There was no outright winner in absolute terms, the main point being that all methods can be computed in real and possibly comparable times. For the EM-algorithm a sample was discarded if there was failure to converge within 1000 iterations with a tolerance of 10^{-5}. Interestingly the majority of failures for the L_2 estimator were for the reason that the nonlinear equation successfully stopped at a point outside the boundary of the parameter space. Either that or the algorithm reached its maximum number of iterations. See Table 4. Results for increasing sample sizes, $n = 500$ say, show much fewer instances of converging to a value outside the boundary of the parameter space, as one would expect with a root-n consistent estimator.

The comparison of variances in Table 5 is, as alluded to earlier, motivated by the relative unbiasedness of the L_2 estimator. For instance see Tables 10 and 12. Summary measures of variance using the determinant and the trace of the sample covariance matrix calculated from the respective sets of 100 successful parameter estimates obtained from simulated samples are then used. Table 5 shows the superiority of the L_2 estimator over that of the EM algorithm, when the model mixtures of two normal distributions are fitted, except for parameter set (5) and parameter set (7). A closer examination of the estimates for parameter set (5) reveals that the L_2 estimator on a few occasions tended to diverge. To illustrate this see the resultant box and whisker plot of the 100 successful estimates of the second scale parameter σ_2 for this two component model. It appears that only a few large values tended to increase the mean squared error and also generalized variance and total variation of estimates for this parameter configuration. See Figure 1. For parameter set (7), when component distributions are well separated, it is perhaps not surprising that the EM_N is potentially superior.

To complete the picture estimated mean and mean squared errors of individual parameter estimates are supplied in the Appendix. These corroborate the story of robustness and efficiency of the L_2 results discussed above.

2.2 Fitting the EM algorithm using t distributions at the model mixture of normals

An original piece of software for fitting mixtures of t distributions is the EMMIX software of McLachlan et al. (1999). According to Lee and McLachlan (2014), Wang et al. (2009) implemented the fitting of multivariate skew t distributions in R using the package EMMIXskew, R Development Core Team (2014). The package allows fitting of t distributions which are implemented in this paper, even though the generated data are normal. We refer to the fit that is obtained using input parameters of 4 degrees of freedom as EM_T4, as indicated earlier. It should be noted that the package adapts the degrees of freedom for each component so that

Table 5 Resultant determinant(Det) and Trace of covariance matrix of estimates when fitting mixtures of two normal distributions. More efficient values are given in bold for each parameter set.

$n = 200$	L_2		EM_N		EM_T4	
	Det	Trace	Det	Trace	Det	Trace
(1)	$\mathbf{5.95\times10^{-10}}$	**0.183**	2.67×10^{-6}	1.204	9.38×10^{-7}	0.708
(2)	$\mathbf{2.11\times10^{-8}}$	**0.341**	4.83×10^{-7}	1.560	2.19×10^{-5}	3.274
(3)	$\mathbf{1.14\times10^{-7}}$	**0.448**	2.25×10^{-6}	1.193	9.79×10^{-5}	3.016
(4)	$\mathbf{3.80\times10^{-10}}$	**0.176**	6.45×10^{-6}	1.320	6.65×10^{-7}	0.761
(5)	8.86×10^{-5}	9.548	$\mathbf{1.94\times10^{-5}}$	**1.409**	5.50×10^{-4}	2.758
(6)	$\mathbf{1.70\times10^{-9}}$	**0.300**	3.15×10^{-7}	1.235	1.13×10^{-7}	0.773
(7)	1.18×10^{-9}	0.236	$\mathbf{6.34\times10^{-10}}$	**0.181**	8.67×10^{-8}	1.086
$n = 500$						
(1)	$\mathbf{5.61\times10^{-12}}$	**0.103**	6.14×10^{-7}	1.124	7.12×10^{-10}	0.298
(2)	$\mathbf{2.85\times10^{-10}}$	**0.141**	1.43×10^{-9}	0.362	8.00×10^{-7}	2.328
(3)	$\mathbf{1.54\times10^{-9}}$	**0.217**	2.20×10^{-8}	0.403	8.46×10^{-6}	1.660
(4)	$\mathbf{4.27\times10^{-12}}$	**0.157**	2.47×10^{-7}	0.852	2.94×10^{-9}	0.380
(5)	2.02×10^{-6}	2.205	$\mathbf{1.46\times10^{-9}}$	**0.204**	7.23×10^{-7}	0.947
(6)	$\mathbf{4.40\times10^{-11}}$	**0.182**	2.38×10^{-9}	0.670	1.33×10^{-9}	0.410
(7)	9.17×10^{-12}	**0.0733**	$\mathbf{4.39\times10^{-12}}$	0.076	3.55×10^{-10}	0.333

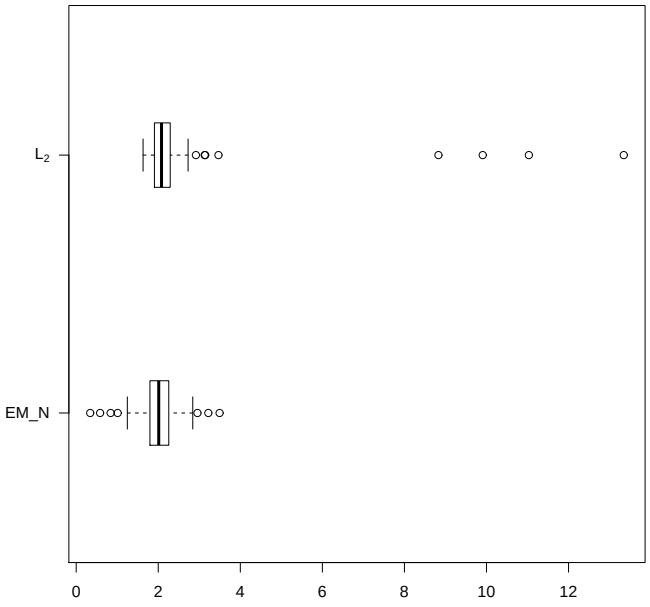

Fig. 1 Box-and-whisker plots of L_2 and EM estimates of σ_2 for parameter set (5)

even though mixtures of t distributions with 4 degrees of freedom are initially fitted the algorithm finally prints out an estimated degrees of freedom for each component. Thus if as in our main examples the data are from normal components then typically the degrees of freedom for each estimated component is relatively large, greater than 15 on average for $n = 200$ when the algorithm converges, thus corroborating the fact that the estimated components tend towards normality as should be expected. As n increased to 500 the average estimated degrees of freedom was more often greater than 20, as expected the degrees of freedom increasing to adapt to what are normal components. We do not report individual statistics on the estimated degrees of freedom here. Setting the tolerance of the algorithm to be 10^{-5} and a maximum number of iterations equal to 1,000 and an initial degrees of freedom of 4 we get in a separate run the last two columns of Table 5. Here it is seen for the seven parameter sets the L_2 -estimator has superior performance to the EM_T4 estimator in all bar parameter set (5), albeit there were fewer problems with convergence for the EM algorithm fitting t distributions (when in fact the data are normal mixtures). While it may be useful also to fit skew normal and skew t distributions using the more recent package EMMIXuskew of Lee and McLachlan (2014), see also Lee and McLachlan (2013), we do not do so here since adding more parameters is likely to increase the variance of the fitted parameter estimates for the parameters of interest, and it is not clear that fitting mixtures of normals should be enhanced by fitting skew distributions. Nevertheless further evidence may surface in future empirical studies.

2.3 Comparison Simulations when Fitting Mixtures of Three Component Normal Distributions

There is no essential difficulty in fitting more than two component normals; we give here some examples of fitting mixtures of normal distributions with three components. The parametric models fitted are given in Table 6 and results based on samples where both L_2 and EM algorithms converged successfully within the boundary of the parameter space. Again the superiority of the L_2 estimator is mainly evident as seen in Table 7. A noted exception is for $n = 500$ and parameter set (8), where apparently the third component fitted had for one estimate only out of the one hundred estimates obtained a mean μ_3^* that was extremely negative and a standard deviation σ_3^* that was unusually large. To demonstrate this we plotted here the box and whisker plots of the estimates of these parameters for both the L_2 and the EM_N algorithm. See Figures 2 and 3.

It is clear that for this parameter set (8) and the L_2 estimator one response explodes the estimated mean and also mean squared errors (see appendix) and consequently also the estimates of generalized variance and Trace of the covariance matrix of estimates. On the other hand and in the main, L_2 estimates of all parameters had either comparable or much improved measures of variation when compared to either fitting the model mixture of three normal distributions using the EM algorithm or in fact fitting a mixture of three t distributions when a mixture of normals is used to generate the data. With the extra component the number of samples where the L_2 estimator was outside the boundary of the parameter space increased, but this decreases with increasing sample size, as noted

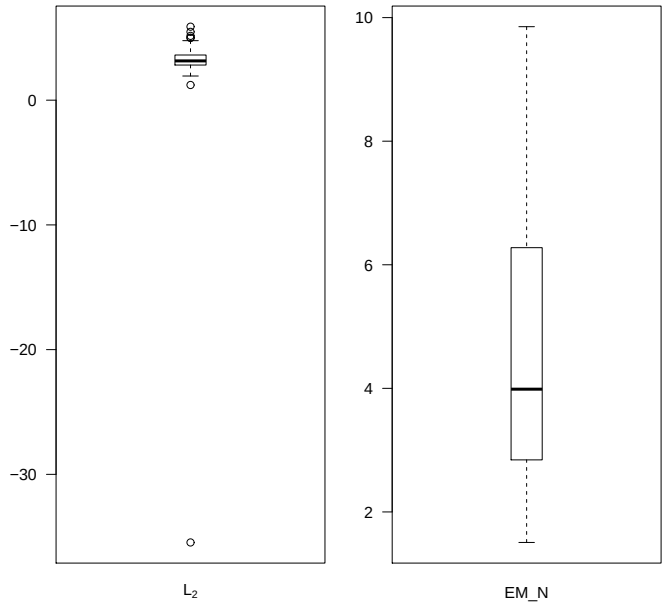

Fig. 2 Box-and-whisker plots of L_2 and EM estimates of μ_3 for parameter set (8)

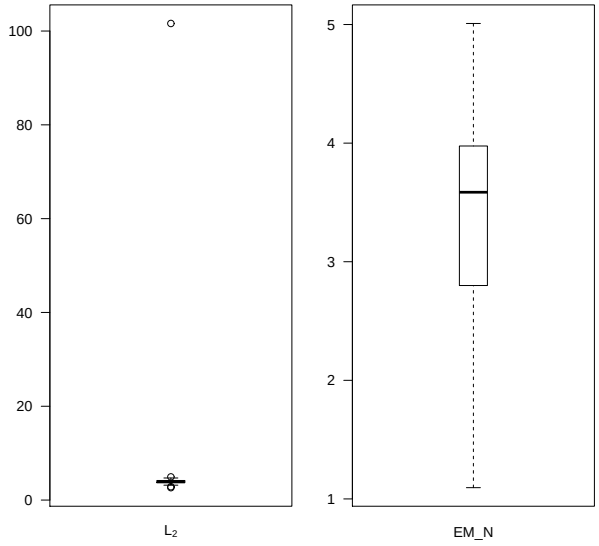

Fig. 3 Box-and-whisker plots of L_2 and EM estimates of σ_3 for parameter set (8)

Table 6 Parameter sets in a mixture of three normal distributions.

	(8)	(9)	(10)	(11)	(12)
ϵ_1	0.30	0.30	0.70	0.40	0.40
ϵ_2	0.50	0.50	0.20	0.30	0.30
μ_1	-3.00	-1.00	-5.00	-1.00	-1.00
μ_2	0.00	0.00	0.00	0.00	0.00
μ_3	3.00	1.00	5.00	1.00	1.00
σ_1	1.00	1.00	1.00	1.00	2.00
σ_2	2.00	1.00	2.00	2.00	1.00
σ_3	4.00	2.00	1.00	1.00	1.00

Table 7 Resultant determinant(Det) and Trace of covariance matrix of estimates when fitting mixtures of three normal distributions. More efficient values are given in bold for each parameter set.

	L_2		EM_N		EM_T4	
$n = 200$	Det	Trace	Det	Trace	Det	Trace
(8)	$\mathbf{9.21\times10^{-10}}$	**5.043**	3.34×10^{-9}	9.450	4.09×10^{-6}	23.510
(9)	$\mathbf{2.02\times10^{-14}}$	**0.599**	9.28×10^{-9}	4.305	1.16×10^{-7}	5.113
(10)	8.55×10^{-14}	1.724	$\mathbf{4.70\times10^{-15}}$	**0.915**	1.84×10^{-12}	5.112
(11)	$\mathbf{2.29\times10^{-12}}$	**0.795**	8.54×10^{-7}	7.224	3.60×10^{-7}	4.019
(12)	$\mathbf{1.43\times10^{-13}}$	**0.574**	3.39×10^{-9}	3.708	2.57×10^{-9}	3.857
$n = 500$						
(8)	1.42×10^{-10}	111.600	$\mathbf{3.10\times10^{-12}}$	**6.391**	2.29×10^{-9}	22.480
(9)	$\mathbf{1.32\times10^{-16}}$	**0.543**	2.90×10^{-9}	4.204	1.10×10^{-10}	2.292
(10)	1.56×10^{-17}	0.604	$\mathbf{4.20\times10^{-18}}$	**0.406**	7.57×10^{-16}	2.208
(11)	$\mathbf{5.48\times10^{-14}}$	**0.549**	9.38×10^{-9}	5.115	6.47×10^{-11}	1.712
(12)	$\mathbf{7.76\times10^{-16}}$	**0.432**	1.65×10^{-10}	2.254	4.52×10^{-11}	1.950

Table 8 Number of samples that failed.

		Mixtures of 3 Normals				
		(8)	(9)	(10)	(11)	(12)
$n = 200$	L_2	20	42	9	77	58
$n = 200$	EM_N	2	21	0	15	18
$n = 200$	EM_T4	0	3	0	0	1
$n = 500$	L_2	3	17	4	34	20
$n = 500$	EM_N	1	65	0	46	52
$n = 500$	EM_T4	0	3	0	1	1

previously. See Table 8.

3 Effects of Single Point Contamination

The L_2 estimator is known to have a bounded form of what is called an influence function (Clarke and Heathcote (1994)), whereas the MLE has an unbounded influence function. As a consequence there can only be bounded influence of a single point or observation on the L_2-estimate whereas for the MLE estimate a single point can have unbounded influence. To demonstrate and illustrate this

we generated a single sample of 200 observations from a mixture of two normal distributions with parameter set ($\epsilon_1 = 0.5,\ \mu_1 = -2,\ \mu_2 = 2,\ \sigma_1 = 1,\ \sigma_2 = 1$) and observe the effect of one extra observation taking on values between -20 and 20. Figure 4 shows the consequence of shifting just one observation in a perturbed manner on the estimated parameters. This graphic illustration should warn those who use classical estimates of parameters from distributions with exponentially decreasing tails to be wary of their data.

4 Real Life Examples

Example 1: This example considers fitting univariate mixtures of normals to the daily returns of the company Bethlehem Steel from October 3^{rd} 1984 until October 22^{nd} 1990. Klar and Meintanis (2005) note the observations corresponding to the October "crash" period (September 23^{rd} to November 3^{rd}1987) which were thought to be extreme and were subsequently removed. Here we compare the L_2 estimate $\tilde{\boldsymbol{\theta}}^*$ and the MLE $\hat{\boldsymbol{\theta}}$ with and without the October "crash" period observations. In addition a suspected outlier corresponding to July 30^{th} 1986 is also removed. This judgement is based on the box and whisker plots and histogram provided in Figure 5. Two component mixture models were fitted, with starting values for the L_2 estimate being chosen by trial and error. The resulting robust estimate was then used as a starting value for the EM algorithm used to calculate the MLE. Table 9 reflects the relative stability of the L_2 estimates compared to the high sensitivity of the MLE estimates to the presence of the data in the "crash" and also the outlying value.

Table 9 Estimated parameter sets in a mixture of two normal distributions,using original data of Klar and Meintanis (2005)

Estimator	$\hat{\epsilon}_1$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
L_2	0.579	-0.289	0.524	1.613	3.626
EM_N	0.899	-0.045	0.570	2.146	6.801
	Original Data				
L_2	0.532	-0.328	0.459	1.541	3.379
EM_N	0.847	-0.106	0.758	2.024	5.173
	Data Without "crash"				
L_2	0.530	-0.335	0.473	1.541	3.360
EM_N	0.760	-0.185	0.759	1.880	4.287
	Data Without "crash" or Outlier				

Example 2: The data consists of heart weight measurements for a sample of male and female cats used for digitalis experiments. This data set was used by Fisher (1947). The indicator variable corresponding to sex has been removed in order to pose the problem as a missing data problem involving mixtures. A comparison is then made between maximum likelihood and the L_2 minimum distance estimator. With the assumption that the component distributions are normal and with the sex of the cats known, the estimated parameter vector is simply,

$$(0.674,\ 11.3,\ 9.2,\ 2.54,\ 1.36)$$

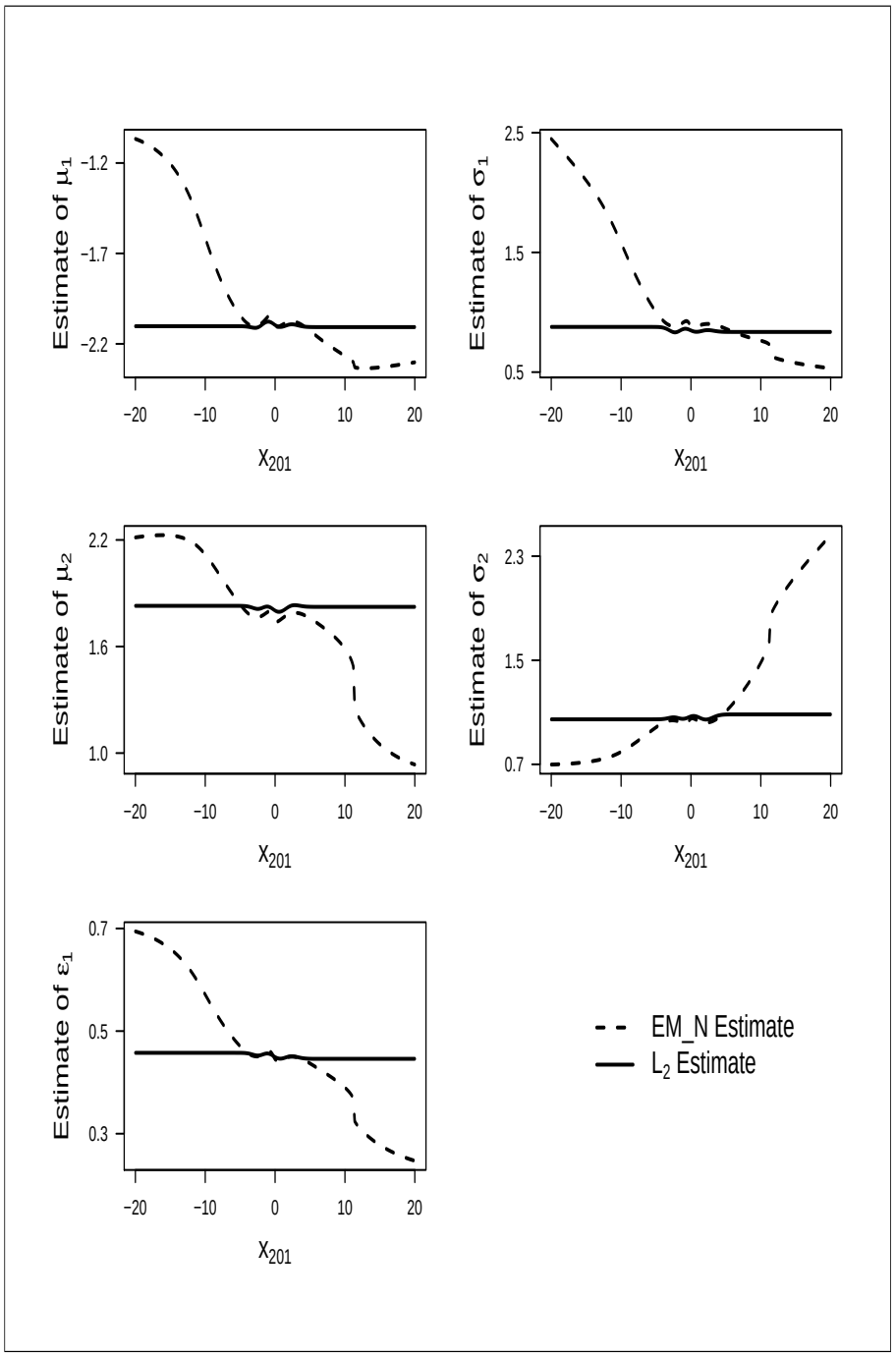

Fig. 4 Comparison between the L_2 estimate and EM_N estimate given contamination in the final observation when $k = 2$

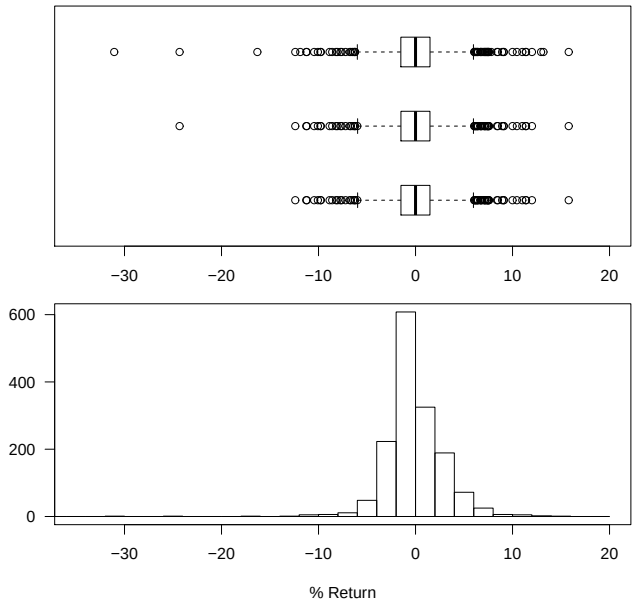

Fig. 5 Boxplots of %Returns of Bethlehem Steel are, from top to bottom, full data set and data set without 'crash' and data set without 'crash' and without 'outlier' and below that is the histogram of full data set.

The mixing proportion of 0.674 corresponds to the proportion of male cats in the sample. In addition $11.3, 9.2, 2.54$ and 1.36 correspond to the individual component sample estimates of $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ respectively. For instance $(11.3, 2.54)$ is the estimate of (μ_1, σ_1) based on male cats only. Now the corresponding robust L_2 estimate from the mixture of two normals assuming cats are not classified according to sex is then

$$\tilde{\boldsymbol{\theta}}^{*} = (0.533,\ 11.84,\ 9.18,\ 2.35,\ 1.37)$$

The maximum likelihood estimate is,

$$\hat{\boldsymbol{\theta}} = (0.483,\ 12.08,\ 9.27,\ 2.45,\ 1.39)$$

For this data, restricting attention to the four individual location and scale parameter estimates from joint estimation using either the L_2 or the EM_N, the robust estimates of location outperform the EM_N estimates in that they are closer to the initial MLE estimates of location assuming populations are identified, while scale estimates are comparable. This is also the case when comparing say the L_2 and EM_N with a classical robust location and scale estimate such as Huber's minimax solution while employing the MAD estimate for scale. See Huber and Ronchetti (2009, pp 74-5,106) with the tuning constant $k = 1.345$ chosen to give 95% efficiency gives

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{H}} = (0.674,\ 11.22,\ 9.19,\ 2.67,\ 1.48).$$

While point estimates have been the subject of this paper the L_2 estimator lends itself to relatively easy estimates of spread. Given a parent sample vector $(x_1, \ldots, x_n)$, from which an estimate $\tilde{\boldsymbol{\theta}}^*$ is calculated and defining $\tilde{\boldsymbol{\theta}}^*_{(\boldsymbol{i})}$ to be the statistic derived from the sub-sample $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ then the multivariate jackknife variance covariance estimator of $\tilde{\theta}^*$ is given by

$$V_{jack} = \frac{n-1}{n} \sum_{i=1}^{n} [\tilde{\boldsymbol{\theta}}^*_{(\boldsymbol{i})} - \bar{\tilde{\boldsymbol{\theta}}}^*][\tilde{\boldsymbol{\theta}}^*_{(\boldsymbol{i})} - \bar{\tilde{\boldsymbol{\theta}}}^*]^T$$

The solutions $\tilde{\boldsymbol{\theta}}^*_{(\boldsymbol{i})}$ are arrived at by starting the nonlinear equation solving algorithm iterations from the robust starting point $\tilde{\boldsymbol{\theta}}^*$. For this data the estimate of variance covariance matrix is

$$\begin{bmatrix} 0.031 & -0.087 & -0.056 & -0.0096 & -0.039 \\ -0.087 & 0.313 & 0.153 & 0.039 & 0.108 \\ -0.056 & 0.153 & 0.178 & 0.047 & 0.088 \\ -0.0096 & 0.039 & 0.047 & 0.063 & 0.015 \\ -0.039 & 0.108 & 0.088 & 0.015 & 0.084 \end{bmatrix}$$

In view of the discussion in the section "Fréchet Differentiability and the Jackknife" in Clarke (2000,p.472-473) and the Fréchet differentiability of the L_2 estimator established in Clarke and Heathcote (1994), the above is a consistent and robust estimate of variance of the L_2 estimate. Bootstrap estimates are also possible using the L_2-estimate. Proponents of the EM_N estimate consider bootstrapped comparisons with the standard information based method in Basford et al. (1997) for example.

5 Conclusion

A criticism of the EM algorithm has been that it is slow to converge. The L_2 algorithm as implemented here has comparable convergence properties as exhibited in Table 3. Nevertheless the benefits in terms of reduced variance and approximate unbiasedness favor the L_2 estimator in the main. It should be realized that the efficiency of the L_2 method is parameter dependent, but we have given here practical examples which support the L_2 method in particular when all $3k-1$ parameters are estimated. The effect of single point contamination on the MLE compared with the L_2 method is graphic. The L_2 method is known to be robust in neighborhoods of the true model distribution, including when data are heavy tailed. Moreover when implemented at the model mixtures of normal distributions it is broadly speaking more efficient than when implementing the MLE using the EM algorithm either assuming the correct model (using EM_N) or assuming t distributions (using EM_T4). Finally, there is in principle an extension of the L_2 minimum distance method to the multivariate data case which may be of further interest. This is intended as a subject for future study.

Table 10 Average of 100 successful estimates for models (1)-(7). Individual values closest to the true parameter indicating less sample bias are in bold.

	$n = 200$	ϵ_1	μ_1	μ_2	σ_1	σ_2
(1)	True	0.750	-1.000	0.000	1.000	1.000
	L_2	0.668	**-1.069**	**-0.072**	**0.995**	**1.038**
	EM_N	**0.693**	-1.105	0.279	0.879	0.752
(2)	True	0.500	-1.000	0.000	1.000	2.000
	L_2	0.558	-1.007	**0.185**	**1.024**	**2.001**
	EM_N	**0.516**	**-1.006**	0.368	0.928	1.859
(3)	True	0.750	-1.000	0.000	2.000	1.000
	L_2	0.620	**-1.212**	-0.0718	**2.067**	1.138
	EM_N	**0.654**	-1.331	**-0.014**	1.905	**0.907**
(4)	True	0.500	1.000	0.000	1.000	1.000
	L_2	0.518	**0.969**	**-0.021**	**1.032**	**1.036**
	EM_N	**0.506**	1.186	-0.153	0.824	0.821
(5)	True	0.500	0.000	0.000	1.000	2.000
	L_2	0.651	**0.007**	0.144	1.116	2.743
	EM_N	**0.501**	-0.019	**-0.019**	**0.887**	**1.954**
(6)	True	0.500	-1.500	0.000	1.000	1.000
	L_2	**0.502**	**-1.436**	**-0.002**	**1.051**	**1.044**
	EM_N	0.495	-1.632	0.149	0.902	0.884
(7)	True	0.500	-5.000	0.000	1.000	2.000
	L_2	**0.494**	-4.989	**-0.015**	**0.989**	**1.996**
	EM_N	0.490	**-5.007**	-0.030	0.981	1.991

Table 11 Mean Squared Errors from 100 successful estimates for models (1)-(7). Values with less mean squared error for individual parameters, pointing to more efficient parameter estimates, are in bold.

	$n = 200$	ϵ_1	μ_1	μ_2	σ_1	σ_2
(1)	L_2	**0.0516**	**0.0471**	**0.0783**	**0.0058**	**0.0168**
	EM_N	0.0801	0.1936	0.8044	0.0655	0.2173
(2)	L_2	**0.0340**	**0.0307**	**0.2203**	**0.0481**	**0.0424**
	EM_N	0.0454	0.0478	1.3840	0.07343	0.1551
(3)	L_2	**0.0576**	**0.2350**	**0.0766**	**0.0468**	**0.1179**
	EM_N	0.0628	0.7935	0.1607	0.1357	0.1647
(4)	L_2	**0.0543**	**0.0510**	**0.0518**	**0.0098**	**0.0117**
	EM_N	0.0873	0.5256	0.5498	0.1366	0.1288
(5)	L_2	0.0726	**0.0157**	3.5140	**0.0861**	6.3740
	EM_N	**0.0633**	0.0740	**0.8673**	0.1340	**0.2723**
(6)	L_2	**0.0470**	**0.1025**	**0.1161**	**0.0193**	**0.0203**
	EM_N	0.0818	0.5467	0.4839	0.0815	0.0915
(7)	L_2	**0.0026**	0.0284	0.1180	0.0133	0.0723
	EM_N	**0.0026**	**0.0246**	**0.0985**	**0.0112**	**0.0436**

Acknowledgements The authors are indebted to Emeritus Professor C.R. Heathcote, now retired, for his pioneering work on minimum distance estimators. Views expressed in this paper are those of the authors and do not necessarily represent those of the Australian Bureau of Statistics. The authors also acknowledge the helpful suggestions on presentation afforded by two anonymous referees that led to an improved paper.

Appendix: Averages and Mean Squared Errors of Estimates

For completeness we include here averages and mean squared errors for individual parameters for parametric ~~models (1)-(12)~~models (1)-(7) (Tables 10 and 11) and models (8)-(12) (Tables 12 and 13) for the L_2 method and the EM algorithm for the MLE obtained using mixtools, that is EM_N.

Table 12 Average of 100 successful estimates for models (8)-(12). Individual values closest to the true parameter indicating less sample bias are in bold.

	$n = 200$	ϵ_1	ϵ_2	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
(8)	True	0.300	0.500	-3.000	0.000	3.000	1.000	2.000	4.000
	L_2	0.363	0.456	-2.784	**0.255**	**3.870**	1.078	**1.945**	**3.914**
	EM_N	**0.354**	**0.491**	**-2.871**	0.312	5.286	**1.020**	1.781	2.856
(9)	True	0.300	0.500	-1.000	0.000	1.000	1.000	1.000	2.000
	L_2	0.396	0.416	**-0.869**	**0.113**	**1.168**	**1.034**	**1.028**	**2.013**
	EM_N	**0.353**	**0.483**	-1.248	0.248	2.333	0.760	0.777	1.361
(10)	True	0.700	0.200	-5.000	0.000	5.000	1.000	2.000	1.000
	L_2	0.687	0.202	-5.013	-0.279	4.716	0.986	**2.028**	1.136
	EM_N	**0.698**	**0.201**	**-4.988**	**-0.029**	**4.954**	**0.999**	1.825	**0.957**
(11)	True	0.400	0.300	-1.000	0.000	1.000	1.000	2.000	1.000
	L_2	0.466	0.225	**-0.934**	**0.138**	**1.042**	**1.081**	**2.019**	**1.048**
	EM_N	**0.377**	**0.307**	-1.174	0.434	1.095	0.821	1.352	0.824
(12)	True	0.400	0.300	-1.000	0.000	1.000	2.000	1.000	1.000
	L_2	**0.368**	**0.232**	**-1.169**	**-0.155**	**0.883**	**1.977**	**1.054**	**1.019**
	EM_N	0.291	0.385	-1.988	-0.248	1.342	1.633	0.766	0.764
	$n = 500$								
(8)	True	0.300	0.500	-3.000	0.000	3.000	1.000	2.000	4.000
	L_2	**0.304**	**0.502**	**-2.963**	**0.007**	**2.894**	0.985	2.035	4.897
	EM_N	0.328	0.515	-2.908	0.219	4.537	**1.014**	**1.992**	**3.374**
(9)	True	0.300	0.500	-1.000	0.000	1.000	1.000	1.000	2.000
	L_2	0.406	0.4005	**-0.868**	0.138	**1.134**	**1.026**	1.026	**1.996**
	EM_N	**0.307**	**0.529**	-1.144	**0.019**	2.045	0.820	**0.979**	1.614
(10)	True	0.700	0.200	-5.000	0.000	5.000	1.000	2.000	1.000
	L_2	0.690	0.206	**-5.002**	-0.128	4.886	**0.998**	**2.025**	1.049
	EM_N	**0.698**	**0.202**	**-4.998**	**-0.010**	**4.954**	**0.998**	1.936	**0.989**
(11)	True	0.400	0.300	-1.000	0.000	1.000	1.000	2.000	1.000
	L_2	0.422	0.282	**-0.962**	0.049	**0.975**	**1.050**	**2.033**	**1.056**
	EM_N	**0.400**	**0.313**	-1.049	**-0.047**	1.094	0.947	1.749	0.865
(12)	True	0.400	0.300	-1.000	0.000	1.000	2.000	1.000	1.000
	L_2	**0.372**	**0.278**	**-1.169**	**-0.067**	**0.956**	**1.985**	**1.040**	**1.033**
	EM_N	0.342	0.336	-1.579	-0.175	1.127	1.837	0.823	0.860

References

1. Amemiya, T. (1985). *Advanced Econometrics,* Cambridge, Massachusetts: Harvard University Press.
2. Benaglia, T., Chauveau, D., Hunter, D. R. and Young, D. (2009). mixtools: An R package for analysing finite mixture models. *J Statist. Soft.,* **32**, (6) 1-29.
3. Basford, K. E., Greenway, D. R., McLachlan, G. J. and Peel, D. (1997). Standard errors of fitted means under normal mixture. *Comp. Statist.,* **12**, 1-17.
4. Biernacki, C. and Chretien, S. (2003). Degeneracy in the maximum likelihood estimation of univariate gaussian mixtures with em. *Statistics and Probability Letters,* **61**, 373-382.
5. Biernacki, C., Celeux, G. and Govaert, G. (2003). Strategies for getting the highest likelihoods in mixture models. Guest Editors: Böhning and Seidel, W. *Comp. Stat. and Data Analysis* **41**, 561-575.
6. Choi, K. and Bulgren, W. G. (1968). An estimation procedure for mixtures of distributions. *J. R. Statist. Soc., B.,* **30**, 444-460.
7. Clarke, B. R. (1989). An unbiased minimum distance estimator of the proportion parameter in a mixture of two normal distributions. *Statist. Prob. Lett.,* **7**, (4), 275-281.
8. Clarke, B. R. (2000). A review of differentiability in relation to robustness with an application to seismic data analysis. *PINSA, A.,* **66**, 467-482.
9. Clarke, B. R. and Heathcote, C. R. (1978). Comment on "Estimating mixtures of normal distributions and switching regressions" by Quandt, R.E. and Ramsey, J.B.. *J. Am. Statist. Ass.,* **73**, 749-750.

Table 13 Mean Squared Errors from 100 successful estimates for models (8)-(12). Values with less mean squared error for individual parameters, pointing to more efficient parameter estimates, are in bold. Values agreeing to the third decimal place are considered tied and subsequently both are highlighted in bold.

	$n = 200$	ϵ_1	ϵ_2	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3
(8)	L_2	0.020	**0.019**	0.215	0.621	**3.970**	0.107	**0.197**	**0.730**
	EM_N	**0.018**	0.027	**0.172**	**0.524**	12.190	**0.080**	0.315	2.728
(9)	L_2	**0.066**	**0.055**	**0.100**	**0.119**	**0.237**	**0.019**	**0.019**	**0.051**
	EM_N	0.075	0.079	0.693	0.491	4.094	0.199	0.162	0.884
(10)	L_2	**0.002**	**0.002**	**0.010**	0.783	0.485	0.007	0.348	0.246
	EM_N	**0.002**	**0.002**	0.011	**0.369**	**0.166**	**0.006**	**0.308**	**0.074**
(11)	L_2	**0.038**	**0.040**	**0.119**	**0.305**	**0.156**	**0.052**	**0.066**	**0.051**
	EM_N	0.057	0.064	0.510	5.281	0.517	0.174	1.020	0.238
(12)	L_2	**0.021**	**0.028**	**0.207**	**0.164**	**0.098**	**0.029**	**0.047**	**0.046**
	EM_N	0.042	0.065	2.889	0.499	0.753	0.453	0.191	0.194
	$n = 500$								
(8)	L_2	**0.006**	**0.009**	**0.048**	**0.221**	15.450	**0.033**	**0.074**	95.450
	EM_N	0.008	0.014	0.067	0.316	**7.328**	0.036	0.151	**1.216**
(9)	L_2	0.081	0.063	**0.109**	**0.139**	**0.175**	**0.009**	**0.011**	**0.026**
	EM_N	**0.038**	**0.038**	0.405	0.147	4.109	0.150	0.088	0.481
(10)	L_2	**0.001**	**0.001**	**0.004**	0.201	0.117	0.003	0.226	0.075
	EM_N	**0.001**	**0.001**	**0.004**	**0.116**	**0.060**	**0.002**	**0.191**	**0.032**
(11)	L_2	**0.028**	**0.035**	**0.065**	**0.143**	**0.094**	**0.039**	**0.116**	**0.032**
	EM_N	0.055	0.054	0.259	3.641	0.426	0.109	0.456	0.159
(12)	L_2	**0.011**	**0.025**	**0.164**	**0.125**	**0.084**	**0.020**	**0.018**	**0.017**
	EM_N	0.029	0.041	1.589	0.278	0.294	0.179	0.186	0.096

10. Clarke, B. R. and Heathcote, C. R. (1994). Robust estimation of k-component univariate normal mixtures. *Ann. Inst. Statist. Math.,* **46**, 83-93.
11. Clarke, B. R. and Futshik, A. (2007). On the convergence of Newton's method when estimating higher dimensional parameters. *J. Multiv. Analysis,* **98**, 916-931.
12. Cutler, A. and Cordiero-Braña, O. I. (1996). Minimum Hellinger distance estimation for finite mixture models. *J. Am. Statist. Ass.,* **91**, 1716-1723.
13. Dempster, A. P., Laird, N. M and Rubin, D.P. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc., B.,* **39**,1-38.
14. Depraetere, N. and Vandebroek, M. (2014). Order selection in finite mixtures of linear regressions. *Statistical Papers,* **55**, 871-911.
15. Fisher, R.A. (1947). The analysis of covariance method for the relation between a part and the whole. *Biometrics,* **3**, 65-68.
16. Fryer, J. G. and Robertson, C. A. (1972). A comparison of some methods for estimating mixed normal distributions. *Biometrika,* **59**, 639-648.
17. Hasselman, B. (2013). nleqslv: Solve systems of non linear equations. R package version 2.0.
18. Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics., 2nd edn.* Hoboken, New Jersey: Wiley.
19. Klar, B. and Meintanis, S. G. (2005). Tests for normal mixtures based on the empirical characteristic function. *Comp. Statist. Data Analysis,* **49**, 227-242.
20. Lee, S. X. and McLachlan, G.J. (2013). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification,* **7**, (3), 241-266.
21. Lee, S. X. and McLachlan, G.J. (2014). EMMIXuskew: An R package for fitting mixtures of multivariate skew t-distributions via the EM algorithm. *J. Statist. Soft.,* **55**, (12), 1-22.
22. Macdonald, P. D. M. (1971). Comment on a paper by Choi,K. and Bulgren, W.G.. *J. R. Statist. Soc., B.,* **33**, 326-329.
23. McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions., 2nd edn.* New York: Wiley.
24. McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.,* New York: Wiley.
25. McLachlan, G. J., Peel, D., Basford, K. E. and Adams, P. (1999). The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software,* **4** No.

2
26. Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London, A., (1887-1895),* **185**, 71-110.
27. Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing,* **10**, 339-348.
28. Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Am. Statist. Ass.,* **73**, 730-738.
29. R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.r-project.org/.
30. Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review,* **26**, 195-239.
31. Seidel, W., Mosler, K. and Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Ann. Inst. Statist. Math.,* **52**, 481-487.
32. Seidel, W. and Ševčíková, H. (2004). Types of likelihood maxima in mixture models and their implication on the performance of tests. *Ann. Inst. Statist. Math.,* **56**, 631-654.
33. Tan, W. Y. and Chang,W. C. (1972). Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *J. Am. Statist. Ass.,* **67**, 702-708.
34. Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.
35. Wang, K., McLachlan,G. J., Ng,A., Peel,D. (2009). EMMIX-skew EM Algorithm for Mixture of Multivariate Skew Normal/t Distributions. EMMIX was originally written in Fortran by David Peel, R package version 1.0.20 *URL http://www.maths.uq.edu.au/ gjm/mix_soft/EMMIX-skew*
36. 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. Am. Statist. Ass.,* **79**, 590-598.
37. Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.,* **11**, 95-103.