

Estimation in Mixtures: The Rectangular Case

G. Lavanya¹, V.V. HaraGopal², S.N. Narahari Pandit³

¹Kakatiya University, Warangal,
^{2,3}Osmania University, Hyderabad

Abstract: In this paper we consider the mixture of two Rectangular Distributions. Methods used for estimating the parameters are (i) Moment ratio method (ii) Extreme value estimation (\hat{b}), this problem has been addressed to some extent by Pavan Kumar (2003), in the context of obtaining stopping rule to estimate population minimum in sampling with replacement from a finite population. Extreme value estimate (\hat{b}) method relatively better than moment ratio method, based on simulation, the percentage of acceptable estimators obtained in samples.

1. Introduction

Many studies have been reported dealing with mixture of populations. However, most of them seem to discuss mainly theoretical aspects, of estimation of parameters rather than consider the practical goodness of estimates obtained by the approaches dealt within these studies (Rider (1961), Tallis and Light (1968), Blischke (1962), Craigmile and Titterington (1997), Hussain and Liu (2009)). Further, they seem to assume that samples (from which parameters are to be estimated) are actually mixtures of a given family of distributions and do not consider whether the sample itself can allow one to check whether it has come from a single population or from a mixture - population.

The present study tries to address these questions, particularly in the context of two component rectangular mixture population.

1.1. Application of Finite Mixture Distribution

Mixtures of distributions arise frequently both in biological and physical sciences, reliability in life testing, engineering, medical and social sciences. Here is an example of mixtures.

In fishery biology: It is often desired to measure certain characteristics in natural population of fish. For this purpose, samples of fish are taken and the desired trait is measured for each fish in the sample. However, many characteristics vary markedly with age of the fish. Then the trait has a distinct distribution for each age group so that the population has a mixture of distributions (Keewhan Choi and Bulgren (1968)).

1.2. Review of Literature

For completeness we mention below some studies in this regard and make a few observations on the same.

Rider (1961), considered the problem of mixture of two exponential distributions. He noted that method of moments gives unacceptable estimates; however the parameters 'a' and 'b' of the constituent distributions are not equal ($a \neq b$) the estimates obtained are consistent. Also, he found variance of the asymptotic distributions of \hat{a} and \hat{b} . He also, finds out that a chi-square test for single exponential can be misleading; a null hypothesis that data are from a single exponential often get accepted, when a chi-square test for goodness of fit is carried out. Tallis and Light (1968) use fractional moments for estimates of parameters of mixed exponential distribution, and compare method of moments with maximum likelihood method and observes that maximum likelihood method involves much larger amount of calculation without increasing in acceptability of estimates. Blischke (1962) considers mixture of two binomials with same sample size 'n' but with different parameters p_1 and p_2 and with mixture parameter p and $q=1-p$ and points out some theoretical anomalies about relative efficacy of estimators. Day (1969) discusses the method of moments, minimum chi-square and Bayes estimators' mixture of two Normal, multivariate populations. He

also points out that they appear greatly inferior to maximum likelihood estimators. Craigmile and Titterington (1997) deal with 2 and 3 component mixture of Rectangular distributions. Method of moments and maximum likelihood was discussed.

Particularly interesting is the recent work of Hussain and Liu (2009) which utilizes the L - moment approach to estimation of parameters.

2. The Present Approach

The present approach is expected to be useful particularly where the range of the variate concerned is doubly finite. In particular we consider the case of rectangular two- component mixtures.

Functions of moments are used to link up the theory with the observed sample moments and ratios (Craigmile and Titterington (1997), Hussain and Liu (2009)) of these functions are used to explicitly compute the parameter estimates. Since ratios of sample moments are involved in the process, the estimates are likely to exhibit rather wide fluctuations around the true parameter values, particularly in small samples.

Let 'a' and 'b' > a, be the lower and upper limits of the range of a random variate. Thus, $\hat{a} = \min\{x_i, i=1,2,\dots, n\}$ and $\hat{b} = \max\{x_i, i=1,2,\dots, n\}$ are consistent but biased estimators of 'a' and 'b', relative bias can be (heuristically) assumed to be order $1/(n+1)$ and here asymptotic unbiased estimation of a and b can be as $\hat{a} = \min\{x_i, i=1,2,\dots, n\} - 1/(n+1)$ and $\hat{b} = \max\{x_i, i=1,2,\dots, n\} + 1/(n+1)$.

The 'range parameters' of rectangular mixture population were revised are finite which can be efficiently estimated by the sample extreme values and hence, estimates so, obtained can be plugged into the formulas for estimation, thus reducing the 'estimate' fluctuation. This has been found to be the case, as revealed by the simulation studies (see Appendix).

3. The paper is divided into two sections

Section A: Dealing with the simpler case, where the mixture components have the ranges $0 \leq x_1 \leq a, 0 \leq x_2 \leq b, (b > a)$ and mixing ratio p:q, $p+q=1$ It may be noted that this case does not seem to have been considered so far.

Explicit formulas linking functions of moments with simple explicit functions of parameters are obtained and explicit expressions for parameter estimates in terms of these ratios of moment's functions are obtained.

This is followed by a simulational study to examine the practical utility of the approach.

For a few selected sets of parameter values a, b and p ($q=1-p$), number of samples $n_s = 100$, of size $n = (20, 50, 100)$ are generated and estimates from these samples are computed.

As already noted, often one finds the observed estimates as unreasonable, (a, b being negative or p not being in the range $(0 \leq p \leq 1)$) For the same samples, the estimates are computed using the formulas directly and also by modifying them by first estimating b ($>a$) by sample maximum (with a correction for bias, heuristically obtained). It is found that in the latter case, estimates will be more likely to be acceptable.

Section B: Deals with the more general 5 - parameter case, where the two-component population have range parameters as $a_1 \leq x_1 \leq b_1, a_2 \leq x_2 \leq b_2$ and mixing ratio p:q .

Three cases arise in this context: The distribution may have:

- (1) gap ($a_1 < b_1 < a_2 < b_2$)
- (2) overlap ($a_1 < a_2 < b_1 < b_2$) and
- (3) Imbedding one range within the other ($a_1 < a_2 < b_2 < b_1$)

Here a_1 and maximum of (b_1, b_2) are estimated by the sample minimum and maximum respectively, quite efficiently. However, estimating the other parameters is more complicated.

However, attempts are made to distinguish between the three possibilities (of a gap, overlap and imbedding) on the basis of sample itself this may make further study of estimation easier. These also are explored through simulation.

4. Some theoretical results

4.1 Moment Ratio Method:

Firstly, we consider the case of range with common origin ‘0’. ($0 \leq x_1 \leq a, 0 \leq x_2 \leq b, p:q$).

Let $f_1(x)$ and $f_2(x)$ be the densities of two populations, with mixing parameter $p:q$. Then $f(x) = pf_1(x) + qf_2(x)$. Where m_k' are k^{th} raw moments. Hence, in the present case, the raw moments of the mixture are

$$m_k' = \frac{pa^k + qb^k}{k + 1}, k = 0, 1, 2, 3, \dots, r \tag{1}$$

Then, four raw moments are

$$pa + qb = 2m_1' \tag{1a}$$

$$pa^2 + qb^2 = 3m_2' \tag{1b}$$

$$pa^3 + qb^3 = 4m_3' \tag{1c}$$

$$pa^4 + qb^4 = 5m_4' \tag{1d}$$

Hence,

$$s = 3m_2' - (2m_1')^2 = pq(a-b)^2 \tag{2a}$$

$$s_1 = 4m_3' - 6m_2'm_1' = pq(a-b)^2(a+b) \tag{2b}$$

$$s_2 = 5m_4' - 8m_3'm_1' = pq(a-b)^2(a^2 + b^2 + ab) \tag{2c}$$

Here we consider $s, s_1,$ and s_2 are functions of raw moments.

From which it follows that

$$a + b = \frac{s_1}{s} \tag{3a}$$

$$a - b = \pm \sqrt{4\left(\frac{s_2}{s}\right) - 3\left(\frac{s_1}{s}\right)^2} \tag{3b}$$

Solving (3a) and (3b) using the observed (that is sample) moments, one gets the estimates as

$$\hat{a} = \frac{\left(\frac{s_1}{s}\right) \pm \sqrt{4\left(\frac{s_2}{s}\right) - 3\left(\frac{s_1}{s}\right)^2}}{2} \text{ and } \hat{b} = \left(\frac{s_1}{s}\right)^2 - \hat{a} \tag{4a}$$

From equation 4(a)

$$\hat{p} = \frac{(2m_1' - \hat{b})}{\hat{a} - \hat{b}}$$

Replacing the right side of these equations by the corresponding moment estimates based on sample values, one gets explicit estimates expression / value, for that sample.

However, expression for $a-b$ involve square roots and two possible solutions for $\hat{a} - \hat{b}$, which may lead to negative values of a and b and also, possibly complex numbers as the estimates.

4.2 Extreme Value Estimation (\hat{b})

This problem is considerably mitigated by first estimating $b(>a)$ by $\hat{b} = \max(x) + \frac{1}{(n+1)}$ the correction factor $\left(\frac{1}{(n+1)}\right)$ being greater than 1 since the $\max(x)$ will necessarily under-estimate b .

Using this estimate of b in subsequent steps of estimation, considerably reduces the likelihood of getting unacceptable estimates ‘ a ’ and ‘ b ’.

Here, we can use first two raw moments. Substituting (\hat{b}) in the above 1(a), 1(b), we get other two estimates (\hat{a}) and (\hat{p}).

We now present simulation results for a few parameter values. For number of samples $ns = 100$ samples of sizes, $n=20, 50, 100$ are generated for each parameter set, the mean, standard deviation,

minimum, maximum of the 'ns' estimates for each parameters are reported along with the number of percentage of acceptable estimates, by both these methods.

We shall now present the more general 5-parameter case.

Here also, we generate samples from population with parameters $a_1 < b_1 < a_2 < b_2$ with $p=0.1:0.2:0.9$. However, for each set $\{a_1, b_1, a_2, b_2\}$ we generate three mixture sample sets considering to the three cases of gap $[(a_1, b_1), (a_2, b_2)]$, overlap $[(a_1, a_2) \text{ and } (b_1, b_2)]$ and imbedding $[(a_1, b_2) \text{ and } (b_1, a_2)]$.

We compute $\hat{a}_1 = \min(x) - \frac{1}{(n+1)}$, $\hat{b}_2 = \max(x) + \frac{1}{(n+1)}$ and first two raw moments and

standard deviation for each of these 'ns' samples. Table 2(a) presents first two raw moments and standard deviation of a_1, b_1, a_2, b_2 for different p values. This helps in arriving at some heuristic, to decide on the types (relative range position) of the two component population giving rise to the mixture.

5. Simulation studies

Table 1(a): gives summary of simulation results in the case of mixture with $a=2$, $b=3$, $n=20$, $ns=100$, $p=0.1:0.2:0.9$. Table 1(b) where $a=2$, $b=3$, $p=0.1:0.2:0.9$ gives the case of sample size $n=50, ns=100$, while table 1(c) $a=2$, $b=4$, where $p=0.3:0.1:0.7$, gives the result for sample size $n=100, ns=100$. Summary statistics namely the mean, standard deviation, minimum, maximum of estimates for a, b and p (i) moments ratio method, (ii) Extreme value estimation ' \hat{b} '. first and then estimating the other two parameters' are given for the five chosen values of 'p'. Also are given the percentages of samples giving acceptable estimates of the parameters (by the two approaches). Thus, for instance from table 1(b) one has for $p=0.3$, 38% only of the samples give acceptable estimates. For a, b as well as p while the corresponding figures, for the second approach are 65%, 65% and 37%. In general the estimates $(\hat{a}, \hat{b}, \hat{p})$, appear to be generally much nearer to the true values in the latter approach. As seen from table 1(c) we the sample size is 100, the performance of moment ratio approach remains quite poor, while the other approach is on the whole very satisfactory. However even in the latter approach there are a few samples which give absurd results, though the number of such samples is quite small. One possible cause of these results, may be due to these samples containing outliers or due to bunching in the sampling processes.

5.1 Simulation Results: (5 – parameter case)

Simulation is carried out for parameters, $a_1=2$, $b_1=3$, $a_2=4$, $b_2=5$. Number of samples $ns = 100$ samples, sample size $n=50$ were generated for the three cases and for $p=0.1:0.2:0.9$ as well as for the case of single rectangular with range ' a_1 ' to ' b_2 '. Theoretical moments were computed and corresponding sample moments obtained by simulation also are computed. Results are given in Table (2a). It was hoped that these results will suggest methods for distinguishing between three possible range relationships by comparing the moments for the single population case with ' a_1 ', ' b_2 ' given and the three mixture cases. However, investigation so far has not shown any definite suggestion. In this regard except that the second raw moment appears considerably larger in the cases (i) Overlap and (ii) gap, while it can be large when 'p' is small, while it could become smaller where 'p' is larger in the case of 'overlap' 'gap'. However, as only to be expected estimation of ' a_1 ' and ' b_2 ' by sample minimum and maximum gives very satisfactory results.

6. Conclusions

1. Even when sample size is relatively small the extreme value estimation approach gives more encouraging result than the moment ratio approach. While in extreme value estimation (\hat{b}) the estimators appear to be much nearer to the true values than the moment ratio method.

2. However, possibility of getting unacceptable estimates cannot be ruled out. In such cases an in depth analysis of the sample, (like studying the structure of first and second difference of order statistics and bootstrapping approach) may be fruitful.
3. In the case of 5 – parameter problems this approach obviously gives very good estimates of ‘a₁’, ‘b₂’, but even the possibility of distinguishing between the three cases is still not answered. However, simulation studies make one feel that this present approach can give more acceptable solutions, to the problem of estimating mixture population parameters.

From table 2(a) it is clear that sample has come from mixture only. In single rectangular first two moments and standard deviation having variation than the mixture of two rectangular (3 types).

4. Since the range parameters can be estimated very satisfactorily one may refer the 5 parameter case to an equivalent 3 parameter case by changing the observation x_i to $y_i = (x_i - a) / (b - a)$ and estimate from the intersection parameter and the mixture parameter ‘p’ for the sample values and convert these interior parameters to back to the case of ‘x’ variate.

Appendix

Table-1(a)

Parameter Estimates of Mixture of two Rectangular Distributions (0->a) & (0->b), a=2, b=3, p=0.1:0.2:0.9, sample size n=20, no of samples (ns) =100										
p	Moment's Ratio				% acceptable estimates	Extreme value estimate (b)				% acceptable estimates
	mean	std	min	max		mean	std	min	max	
p=0.1										
\hat{a}	0.7519	0.4875	-0.9467	2.2065	32	1.3388	0.7733	-73.9932	610.5637	55
\hat{b}	3.1509	0.3821	1.5117	3.9334	32	2.9824	0.1820	2.2624	3.1454	55
\hat{p}	0.1939	0.1306	-2.7984	0.4983	32	0.2816	0.1529	-12.2120	6.4717	15
p=0.3										
\hat{a}	0.8065	0.4702	-0.8081	2.0831	30	1.3744	0.6057	-37.8810	341.0821	54
\hat{b}	3.1602	0.3798	1.2568	4.2655	30	2.9342	0.2362	2.0089	3.1454	54
\hat{p}	0.3075	0.1579	-1.8679	0.5917	30	0.3299	0.1385	-6.7881	13.6276	20
p=0.5										
\hat{a}	0.6553	0.4480	-0.5084	2.2762	41	1.4684	0.6714	-14.8261	23.1172	65
\hat{b}	3.1297	0.5569	1.3100	5.6229	41	2.8904	0.2327	2.0079	3.1445	65
\hat{p}	0.3209	0.1067	-3.2101	0.5782	41	0.3921	0.1963	-3.2646	2.6373	34
p=0.7										
\hat{a}	0.7258	0.4430	-2.3191	2.3510	37	1.6681	0.6416	-85.3816	11.4712	74
\hat{b}	3.0178	0.5185	1.5632	5.2384	37	2.7951	0.3482	1.9453	3.1432	74
\hat{p}	0.3695	0.1745	-1.0345	0.8897	37	0.5771	0.2256	-2.5872	7.5992	47
p=0.9										
\hat{a}	0.5369	0.3429	-1.0404	1.6802	20	1.5476	0.7721	-8.6046	50.8312	78
\hat{b}	2.8828	0.8001	0.6804	5.0153	20	2.3397	0.4163	1.6678	3.1108	78
\hat{p}	0.4063	0.1588	-4.6479	0.6322	20	0.4695	0.2687	-22.4774	106.4062	36

Table-1(b)

Parameter Estimates of Mixture of two Rectangular Distributions (0-->a) & (0-->b), a=2, b=3, p=0.1:0.2:0.9, sample size n=50, no of samples (ns) =100										
p=0.1	Moment s Ratio				% acceptable estimates	Extreme value estimate (\hat{b})				% acceptable estimates
	mean	std	min	max		mean	std	min	max	
\hat{a}	0.7481	0.5011	-2.9709	2.2220	22	1.5644	0.7603	-63.8653	127.986	58
\hat{b}	3.1896	0.3942	1.1218	4.6474	22	2.9888	0.0784	2.7296	3.0590	58
\hat{p}	0.1599	0.1122	-6.2551	0.4220	22	0.2762	0.2861	-2.7852	2.1368	19
p=0.3										
\hat{a}	0.7626	0.4093	-7.2228	1.8306	38	1.6504	0.6610	-30.3298	23.6322	65
\hat{b}	3.1385	0.3238	1.8412	10.2170	38	2.9791	0.0738	2.7296	3.0588	65
\hat{p}	0.2580	0.1313	-0.7436	0.6065	38	0.3138	0.1638	-6.9446	4.4906	37
p=0.5										
\hat{a}	0.9096	0.4875	-4.7760	1.9928	47	1.8023	0.4680	-123.5460	17.2694	80
\hat{b}	3.1420	0.3586	2.2325	7.5100	47	2.9383	0.0973	2.6874	3.0588	80
\hat{p}	0.3620	0.1186	-0.4609	0.7157	47	0.4595	0.1830	-15.2355	1.2987	62
p=0.7										
\hat{a}	0.9353	0.5290	-2.8318	2.1264	46	1.9078	0.4672	0.7822	4.5752	96
\hat{b}	3.0336	0.3706	1.6793	6.3080	46	2.8837	0.1803	2.0446	3.0590	96
\hat{p}	0.4176	0.1424	-0.9976	0.8192	46	0.5980	0.1744	-1.7401	31.6168	75
p=0.9										
\hat{a}	0.7405	0.4598	-1.8940	1.7099	40	1.7475	0.5882	-24.5125	8.0062	83
\hat{b}	2.7766	0.5455	0.7725	4.4810	40	2.5633	0.3561	1.9467	3.0588	83
\hat{p}	0.3794	0.1616	-4.5247	0.6998	40	0.6057	0.2710	-23.7962	7.6527	46

Table-1(c)

Parameter Estimates of Mixture of two Rectangular Distributions (0-->a) & (0-->b), a=2, b=4, p=0.3:0.1:0.7, Number of samples (ns) =100, sample size (n) =100

p=0.3	Moment Ratio				% acceptable estimates	Extreme value estimate (b)				% acceptable estimates (a, b & p)
	mean	std	min	max		mean	std	min	max	
\hat{a}	1.5294	0.4293	0.4380	2.6262	61	1.9800	0.2764	1.4665	3.0123	100
\hat{b}	4.0891	0.2293	3.4620	4.5250	61	3.9001	0.1193	3.4428	4.0398	100
\hat{p}	0.6095	0.0994	0.3589	0.9789	61	0.6879	0.1009	0.4283	0.9804	99
p=0.4										
\hat{a}	1.4776	0.4194	0.6427	2.5976	59	2.0203	0.2984	1.4165	2.9020	100
\hat{b}	4.1146	0.2047	3.4261	4.4930	59	3.9506	0.0858	3.5984	4.0378	100
\hat{p}	0.5315	0.1068	0.3574	0.8207	59	0.6012	0.1120	0.3808	0.8945	99
p=0.5										
\hat{a}	1.6230	0.4843	0.8304	2.8170	71	1.9421	0.3753	1.1318	3.1448	100
\hat{b}	4.0625	0.2111	3.4248	4.4045	71	3.9651	0.0763	3.6703	4.0396	100
\hat{p}	0.4614	0.1007	0.1777	0.6999	71	0.5018	0.1291	0.1872	0.9684	100
p=0.6										
\hat{a}	1.6086	0.5376	0.6797	2.9637	63	2.0441	0.5128	1.0749	3.5099	99
\hat{b}	4.0588	0.2545	3.2427	4.5963	63	3.9803	0.0524	3.8189	4.0398	99
\hat{p}	0.3746	0.1171	0.1163	0.7490	63	0.4247	0.1307	0.1236	1.0086	99
p=0.7										
\hat{a}	1.6129	0.5872	0.5246	3.2331	58	2.0047	0.6162	0.2876	3.6996	98
\hat{b}	4.0536	0.2885	3.2120	4.5421	58	3.9798	0.0551	3.7665	4.0398	98
\hat{p}	0.2984	0.1119	0.0521	0.6370	58	0.3267	0.1394	0.0541	0.8182	94

Table-2(a)

Mixture of two Rectangular Distributions with Parameters (a_1, b_1, a_2, b_2 & p)										
Three types (i) gap (ii)overlap (iii) Imbedding vs single population (a_1, b_2)										
(a) Single Distribution vs (b) Three types of Mixture Distributions										
$a_1=2, b_1=3, a_2=4, b_2=5, p=0.1:0.2:0.9$, sample size ' n '=50, number of samples ' ns '=100										
	Theoretical			Simulation			Extreme value Estimates			
a) single	3.5	13	0.87	3.51	13.09	0.86	\hat{a}_1		\hat{b}_2	
b) 'p' values	m_1'	m_2'	std	m_1'	m_2'	std	m_1'	std	m_1'	Std
i) 0.1	3.5	12.40	0.39	3.50	12.40	0.39	2.14	0.17	5.08	0.02
0.3	3.7	14.23	0.74	3.70	14.23	0.74	2.31	0.32	5.05	0.04
0.5	3.5	12.83	0.76	3.50	12.83	0.76	2.43	0.34	4.57	0.35
0.7	3.3	11.43	0.74	3.30	11.43	0.74	2.02	0.06	5.06	0.02
0.9	3.1	10.03	0.65	3.10	10.03	0.65	2.09	0.13	5.04	0.05
ii) 0.1	4.3	18.93	0.67	4.30	18.93	0.67	2.16	0.20	4.88	0.24
0.3	3.9	16.13	0.96	3.90	16.13	0.96	1.99	0.03	5.06	0.03
0.5	3.5	13.33	1.04	3.50	13.33	1.04	2.03	0.06	5.02	0.08
0.7	3.1	10.53	0.96	3.10	10.53	0.96	2.08	0.13	4.98	0.09
0.9	2.7	7.73	0.67	2.70	7.73	0.67	1.98	0.02	5.03	0.06
iii) 0.1	3.5	12.40	0.39	3.50	12.40	0.39	2.00	0.04	4.72	0.29
0.3	3.5	12.53	0.53	3.50	12.53	0.53	2.03	0.07	5.02	0.08
0.5	3.5	12.67	0.65	3.50	12.67	0.65	1.98	0.02	4.93	0.15
0.7	3.5	12.80	0.74	3.50	12.80	0.74	2.00	0.04	4.72	0.29
0.9	3.5	12.93	0.83	3.50	12.93	0.83	2.02	0.06	5.03	0.06

References

- [1] Blischke W.R(1962): Moment Estimators for the Parameters of a Mixture of Two Binomial Distributions Ann. Math. Statist. 33, 2, pp. 444 - 454.
- [2] Day N.E (1969): Estimating the Components of a Mixture of Normal Distributions. Biometrika, 56, 3, pp. 463 - 474.
- [3] Hussein. A: Liu. J(2009): Parametric Estimation of Mixtures of Two Uniform Distributions: Journal of Statistical Computation and Simulation, 79, 4, pp. 395 - 410.
- [4] Keewhan Choi and Bulgren W. G. (1968): An Estimation Procedure for Mixtures of Distributions. Journal of the Royal Statistical Society. Series B (Methodological), 30, 3, pp.444 – 460.
- [5] Pavan Kumar(2003):An Empirical Investigation to Some Problems in Minimum Value estimation and one sided Probabilities :PhD thesis submitted to Osmania University.
- [6] Peter F. Craigmile: Titterington D.M(1997): Parameter Estimation for Finite Mixtures of Uniform Distributions Communications in Statistics. 26, pp. 1981-1995.
- [7] Rider P.R(1961): The Method of Moments Applied to a Mixture of Two Exponential Distributions. Ann. Math. Statist. 32, pp.143-147.
- [8] Tallis G.M and Light .R (1968): The Use of Fractional Moments for Estimating the Parameters of a Mixed Exponential Distributions. Technometrics. 10, 1, pp.161 – 175.