Journal of Computer Science & Computational Mathematics, *Volume 11, Issue 1, March 2021*
DOI: 10.20967/jcscm.2021.01.002

Application of Bivariate Beta Mixture Distribution in Simulation Data Proportion with High Correlation

Nurvita Trianasari[1,2,*], I Made Sumertajaya[2], Erfiani[3], I Wayan Mangku[4]

[1]Telkom of Economics and Business School, Telkom University, Bandung, Indonesia
[2]Department of Statistics, Student of Doctoral Program at Sekolah Pascasarjana-IPB University
[3]Department of Statistics, IPB University, Bogor, Indonesia
[4]Department of Mathematics, IPB University, Bogor, Indonesia
[*]Corresponding author email: nurvitatrianasari@telkomuniversity.ac.id

Abstract: Cluster analysis is a multivariate analysis that aims to cluster objects or data so that objects or data that are in the same cluster have relatively more homogeneous properties than objects or data in different clusters. Probabilistic clustering method is often based on the assumption that data comes from a mixture of distributions, for examples Poisson, normal, lognormal, and Erlang. Thus the probabilistic clustering problem is transformed into a parameter estimation problem because the data is modeled by a cluster of mixture distribution. Data points that have the same distribution can be defined as one cluster. In this paper the distribution of bivariate beta mixtures for bivariate cases will be applied to the data on the proportion of data simulation with high correlation based on the results of the analysis on the this paper. While the ICL value of BIC in value in one cluster. Then it can be concluded that occur in Mixture 2 clusters.

Keywords: bivariate beta mixture model, integrated classification likelihood estimation Bayesian.

1. Introduction

The concept of cluster formation includes hierarchical methods, non-hierarchical methods and clustering methods that are probable (probabilistic clustering). In addition to the hierarchical and non-hierarchical methods outlined above, there is other method that is often used, namely the clustering method that has the opportunity to determine the optimal number of groups based on the distribution of the data. This clustering method is called a probabilistic clustering technique which assumes that the data follows a certain distribution. Probabilistic methods have the potential to be widely used in a variety of applications such as market segmentation, image segmentation [1] and [2], handwriting recognition [3], and document clustering [4]. This clustering method has the opportunity to try to optimize the suitability of the observed data with mathematical models using a probabilistic approach [5]. This method is often based on the assumption that data comes from a mixture of probability distributions for example Poisson, normal, lognormal, and Erlang. Thus the clustering problem is transformed into a parameter estimation problem because the data is modeled by a cluster of mixture distribution. Data points that have the same distribution can be defined as groups.

In [6], discusses the distribution of beta mixtures of multiple variables where the parameter estimation use the EM algorithm and the determination of the optimal number of groups using the integrated classification likelihood (Bayesian information criterion) determinant method. This distribution is applied to identify users on the community question answering site (CQA). In this paper the distribution of beta mixtures for single variable cases will be applied to the data on the proportion of data simulation with high correlation.

2. Beta Distribution

Let Y be a random variable having beta distribution with the parameters α and β, where $-\infty < \alpha < \infty$ and $-\infty < \beta < \infty$. The density function of this random variable is:

$$g\left(y|\alpha,\beta\right) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}; \quad 0 < y < 1. \tag{1}$$

The mean and variances of this random variable are :

$$E\left(Y\right) = \frac{\alpha}{\alpha+\beta},$$

and

$$Var\left(Y\right) = \frac{\alpha\beta}{\left[\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)\right]}.$$

3. The Bivariate Beta Mixture Model

In [6] used the i, x_i, $i = 1,2,\cdots,n$, observation data to form a mixture density function

$$f(x_i|\boldsymbol{\alpha},\boldsymbol{a},\boldsymbol{b}) = \sum_{c=1}^{C} \alpha_c f_c(x_i|a_c,b_c) \tag{2}$$

where $\boldsymbol{\alpha} = \{\alpha_1,\, \alpha_2,\, \ldots, \alpha_C\},\ \sum_{c=1}^{C}\alpha_c = 1\,;\ \alpha_c > 0$ express the mixture coefficient; C denotes the number of groups in the mix; f_c denotes the density function for each c-th cluster; $\boldsymbol{a} = \{a_1,\ a_2,\ \dots, a_C\}$ and $\boldsymbol{b} = \{b_1,\ b_2,\ \dots, b_C\}$ where a_C and b_C represent c-cluster parameters.

The density function of the beta distribution of a single variable for the c-class mix of beta is defined as

$$f_c\left(x_i|a_c,\ b_c\right) = \frac{\Gamma(a_c+b_c)}{\Gamma\left(a_c\right)\Gamma\left(b_c\right)} x_i^{a_c-1}\left(1-x_i\right)^{b_c-1} \qquad (3)$$

where $\Gamma(.)$ states the gamma function which is defined as

$$\Gamma\left(y\right) = \int_0^\infty t^{y-1}e^{-t}dt; \quad t > 0.$$

4. Maximum Likelihood Estimation for the Bivariate Beta Mixture Model

The parameters of the bivariate BMM can be estimated using maximum likelihood estimation. Suppose that Θ=$\{\alpha_1,\ \alpha_2,\dots,\ \alpha_C;\ a_1,\ a_2,\dots,\ a_C;\ b_1,\ b_2,\dots,\ b_C\}$ represents the set of unknown mixture parameters of the model and $x = \{x_1,\ x_2,\dots,\ x_n\}$ represent the set of the normalized feature vectors. Therefore, the likelihood function corresponding to C components of the mixture can be expressed as [6]

$$L\left(X|\Theta\right) = \prod_{i=1}^{n} f\left(x_i|\boldsymbol{\alpha},\boldsymbol{a},\boldsymbol{b}\right) = \prod_{i=1}^{n}\sum_{c=1}^{C}\alpha_c f_c\left(x_i|a_c,b_c\right). \qquad (4)$$

The expectation maximization (EM) algorithm is used to estimate the mixture model parameters for maximum likelihood in which each user's feature vector x_i is assigned to C dimensional indication vector $\boldsymbol{z}_i = \left(z_{i1},\ z_{i2},\ \dots,\ z_{iC}\right)^T$ such that

$$z_{ic} = \begin{cases} 1 & ;\text{ If } x_i \text{ belongs to the component } c \\ 0 & ;\text{ otherwise.} \end{cases} \qquad (5)$$

Suppose that $Z = \{z_1, z_2, \dots, z_n\}$ denote the set of indication vectors for set of users' $X = \{x_1, x_2, \dots, x_n\}$. The likelihood function of the data set is given by

$$L\left(X,Z|\Theta\right) = \prod_{i=1}^{n}\prod_{c=1}^{C}\left[\alpha_c f_c\left(x_i|a_c,\ b_c\right)\right]^{z_{ic}}. \qquad (6)$$

Next, take the logarithm of the likelihood function, which is given by

$$\log\left(L\left(X,Z|\Theta\right)\right) = \sum_{i=1}^{N}\sum_{c=1}^{C} z_{ic}\log\left[\alpha_c f_c\left(x_i|a_c,b_c\right)\right]. \qquad (7)$$

Now, the estimation of Θ is done through EM algorithm with number of iterations $I = \{0,1,2,\dots\}$ between the expectation and maximization steps so as to a sequence estimate $\left\{\widehat{\Theta}\right\}^{(I)}$ until the change in the value of the log-likelihood function expressed in equation Eq. (7) is negligible.

Expectation step: the indication for the c-component of feature vectors replaced its expectations as follows

$$z_{ic}^{(I)} = E\left[z_{ic}|x,\Theta\right] = \frac{\widehat{\alpha}_c^{(I)} f_c\left(x_i|\hat{a}_c,\ \hat{b}_c\right)}{\sum_{k=1}^{C}\widehat{\alpha}_k^{(I)} f_k\left(x_i|\hat{a}_c,\ \hat{b}_c\right)}. \qquad (8)$$

Maximization steps : the set of unknown parameters Θ=$\{\alpha_1,\alpha_2,\dots,\alpha_C;a_1,a_2,\dots,a_C;b_1,b_2,\dots,b_C\}$ of the mixture model are calculated using the estimated z_{ic} values in the expectation step. The mixing coefficients of the model are calculated as

$$\widehat{\alpha}_c^{(I+1)} = \frac{\sum_{i=1}^{n}\hat{z}_{ic}^{(I)}}{n}; \quad c = 1,2,\dots,C. \qquad (9)$$

The gradient derivative of the expectation of the log-likelihood of the dataset a_c and b_c and equated to zero, which is used to find the value $\hat{a}_c$, $\hat{b}_c$ that maximizes the likelihood as follows

$$\begin{bmatrix} \frac{\partial E\left[\log\left(L\left(X,Z|\Theta\right)\right)\right]}{\partial a_c} \\ \frac{\partial E\left[\log\left(L\left(X,Z|\Theta\right)\right)\right]}{\partial b_c} \end{bmatrix} = \mathbf{0} \qquad (10)$$

where,

$$\frac{\partial E\left[\log\left(L\left(X,Z|\Theta\right)\right)\right]}{\partial a_c} = \sum_{i=1}^{n}\hat{z}_{ic}\left[\frac{\Gamma'(a_c+b_c)}{\Gamma(a_c+b_c)} - \frac{\Gamma'(a_c)}{\Gamma(a_c)} + \log(x_i)\right] \qquad (11)$$

and

$$\frac{\partial E\left[\log\left(L\left(X,Z|\Theta\right)\right)\right]}{\partial b_c} = \sum_{i=1}^{n}\hat{z}_{ic}\left[\frac{\Gamma'(a_c+b_c)}{\Gamma(a_c+b_c)} - \frac{\Gamma'(b_c)}{\Gamma(b_c)} + \log\left(1-x_i\right)\right]. \qquad (12)$$

From equations Eq. (11) and Eq. (12), equation Eq. (10) can be represented as follows

$$\begin{bmatrix} \sum_{i=1}^{n}\hat{z}_{ic}\left[\psi\left(a_c+b_c\right) - \psi\left(a_c\right) + \log(x_i)\right] \\ \sum_{i=1}^{n}\hat{z}_{ic}\left[\psi\left(a_c+b_c\right) - \psi\left(b_c\right) + \log(1-x_i)\right] \end{bmatrix} = \mathbf{0} \qquad (13)$$

where $\psi(.)$ represents the digamma function defined as $\psi\left(\lambda\right) = \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}$. An exact solution to equation Eq. (13) as the digamma function is defined through integration. Therefore, the Newton-Raphson (a tangent method for root finding) is used to estimate parameter a_c and b_c iteratively as

$$\begin{bmatrix} a_c^{(I+1)} \\ b_c^{(I+1)} \end{bmatrix} = \begin{bmatrix} a_c^{(I)} \\ b_c^{(I)} \end{bmatrix} - \begin{bmatrix} \frac{\partial E[\log(L(X,Z|\Theta))\ ]}{\partial a_c} \\ \frac{\partial E[\log(L(X,Z|\Theta))\ ]}{\partial b_c} \end{bmatrix} x \begin{bmatrix} \frac{\partial^2 E[\log(L(X,Z|\Theta))\ ]}{(\partial a_c)^2} & \frac{\partial^2 E[\log(L(X,Z|\Theta))\ ]}{\partial a_c \partial b_c} \\ \frac{\partial^2 E[\log(L(X,Z|\Theta))\ ]}{\partial b_c \partial a_c} & \frac{\partial^2 E[\log(L(X,Z|\Theta))\ ]}{(\partial b_c)^2} \end{bmatrix}^{-1} \qquad (14)$$

where

$$\frac{\partial^2 E\left[\log\left(L\left(X,Z|\Theta\right)\right)\right]}{\left(\partial a_c\right)^2} = \sum_{i=1}^{n} \hat{z}_{ic}\left[\psi'\left(a_c+b_c\right)-\psi'\left(b_c\right)\right] \tag{15}$$

where $\psi'(.)$ is a tri-gamma function. The initial value of $a_c^{(0)}$ and $b_c^{(0)}$ needed to start the iteration process expressed in equation Eq. (14) is done through estimating the moment of beta distribution. The moment estimates $a_c^{(0)}$ and $b_c^{(0)}$ are defined as

$$\hat{a}_c^{(0)} = \overline{\mu}_c\left[\frac{\overline{\mu}_c(1-\overline{\mu}_c)}{\sigma_c^2}-1\right] \tag{16}$$

$$\hat{b}_c^{(0)} = (1-\overline{\mu}_c)\left[\frac{\overline{\mu}_c(1-\overline{\mu}_c)}{\sigma_c^2}-1\right] \tag{17}$$

where $\overline{\mu}_c$ is the sample mean and σ_c^2 is the sample variance of the feature value corresponding to D-dimension of the feature vectors and belongs to the C component of beta distribution. The Newton-Raphson algorithm converges when the change in values of estimates $\hat{a}_c$ and $\hat{b}_c$ is less than a small positive value ξ.

The maximum possible estimate of the beta distribution parameters can be done using the EM algorithm. The EM algorithm depends on initialization, Fuzzy C-Means (FCM) is used to initialize. First, the data set $(x_1,x_2,\cdots,x_n)$ is partitioned into a C cluster. Next, the parameters of each component of the dataset is estimated using the method of moment of the beta distribution and setting them as initial parameters which is required in EM algorithm.

5. Estimating the Number of Components in the Mixture

Various approaches have been proposed to estimate the number of components in mixture model. In [6] use deterministic approach based on EM algorithm to obtain a range of values for $C=1,2,\ldots,C_{\max}$ which is assumed to have optimal value of C. The number of components is selected according to the following criteria

$$\hat{C} = \mathrm{argmin}_C\left\{MSC\left(\widehat{\Theta}\left(C\right),C\right),C=1,2,\ldots,C_{\max}\right\} \tag{18}$$

where $\widehat{\Theta}(C)$ is an estimate of the mixture parameters assuming that it has C components, and MSC $\left(\widehat{\Theta}\left(C\right),C\right)$ is the model selection criterion. In [6], ICL-BIC is used as the model selection criterion defined as follows

$$\mathrm{ICL}-\mathrm{BIC}\left(C\right) = -2\log L_C + p\log\left(n\right) - 2\sum_{i=1}^{n}\sum_{c=1}^{C} z_{ic}\log z_{ic} \tag{19}$$

where L_C is the logarithm for getting the maximum likelihood solution of the beta mixture model and p is the number of estimated parameters. The detailed procedure for estimating the optimal number of beta components in the mixture of the dataset is illustrated in the following algorithm.

Algorithm 1 Estimating the number of components in the beta mixture

```
Input: $(x_1,x_2,\cdots,x_n)$ and $C_{\max}$
Output: The optimal number of components C representing beta mixture
Begin
for $C=1$ to $C_{\max}$ do
    if $C=1$ then
        Estimates that each parameter pair $\left\{\hat{a}_c,\ \hat{b}_c\right\}$ using equation Eq. (14)
        Compute the value of ICL-BIC (C) using equation Eq. (19)
    else
        Initialize EM algorithm using FCM clustering algorithm; alternate the following two steps to estimate the mixture parameters as:
        E-Step: Compute $z_{ic}^{(I)}$ using equation Eq. (8)
        M-Step:
          (1) Estimate the mixing coefficients using equation Eq. (9)
          (2) Estimate $\left\{\hat{a}_c,\ \hat{b}_c\right\}$ using equation Eq. (14)
        Repeat E-Step and M-Step until the change in equation Eq. (7) is negligible
        Compute the value of ICL-BIC $(C)$ using equation Eq. (19)
    end if
end for
Select $\hat{C}$ such that $\hat{C} = argmin_C\left\{\mathrm{ICL}-\mathrm{BIC}\left(C\right),C=1,2,\ldots,C_{\max}\right\}$
End
```

6. Application

6.1 Data

The data that will be used to apply the distribution of the beta mixture of two variables is the simulation data generated by the MATLAB software from the distribution of the beta mixture of two variables. Simulation data obtained from the generation of data through MATLAB software from the distribution of beta mix two variables with the number of groups 2 and sample size 300. The parameters of the distribution are $\alpha_1 = 0.25$, $\alpha_2 = 0.75$, $a_1 = 20$, $a_2 = 30$, $b_1 = 25$, and $b_2 = 30$, $c_1 = 35$, and $c_2 = 6$.

6.2 Methods

To do the modeling of proportion data it is necessary to do the following steps:

1. Make a statistical summary for simulation data.
2. Make a histogram for simulation data to see the shape of the distribution of simulation data.
3. Make a scatter diagram for the simulation data to see the number of groups in the simulation data.
4. Modeling one proportion variable to find out the best model for one proportion variable based on the minimum ICL-BIC value.
5. Model the data of two proportional variables by the mixed beta distribution of two variables to see the best model for data of two proportional variables based on the minimum ICL-BIC value.

6.3 Result and Discussion

Table 1 presents a summary of statistics for simulation data obtained from generating data through MATLAB software. Noted: variable 1 is the similar with peubah 1 and variable 2 is the similar with peubah 2.

Table 1. Summary of Statistics Simulation Data

Statistics	Variable 1	Variable 2
Mean	0.6896	0.7088
Median	0.7948	0.7965
Standard Deviation	0.2224	0.1979
Minimum	0.2041	0.2738
Maximum	0.9550	0.9549
Pearson Correlation	0.9760	

Figure 1 and Figure 2 each display a histogram for variable 1 and 2. Based on Figure 1 and Figure 2, it can be seen that the variable 1 data and the variable 2 data each have 2 modes. The results of distribution testing using the Kolmogorov-Smirnov distribution suitability test through MATLAB software show that the standard beta distribution is not suitable for modeling variable 1 and variable 2 data (Appendix 2). The distribution of the beta mixture of one variable can be tried to match the data case, because there are more than one mode.

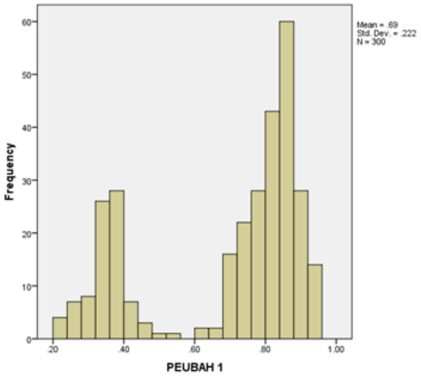

Figure 1. Variable 1 Histogram Data

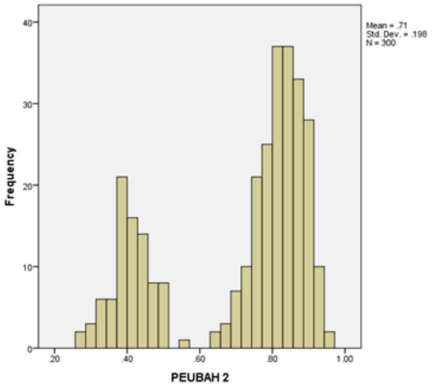

Figure 2. Variable 2 Histogram Data

Figure 3 displays the scatter diagram for simulation data. Based on Figure 3, it can be seen that the scattering of data forms as many as 2. groups. Therefore, for simulation data, we can try to match the distribution of the beta mixture of two variables.

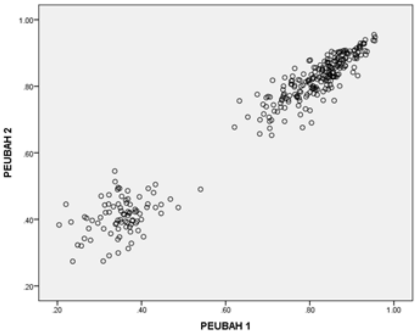

Figure 3. Variable 1 and Variable 2 Scattered Data

Table 2. Variable 1 Data Modeling Result

Number of Cluster	ICL-BIC Value	AIC Value
1	-162.2994	-169.7070
2	-405.2062	-429.0855
3	NaN	NaN
4	-21.7745	-246.3196

Table 2 shows the results of modeling for data on variable 1. The Matlab program for modeling is presented in

Appendix 1. It can be seen that the best model for data variable 1 is the mixed beta model of one variable with the number of groups is 2, because it has ICL-BIC values as well as the smallest AIC. This is consistent with the histogram data description. The estimated parameters for the model are $\alpha_1 = 0.72$, $\alpha_2 = 0.28$, $a_1 = 23.88$, $a_2 = 24.64$, $b_1 = 5.16$, and $b_2 = 45.61$. Figure 4 shows the concentration curve function of the beta mixed model probability for variable 1 data.

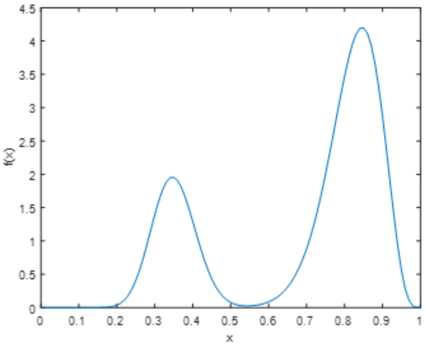

Figure 4. Concentration Function Curve of Beta Mixed Model Opportunity for Variable 1 Data

Table 3. Results of Data Modeling on Variable 2

Number of Cluster	ICL-BIC Value	AIC Value
1	-220.2678	-227.6754
2	-445.5817	-468.4450
3	-152.4188	-325.3160
4	-77.6270	-292.0483

Table 3 shows the results of modeling for data on variable 2. The Matlab program for modeling is presented in Appendix 1. It can be seen that the best model for two variable data is the mixed beta model of one variable with the number of groups is 2, because it has ICL-BIC values as well as the smallest AIC. This is consistent with the histogram data description. The estimated parameters for the model are $\alpha_1 = 0.28$, $\alpha_2 = 0.72$, $a_1 = 32.17$, $a_2 = 28.20$, $b_1 = 46.31$, and $b_2 = 5.90$. Figure 5 displays the density of the beta mix model function curve for the two variable data.

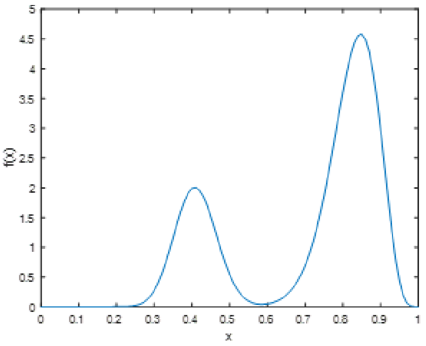

Figure 5. Curve Concentration Function Probabilities for Beta Mixed Models for Variable 2 Data

Table 4 displays the results of modeling for the data of two variables of the results of the simulation using the model developed in this paper. It can be seen that the best model for the two variables data is the mixed beta model of two variables with the number of groups is 2, because it has the smallest ICL-BIC value. This is consistent with the description of the two scatter data scatter diagrams. The estimated parameters for the model are presented in Table 5. Whereas the image of the opportunity density function curve is presented in Figure 6.

Table 4. Results of Data Analysis of Two Variables for Simulation Data

Number of Cluster	ICL-BIC Value
1	-741.0634
2	-1.544,8087
3	-1.319,6619
4	NaN
5	NaN
6	NaN

Table 5. Estimated Parameters of the Two Variable Beta Mixed Model for Simulation Data

j	$\hat{\alpha}_j$	$\hat{a}_j$	$\hat{b}_j$	$\hat{c}_j$
1	0.7167	28.8823	29.4343	6.1433
2	0.2833	24.0343	30.7501	44.2404

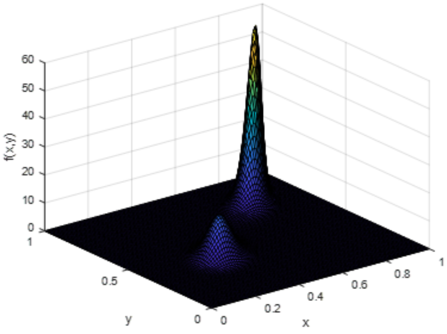

Figure 6. Curve Concentration Function Opportunity for Beta Mixed Model for Simulation Data

As a comparison, Table 6 presents the results of modeling simulation data using a model developed in [6] which assumes that each random variable is mutually independent. It can be seen that the best model is a mixed beta model with the number of groups is 2, because it has the smallest ICL-BIC value. The results are the same as data modeling using the two variable mix beta model developed in this paper However, the ICL-BIC value for the model developed in this paper is smaller than the ICL-BIC value for the model developed in [6]. Thus a suitable model for the two variables of simulation results is a mixed beta model developed in this paper. This happens because the correlation value is high in the amount of 0.9760. Table 6.3 presents the estimated parameters of the model, while the image of the density function curve is presented in Figure 7.

Table 6. Results of Two Variable Data Modeling for Simulation Data Using the Model [6]

Number of Cluster	ICL-BIC Value
1	-361,4910
2	-1.222,9046
3	-1.184,6463
4	-1.079,1587
5	NaN
6	NaN

Table 7. Estimated Parameters of [6]

Beta Mixture Model. For Simulation Data					
j	$\widehat{\alpha}_j$	$\hat{a}_{1j}$	$\hat{a}_{2j}$	$\hat{b}_{1j}$	$\hat{b}_{2j}$
1	0.2833	23.2987	32.0867	42.8686	46.1708
2	0.7167	24.5835	28.3154	5.2859	5.9240

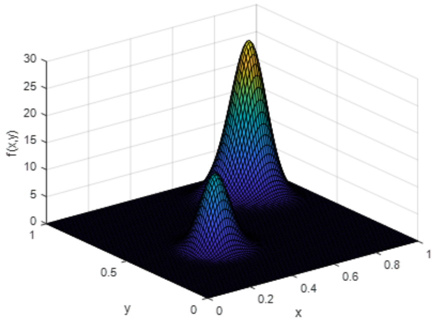

Figure 7. Curve Concentration Function Opportunity of Beta 's Mixed Model for Simulation Data

6.4 Conclusions

The suitable model for data proportion of variables 1 and variable 2 of the simulation results respectively is the beta mixture model of one variable with the number of groups is 2. The suitable model for the data of two variables is the proportion of simulation results is the beta mixture model developed in this paper, with many the group is 2, because the correlation value is large.

References

[1] K. Blekas, A. Likas, N. P. Galatsanos, and I. E. Lagaris, "A spatially constrained mixture model for image segmentation," *IEEE transactions on Neural Networks*, vol. 16, no. 2, pp. 494–498, 2005.

[2] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, vol. 2. IEEE, 1999, pp. 246–252.

[3] M. Revow, C. K. Williams, and G. E. Hinton, "Using generative models for handwritten digit recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 18, no. 6, pp. 592–606, 1996.

[4] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning*, vol. 42, no. 1, pp. 177–196, 2001.

[5] C. C. Aggarwal and C. R. D. Clustering, "Algorithms and applications," 2014.

[6] T. P. Sahu, N. K. Nagwani, and S. Verma, "Multivariate beta mixture model for automatic identification of topical authoritative users in community question answering sites," *IEEE Access*, vol. 4, pp. 5343–5355, 2016.