Proceedings of The 9th Islamic Countries Conference on Statistical Sciences 2007
ICCS-IX 12-14 Dec 2007

Mixture proportion in view of statistical evidences

M. Arashi and M. Emadi
School of Mathematical Sciences, Ferdowsi University of Mashhad
P.O. Box 1159-9177948953, Mashhad, IRAN

Abstract

In some practical inferential situations, it is needed to mix some adequate sorts of distributions to fit a robust model for multimodal observations. In this paper, we study the behavior of mixture proportion in a mixture of two asymmetric normal distributions with the interpretation of data as evidences. In this approach, for visualizing and understanding model of interest, the profile likelihood has been used to eliminate the nuisance parameter.

Key words and phrases: Statistical evidence, Likelihood ratio, Profile likelihood, Asymmetric normal type I, Finite mixture model.

1 Introduction and Preliminaries

Introduction and summary mixtures of normal distributions have a long history in statistics, dating back to the late 19th century and the writings of Newcomb (1886) and Pearson (1894). Since then, they appear as models in diverse areas of applied research. However, even in the simplest of cases, the two-component normal mixture, one encounters serious theoretical as well as computational difficulties when attempting to perform basic statistical analysis such as parameter estimation and goodness-of-fit. Following Rao's (1948) paper likelihood estimation appears not to have been pursued further until Hasselblad (1966, 1969) addressed the problem, initially for a mixture of g univariate normal distributions with equal variances. The likelihood approach to the fitting of mixture models, in particular normal mixtures, has since been utilized by several authors, including dick and Bowden (1973), Hosmer (1973a and b, 1974, 1978), O'neill (1978), Ganesalingam and Mclachlan (1978,1979a, 1980a), and Aitkin (1980a).

In many common statistical problems, we encounter observations with more than one mode. Particularly, when we focus on what the data say, it is important to

choose a robust model which fits well. Then, we mix two or more suitable distributions for reasonable justification of observables. Because of applicability of normal distribution, in this approach, we use mixture of two asymmetric normal distributions and study the behavior of mixture proportion in order to distinguish accurate model for bimodal observations.

The asymmetric normal is a class of distributions that includes the normal one as a special case, and skew normal as a special subclass. A random variable X is said to be asymmetric normal type I with location, scale and shape parameters μ, σ, ζ respectively, denoted by $X \sim ANI(\mu, \sigma, \zeta)$, and its probability density function (pdf) is

$$f_X(x; \mu, \sigma, \zeta) = \sigma^{-1} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\, G\left(\frac{x-\mu}{\sigma}, \zeta\right)\right], \tag{1.1}$$

where $G(x, \zeta) = \exp[\frac{-2\zeta}{\pi} \arctan(x)]$, and $\frac{2}{\pi} \arctan(x)$ is cauchy pdf. See Evans et al (2000).

For $\zeta = 0$ it coincides with Standard Normal distribution, and for $\zeta > 0$, it has right skewness and for $\zeta < 0$ it has left skewness.

The mixture of two ANIs with different shape parameters, which lead skewness of opposite sites, is given by

$$g(x; \mu, 1, c, \phi) \;=\; \phi ANI(\mu, 1, \zeta_1) + (1-\phi) ANI(\mu, 1, \zeta_2), \tag{1.2}$$

where $\zeta = (\zeta_1, \zeta_2)$ and $0 \leq \phi \leq 1$.

Data drawn from a statistical model, make realistic and available evidence in interpreting and examining the distribution. In some sense, given body of data represent evidence supporting statistical hypotheses about parameters of the model against another. Criteria use in competition between one hypothesis about the parameter of interest against another is likelihood performance. Emadi and Arghami (2003), Emadi et al (2005) and Arashi and Emadi (2006) have studied some measures of support for statistical hypotheses. An interesting question is how a number of observations verify the mixture of normal distributions , in terms of the amount of statistical evidence they provide about the unknown parameter(s). We use the probabilities of observing strong misleading evidence and weak evidence for the numbers of iid observation. We assume that f_i is the probability density function of a continuous random variable X under simple hypotheses H_i, $\quad (i = 1, 2)$. Suppose we can observe the sequence of iid observations X_1, X_2, ..., where each is distributed as X.

Let η be any measure of support for one hypothesis against another with values in the unit interval. Then the probabilities of observing strong misleading evidence

under H_1, H_2 are $M_1 = P_1(\eta \leq 1-c) = K_1(1-c)$ and $M_2 = P_2(\eta \geq c) = 1-K_2(c)$, respectively, and the probabilities of weak evidence under H_1 and H_2 are $W_1 = P_1(1-c < \eta < c) = K_1(c) - K_1(1-c)$ and $W_2 = P_2(1-c < \eta < c) = K_2(c) - K_2(1-c)$, respectively (see, Royall (2000)). Here c, $0.5 \leq c < 1$, is a threshold of strong evidence, and K_1 and K_2 are cdf's of η under H_1 and H_2, respectively. We argue that since both misleading and weak evidence are undesirable, and obtaining strong misleading evidence is more important than obtaining just weak evidence, a pre experimental measure desirability of a measure of evidence can be taken to be

$$e(\eta) \;=\; 1 - \int_{0.5}^{1} \{\gamma[M_1(t) + M_2(t)] + W_1(t) + W_2(t)\}\, dt. \tag{1.3}$$

where $\gamma \geq 1$ and M_i and W_i are respectively the probabilities of strong misleading evidence and weak evidence under H_i, $(i = 1,\ 2)$. The following theorem gives $e(\eta)$ in terms of K_1 and K_2.

Theorem 1. (Emadi et al 2005)Under the assumptions of this section, we have

$$e(\eta) = \int_0^1 (K_2(t) - K_1(t))dt + (2-\gamma)\left(\frac{1}{2} + \int_0^{0.5} K_1(t)dt - \int_{0.5}^1 K_2(t)dt\right).$$

Corollary 1. Under the assumptions of theorem 1, for $\gamma = 2$ we have

$$\begin{aligned} e(\eta) &= \int_0^1 [K_2(t) - K_1(t)]dt \\ &= E_{H_1}(\eta) - E_{H_2}(\eta). \end{aligned}$$

It is interesting to note that for $\gamma = 2$, $e(\eta)$ (which was introduced and used by Emadi and Arghami (2003)) has another interpretation, this being the area (with unit square) between the curves of $K_1(t)$ and $K_2(t)$.
Let λ be the likelihood ratio for the competing hypotheses $H_1 : \theta = \theta_1$ and $H_2 : \theta = \theta_2$ so that

$$\begin{aligned} \lambda &= \frac{L_1}{L_2} \\ &= \prod_{i=1}^{n} \frac{f_1(X_i)}{f_2(X_i)}. \end{aligned} \tag{1.4}$$

Through out the paper we shall use $\eta = \lambda/(\lambda+1)$ as a measure of support for H_1 against H_2.

2 Model Selection

In this section we compute $M1$, $M2$, $W1$ and $W2$ for two groups of hypotheses about the mixture proportion ϕ in making evidential selection of true model.
Impossibility of analytical derivatives of latter results based on the mixture model (1.2), via a simulation, we make decision for the following two special groups of hypotheses.

$$Group1 = \begin{cases} H_1 : \phi = 0.5 \\ H_2 : \phi = 0.75 \end{cases} \quad \text{and} \quad Group2 = \begin{cases} H_1 : \phi = 0.5 \\ H_2 : \phi = 0.25 \end{cases} \tag{2.1}$$

Without loss of generality, let $\sigma = 1$ and thought the best values for better adoption of mixture model due to the data set, let $\zeta_1 = 2$ and $\zeta_2 = -2$ in (1.2). Then the model of interest is given by

$$\begin{aligned} g(x;\mu,1,\zeta,\phi) &= \phi \exp\{-(x-\mu)^2 \exp[\frac{-4}{\pi}\arctan(x-\mu)]\} \\ &\quad +(1-\phi)\exp\{-(x-\mu)^2 \exp[\frac{4}{\pi}\arctan(x-\mu)]\}. \end{aligned} \tag{2.2}$$

It is desirable to select accurate model just by recognizing how amount of weight can be given to each ANI. In other words, in order to understand what the data say about the model by (1.4), we have to look at likelihood function when the parameter space has two dimensions, (μ, ϕ). Then it is not as easy to appreciate the likelihood function. Our problem is that we want to represent, interpret and report the evidences about ϕ alone, not for μ. So μ is a nuisance parameter.
By the form of the model (2.2), one adequate technique to remove μ from the model, is using profile likelihood. See Royall (1999).

In n random variables drawn from the model (2.2), profile likelihood for known ζ, is given by

$$\begin{aligned} L_p(\phi) &= Max_\mu L(\phi,\mu) \\ &= Max_\mu \prod_{i=1}^{n} g(x_i;\mu,1,\zeta,\phi), \end{aligned} \tag{2.3}$$

where $L(\phi,\mu)$ is the likelihood function.

The function in (2.3) can not be generally obtain in a closed form. Then, in such situations, one can use numerical techniques.

3 Simulation

For each group of hypotheses, n=12, 20 and 60 random variables are taken from the model (2.2). The function in (2.3) is computed and the whole process repeated r=1000 times. Values of $M1$, $M2$, $W1$ and $W2$ are achieved. We have used packages Maple9.5 and Minitab14 to do numeric computations.
Note that, in general, the data set can not obtain from the model (2.2). Thus the data arise from symmetric normal distributions with different location parameters in order to use in the model (2.2). The method of sampling is very important. It changes due to hypotheses as follows.
1) For $\phi = 0.5$, we take one digit in random form $\{0, 1\}$. If it obtains 0, we will take one random sample from $N(0, 1)$ otherwise from $N(4, 1)$.
2) For $\phi = 0.75$, we take one digit in random form $\{0, 1, 2, 3\}$. If it obtains 0, we will take one random sample from $N(4, 1)$ otherwise from $N(0, 1)$.
3) For $\phi = 0.25$, we take one digit in random form $\{0, 1, 2, 3\}$. If it obtains 0, we will take one random sample from $N(0, 1)$ otherwise from $N(4, 1)$.
The graphical results are given in Figures 1 and 2.

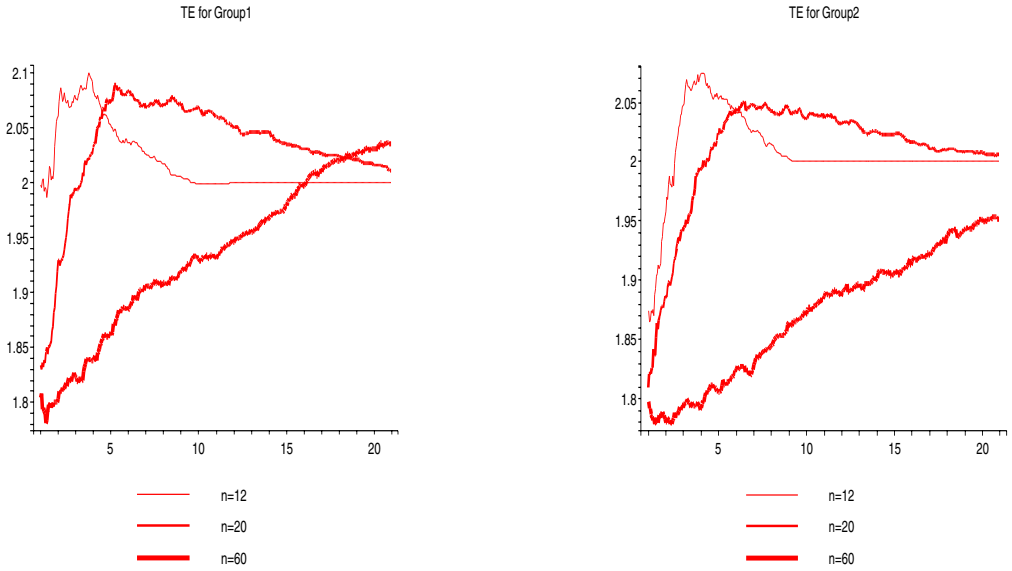
4 Concluding Remarks

(1) If $k \geq 8 \Rightarrow M1 \ll 0.14$; so the probability of observing strong misleading of H_1 when H_1 is true is quite low. In the other word, In 1000 times repetition of process, when the hypothesis H_1 is true, 14 times the data have strong misleading evidence from H_1 or the probability of strong misleading evidence is at most 0.14.

(2) If $k \geq 8 \Rightarrow M2 \approx 0$; so the probability of observing strong of H_1 when H_2 is true is approximately equal to zero. In the other word, when the hypothesis H_2 is true the data could not have strong evidence from H_1; which means that the probability of weak evidence or strong misleading evidence is quite high.

(3) If $k \geq 8 \Rightarrow W1 \gg 0.84$; so the probability of weak evidence from hypothesis H_1 when it is true is high.

(4) If $k \geq 8 \Rightarrow W2 \gg 0.96$; so the probability of weak evidence from hypothesis H_2 when it is true is very high.



(5) Emadi et al (2005) introduced the measure of Total Error (TE) or the error of evidential inference which is followed as

$$TE \quad = \quad 2(M1 + M2) + (W1 + W2).$$

Note that the above statement is a special form of $e(\eta)$ introduced in 1.3.
In the table below the average of TE is computed for each group of hypotheses

	Group1	Group2
n=12	2.0	2.0
n=20	2.0	2.0
n=60	1.9	1.8

(4.1)

In details one could obtain the following plots for different values of k.

We can coclude that when the number of observations (n), increases, the value of TE for each k in the average of 1000 times repetition decreases. When n gets bigger the exact value of k where the TE goes the be smaller increases; and as if its value is smaller it is better, we can conclude that as n increases the TE of evidential analysis decreases.

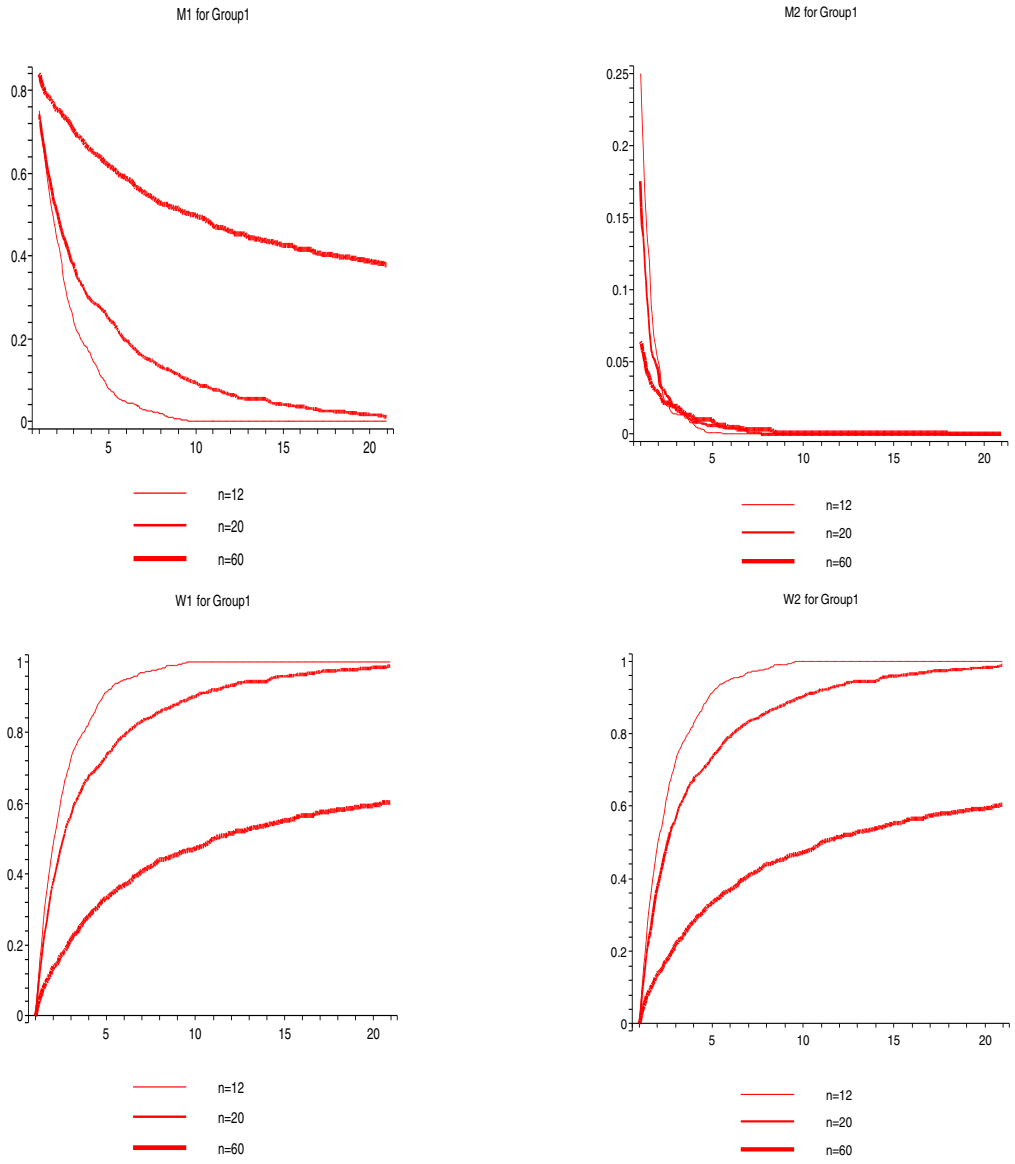


Figure 1: Plots for Group1

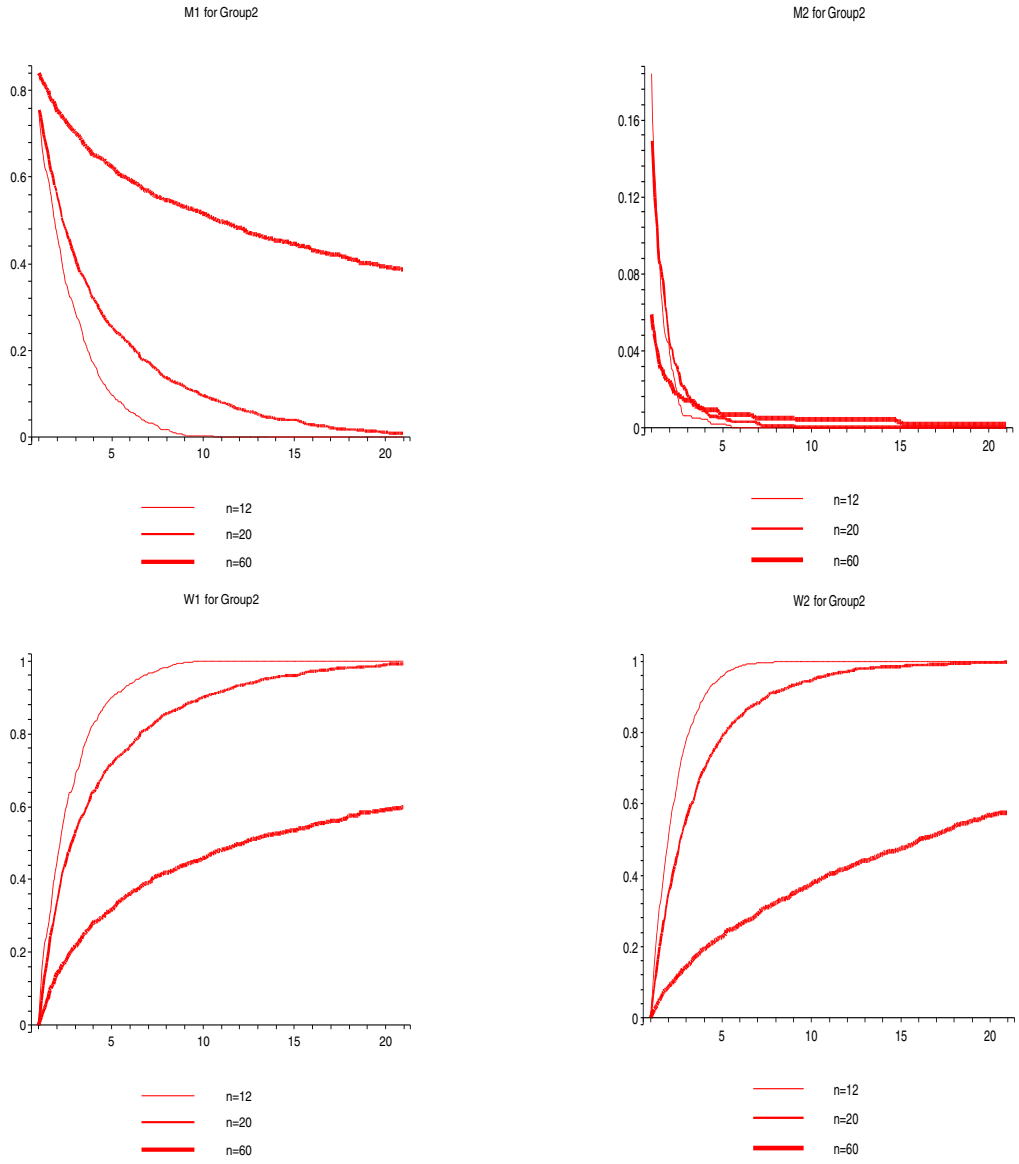


Figure 2: Plots for Group2

References

Arashi, M. and Emadi, M. (2006). Evidential inference based on record data and inter-record times. *Statistical Papers*, Online First.

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence, *Statist. Med.* 21, 2563 - 2599.

Emadi, M. and Arghami, N. R. (2003). Some measures of support for statistical hypotheses. *J. Stat. Theory Appl.*, 2, 165–176.

Emadi, M., Ahmadi, J. and Arghami, N. R. (2005). Comparing of record data and random observation based on statistical evidence *Statistical Papers*.

Evans, M., Hastings, N. and Peacock, B. (2000). *Statistical distribution*, Third Edition, John Wiley and Sons, New York.

Ganesalingam, S. and McLachlan, G. J. (1978). The efficiency of linear discriminant function based on unclassified initial sample, *Biometrika*, 65, 658–662.

Ganesalingam, S. and McLachlan, G. J. (1979a). Small sample results for a linear discriminant function estimated from a mixture of normal populations, *J. Statist. Comput. Simul.*, 9, 151–158.

Ganesalingam, S. and McLachlan, G. J. (1980a). A comparison of the mixture and classification approaches to cluster analysis, *Commun. Statist.- Theor. Meth.*, A9, 923–933.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions, *Trigonometrics*, 8, 431–444.

Hasselblad, V. (1969). Estimation of finite mixtures of distribution from the exponential family, *J. Amer. Statist. Assoc.*, 64, 1459–1471.

Hosmer, D. W. (1973a). On MLE of the parameters of a mixture of two normal distributions when the sample size is small, *Commun. Statist.*, 1, 217–227.

Hosmer, D. W. (1973b). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample., *Biometrics*, 29, 761–770.

Hosmer, D. W. (1974). Maximum likelihood estimates of the parameters of a mixture of two regression lines, *Commun. Statist.*, 3, 995–1006.

Hosmer, D. W. (1978). A use of mixtures of two normal distributions in a classifications problem, *J. Statist. Comput. Simul.*, 6, 281–294.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, Marcel Dekker Inc., New York.

Newcomb, S. (1886). A generalized theory of combination of observations so as to obtain the best result, *Amer. J. Math.*, 8, 343–366.

O'Neill, T. J. (1978). Normal discrimination with unclassified observations, *J. Amer. Statist. Assoc.*, 73, 821–826.

Pearson, K. (1894) Contribution to the mathematical theory of evolution, *Phil. Trans.*, A185, 71–110.

Royall, R. (1999) *Statistical Evidence: A Likelihood Paradigm.* Chapman & Hall, New York.

Royall, R. (2000) On the probability of observing misleading statistical evidence. *J. Amer. Statist. Assoc.* 95, 760–780