ELSEVIER Journal of Computational and Applied Mathematics 84 (1997) 207–217

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

A comparison of some estimators of the mixture proportion of mixed normal distributions[1]

M.C. Pardo

Universitary School of Statistics, Complutense University of Madrid, 28040-Madrid, Spain

Received 29 October 1996; received in revised form 16 June 1997

Abstract

Fisher's method of maximum likelihood breaks down when applied to the problem of estimating the five parameters of a mixture of two normal densities from a continuous random sample of size n. Alternative methods based on minimum-distance estimation by grouping the underlying variable are proposed. Simulation results compare the efficiency as well as the robustness under symmetric departures from component normality of these estimators. Our results indicate that the estimator based on Rao's divergence is better than other classic ones.

Keywords: Minimum-distance estimator; Simulation; Relative efficiency

AMS classification: 62F10; 62F35

1. Introduction

Distributions which result from the mixing of two or more component distributions are designated as "compound" or "mixed" distributions. Such distributions arise in a wide variety of practical situations ranging from distributions of wind velocities to distributions of physical dimensions of various mass-produced items. The moment solution to the problem of estimating the five parameters of an arbitrary mixture of two unspecific normal densities was studied as early as 1894 by Karl Pearson [19]. Yet, despite the fact that many random phenomena have subsequently been shown to follow this distribution, it is only recently that the estimation problem has been seriously reconsidered. Hasselblad [12] seems to have been the first to reopen the question. Since then, the problem has also attracted the attention of Cohen [7], who shows how the computation of Pearson's moment method can be lightened to some extent. Maximum-likelihood estimates computed with all the information available are reported to be the best under all circumstances, however, they plainly misbehave in estimating mixed distributions because the maximum-likelihood function is not a bounded function

[1] This work was supported by Grant DGICYT PB94-0308.

PII S0377-0427(97)00124-6

in this case, Le Cam [15]. Then the maximum-likelihood procedure with original data cannot be universally recommended. So Day [9] and Behboodian [2] find an appropriate local maximum of the likelihood function by using iterative techniques. Fryer and Robertson [10] compare the moment estimates and the multinomial maximum likelihood and minimum χ^2 estimates obtained by grouping the underlying variable. They show that the grouped estimates are more accurate than the moment estimates for most distributions. Recently, Woodward et al. [24, 25] have carried out an interesting comparison between the maximum-likelihood estimator and the minimum distance estimators based on the Cramér–Von Mises and Hellinger distances, respectively.

In this paper we examine the use of minimum-distance estimation based on Burbea and Rao divergence [6] (MR_{ϕ_α}) as an alternative to maximum-likelihood (ML) estimation by grouping the underlying variable in both cases for the estimation of the parameters of the mixture density

$$f(x) = \lambda f_1(x) + (1 - \lambda) f_2(x)$$

when the component distributions in the simulated samples are normal and when they are not.

There is no doubt that the choice of the number of classes, M, to group the underlying variable is an important question. However, in this paper, it is not so important since we only compare estimators obtained by grouping the underlying variable. Therefore, we are interested in studying the behavior of the estimators under the same conditions. In fact, there are no papers related with the mixture of normal distributions which carry out the study of the choice of M. Fryer and Robertson [10] said "The method of grouping was dictated to a large extent by practical considerations, and it is not claimed that the groupings are in any sense optimal". We join with them in that feeling.

Anyway, we suppose that we need the estimations to construct a goodness-of-fit test. There are many papers that study the problem of choosing cells in this situation. One alternative is to do the same partition to estimate the parameters than to test the null hypothesis. This choice is guided by two considerations: the power of the resulting test, and the desire to use the asymptotic distribution of the statistic as an approximation to the exact distribution for sample size n. Mann and Wald [16] initiated the study of the choice of cells in the Pearson test of fit to a continuous distribution. They recommended, first, that the cells be chosen to have equal probabilities under the hypothesized distribution. The advantages of such a choice for the Pearson tests are (1) unbiasedness, (2) maximal power, and (3) empirical studies have shown that the asymptotic distribution of these statistics is a more accurate approximation to the exact distribution. Mann and Wald then made recommendations on the number M of equiprobable cells to be used. They found that for a sample of size n (large) and significance level α, one should use approximately $M = 4\,[2n^2/c(\alpha)^2]^{1/5}$, where $c(\alpha)$ is the upper α-point of the standard normal distribution. Retracting the Mann–Wald calculations using better approximations, as in Schorr [22], confirms that the optimum M is smaller than this value. He recommended to use $M = 2n^{2/5}$.

Another alternative is to consider different M values and to calculate for each one the corresponding estimator. The best M would be that corresponding to the estimator with less bias and mean-squared error.

In Section 2, we provide background material on the minimum R_{ϕ_α}-divergence estimator (MR_{ϕ_α} E). In Section 3 we carry out a simulation study for comparing the ML estimator (MLE), the minimum chi-square estimator (MCSE) and the MR_{ϕ_α} E with different α values for grouped data.

2. The minimum Burbea and Rao distance estimator

Consider the probability densities $f_\theta(x)$ with respect to a σ-finite measure μ on the statistical space $(\mathcal{X},\beta_{\mathcal{X}},P_\theta)_{\theta\in\Theta\subseteq R^{M_0}}$ and a decomposition $\{A_1,\ldots,A_M\}$ of $\mathcal{X}$. Then the formula $P_\theta(A_i)=q_i(\theta)$, $i=1,\ldots,M$ defines a discrete statistical model. Let $X_1,\ldots,X_n$ be a random sample drawn from the previous population and let $\hat{p}_i=n_i/n$ be the A_i relative frequency, $i=1,\ldots,M$. If we are interested in estimating θ by the maximum-likelihood method we have to maximize for fixed $(n_1,\ldots,n_M)$

$$P_\theta(N_1=n_1,\ldots,N_M=n_M)=\frac{n!}{n_1!\ldots n_M!}q_1(\theta)^{n_1}\ldots q_M(\theta)^{n_M},$$

so

$$\log P_\theta(N_1=n_1,\ldots,N_M=n_M)=-nD^{\text{KULLBACK}}(\hat{P},Q(\theta))+l,$$

where $\hat{P}=(\hat{p}_1,\ldots,\hat{p}_M)^t$, $Q(\theta)=(q_1(\theta),\ldots,q_M(\theta))^t$, D^{KULLBACK} the Kullback divergence [14] and l an independent value of θ. Then to estimate θ by the discrete model maximum-likelihood estimator is equivalent to minimize on $\theta\in\Theta\subseteq R^{M_0}$ the Kullback divergence.

Now, then the Kullback divergence is not the unique divergence measure, so we can choose as θ estimator the $\tilde{\theta}$ value which verifies the following:

$$D(\hat{P},Q(\tilde{\theta}))=\inf_{\theta\in\Theta\subseteq R^{M_0}}D(\hat{P},Q(\theta)),$$

D being every divergence measure.

Depending on the divergence measure chosen, you have different estimators. On the one hand, if

$$D(\hat{P},Q(\theta))=n\sum_{i=1}^{M}\frac{(\hat{p}_i-q_i(\theta))^2}{q_i(\theta)}$$

the corresponding $\tilde{\theta}$ is the well-known minimum χ^2 estimator, studied in this context by Fryer and Robertson [10].

On the other hand, if we consider the Burbea and Rao divergence [6],

$$D(\hat{P},Q(\theta))=R_{\phi_\alpha}(\hat{P},Q(\theta))=\tfrac{1}{2}\left(H_{n,\phi_\alpha}(\hat{P})+H_{n,\phi_\alpha}(Q(\theta))\right)-H_{n,\phi_\alpha}\left(\frac{\hat{P}+Q(\theta)}{2}\right),$$

where

$$H_{n,\phi_\alpha}(P)=\begin{cases}\sum_{i=1}^{M}\frac{1}{1-\alpha}(p_i^\alpha-p_i) & \alpha\neq 1,\\ -p_i\ln p_i & \alpha=1\end{cases}$$

is the entropy of degree α due to Havrda and Charvát [13], the corresponding $\tilde{\theta}$ will be called the minimum R_{ϕ_α}-divergence estimator. Rao [21] used the family of ϕ_α-entropies in genetic diversity between populations. In the particular case of $\alpha=2$, we obtain the Gini–Simpson index. This measure of entropy was introduced by Gini [11] and by Simpson [23] in biometry and its properties have been studied by various authors (Bhargava and Doyle [3], Bhargava and Uppulari [4], Agresti and

Agresti [1]). Note that if we consider the Gini–Simpson index, then the associated R_ϕ-divergence is proportional to the square of the Euclidean distance

$$R_{\phi_2}(\pi^1,\pi^2)=\frac{1}{4}\sum_{i=1}^{M}\left(\pi_i^1-\pi_i^2\right)^2.$$

In order to solve the problem to estimate the mixture proportion of mixed-normal distributions, we will define in a convenient way the R_{ϕ_α}-divergence estimator. The following definition was given in Pardo [17].

Definition 1. Let us suppose that n observations are drawn at random and with replacement from a population with statistical space $(\mathcal{X},\beta_{\mathcal{X}},P_\theta)_{\theta\in\Theta\subseteq R^{M_0}}$, the minimum R_{ϕ_α}-divergence estimator of θ is every $\hat{\theta}_{\phi_\alpha}\in\overline{\Theta}$ that verifies

$$R_{\phi_\alpha}(\hat{P},Q(\hat{\theta}_{\phi_\alpha}))=\inf_{\theta\in\Theta\subseteq R^{M_0}} R_{\phi_\alpha}(\hat{P},Q(\theta)),$$

where $\hat{P}$ is the relative frequency vector. So the minimum R_{ϕ_α}-divergence estimator will be $\hat{\theta}_\phi = \arg\inf_{\theta\in\Theta\subseteq R^{M_0}} R_{\phi_\alpha}(\hat{P},Q(\theta))$.

The importance of the family of divergence measures considered in the previous definition can be seen in the aforementioned paper of Burbea and Rao [6]. For example, a surprising result is the fact that the R_{ϕ_α}-divergence is convex on $\Delta_M\times\Delta_M$, where $\Delta_M=\{(p_1,\ldots,p_M)^t/\sum_{i=1}^M p_i=1,\ p_i\geqslant 0,\ i=1,\ldots,M\}$, if only if $\alpha\in[1,2]$ for $M>2$, and if only if $\alpha\in[1,2]$ or $\alpha\in\left[3,\frac{11}{3}\right]$ for $M=2$. This establishes the range of α for which this measure is useful in practical applications. Some important properties of this divergence family can be seen in Pardo and Vajda [18].

Throughout, we assume the model is correct and $M_0<M-1$. Furthermore, we restrict ourselves to unknown parameters θ^0 satisfying the regularity conditions of Birch [5] which are neccesary to prove that the MLE for grouped data is asymptotically distributed as a normal. Consider

$$A(\theta)=\operatorname{diag}\left(Q(\theta)^{(\alpha/2)-1}\right)J(\theta),$$

where

$$J(\theta)=(J_{jr}(\theta))_{\substack{j=1,\ldots,M\\ r=1,\ldots,M_0}}$$

is an $M\times M_0$ Jacobian matrix being

$$J_{jr}(\theta)=\frac{\partial q_j(\theta)}{\partial\theta_r}.$$

Then, assuming that the function $Q:\Theta\to\Delta_M$ has continuous second partial derivatives in a neighborhood of θ^0, the following asymptotic properties were shown in Pardo [17]:

(1)

$$\widehat{\theta}_{\phi_\alpha}=\theta^0+(A(\theta^0)^tA(\theta^0))^{-1}A(\theta^0)^t\operatorname{diag}\left(Q(\theta^0)^{(\alpha/2)-1}\right)(\hat{P}-Q(\theta^0))+o(\|\hat{P}-Q(\theta^0)\|),$$

where $\widehat{\theta}_{\phi_\alpha}$ is unique in a neighborhood of θ^0.

(2)

$$\sqrt{n}(\hat{\theta}_{\phi_\alpha} - \theta^0) \approx \mathrm{N}(0, \Sigma),$$

where

$$\Sigma = B(\theta^0)\,(\mathrm{diag}(Q(\theta^0)) - Q(\theta^0)Q(\theta^0)^{\mathrm{t}})\,B(\theta^0)^{\mathrm{t}}$$

with

$$B(\theta^0) = (A(\theta^0)^{\mathrm{t}}A(\theta^0))^{-1}A(\theta^0)^{\mathrm{t}}\,\mathrm{diag}\left(Q(\theta)^{(\alpha/2)-1}\right).$$

(3)
$Q(\hat{\theta}_\phi)$ is a $\sqrt{n}$-consistent estimator of $Q(\theta^0)$, i.e.,

$$\sqrt{n}\parallel Q(\hat{\theta}_\phi) - Q(\theta^0)\parallel \leqslant \mathrm{O}_p(1).$$

Remark 1. We note that if we consider the R-divergence or equivalently $\alpha \to 1$, we get that

$$\hat{\theta}_{\phi_1} = \theta^0 + (A(\theta^0)^{\mathrm{t}}A(\theta^0))^{-1}A(\theta^0)^{\mathrm{t}}\,\mathrm{diag}\left(Q(\theta^0)^{-1/2}\right)(\hat{P} - Q(\theta^0)) + \mathrm{o}(\parallel \hat{P} - Q(\theta^0)\parallel),$$

where

$$A(\theta) = \mathrm{diag}\left(Q(\theta^0)^{-1/2}\right)J(\theta).$$

and

$$\sqrt{n}(\hat{\theta}_{\phi_1} - \theta^0)^{\mathrm{t}} \to \mathrm{N}(0, I(\theta^0)^{-1}),$$

being $I(\theta)$ the Fisher information matrix. So the $\hat{\theta}_{\phi_1}$ estimator is a BAN (best asymptotically normal) estimator.

In the following section we will present a simulation study to know the behavior of our estimator.

3. Simulation results

In this section we report the results of simulations designed to empirically compare ML, minimum chi-square (MCS), MR_{ϕ_2} and MR_{ϕ_1} estimations of parameters for a mixture of normal, in which we analyze the efficiency as well as the robustness of them. Simulations reported in this section are based on mixing proportions 0.25, 0.5 and 0.75. For each of these mixing proportions, firstly, we considered mixtures of the densities $f_1(x)$ and $f_2(x)$, where $f_1(x)$ is the density for the random variable $X = aY$ and $f_2(x)$ is the density associated with $X = Y + b$ where $a > 0$ and $b > 0$ and the distribution of Y is normal. Secondly, we consider Y as a Student's t with two or four degrees of freedom, or double exponential, to study the robustness under symmetric departures from component normality. Thus, "a" is the ratio of scale parameters which we take to be 1 and $\sqrt{2}$ while "b" was selected to provide the desired overlap between the two distributions. We considered "overlap" [24] as the probability of misclassification using this rule: Classify an observation x as being from population 1 if $x < x_c$ and from population 2 if $x \geqslant x_c$, where x_c is the unique point between μ_1

Table 1
Simulation results for mixtures of normal components

			0.1 Overlap			0.03 Overlap		
λ	a	Estimator	$\widehat{\text{Bias}}$	$\widehat{\text{nMSE}}$	$\widehat{E}$	$\widehat{\text{Bias}}$	$\widehat{\text{nMSE}}$	$\widehat{E}$
0.25	1	MR_{ϕ_2}E	0.070	1.99	1.02	0.042	0.82	1.11
		MR_{ϕ_1}E	0.068	1.89	0.96	0.042	0.81	1.10
		MLE	0.063	1.95		0.037	0.73	
		MCSE	0.060	1.95	1.00	0.036	0.78	1.06
0.5	1	MR_{ϕ_2}E	−0.009	1.30	0.85	−0.008	0.71	0.89
		MR_{ϕ_1}E	−0.007	1.28	0.84	−0.010	0.72	0.90
		MLE	−0.015	1.52		−0.012	0.79	
		MCSE	−0.016	1.52	1.00	−0.011	0.78	0.97
0.25	$\sqrt{2}$	MR_{ϕ_2}E	−0.020	0.874	0.941	0.001	0.571	0.993
		MR_{ϕ_1}E	−0.021	0.937	1.010	0.003	0.562	0.976
		MLE	−0.025	0.928		−0.001	0.575	
		MCSE	−0.023	0.977	1.052	−0.001	0.570	0.991
0.5	$\sqrt{2}$	MR_{ϕ_2}E	−0.099	2.004	0.937	−0.058	1.027	0.951
		MR_{ϕ_1}E	−0.099	2.010	0.940	−0.058	1.027	0.951
		MLE	−0.101	2.137		−0.058	1.079	
		MCSE	−0.098	2.030	0.949	−0.060	1.042	0.965
0.75	$\sqrt{2}$	MR_{ϕ_2}E	−0.166	4.32	0.94	−0.085	1.56	1.07
		MR_{ϕ_1}E	−0.167	4.39	0.96	−0.084	1.48	1.01
		MLE	−0.167	4.57		−0.076	1.46	
		MCSE	−0.173	4.97	1.09	−0.077	1.52	1.03

and μ_2 such that $\lambda f_1(x_c) = (1-\lambda)f_2(x_c)$. The overlaps examined in the current study are 0.03 and 0.1.

For each set of configurations considered, 500 samples of size $n = 100$ were generated from the corresponding mixture distribution, and for each sample considered the ML, MCS, MR_{ϕ_2} and MR_{ϕ_1} estimates were obtained. The iterative procedure proposed in Section 2 was implemented using the IMSL subroutine ZXMIN which minimize a function of various variables. Although all these estimation procedures provide estimators of all five parameters, Tables 1 and 2 present only estimations of λ. In Table 3 estimations for all the parameters are given.

For either of the estimators proposed in the previous section to be used in practice, one must have starting values for the iterative procedures. We chose to obtain these values by employing an ad hoc quasi-clustering technique used by Woodward et al. [24] that is easy to implement. They allow as possible values for the initial estimate of λ only the values $0.1,\ldots,0.9$. For each of these values of λ, the sample is divided into two subsamples, $Y_1, Y_2, \ldots, Y_{n_1}$ and $Y_{n_1+1}, Y_{n_1+2}, \ldots, Y_n$, where Y_i is the ith order statistic and n_1 is "$n\lambda$" rounded to the nearest integer. So, $\widehat{\lambda}$ is that value at which $\lambda(1-\lambda)(m_1 - m_2)^2$ is maximized, $\widehat{\mu}_1 = m_1$, $\widehat{\mu}_2 = m_2$, $\widehat{\sigma}_1^2 = ((m_1 - r_1^{(0.25)})/0.6745)^2$ and $\widehat{\sigma}_2^2 = ((r_2^{(0.75)} - m_2)/0.6745)^2$, where m_j is the sample median of the jth subsample and $r_j^{(q)}$ is the qth percentile from the jth subsample.

In Table 1 we present summary results of the simulation comparing the performance of the estimators for mixtures of normal components. Estimates of the bias and mean-squared error (MSE)

Table 2
Simulation results for mixtures of nonnormal components

			0.1	Overlap		0.03	Overlap	
λ	a	Estimator	$\widehat{\text{Bias}}$	$n\widehat{\text{MSE}}$	$\widehat{E}$	$\widehat{\text{Bias}}$	$n\widehat{\text{MSE}}$	$\widehat{E}$
		Double exponential components						
0.25	1	$MR_{\phi_2}E$	0.059	1.069	1.066	0.061	0.861	1.194
		$MR_{\phi_1}E$	0.062	1.103	1.100	0.064	0.903	1.253
		MLE	0.053	1.002		0.053	0.721	
		MCSE	0.053	0.986	0.9837	0.051	0.726	1.007
0.5	1	$MR_{\phi_2}E$	−0.001	0.794	1.106	0.0008	0.400	1.023
		$MR_{\phi_1}E$	−0.002	0.748	1.041	−0.001	0.416	1.063
		MLE	−0.003	0.718		−0.001	0.391	
		MCSE	−0.003	0.718	1.000	0.001	0.415	1.060
0.25	$\sqrt{2}$	$MR_{\phi_2}E$	0.012	0.702	1.074	0.039	0.572	1.022
		$MR_{\phi_1}E$	0.016	0.690	1.057	0.041	0.607	1.086
		MLE	0.013	0.653		0.035	0.559	
		MCSE	0.012	0.632	0.967	0.034	0.560	1.001
0.5	$\sqrt{2}$	$MR_{\phi_2}E$	−0.056	1.178	1.146	−0.032	0.588	1.066
		$MR_{\phi_1}E$	−0.052	1.071	1.042	−0.031	0.577	1.045
		MLE	−0.047	1.027		−0.029	0.552	
		MCSE	−0.053	1.128	1.098	−0.026	0.566	1.026
0.75	$\sqrt{2}$	$MR_{\phi_2}E$	−0.106	1.950	0.927	−0.090	1.392	1.290
		$MR_{\phi_1}E$	−0.108	2.027	0.963	−0.087	1.366	1.266
		MLE	−0.098	2.103		−0.073	1.079	
		MCSE	−0.090	1.859	0.883	−0.074	1.133	1.049
		$t(4)$ components						
0.25	1	$MR_{\phi_2}E$	0.084	2.608	0.951	0.051	0.941	1.039
		$MR_{\phi_1}E$	0.085	2.673	0.975	0.051	0.949	1.048
		MLE	0.085	2.742		0.047	0.905	
		MCSE	0.077	2.478	0.903	0.047	0.900	0.994
0.5	1	$MR_{\phi_2}E$	−0.005	1.548	0.888	−0.006	0.792	0.924
		$MR_{\phi_1}E$	−0.003	1.529	0.877	−0.006	0.778	0.908
		MLE	−0.008	1.742		−0.007	0.856	
		MCSE	−0.009	1.791	1.028	−0.008	0.857	1.000
0.25	$\sqrt{2}$	$MR_{\phi_2}E$	−0.002	0.962	0.940	0.010	0.559	1.011
		$MR_{\phi_1}E$	−0.005	0.946	0.925	0.008	0.516	0.934
		MLE	−0.009	1.022		0.005	0.552	
		MCSE	−0.009	0.983	0.962	0.006	0.559	1.011
0.5	$\sqrt{2}$	$MR_{\phi_2}E$	−0.094	2.290	0.892	−0.062	1.102	0.912
		$MR_{\phi_1}E$	−0.096	2.316	0.903	−0.063	1.145	0.948
		MLE	−0.099	2.565		−0.058	1.208	
		MCSE	−0.097	2.516	0.981	−0.057	1.121	0.928

Table 2. (Contd.)

λ	a	Estimator	0.1 Overlap $\widehat{\text{Bias}}$	nMSE	$\widehat{E}$	0.03 Overlap $\widehat{\text{Bias}}$	nMSE	$\widehat{E}$
0.75	$\sqrt{2}$	$MR_{\phi_2}E$	−0.182	5.257	0.923	−0.104	1.894	0.986
		$MR_{\phi_1}E$	−0.182	5.375	0.944	−0.106	2.037	1.061
		MLE	−0.184	5.692		−0.097	1.920	
		MCSE	−0.181	5.567	0.978	−0.095	1.854	0.965
				$t(2)$ components				
0.25	1	$MR_{\phi_2}E$	0.113	6.633	0.991	0.063	1.515	1.030
		$MR_{\phi_1}E$	0.113	6.657	0.995	0.063	1.551	1.055
		MLE	0.109	6.689		0.059	1.470	
		MCSE	0.108	6.629	0.991	0.057	1.448	0.985
0.5	1	$MR_{\phi_2}E$	−0.011	3.594	0.900	−0.002	1.150	1.064
		$MR_{\phi_1}E$	−0.011	3.633	0.910	−0.003	1.067	0.988
		MLE	−0.014	3.992		−0.005	1.080	
		MCSE	−0.017	3.863	0.967	−0.005	1.120	1.036
0.25	$\sqrt{2}$	$MR_{\phi_2}E$	0.035	4.372	0.945	0.012	0.852	0.982
		$MR_{\phi_1}E$	0.034	4.566	0.987	0.008	0.864	0.996
		MLE	0.038	4.622		0.011	0.867	
		MCSE	0.032	4.659	1.007	0.013	0.787	0.908
0.5	$\sqrt{2}$	$MR_{\phi_2}E$	−0.117	4.255	0.894	−0.070	1.697	0.889
		$MR_{\phi_1}E$	−0.117	4.366	0.917	−0.069	1.675	0.878
		MLE	−0.124	4.759		−0.070	1.907	
		MCSE	−0.124	4.810	1.010	−0.069	1.745	0.914
0.75	$\sqrt{2}$	$MR_{\phi_2}E$	−0.225	10.760	0.966	−0.122	3.202	0.991
		$MR_{\phi_1}E$	−0.224	10.580	0.950	−0.120	3.258	1.008
		MLE	−0.231	11.131		−0.115	3.230	
		MCSE	−0.232	11.259	1.011	−0.112	3.129	0.968

based on the simulations are given by

$$\widehat{\text{bias}} = \frac{1}{n_s}\sum_{i=1}^{n_s}(\widehat{\lambda}_i - \lambda)$$

and

$$\widehat{\text{MSE}} = \frac{1}{n_s}\sum_{i=1}^{n_s}(\widehat{\lambda}_i - \lambda)^2,$$

where n_s the number of samples and $\widehat{\lambda}_i$ denotes an estimate of λ for the ith sample. It should be noted that $n\widehat{\text{MSE}}$ is the quantity actually given in the tables. We also table empirical measures of

Table 3
Estimated relative efficiencies of the MR_{ϕ_2} E, MR_{ϕ_1} E and MCSE relative to the MLE for mixture model parameters

			0.1					0.03				
λ	a	Estimator	μ_1	σ_1	μ_2	σ_2	λ	μ_1	σ_1	μ_2	σ_2	λ
			Normal									
0.25	1	$MR_{\phi_2}E$	0.97	0.87	0.99	1.72	1.02	0.95	0.88	1.11	1.05	1.11
		$MR_{\phi_1}E$	0.98	0.93	0.98	1.06	0.96	0.99	0.93	1.07	1.05	1.10
		MCSE	1.06	1.00	0.97	1.01	1.00	1.01	1.03	1.03	0.98	1.06
0.5	1	$MR_{\phi_2}E$	0.97	0.95	0.90	0.90	0.85	1.00	0.97	0.89	0.86	0.89
		$MR_{\phi_1}E$	0.96	0.99	0.82	0.89	0.84	1.05	1.00	0.86	0.92	0.90
		MCSE	1.02	1.07	1.01	0.96	1.00	1.00	1.00	0.97	0.95	0.97
0.25	$\sqrt{2}$	$MR_{\phi_2}E$	0.95	0.84	1.02	1.05	0.94	0.94	0.80	1.03	1.06	0.99
		$MR_{\phi_1}E$	0.96	0.84	1.04	1.04	1.01	0.92	0.79	1.04	1.06	0.97
		MCSE	0.98	0.98	1.02	0.99	1.05	0.98	0.92	0.96	1.02	0.99
0.5	$\sqrt{2}$	$MR_{\phi_2}E$	0.97	0.83	0.81	0.83	0.93	1.03	1.08	0.85	0.81	0.95
		$MR_{\phi_1}E$	0.97	0.83	0.82	0.76	0.94	1.03	1.11	0.88	0.83	0.95
		MCSE	1.01	0.95	0.98	0.98	0.94	1.02	1.12	0.92	0.90	0.96
0.75	$\sqrt{2}$	$MR_{\phi_2}E$	0.94	0.92	0.92	0.90	0.94	1.08	1.02	0.82	0.69	1.07
		$MR_{\phi_1}E$	0.95	0.92	0.90	0.85	0.96	1.06	0.95	0.79	0.69	1.01
		MCSE	0.94	0.93	0.97	0.97	1.09	1.02	1.00	0.95	0.97	1.03
			$t(4)$									
0.25	1	$MR_{\phi_2}E$	1.19	0.80	1.07	0.97	0.95	0.82	0.70	1.13	1.10	1.03
		$MR_{\phi_1}E$	0.98	0.76	1.16	0.95	0.97	0.78	0.77	1.15	1.10	1.04
		MCSE	1.02	1.01	1.06	0.95	0.90	1.03	0.90	1.00	1.02	0.99
0.5	1	$MR_{\phi_2}E$	0.94	1.04	0.95	0.80	0.88	1.03	1.08	0.85	0.81	0.92
		$MR_{\phi_1}E$	0.94	1.01	0.88	0.90	0.87	1.03	1.11	0.88	0.83	0.90
		MCSE	0.99	1.06	1.03	1.00	1.02	1.02	1.12	0.92	0.90	1.00
0.25	$\sqrt{2}$	$MR_{\phi_2}E$	0.96	0.83	1.05	0.91	0.94	0.83	0.93	0.98	1.04	1.01
		$MR_{\phi_1}E$	0.96	0.81	1.07	0.95	0.92	3.14	0.91	0.96	0.97	0.93
		MCSE	0.99	0.97	1.02	0.96	0.96	0.91	1.00	0.89	0.97	1.01
0.5	$\sqrt{2}$	$MR_{\phi_2}E$	0.93	0.95	0.89	0.81	0.89	1.01	0.89	0.85	0.68	0.91
		$MR_{\phi_1}E$	0.93	1.02	0.94	0.89	0.90	1.00	0.96	0.81	0.69	0.94
		MCSE	0.99	1.06	0.96	0.96	0.98	0.95	0.90	0.97	0.89	0.92
0.75	$\sqrt{2}$	$MR_{\phi_2}E$	0.98	0.90	0.84	0.86	0.92	1.08	1.07	0.76	0.66	0.98
		$MR_{\phi_1}E$	1.00	0.91	0.88	0.91	0.94	1.09	1.08	0.87	0.82	1.06
		MCSE	0.97	1.06	1.02	1.08	0.97	0.99	1.00	0.95	1.09	0.96

the relative efficiencies of the MCSE, $MR_{\phi_1}E$ and $MR_{\phi_2}E$ with the MLE, i.e.,

$$\hat{E} = \frac{\widehat{\mathrm{MSE}}(\neq \mathrm{MLE})}{\widehat{\mathrm{MSE}}(\mathrm{MLE})}.$$

Analyzing the results of Table 1, we find that the estimated bias and the $\widehat{\mathrm{MSE}}$ associated with the $MR_{\phi_2}E$ are generally smaller than those for the MLE and the MCSE.

In Table 2 we display the results for the nonnormal components. In the case of double-exponential components is not clear which is the best because, in general, the MCSE has less $\widehat{\text{Bias}}$ than the others but the MLE has less $\widehat{\text{MSE}}$. However, in the case of Student's t components are very clear than the $\text{MR}_{\phi_\lambda}\text{E}$ is the best. It has less $\widehat{\text{MSE}}$ and $\widehat{\text{Bias}}$ than the others. The superiority of it is even clear for $t(2)$ components, i.e., when the departure from normality is more extreme. In this setting the performance of the MLE and MCSE further deteriorates with respect to that of the $\text{MR}_{\phi_\lambda}\text{E}$.

Although our emphasis here has been on the estimation of the mixing proportion, the estimation routines used here obtain estimation for all five of the parameters. So, it seems obvious to question about whether the results shown for λ are similar for the rest of the parameters μ_1, σ_1, μ_2 and σ_2. In Table 3 we display empirical relative efficiencies for all the parameters for normal and $t(4)$ mixtures. From the table we see that the results for the other parameters also exhibited patterns similar to those shown in Tables 1 and 2, i.e., the $\text{MR}_{\phi_\lambda}\text{E}$ is a very attractive alternative to both the MLE and the MCSE.

4. Concluding remarks

Our results indicate that the $\text{MR}_{\phi_\lambda}\text{E}$ is better than the MLE and MCSE at the true model and under the Students' t components. While all of them perform comparably under the double-exponential components. As would be expected, the performance of the estimators declines as the overlap between the two components increases.

References

[1] A. Agresti, R.F. Agresti, Statistical analysis of qualitative variation, K.E. Schussler (Ed.,), Soc. Methodology, (1978) pp. 204–237.
[2] J. Behboodian, On a mixture of normal distributions, Biometrika 57 (1970) 215–217.
[3] T.N. Bhargava, P.H. Doyle, A geometric study of diversity, J. Theoret. Biol. 43 (1974) 241–251.
[4] T.N. Bhargava, V.R.R. Uppuluri, On an axiomatic derivation of Gini diversity, with applications, Metron 30-VI (1975) 1–13.
[5] M.W. Birch, A new proof of the Pearson–Fisher theorem, Ann. Math. Statist. 35 (1964) 817–824.
[6] J. Burbea, C.R. Rao, On the convexity of some divergence measures based on entropy functions, IEEE Trans. Inform. Theory 28 (1982) 489–495.
[7] A.C. Cohen, Estimation in mixtures of two normal distributions, Technometrics 9 (1967) 15–28.
[8] I. Csiszár, A class of measures of informativity of observation channels, Period. Math. Hungar. 2 (1972) 191–213.
[9] N.E. Day, Estimating the components of a mixture of normal distributions, Biometrika 56(3) (1969) 463–474.
[10] J.G. Fryer, C.A. Robertson, A comparison of some methods for estimating mixed normal distributions, Biometrika 59(3) (1972) 639–648.
[11] C. Gini, Variabilitá e mutabilitá, Studi Economico-Giuridici della Facolta di Giiurisprudenza dell Universitá di Cagliari, a III, PArte II, (1912).
[12] V. Hassenblad, Estimation of parameters for a mixture of normal distributions, Technometrics 8 (1966) 431–434.
[13] M.E. Havrda, F. Charvát, Quantification method of classification processes: concept of structural α-entropy, Kybernetika 3 (1967) 30–35.
[14] S. Kullback, R. Leibler, On information and sufficiency. Ann. Math. Statist. 22 (1951) 79–86.
[15] L. Le Cam, Maximum likelihood: An introduction, Internat. Statist. Rev. 58(2) (1990) 153–171.

[16] H.B. Mann, A. Wald, On the choice of the number of class intervals in the application of the chi-square test, Ann. Math. Statist. 13 (1942) 306–317.
[17] M.C. Pardo, Asymptotic behaviour of an estimator based on Rao's divergence, Kybernetika 33 (1997).
[18] M.C. Pardo, I. Vajda, About distances of discrete distributions satisfying the data proccesing theorem of information theory, Trans. IEEE Inform. Theory 43(4) (1997) 1288–1293.
[19] K. Pearson, Contributions to the mathematical theory of evolution, Philos. Trans. Roy. Soc. Ser.A 185 (1894) 71–110.
[20] K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philos. Mag. 50 (1900) 157–172.
[21] C.R. Rao, Diversity and dissimilarity coefficients: an unified approach, J. Theoret. Pop. Biol. 21 (1982) 24–43.
[22] A.R. Schorr, On the choice of the class intervals in the application of the chi-square test, Math. Oper. Forsch. u. Statist. 5 (1974) 357–377.
[23] E.H. Simpson, Measurement of diversity, Nature 163 (1949) 688.
[24] W.A. Woodward, W.C. Parr, W.R. Schucany, H. Lindsay, A comparison of minimum distance and maximum likelihood estimation of a mixture proportion, J. Amer. Statist. Assoc. 79 (1984) 590–598.
[25] W.A. Woodward, P. Whitney, P.W. Eslinger, Minimum Hellinger distance estimation of mixture proportions, J. Statist. Plann. Inference 48 (1995) 303–319.