Far East J.Theo.Stat. 20(2) (2006), 217-240.

MAXIMUM LIKELIHOOD ESTIMATORS FOR NORMAL AND GAMMA MIXTURES

K.O. BOWMAN[1] and L.R. SHENTON[2]

[1]Computational Sciences and Engineering Division
Oak Ridge National Laboratory
P.O.Box 2008, 4500N, MS-6191
Oak Ridge, TN 37831-6191, USA
bowmanko@ornl.gov
[2]Department of Statistics
University of Georgia
Athens, Georgia 30602, USA

September 25, 2006

Abstract

First of all we consider iterative schemes to discover solutions for maximum likelihood estimators for normal mixtures, and gamma mixtures. From our examples, the iterative scheme works well for the normal, but poorly for the gamma mixtures. Second, we set up short tabulations of the low order asymptotic moments (bias, variances to order N^{-1}, and N^{-2}, skewness to order $1/\sqrt{N}$, N being the sample size). These moments are derived from a Maple code interpretation due to Bowman and Shenton (Bias, variance, skewness and kurtosis of maximum likelihood estimator using Maple, *Far East J.Theo.Stat.* 17(2), (2005), 137-195). In conclusion this paper supplies further evidence for the correctness of the Maple code.

Keywords and phrases: Asymptotic skewness, asymptotic variance, iterative schemes.

1 Introduction

The two component normal mixture distribution involves 5 parameters, two means, two standard deviations and a proportion. Karl Pearson (1894) studied this distributional model since fitting a single normal component was found to be inadequate. Pearson used the method of moments focusing on the three sample cumulants κ_3, κ_4, and κ_5, relating these to the estimators of the 5 parameter normal mixture involves tedious elementally mathematics. So the Pearson study became less interesting. Later work, in particular, involved steepest descent and Newton-Raphson approaches initiated by Hald (1952) and Hasselblad (1966) and using the maximum likelihood approaches.

We have advanced the subject in two directions. First we have a Maple code which produces low order moments of maximum likelihood estimators, only requiring a definition of the basic probability function (Bowman and Shenton, 2005). This code is based on our study (Shenton and Bowman, 1977) which sets up a Taylor series for the logarithms of the likelihood function. Several examples verifying the formula are available. Since several parameters may be involved, integration problems in the Maple code may be a problem. Secondly we are now introducing an iterative scheme which applies to the general case of s component ($3s - 1$ parameters). The iterative scheme is based on the logarithmic derivatives of the probability function with respect to the parameters (means, standard deviations or variances, proportions). In particular this derivative with respect to the proportion parameter plays a dominant role leading to a criterion for convergence. The iterative scheme is simple in form, easy to relate to a computer. Examples included relate to biological data, percentage ash content in peat, and release time of a relay. The ash content in peat data provides a remarkable example for which the χ^2-value take the value 5.74 (no tail grouping being involved). There is a one in a hundred chance of a lower χ^2 value. When grouping is involved the χ^2 value is 5.41, with a very satisfactory probability of acceptance. We note that a χ^2 for this case had been computed by Hasselblad, but otherwise these appears to be no case in the literature showing the complete table of goodness of fit and theory vs observed. Such a marvelous example of statistical model building has been ignored. The actual source of the ash content data is not given in Hald. However, he does suggest that since a two component normal mixture does give a good fit, then there may be two sources for the samples of peat drawn.

It is assumed that the data base is divided into a set of equal intervals, the frequencies in order being n_1, n_2, $\cdots$, n_x, $\cdots$ for a sample of size $N = n_1 + n_2 + \cdots$. The variate n_x is the ordinate at the center of the xth interval. In a sense when, grouping of data is appropriate, we are looking at the discrete case.

Another widely used continuous distribution is the gamma with mean $a\rho$ and variance $a^2\rho$. In section 4 we give expressions for the low order moments of maximum likelihood estimators.

2 Maximum likelihood equations for normal mixtures

$$P(X = x) = \pi_1 P_1 + \pi_2 P_2 = P(x) = P \tag{1}$$

and

$$P_i = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2\sigma_i^2}(x-\lambda_i)^2}$$

where $0 < \pi_i < 1$, $\sigma_i > 0$, $i = 1, 2$, $-\infty < x < \infty$.

Now for $i = 1$,

$$\frac{\partial \ln P(x)}{\partial \pi_1} = \frac{(P_1 - P_2)}{P} \tag{2}$$

and

$$\pi_1 \frac{\partial \ln P(x)}{\partial \pi_1} = \frac{(\pi_1 P_1 - \pi_1 P_2)}{P} = \frac{P - P_2}{P}.$$

Hence $\frac{\partial \ln P(x)}{\partial \pi} = 0$ leads to

$$\begin{cases} \sum \frac{n_x}{N} \frac{\hat{P}_2}{\hat{P}} = 1 \\ \sum \frac{n_x}{N} \frac{\hat{P}_1}{\hat{P}} = 1 \end{cases} \tag{3}$$

for maximum likelihood estimators $\hat{\pi}_1$, $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\sigma}_1$, $\hat{\sigma}_2$.

Now following the description in section 2 of Bowman and Shenton (2006) we have

$$_{new}\pi_1 = \pi_1 \sum \frac{n_x}{N} \frac{\tilde{P}_1}{\tilde{P}}$$

and $_{new}\pi_2 = 1 -_{new} \pi_1$. Moreover

$$\frac{\partial P_1}{\partial \lambda_1} = P_1 \frac{\partial \ln P_1}{\partial \lambda_1} = \pi_1 P_1 \left\{ \frac{(x - \lambda_1)}{\sigma_1^2} \right\}$$

and

$$\frac{\partial \ln P(x)}{\partial \lambda_1} = \frac{\pi_1 P_1 \frac{(x-\lambda_1)}{\sigma_1^2}}{P}$$

so leading to

$$_{new}\hat{\lambda_1} = \sum \frac{x n_x \tilde{P}_1}{\sigma_1^2 \tilde{P}} / \sum \frac{n_x}{N} \frac{\tilde{P}_1}{\tilde{P}}.$$

Now the denominator will ultimately reach unity from (3), if there is convergence, so we use

$$_{new}\hat{\lambda}_1 = \sum \frac{x n_x}{N} \frac{\tilde{P}_1}{\tilde{P}}.$$

Similarly

$$_{new}\hat{\lambda}_2 = \sum \frac{x n_x}{N} \frac{\tilde{P}_2}{\tilde{P}}.$$

For the σ's we consider

$$\ln P_1 \left\{ -\ln(\sqrt{2\pi}) - \ln \sigma_1 - \frac{(x - \lambda_1)^2}{2\sigma^2} \right\},$$

and

$$\frac{\partial \ln P}{\partial \sigma_1} = P_1 \left\{ -\frac{1}{\sigma_1} + \frac{(x-\lambda_1)^2}{\sigma^3} \right\} / P$$

so that

$$_{new}\hat{\sigma}_1^2 = \sum \frac{n_x}{N} \frac{(x-\lambda_1)^2 \tilde{P}_1}{\tilde{P}}.$$

Similarly

$$_{new}\hat{\sigma}_2^2 = \sum \frac{n_x}{N} \frac{(x-\lambda_2)^2 \tilde{P}_2}{\tilde{P}}.$$

The notation here uses $\tilde{P}_1$, $\tilde{P}_2$, and $\tilde{P}$ to refer to the parameters π_1, λ_1, λ_2,σ_1, σ_2 brought up to date with the phase of the cycle involved.

For the 3 component normal mixture, parameters are λ_1, σ_1, λ_2, σ_2, λ_3, σ_3, π_1, and π_2. We have

$$_{new}\hat{\pi}_1 = \pi_1 \sum \frac{n_x}{N} \frac{\tilde{P}_1}{\tilde{P}},$$

using this find

$$_{new}\hat{\pi}_2 = \pi_2 \sum \frac{n_x}{N} \frac{\tilde{P}_2}{\tilde{P}},$$

$$\begin{aligned}
_{new}\hat{\lambda}_1 &= \sum \frac{x n_x}{N} \frac{\tilde{P}_1}{\tilde{P}}, \quad & _{new}\hat{\sigma}_1^2 &= \sum \frac{n_x}{N} (x-\hat{\lambda}_1)^2 \frac{\tilde{P}_1}{\tilde{P}}, \\
_{new}\hat{\lambda}_2 &= \sum \frac{x n_x}{N} \frac{\tilde{P}_2}{\tilde{P}}, \quad & _{new}\hat{\sigma}_2^2 &= \sum \frac{n_x}{N} (x-\hat{\lambda}_2)^2 \frac{\tilde{P}_2}{\tilde{P}}, \\
_{new}\hat{\lambda}_3 &= \sum \frac{x n_x}{N} \frac{\tilde{P}_3}{\tilde{P}}, \quad & _{new}\hat{\sigma}_3^2 &= \sum \frac{n_x}{N} (x-\hat{\lambda}_3)^2 \frac{\tilde{P}_3}{\tilde{P}}.
\end{aligned}$$

The cycle is continued until satisfactory convergence is apparent. We now give several examples of mixture distributions, the first and second from Karl Pearson.

3 Examples

3.1 Pearson's *Breadth of "Forehead" of Crabs*

Pearson's (1984, p85) best fit (moment estimators) for the mixture of two normal distributions is given on (p.88) and is used as initial input to our maximum likelihood estimators,

Table 1 Maximum likelihood estimators and their asymptotic moments
($n = 1000$)

	Moment estimator	mle 500 cycles	Bias	σ	$\sqrt{\beta_1}$
π_1	0.4145	0.43	-0.28	0.30	-4.06
π_2	0.5855	0.57			
λ_1	13.282	13.56	-4.19	3.19	-6.28
λ_2	19.289	19.27	0.99	0.65	-6.22
σ_1	4.4685	4.58	-1.68	1.08	-7.40
σ_2	3.1154	3.15	0.03	0.43	-0.36

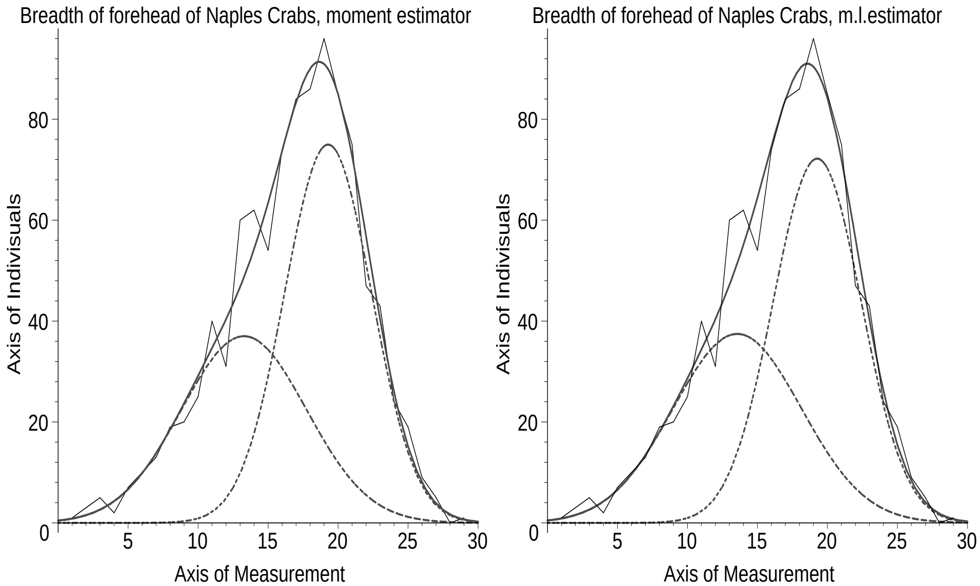

Figure 1: Data and fits of Pearson example

Discussion: Some details of low-order moments are given in Table 1. Convergence is very slow, 500 cycles being used. The asymptotic low-order moments of maximum likelihood estimators were computed by using the Maple code (see Bowman and Shenton, 2005).

The fits of using moment estimators and maximum likelihood estimators (Table 2) are almost identical as shown in the Figure 1. The goodness of fits test values are 21.21 and 20.47, with degree of freedom 23. The probability of a larger value of χ^2, from a table, is 70% approximately. Note that bias values are relatively small, but in all cases except the estimator $\hat{\sigma}_2$, the skewness is large, in values lying between 4 and 7.

Table 2. Data frequency and fits of Pearson example

Abscissa	Ordinates	Pearson fits	Maximum likelihood fits
1	1	0.85	0.87
2	3	1.53	1.55
3	5	2.62	2.63
4	2	4.28	4.24
5	7	6.65	6.53
6	10	9.82	9.60
7	13	13.82	13.47
8	19	18.52	18.05
9	20	23.71	23.17
10	25	29.15	28.64
11	40	34.66	34.34
12	31	40.37	40.38
13	60	46.70	47.13
14	62	54.26	55.09
15	54	63.42	64.45
16	74	73.68	74.62
17	84	83.39	83.93
18	86	90.00	89.97
19	96	90.96	90.42
20	85	85.00	84.21
21	75	72.81	72.09
22	47	56.87	56.44
23	43	40.37	40.29
24	24	25.99	26.17
25	19	15.17	15.46
26	9	8.02	8.30
27	5	3.84	4.06
28	0	1.67	1.81
29	1	0.66	0.74
Total	1000	998.79	998.65
Test		21.21	20.47

3.2 Pearson's *Palaemon serratus measurements*

See Pearson (1984) on page 101, beginning with the data sets

Table 3 Palaemon serratus - Measurements in 998 specimens (adult female) from penultimate to hindmost tooth on the carapace.

Entry interval	Number of specimens	Entry interval	Number of specimens
1 (27)	1	23 (49)	25
2 (28)	0	24 (50)	17
3 (29)	0	25 (51)	11
4 (30)	0	26 (52)	8
5 (31)	1	27 (52)	4
6 (32)	0	28 (53)	1
7 (33)	3	29 (54)	0
8 (34)	3	30 (55)	0
9 (35)	4	31 (56)	1
10 (36)	11	32 (57)	1
11 (37)	24	33 (58)	0
12 (38)	38	34 (59)	0
13 (39)	56	35 (60)	0
14 (40)	80	36 (61)	0
15 (41)	105	37 (62)	0
16 (42)	121	38 (63)	0
17 (43)	117	39 (64)	1
18 (44)	108	40 (65)	0
19 (45)	77	41 (66)	0
20 (46)	69	42 (67)	0
21 (47)	62	43 (68)	1
22 (48)	48		

Pearson pointed out the "**gigantic**" values at 65 and 69. The data is given graphically in Figure 2.

For solutions, Pearson finds $\sigma_1 = 3.5595$, $\sigma_2 = 5.7626\sqrt{-1}$, i.e. no solution for the two component model. We used as initial values for the iterative solution of maximum likelihood estimators $\hat{\lambda_1}$, $\hat{\lambda}_2$, $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\pi}_1$. See Table 4.

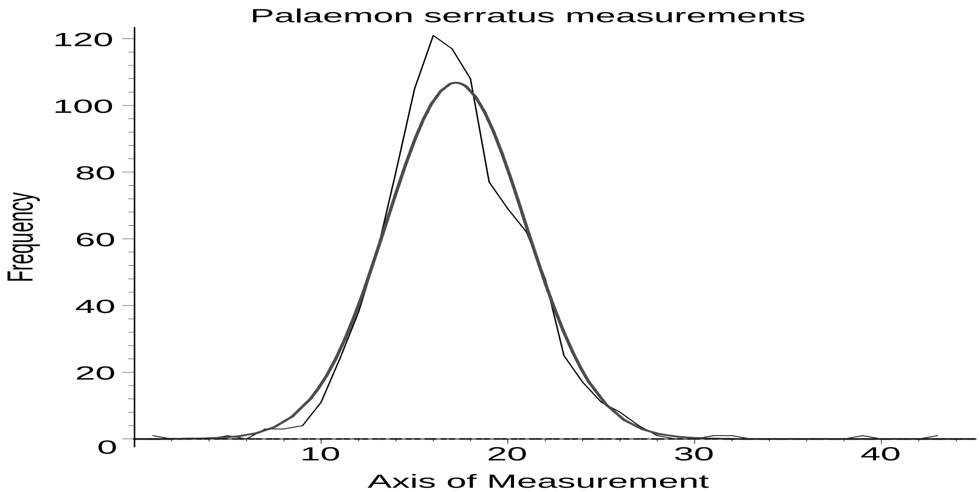

Figure 2: Data and fits of Palaemon serratus measurement

Table 4 Maximum likelihood estimators

	λ_1	λ_2	σ_1	σ_2	π_1	π_2
Moment estimator	16.0440	21.5023	3.3897	8.9330	0.9730	0.0170
mle 10 cycles	17.2076	1.0010	3.7239	$\underline{0.0010}$	0.9990	0.0010

The underscored variance term shows there is no two component normal mixture solution. The values $\lambda_1 = 17.21$, and $\sigma_1 = 3.72$ are suggested for a one component normal model.

3.3 Data from Hald (1952, p156) and studied by Hasselblad (1966)

The data has sample size 430 and is widely referred to in mixture distribution literature (Everitt and Hand, Johnson, Kotz, and Kemp). Hasselblad used steepest descent, and Newton-Raphson for solutions. He worked on IBM 7094 with the Fortran IV code, stating that 1/10 second computer time was needed for steepest descent, one second for Newton-Raphson.

With our iterative approach we used the fact that for maximum likelihood estimates $\sum \frac{n_x}{N}\frac{P_1}{P} = 1$ as a criterion for convergence. The sequence of approximations is shown in Table 5. It turns out that the solution by Bowman and Shenton for this ash content data agrees more or less perfectly with Hasselblad; the comparison with Hasselblad and low order moments of maximum likelihood estimators computed by Maple code are given in Table 5.

The moments of the data are mean=6.4465, standard deviation=2.2013, $\sqrt{b_1} = -0.3916$, and $b_2 = 2.5088$. This is the case which clearly shows the distribution has two components. λ_1 is near 3.25, λ_2 is near 7.25. We can take σ_1^2 to be 1.0 and σ_2^2 to be 2.0, and π to be 0.2 for initial estimates for the cycles of iteration scheme.

Table 5 The iteration solution sequence for Hald's ash content data

	$\sum \frac{n_x}{N}\frac{P_1}{P}$	λ_1	λ_2	σ_1^2	σ_2^2	π
1	1.0803	3.4610	7.1929	1.0720	2.1748	0.21606
2	1.0230	3.3511	7.2996	1.1787	2.2505	0.22103
3	1.0074	3.3070	7.3373	1.1429	2.2193	0.22267
4	1.0027	3.2823	7.3529	1.1053	2.1943	0.22328
5	0.9994	3.2622	7.3619	1.0777	2.1834	0.22316
6	0.9973	3.2464	7.3658	1.0571	2.1807	0.22256
7	0.9963	3.2345	7.3660	1.0414	2.1825	0.22173
8	0.9959	3.2258	7.3641	1.0296	2.1866	0.22083
9	0.9961	3.2197	7.3610	1.0208	2.1916	0.21995
10	0.9964	3.2156	7.3576	1.0143	2.1967	0.21917
11	0.9969	3.2129	7.3542	1.0096	2.2014	0.21850
12	0.9975	3.2112	7.3511	1.0063	2.2055	0.21794
13	0.9980	3.2102	7.3484	1.0040	2.2089	0.21750
14	0.9984	3.2097	7.3462	1.0025	2.2117	0.21715
15	0.9988	3.2095	7.3444	1.0016	2.2140	0.21689
16	0.9991	3.2095	7.3430	1.0010	2.2157	0.21669
17	0.9993	3.2096	7.3420	1.0007	2.2170	0.21655
18	0.9995	3.2097	7.3412	1.0006	2.2179	0.21645
19	0.9997	3.2099	7.3406	1.0005	2.2186	0.21638
20	0.9998	3.2100	7.3402	1.0005	2.2191	0.21633
21	0.9999	3.2102	7.3399	1.0006	2.2194	0.21630
22	0.9999	3.2103	7.3397	1.0007	2.2196	0.21628
23	1.0000	3.2104	7.3396	1.0007	2.2198	0.21627
24	1.0000	3.2105	7.3395	1.0008	2.2198	0.21626
25	1.0000	3.2105	7.3395	1.0008	2.2198	0.21626
26	1.0000	3.2106	7.3394	1.0009	2.2199	0.21626
27	1.0000	3.2106	7.3394	1.0009	2.2199	0.21626
28	1.0000	3.2106	7.3394	1.0009	2.2199	0.21626
29	1.0000	3.2106	7.3394	1.0010	2.2199	0.21626
30	1.0000	3.2106	7.3394	1.0010	2.2199	0.21627
31	1.0000	3.2107	7.3394	1.0010	2.2199	0.21627
32	1.0000	3.2107	7.3394	1.0010	2.2199	0.21627
33	1.0000	3.2107	7.3394	1.0010	2.2199	0.21627
34	1.0000	3.2107	7.3394	1.0010	2.2199	0.21627
35	1.0000	3.2107	7.3394	1.0010	2.2199	0.21627

Table 6. Solutions by Hasselblad and Bowman & Shenton, and asymptotic low order moments of maximum likelihood estimators

	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\pi}_1$
Hasselblad	3.210	7.339	1.000	2.220	0.2162
Hald	3.1	7.2	0.64	2.25	0.20
B & S	3.2107	7.3395	1.0010	2.2198	0.2163
	Low order moments of mle $N = 430$				
Bias	0.0107	0.0055	-0.0058	-0.0069	0.0020
s.d.	0.1954	0.1219	0.1321	0.0911	0.0325
$\sqrt{\beta_1}$	0.4217	-0.0115	0.4504	0.0903	0.2528

Note: The sample size is used for these moments. The smallness of the asymptotic skewness values is remarkable.

The fit of the theoretical distribution (two component normal mixture) in Table 7 is so good, perhaps too good. Take a look at the penultimate values of ash content, where discrepancies are usually discussed. For $x = 0.75\%$ content, the comparison is 1 against 0.9; for $x = 10.75\%$ content the comparison is 4 against 3.29. Using the frequencies as they are stated in Table 7, yields a χ^2 value of 5.74 with degree of freedom $\nu = 23 - 1 - 5 = 17$. There is therefore a one in a hundred chance of a value less than 5.74. If the first three, and last two frequencies are combined then the degrees of freedom are reduced but the final assessment of goodness of fit is about the same as the unabridged.

It seems to us quite remarkable that although Hasselblad's study is often quoted as a reference, we have not found one which deploys the goodness of fit table, surely a grave omission.

Table 7. The ash content data frequency and fits

Ash percent	Frequency	Fit by BS	Fit by Hasselblad
0.25	1	0.23	0.23
0.75	1	0.90	0.90
1.25	2	2.73	2.72
1.75	5	6.43	6.43
2.25	12	11.83	11.83
2.75	18	17.07	17.08
3.25	20	19.57	19.57
3.75	19	18.51	18.51
4.25	16	16.07	16.06
4.75	14	15.64	15.64
5.25	20	19.22	19.20
5.75	25	26.28	26.29
6.25	35	34.72	34.73
6.75	43	41.76	41.76
7.25	48	45.04	45.05
7.75	45	43.44	43.44
8.25	35	37.43	37.43
8.75	26	28.82	28.82
9.25	17	19.83	19.82
9.75	13	12.19	12.18
10.25	9	6.69	6.69
10.75	4	3.29	3.28
11.25	2	1.44	1.44
Total	430	429.11	429.11
Test (df14)		5.47	5.48
Test (df17)		5.74	5.74

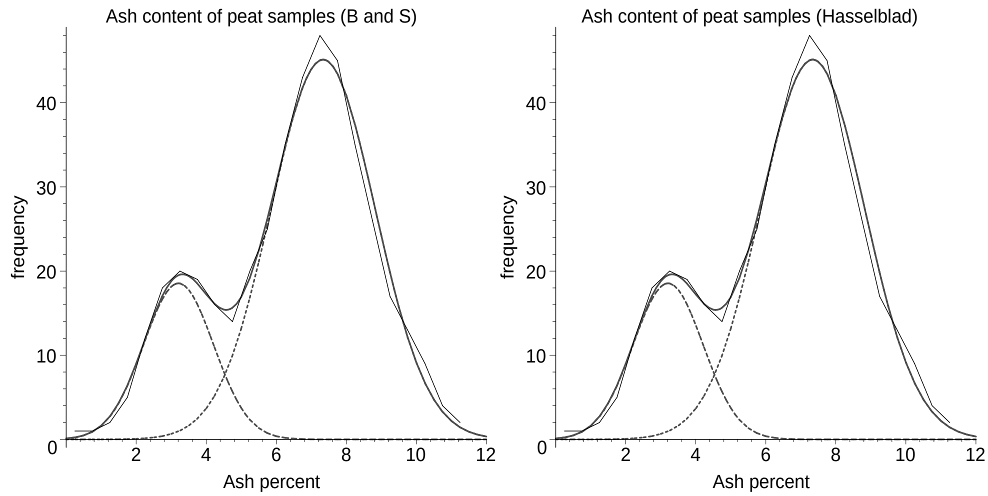

Figure 3: Data and fits of Ash contents

3.4 Hald's (1952, p155), 466 measurement of the release time of a relay

From the histogram of the release time data, Hald suggest that a two component normal mixture should be appropriate with $\sigma_1 = \sigma_2$ approximately, $\pi_1 = \pi_2$ approximately, and $\lambda_2 - \lambda_1 = 0.08$ second.

Table 8 The distribution of 466 measurements of release time of a relay and fit

	Frequency	mle fit		Frequency	mle fit
1.00	1	.57	1.12	12	14.92
1.01	0	1.99	1.13	28	21.15
1.02	2	5.68	1.14	20	28.14
1.03	20	13.18	1.15	27	32.17
1.04	23	24.85	1.16	39	31.05
1.05	49	38.12	1.17	30	25.21
1.06	41	47.56	1.18	14	17.20
1.07	43	48.32	1.19	8	9.87
1.08	39	40.12	1.20	6	4.76
1.09	27	27.75	1.21	1	1.93
1.10	21	17.41	1.22	0	0.66
1.11	14	12.98	1.23	1	0.19
Total				466	465.78

The moments of the data are mean=1.1022, standard deviation=0.0488, $\sqrt{b_1} = 0.2846$, and $b_2 = 1.84$. We used initial input $\lambda_1 = 1.07$, $\lambda_2 = 1.16$, $\sigma_1 = 0.05$, $\sigma_2 = 0.03$, and $\pi = 0.6$ from the plot of data points.

Our maximum likelihood estimators, and their low asymptotic moments are shown in the Table 9.

Table 9 Maximum likelihood estimators and their asymptotic moments

	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\pi}$
m.l.e.	1.0657	1.1530	0.0220	0.0240	0.5818
Bias	-0.0000	-0.0000	-0.0001	-0.0001	-0.0003
s.d.	0.0017	0.0023	0.0013	0.0018	0.0269
$\sqrt{\beta_1}$	0.0667	-0.1662	0.1482	0.2265	-0.0376

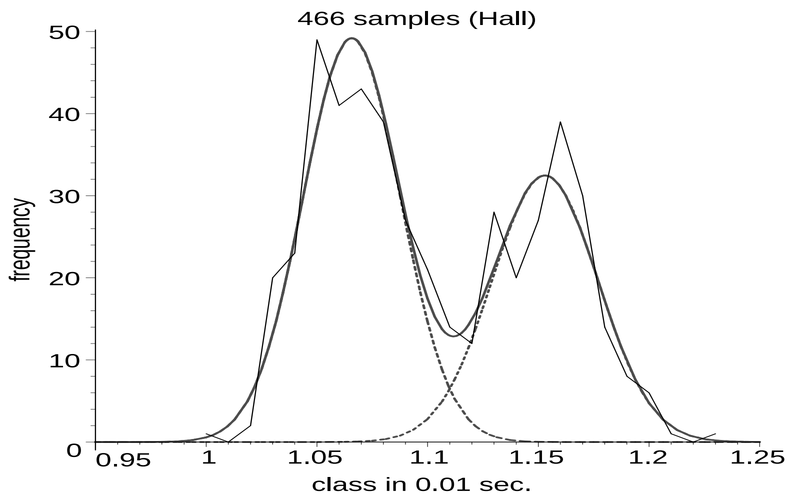

Figure 4: Data and fits of Hald's example

Goodness of fit χ^2-test is 28.64 with degree of freedom 18 and $\chi^2_{\alpha=0.5} = 28.87$.

We quote from Hald (1952, p155) "A more detailed examination of the relay led to the following explanation of the heterogeneity: When the current through the coil of the relay is closed, a worm, which is connected with an induction break disc, meshes with a toothed sector, which causes the release key to close. The above difference of 0.08 second corresponds to two positions, separated from each other by one tooth. If the relative position of the worm and the toothed sector has not been adjusted with extreme accuracy, which of these two positions is taken up depends on the position at which the induction break disc has stopped at the end of the previous operation. The relay was adjusted, and a new distribution was observed in which by far the greater number of the measurements belonged to the distribution with the larger mean. Further adjustment of the relay would no doubt have removed the heterogeneity completely."

4 Maximum likelihood equations for gamma mixtures

$$P(X = x) = \pi_1 P_1(x; a_1, \rho_1) + \pi_2 P_2(x; a_2, \rho_2) = P(x) = P \tag{4}$$

and

$$P_i = \frac{e^{-x/a_i}(x/a_i)^{\rho_i - 1}}{a_i \Gamma(\rho_i)}$$

where $0 < \pi_i < 1$, $\rho_i > 0$, $a_i > 0$, $i = 1, 2$, $0 < x < \infty$.

For the iterative maximum likelihood solutions: given an initial set $(a_1, \rho_1, a_2, \rho_2, \pi_1)$

$$_{new}\hat{\pi}_1 = \pi_1 \sum_{x=0}^{n} \frac{n_x}{N} \frac{\tilde{P}_1}{\tilde{P}},$$

so $_{new}\hat{\pi}_2 = 1 -_{new} \hat{\pi}_1$;

$$_{new}\hat{a}_1 = \frac{1}{\rho_1} \sum_{x=0}^{n} x \frac{n_x}{N} \frac{\tilde{P}_1}{\tilde{P}}, \qquad _{new}\hat{a}_2 = \frac{1}{\rho_2} \sum_{x=0}^{n} x \frac{n_x}{N} \frac{\tilde{P}_2}{\tilde{P}}.$$

Now for $\hat{\rho}_i$ solve the equations

$$\psi(_{new}\hat{\rho}_1) = \sum_{x=0}^{n} \frac{n_x}{N} \ln\left(\frac{x}{a_1}\right) \frac{\tilde{P}_1}{\tilde{P}}, \qquad \psi(_{new}\hat{\rho}_2) = \sum_{x=0}^{n} \frac{n_x}{N} \ln\left(\frac{x}{a_2}\right) \frac{\tilde{P}_2}{\tilde{P}}. \tag{5}$$

In connection with these two equations, note that

$$\int_0^\infty e^{-x/a}(x/a)^{\rho-1} dx = a\Gamma(\rho), \qquad (a, \rho > 0)$$

so that differentiating with respect to ρ (when valid) leads to

$$E \ln(x/a) = \psi(\rho),$$

which supports the sample expansion in (5). This iterative scheme, in our experience only, produces convergence when the initial input is very close to the exact solution.

Examples:

We have found a study of maximum likelihood methods for two component gamma mixtures (location known) in Kanno (1982). The author uses Newton-Raphson to search for a solution, noting that initial values for the 5 parameters π_1, a_1, ρ_1, a_2, ρ_2 play an important role.

Kanno's Table 4 (our Table 10), using a random number generator, creates samples of sizes $N = 120$, $N = 250$, and $N = 500$, for three sets of values of the 5 parameters. The results are based on 100 sets of data for each example.

Table 10 Means and Variances of ML-estimates obtained by the Switching Technique

(a) $\pi_1 = 0.6,\ a_1 = 1.0,\ a_2 = 3.0,\ \rho_1 = 10.0,\ \rho_2 = 10.0$

Sample Size		Mean	Variance	Convergence Ratio*
$n = 120$	π_1	0.612814	0.003202	
	a_1	0.983195	0.058104	
	a_2	2.656232	0.516423	84/100
	ρ_1	10.843501	5.873739	
	ρ_2	12.373497	14.241284	
$n = 250$	π_1	0.600321	0.001453	
	a_1	0.972326	0.030718	
	a_2	2.856105	0.359719	92/100
	ρ_1	10.602929	2.675162	
	ρ_2	11.069730	7.065392	
$n = 500$	π_1	0.599084	0.000991	
	a_1	0.979151	0.017180	
	a_2	2.978927	0.220518	98/100
	ρ_1	10.366122	1.458317	
	ρ_2	10.341305	3.369693	

*The convergence ratio gives the number of times the iteration converged over the number of cases examined. For definition of switching see Kanno (1982).

(b) $\pi_1 = 0.4,\ a_1 = 1.0,\ a_2 = 2.0,\ \rho_1 = 8.0,\ \rho_2 = 12.0$

Sample Size		Mean	Variance	Convergence Ratio
$N = 120$	π_1	0.420285	0.003288	
	a_1	1.011345	0.162246	
	a_2	1.786265	0.084904	85/100
	ρ_1	9.037800	9.081267	
	ρ_2	14.414256	12,762469	
$N = 250$	π_1	0.405259	0.000845	
	a_1	0.970729	0.052691	
	a_2	1.886898	0.084904	91/100
	ρ_1	8.694442	3.675052	
	ρ_2	13.120551	5.645092	
$N = 500$	π_1	0.405259	0.000845	
	a_1	0.996611	0.023584	
	a_2	1.953517	0.050870	94/100
	ρ_1	8.237146	1.366583	
	ρ_2	12.482812	2.453989	

(c) $\pi_1 = 0.3,\ a_1 = 1.0,\ a_2 = 1.5,\ \rho_1 = 8.0,\ \rho_2 = 12.0$
(sample size $N = 500$)

	Mean	Variance	Convergence Ratio
	Complete NR-method		
π_1	0.397078	0.030406	
a_1	1.370825	0.342409	
a_2	1.372027	0.163267	30/100
ρ_1	7.146651	2.084019	
ρ_2	13.934228	8,870050	
	Switching Technique		
π_1	0.405259	0.000845	
a_1	1.291289	0.052691	
a_2	1.886898	0.080272	42/100
ρ_1	6.973615	1.002152	
ρ_2	14.136263	4.443573	

In this case, the convergence ratio of the NR-method increased to 30/100 compared with 1/100 for case (1), and for 10/100 for case (2), but it was less than that of switching technique. From the sample means and variances in Table 4(c), it seems that the complete NR-method is not better than the switching technique even if the difference between two scale parameters is fairly small.

In Table 11 we list the asymptotic low oder maximum likelihood moments, with a comparison to Kanno's findings in the case of bias and variance (first order terms). Briefly the notation for a statistic t is:

$$\begin{array}{lll} \text{Bias} & E(\hat{t} - t) \sim \mu'_{11}/N + \mu'_{12}/N^2 + \cdots & (N \to \infty) \\ Var(\hat{t}) & \mu_2(\hat{t}) \sim \mu_{21}/N + \mu_{22}/N^2 + \cdots & \\ \text{Skewness} & \sqrt{\beta_1} \sim \mu_3/\mu_2^{3/2} \quad \text{location and scale free} & \end{array}$$

and $\sqrt{\beta_{11}}$ is the coefficient of $1/\sqrt{N}$ in the skewness. For the three sample values the obvious choice for an asymptotic study is to take the case $N = 500$. For three cases involving 3 sets of 5 parameters (Table 10 and Table 11) compares the results of Kanno with our asymptotic values derived from our Maple code approach, the comparisons relating to relative bias, and first order variance; in addition we record the asymptotic standard deviation, and asymptotic skewness for the 5 estimators $\hat{a}_1$, $\hat{\rho}_1$, $\hat{a}_2$, $\hat{\rho}_2$ and $\hat{\pi}_1$.

Table 11 Asymptotic low order moments of 5 parameters with sample size $N = 500$

Case		$\hat{a}_1$	$\hat{\rho}_1$	$\hat{a}_2$	$\hat{\rho}_2$	$\hat{\pi}_1$
Case 1	$\mu'_{11}/(N\mu'_1)$	-0.0021	0.0150	0.0014	0.0271	-0.0018
	Kanno	-0.0209	0.0366	-0.0070	0.0341	0.0015
	μ_{21}/N	0.0152	1.1611	0.2402	3.2846	0.0009
	Kanno	0.0172	1,4583	0.2205	3.3697	0.0010
	s.d.	0.1233	1.0776	0.4901	1.8123	0.0302
	$\sqrt{\beta_{11}}/\sqrt{N}$	0.3300	0.4071	0.5928	0.3534	-0.1591
Case 2	$\mu'_{11}/(N\mu'_1)$	0.0010	0.0237	-0.0027	0.0195	0.0008
	Kanno	-0.0034	0.0296	-0.0233	0.0411	0.0133
	μ_{21}/N	0.0289	1.2791	0.0617	2.6237	0.0009
	Kanno	0.0236	1.3666	0.0509	2.4540	0.0008
	s.d.	0.1700	1.1310	0.2484	1.6198	0.0304
	$\sqrt{\beta_{11}}/\sqrt{N}$	0.5499	0.4949	0.3339	0.4027	0.0413
Case 3	$\mu'_{11}/(N\mu'_1)$	0.0394	0.0576	-0.0094	0.0462	0.0410
	Kanno	0.2913	-0.1283	-0.1093	0.1780	0.2063
	μ_{21}/N	0.1190	3.6842	0.0598	5.4897	0.0059
	Kanno	0.0903	1.0022	0.0312	4.4436	0.0053
	s.d.	0.3450	1.9194	0.2446	2.3430	0.0771
	$\sqrt{\beta_{11}}/sqrtN$	1.3865	0.7212	0.1966	0.9113	0.9878

Discussion:

(i) The relative bias (μ'_{11}/μ'_1) is in general quite small, an exception being the one of the proportion π_1.

(ii) The variances are in good agreement except for case 3 and $\mu_{21}(\hat{\rho}_1)$.

(iii) Asymptotic skewness is in general less than unity in value, a quite acceptable result in practice. There is an exception for case 3, for which $\sqrt{\beta_{11}(\hat{a}_1)} = 1.4$, and $\sqrt{\beta_{11}(\hat{\pi}_1)} = 0.99$.

(iv) For the three cases graphs of the sample data against theory are exhibited. Case 2 is clearly a bimodal distribution, whereas for case 1 the second component is muted.

5 Conclusion

In this paper the Maple code implementation of low order asymptotic moments of maximum likelihood estimators applied to normal and gamma mixture distributions receives further support from the examples studied. Including starting values in iterative solutions play an important role.

In a recent paper by Karlis and Xekalaki (2005), the references cite 155 papers. Our work suggests one direction for further research.

References

Bowman, K.O. and Shenton, L.R. (2005), Bias, variance, skewness and kurtosis of maximum likelihood estimators using Maple, *Far East Journal of Theoretical Statistics*, 17(2), 137-195.

Bowman, K.O. and Shenton, L.R. (2006), Binomial and Poisson mixtures, maximum likelihood, and Maple code, *Far East Journal of Theoretical Statistics*, 20(1), 73-95.

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

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

Hasselblad, V. (1966). Estimation of Parameters for a Mixture of Normal Distributions, *Technometrics*, 8(3), 431-444.

Johnson, N.L., Kotz, S. and Kemp, A.W. (1993), *Univariate Discrete Distributions*, John Wiley & Sons, Inc., New York.

Kanno, R. (1982), Maximum likelihood estimation of parameters for a mixture of two gamma distributions, *Rep.Stat.Appl.Res., JUSE*, 29(3), 14-24.

Karlis, D. and Xekalaki, E. (2005), Mixed Poisson distributions, *International Statistical Review*, 73(1), 35-58.

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

Shenton, L.R. and Bowman, K.O. (1977). *Maximum Likelihood estimation in small Samples*, Charles Griffin and Co., Ltd.

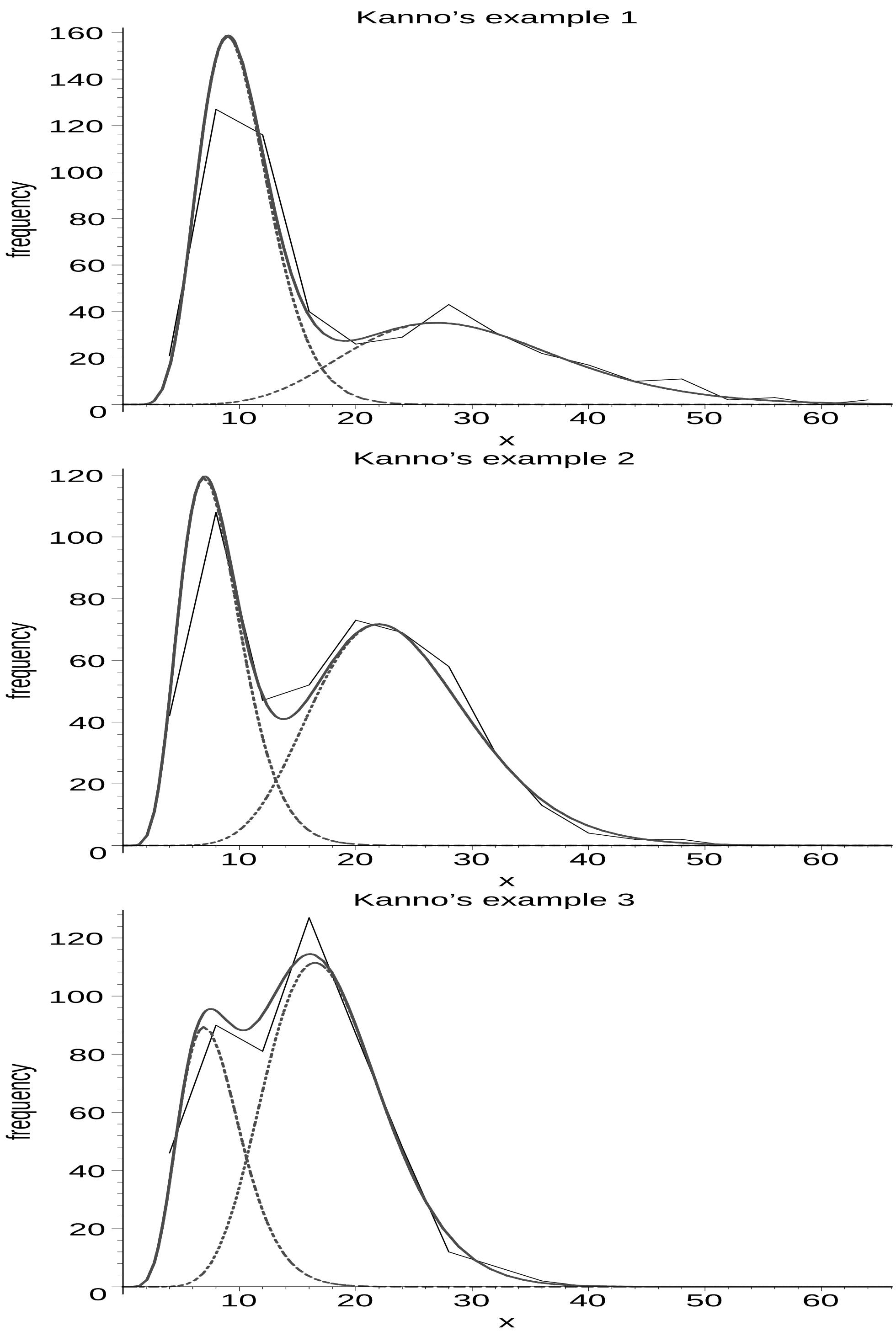

Figure 5: Gamma mixture of Kanno's 3 examples