A representation of random variables for finite mixture model based on combinatorial form

Supawan Khotama and Watcharin Klongdee*
Mathematics Department, Khon Kaen University, Khon Kaen, 40002, Thailand,
supawan.k@kkumail.com and kwatch@kku.ac.th
*(*Corresponding Author)*

Abstract

A formula of generating random variable for finite mixture model is proposed. The main contribution of the work is a representation of random variable for finite cdf mixture model. We illustrate the generating random variable from the four components including a mixture of normal distribution, logistic distribution, log-normal distribution and gamma distribution in case of the number of the random variable is different, which present both the density and the cumulative probability and compare with mixture distribution. The results show that the more numbers of the random variable, the more the density and the cumulative probability are at the similar values more than small amount of number of the random variable.

Keywords: mixture model, random variable, mixing weight, generating

1. Introduction

A finite mixture model (FMM) is a family of probability density function (pdf) which is a convex combination of finite pdfs, two or more pdfs. By convex combination property, FMM is a generalization or approximation of each individual pdf. Accordingly, FMMs are a flexible tool for analyzing and explaining the complex data. The applications of FMMs were found in several research fields, especially in statistical analysis and machine learning, such as modeling, clustering, classification, and segmentation, see [5], [1], and [3]. Also, many researches propose random variable of mixture model in several formats. For example, Reference [2] described random variable as a mixture of two normal distributions.

In this section, we shall describe the definitions, notations and assumptions of FMM. Aftermost, we shall present the main theorem of this paper. In this paper, we assume that all random variables are defined in a probability space $(\Omega, \mathcal{F}, \Pr)$. Let X be a continuous random variable and $f_1, f_2, \dots, f_n$ be probability density functions of random variables $X_1, X_2, \dots, X_n$, corresponding to the parameter vectors $\Theta_1, \Theta_2, \dots, \Theta_n$, respectively. The random variable X is said to arise from a finite mixture model if it has a pdf of the form

$$f(x;\Theta) = \sum_{k=1}^{n} p_k f_k(x;\Theta_k), \tag{1}$$

where $\Theta = (\Theta_1, \Theta_2, \dots, \Theta_n, p_1, p_2, \dots, p_n)$ is a vector of parameters and $p_k \geq 0$ for $1 \leq k \leq n$ and $\sum_{k=1}^{n} p_k = 1$. The number n is called the mixture component, p_i is called the mixing weight of i^{th} component. Usually, FMM can be characterized in convex combination form of cumulative distribution functions (cdfs), $F_1, F_2, \dots, F_n$ corresponding to the pdfs $f_1, f_2, \dots, f_n$, respectively. Consequently, (1) is equivalent to the following form

$$F(x;\Theta) = \sum_{k=1}^{n} p_k F_k(x;\Theta_k). \tag{2}$$

Next, we recall an important property mixture model, see [4]. Let χ_A denote a indicator function of event A, i.e., $\chi_A(x) = 1$ if $x \in A$, or $\chi_A(x) = 0$ if $x \notin A$. A discrete random variable Δ is given by $\Pr(\Delta = k) = p_k$. We found that a random variable X is defined by

$$X = \sum_{k=1}^{n} \chi_{\{\Delta=k\}} X_k, \tag{3}$$

a random variable X is a random variable of mixture model (1) such that Δ is independent of the random variables $X_1, X_2, \ldots, X_n$ with distribution functions $F_{X_i} = F_i$.

Our goal is to propose a representation of random variable for finite mixture model which differ from (3) under the following assumption.

Assumption A: Let $n \geq 2$ be an integer. A discrete random variable Δ such that $\Pr(\Delta = k) = p_k$, and the finite sequence of random variable $\{X_i : 1 \leq i \leq n\}$ are independent.

Under Assumption A, we obtain that for each $x \in \mathbb{R}$ and for some $k \in \{1,2,3 \ldots, n\}$,

$$\Pr(\Delta = k, h(X_1, X_2, \ldots, X_n) \leq x) = \Pr(\Delta = k)\Pr(h(X_1, X_2, \ldots, X_n) \leq x)$$

for all a measurable function h. This leads to the following theorem.

Theorem 1: Let Assumption A hold, and $p_1, p_2, \ldots, p_n$ be nonnegative real numbers such that $\sum_{k=1}^{n} p_k = 1$. If random variable X has the cdf in the form $F(x;\Theta) = \sum_{k=1}^{n} p_k F_k(x;\Theta_k)$, when $F_k(x;\Theta_k)$ is cdf of $X_k, k = 1,2,3,\ldots,n$, then

$$X = \frac{1}{(n-1)!} \sum_{m=1}^{n} (-1)^{n+m} \binom{n-1}{m-1} \prod_{\substack{j=1 \\ j \neq m}}^{n} (\Delta - j) X_m \tag{4}$$

is distributional equality.

2. Proof of Theorem

Proof: Let X be random variable which has the cdf in the form

$$F(x;\Theta) = \sum_{k=1}^{n} p_k F_k(x;\Theta_k), x \in \mathbb{R} \tag{5}$$

For each $x \in \mathbb{R}$ and $k = \{1,2, \ldots, n\}$, we have

$$\Pr\left(\frac{1}{(n-1)!} \sum_{m=1}^{n} (-1)^{n+m} \binom{n-1}{m-1} \prod_{\substack{j=1 \\ j \neq m}}^{n} (\Delta - j) X_m \leq x \right)$$

$$= \sum_{k=1}^{n} \Pr\left(\frac{1}{(n-1)!} \sum_{m=1}^{n} (-1)^{n+m} \binom{n-1}{m-1} \prod_{\substack{j=1 \\ j \neq m}}^{n} (\Delta - j) X_m \leq x, \Delta = k \right)$$

$$= \sum_{k=1}^{n} \Pr\left(\frac{1}{(n-1)!} \sum_{m=1}^{n} (-1)^{n+m} \binom{n-1}{m-1} \prod_{\substack{j=1 \\ j \neq m}}^{n} (\Delta - j) X_m \leq x \mid \Delta = k \right) \Pr(\Delta = k).$$

Under Assumption A, we assume that Δ is independent of the random variables $X_1, X_2, \ldots, X_n$, then

$$\sum_{k=1}^{n} p_k \Pr\left(\frac{1}{(n-1)!} \sum_{m=1}^{n} (-1)^{n+m} \binom{n-1}{m-1} \prod_{\substack{j=1 \\ j \neq m}}^{n} (k - j) X_m \leq x \right).$$

We obtain that

$$\sum_{k=1}^{n} \Pr\left(\frac{1}{(n-1)!} \left[\sum_{\substack{m=1 \\ m \neq k}}^{n} (-1)^{n+m} \binom{n-1}{m-1} \prod_{\substack{j=1 \\ j \neq m}}^{n} (k-j) X_m + (-1)^{n+k} \binom{n-1}{k-1} \prod_{\substack{j=1 \\ j \neq k}}^{n} (k-j) X_k \right] \leq x \right) p_k$$

$$= \sum_{k=1}^{n} \Pr\left(\frac{1}{(n-1)!} \left[\sum_{\substack{m=1 \\ m \neq k}}^{n} (-1)^{n+m} \binom{n-1}{m-1} (k-k) \prod_{\substack{j=1 \\ j \neq m \\ j \neq k}}^{n} (k-j) X_m + (-1)^{n+k} \binom{n-1}{k-1} \prod_{\substack{j=1 \\ j \neq k}}^{n} (k-j) X_k \right] \leq x \right) p_k$$

$$= \sum_{k=1}^{n} \Pr\left(\frac{1}{(n-1)!} (-1)^{n+k} \binom{n-1}{k-1} (k-1)(k-2) \cdots (k-(k-1))(k-(k+1)) \cdots (k-n) X_k \leq x \right) p_k$$

$$= \sum_{k=1}^{n} \Pr\left(\frac{(k-1)!\,(n-k)!}{(n-1)!} (-1)^{n+k} \binom{n-1}{n-k} (-1)^{n-k} X_k \leq x \right) p_k$$

$$= \sum_{k=1}^{n} \Pr\left(\binom{n-1}{n-k}^{-1} \binom{n-1}{n-k} X_k \leq x \right) p_k$$

$$= \sum_{k=1}^{n} \Pr(X_k \leq x)\, p_k$$

$$= \sum_{k=1}^{n} p_k F_k(x; \Theta_k)$$

Therefore, we conclude that

$$\Pr(X \leq x) = \Pr\left(\frac{1}{(n-1)!} \sum_{m=1}^{n} (-1)^{n+m} \binom{n-1}{m-1} \prod_{\substack{j=1 \\ j \neq m}}^{n} (\Delta - j) X_m \leq x \right).$$

This completes the proof of the theorem.

3. Some examples

Next, we consider the FMM with four components

$$F(x;\Theta) = p_1 F_1(x;\Theta_1) + p_2 F_2(x;\Theta_2) + p_3 F_3(x;\Theta_3) + p_4 F_4(x;\Theta_4), \tag{6}$$

where $F_1(x;\Theta_1), F_2(x;\Theta_2), F_3(x;\Theta_3)$ and $F_4(x;\Theta_4)$ are cdf of random variable X_1, X_2, X_3, and X_4, respectively. By Theorem1, we found that

$$X = \frac{1}{3!}\sum_{m=1}^{4}(-1)^{4+m}\binom{3}{m-1}\prod_{\substack{j=1\\ j\neq m}}^{4}(\Delta - j)X_m \tag{7}$$

is a random variable of cdf $F(x;\Theta)$. Particularly, we set X_1~Normal(0,2), X_2~Logistic(5,1), X_3~ Lognormal(2,1), and X_4~ Gamma(10,2), with mixing weight $p_1 = 0.2, p_2 = 0.35, p_3 = 0.15, and\ p_4 = 0.3$,

$$F(x;\Theta) = 0.2\left(\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x}{\sqrt{4}}\right)\right]\right) + 0.35\left(\frac{1}{1+e^{(5-x)}}\right) + 0.15\left[\frac{1}{2}+\frac{1}{2}\operatorname{erf}\left(\frac{\ln(x)-2}{\sqrt{2}}\right)\right] + 0.3\left(\frac{\gamma(10,\frac{x}{2})}{\Gamma(10)}\right), \tag{8}$$

where $\gamma(k,\frac{x}{\theta})$ is the lower incomplete gamma function, and erf(x) is a special function (non-elementary) of sigmoid shape that occurs in probability.

One important application of Theorem 1 is the generating random numbers with distribution function $F(x;\Theta)$. The simple method of generating random variable is called inversion method, set $x = F^{-1}(u)$ where $u \sim U[0,1]$, see [6]. From (8), we generate found that $F(x;\Theta)$ is rather complicate, and it is not easy to calculate inverse function. From (7), we random numbers for each random variable, and it are easy to random Δ. Hence, the generating random numbers of mixture model from (7) is easy, and the following random result is obtained. Now, we generate random variable X from (7). This generative representation is explicit: $Im(\Delta) \in \{1,2,3,4\}$. We illustrate the density and the cumulative probability derived from both the generating random variable by Theorem 1 and the direct mixture distribution in case of number of generating random variable (n) are 100, 500, 1000, and 10000, respectively.

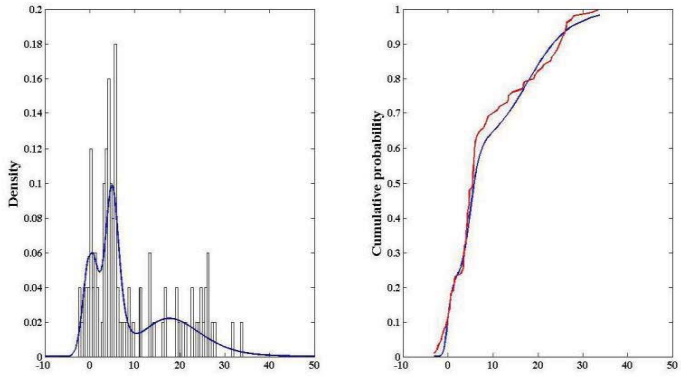

Figure 1: (Left panel :) Comparison of density derived from both the generating random variable and mixture distribution. (Right panel :) Comparison of cumulative probability derived from both the generating random variable and mixture distribution. n=100

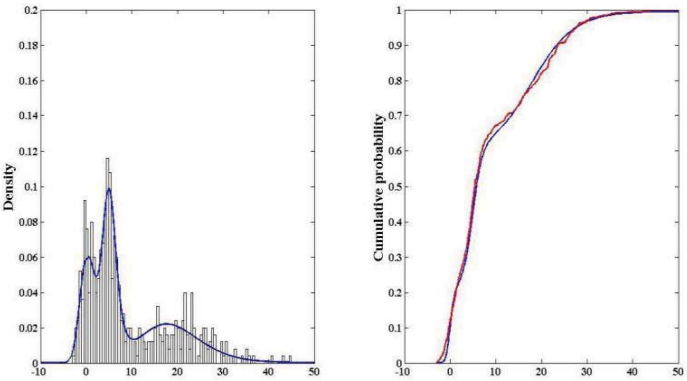

Figure 2: (Left panel :) Comparison of density derived from both the generating random variable and mixture distribution. (Right panel :) Comparison of cumulative probability derived from both the generating random variable and mixture distribution. n=500

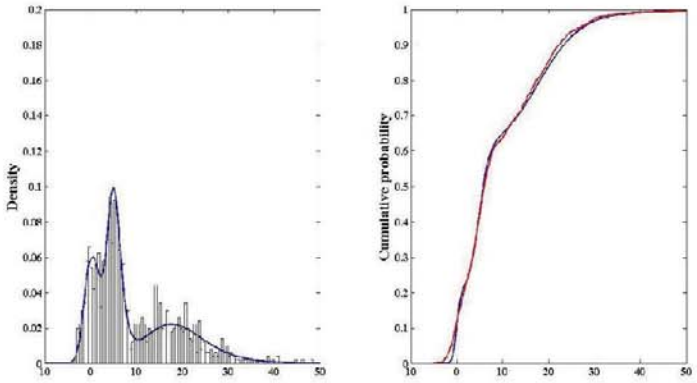

Figure 3: (Left panel :) Comparison of density derived from both the generating random variable and mixture distribution. (Right panel :) Comparison of cumulative probability derived from both the generating random variable and mixture distribution. n=1000

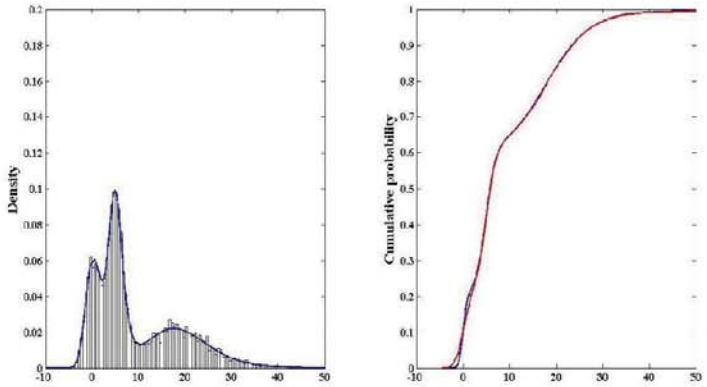

Figure 4: (Left panel :) Comparison of density derived from both the generating random variable and mixture distribution. (Right panel :) Comparison of cumulative probability derived from both the generating random variable and mixture distribution. n=10000

Figure 1 to Figure 4 show that the comparison of density and cumulative probability derived from both the generating random variable by Theorem1 and the directly mixture distribution at n=100, 500, 1000 and 10000, respectively.

Concerning Figure 1 to Figure 4, it is interesting to note that, when there are many number of the generating random variable, the density and the cumulative probability derived from the generating random variable by Theorem1 is similar to the density and the cumulative probability of the direct mixture distribution. Error values show in Tables 1.

Table 1: Statistical errors for different random variable number of random variable and mixture distributions

Number of random variable	Mean Absolute Error (MAE)	Mean Square Error (MSE)
100	0.0337	0.00170
500	0.0179	0.00043
1000	0.0127	0.00025
10000	0.0082	0.00015

We have checked error values of random variable and mixture distribution. Tables 1 shows that the Mean Absolute Error (MAE) and Mean Square Error (MSE) of random variable and mixture distribution in case of number of random variable are 100, 500, 1000, and 10000, respectively. As can be observed, random variable numbers as 10,000 has MAE and MSE least and random variable numbers as 1000, 500, and 100, respectively.

4. Conclusions

A simple formula of the generating random variable for finite cdf mixture model shows the examples of four components including a mixture of normal distribution, logistic distribution, log-normal distribution and gamma distribution. The result shows that if there is more amount of the random variable, it will cause similar value of density and cumulative probability obtained from random variable and mixture distribution more than a small amount of the random variable. In the further research will be using a formula of the generating random variable, which obtained from this research, apply to a real world situation.

5. Acknowledgements

This research was supported by Research Professional Development Project under the Science Achievement Scholarship of Thailand (SAST).

6. References

[1] Farnoosh R, Zarpak B. Image segmentation using Gaussian mixture model. IUST International Journal of Engineering Science. 2008; 19(1-2): 29-32.

[2] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2nd ed. California: Springer; 2008.

[3] Kollu R, Rayapudi SR, Narasimham SVL, Pakkurthi KM. Mixture probability distribution functions to model wind speed distributions. International Journal of Energy and Environmental Engineering. 2012; 3-27.

[4] Mikosch T. Non-Life Insurance Mathematics. 2nd ed. Berlin: Springer; 2008.

[5] Morgan EC, Lackner M, Vogel RM, Baise LG. Probability distribution for offshore wind speeds. Energy Conversion and Management. 2011; 52(2011): 15-26.

[6] Seydel R. Tools for Computation Finance. 2nd ed. Berlin: Springer; 2003.