Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2021, 2021:26

https://doi.org/10.28919/cmbn/5390

ISSN: 2052-2541

# BIVARIATE BETA MIXTURE MODEL WITH CORRELATIONS

NURVITA TRIANASARI[1,2,*], I MADE SUMERTAJAYA[3], ERFIANI[3], I WAYAN MANGKU[4]

[1]Department of Statistics, Student of Doctoral Program at Sekolah Pasca Sarjana-IPB, Bogor, Indonesia

[2]Telkom of Economics and Business School, Telkom University, Bandung, Indonesia

[3]Department of Statistics, IPB University, Bogor, Indonesia

[4]Department of Mathematics, IPB University, Bogor, Indonesia



**Abstract**: The method of clustering is a probabilistic model based on clustering technique. The clustering method is often based on the assumption that data comes from a mixed model. One such mixture model is the beta mixed model. This mixed model can be used for the case of one variable or multiple variables. However, for the mixed beta model of the double variable, each variable is assumed to be independent. In this article, we propose a mixed beta model with correlated variables. The parameter estimation method uses the MLE method via the EM algorithm. While determining the optimal number of clusters using the ICL-BIC criteria. Monte Carlo simulation is used to see the performance of the model.

**Keywords**: probabilistic clustering; MLE method; EM algorithm; ICL-BIC; Monte Carlo simulation.

**2010 AMS Subject Classification:** 68T10.

## 1. INTRODUCTION

Cluster analysis is a double variable analysis which aims to group objects or data so that objects or data that are in the same cluster have relatively homogeneous properties than objects or data that are in different clusters (Johnson et al. 2007; Zickmund et al. 2010).

The concept of forming groups is a hierarchical method, a non-hierarchical method and a probability clustering method. The probability clustering method is a probabilistic model based

---

*Corresponding author

E-mail address: nurvitatrianasari@telkomuniversity.ac.id

Received January 4, 2021

clustering technique which assumes that the data follows a certain distribution. Hordering methods have the opportunity to be widely used in various applications such as market segmentation, image segmentation (Blekas et al. 2005 and Stauffer et al. 1999), handwriting recognition (Revow et al. 1996), and document clustering (Hoffman 2001). The clustering method has the opportunity to try to optimize the compatibility between the data observed with a mathematical model using a probabilistic approach. The method is often based on the assumption that the data comes from a mixture (mixture) of the distribution of opportunities, for example Poisson, beta, normal, lognormal, and Erlang. Thus the clustering problem is transformed into the parameter estimation problem because the data is modeled by a mixed distribution of the cluster. Data points that have the same distribution can be defined as groups. A mixed model with too many clusters might overfit with data, whereas a mixed model with too few clusters is not flexible enough to approach the real model.

Sahu et al. (2016) discusses the mixed model of beta double variable with the estimated parameters using the EM algorithm and the determination of the optimal cluster using the ICL-BIC deterministic method (integrated classification likelihood Bayesian information criterion). Sahu et al. (2016) assumes that there is no correlation between the variables. Until now there has been no research that discusses the mixed model beta two variables that involve correlations between variables. Olkin and Liu (2003) discuss the problem of beta formation of two variables based on the existence of correlations between variables while research related to the mixed beta model of two variables usually assumes no correlation between variables. Related to the results of research by Sahu et al. (2016) and Olkin and Liu (2003), we need another method in determining the optimal optimal number of groups in the mixed beta model of two variables by involving correlations between the variables. The purpose of this article is to discuss the mixed beta model of two variables involving correlation between variables by utilizing the results of research from Sahu et al. (2016) and Olkin and Liu (2003). Monte Carlo simulations will be used to evaluate the performance of the proposed method.

## 2. MIXED BETA TWO VARIABLE MODEL

The density function of opportunities for variables and those that follow the beta mixture distribution of two variables (Olkin and Liu, 2003) are

$$f_{X_1,X_2}(x_1,x_2)=\sum_{j=1}^{k}\alpha_j\frac{x_1^{a_j-1}x_2^{b_j-1}(1-x_1)^{b_j+c_j-1}(1-x_2)^{a_j+c_j-1}}{B(a_j,b_j,c_j)(1-x_1x_2)^{a_j+b_j+c_j}} \tag{1}$$

where $\alpha_j > 0, \alpha_1 + \alpha_2 + \cdots + \alpha_C = 1$ with $a_j > 0$, $b_j > 0$ and $c_j > 0$ for $j = 1, 2, \cdots, k$ are parameters of the beta mixture distribution of two variables, and

$$B\left(a_j, b_j, c_j\right) = \frac{\Gamma(a_j)\Gamma(b_j)\Gamma(c_j)}{\Gamma(a_j + b_j + c_j)}, \text{for } j = 1, 2, \cdots, k.$$

Figure 1 illustrates the opportunity density density curve for the distribution of beta mixes of two variables for various combinations of parameters. The expected value and variety of variables and can be obtained directly from the marginal distribution as follows:

$$E(X_1) = \sum_{j=1}^{k} w_j \frac{a_j}{a_j + c_j},$$

$$Var(X_1) = \sum_{j=1}^{k} w_j \left(\left(\frac{a_j}{a_j + c_j}\right)^2 + \frac{a_j c_j}{\left(a_j + c_j\right)^2 \left(a_j + c_j + 1\right)}\right) - \left(\sum_{j=1}^{k} w_j \frac{a_j}{a_j + c_j}\right)^2,$$

$$E(X_2) = \sum_{j=1}^{k} w_j \frac{b_j}{b_j + c_j}$$

$$Var(X_2) = \sum_{j=1}^{k} w_j \left(\left(\frac{b_j}{b_j + c_j}\right)^2 + \frac{b_j c_j}{\left(b_j + c_j\right)^2 \left(b_j + c_j + 1\right)}\right) - \left(\sum_{j=1}^{k} w_j \frac{b_j}{b_j + c_j}\right)^2,$$

the density function of the opportunity for the distribution of a beta mixture of two variables can be written in another form, namely:

$$f_{X_1,X_2}(x_1, x_2) = \sum_{j=1}^{k} w_j f_{X_1,X_2}^{j}(x_1, x_2)$$

with

$$f_{X_1,X_2}^{j}(x_1, x_2) = \sum_{i=0}^{\infty} d_j A(i) \frac{x_1^{a_j+i-1}(1 - x_1)^{b_j+c_j-1}}{B(a_j + i, b_j + c_j)} \frac{x_2^{b_j+i-1}(1 - x_2)^{a_j+c_i-1}}{B(b_j + i, a_j + c_j}$$

The form of the opportunity density function above can be used to calculate the expected value of two variables $X_1$ and $X_2$, i.e.

$$E\left(X_1^k X_2^l\right) = \int_0^1 \int_0^1 x^k y^l \sum_{j=1}^{k} w_j f_{X,Y}^{j}(x, y) dx dy$$

$$E\left(X_1^k X_2^l\right) = \sum_{j=1}^{k} w_j \int_0^1 \int_0^1 x^k y^l \sum_{j=1}^{k} w_j f_{X,Y}^{j}(x, y) dx dy$$

$$= \sum_{j=1}^{k} w_j \; {}_3F_2\left(a_j + k, b_j + l, s_j; s_j + k, s_j + l; 1\right) \tag{2}$$

where $s_j = a_j + b_j + c_j$

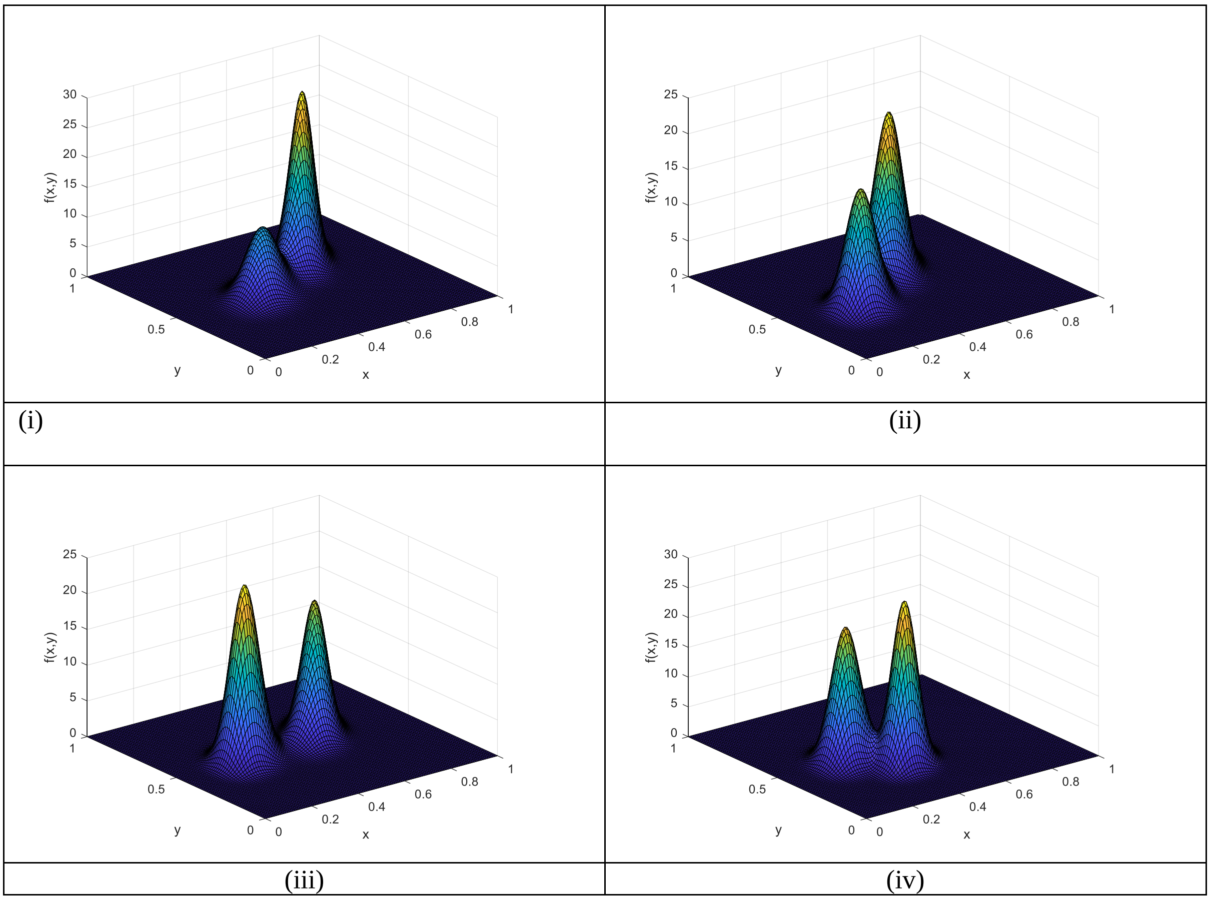


Figure 1. Density Function for Beta Distribution of Two Variable Variants for different value of parameter Combinations: (i) $w_1 = 0{,}35$, $w_2 = 0{,}65$, $a_1 = 15$, $a_2 = 30$, $b_1 = 25$, and $b_2 = 30$, $c_1 = 25$, and $c_2 = 15$; (ii) $w_1 = 0{,}4$, $w_2 = 0{,}6$, $a_1 = 15$, $a_2 = 25$, $b_1 = 25$, and $b_2 = 30$, $c_1 = 35$, and $c_2 = 20$; (iii) $w_1 = 0{,}45$, $w_2 = 0{,}55$, $a_1 = 35$, $a_2 = 15$, $b_1 = 25$, and $b_2 = 35$, $c_1 = 20$, and $c_2 = 35$; (iv) $w_1 = 0{,}5$, $w_2 = 0{,}5$, $a_1 = 35$, $a_2 = 15$, $b_1 = 25$, and $b_2 = 35$, $c_1 = 45$, and $c_2 = 35$

## 3. MLE FOR THE BVARIATE BETA MIXTURE MODEL PARAMETER

In the following, we will look for estimators of the beta distribution parameters of two variables using the maximum likelihood method. Suppose a random sample is sized $n$, i.e. $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from the distribution of a beta mixture of two Olkin-Liu variables such as Equation (1). The realization of the random sample is $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The likelihood function for the random sample is (Olkin and Liu, 2003)

$$L_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\alpha}) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{k} \alpha_j \frac{{x_{1i}}^{a_j-1} {x_{2i}}^{b_j-1} (1 - x_{1i})^{b_j+c_j-1} (1 - x_{2i})^{a_j+c_j-1}}{B(a_j, b_j, c_j)(1 - x_{1i}x_{2i})^{a_j+b_j+c_j}} \right\}$$

with $\boldsymbol{a} = (a_1, a_2, \dots, a_k)$, $\boldsymbol{b} = (b_1, b_2, \dots, b_k)$, $\boldsymbol{c} = (c_1, c_2, \dots, c_k)$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_k)$.

The possibility function is:

$$l_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{k} \alpha_j \frac{{x_{1i}}^{a_j-1} {x_{2i}}^{b_j-1} (1-x_{1i})^{b_j+c_j-1} (1-x_{2i})^{a_j+c_j-1}}{B(a_j, b_j, c_j)(1-x_{1i}x_{2i})^{a_j+b_j+c_j}} \right\}$$

The log-possibility function is

$$l_1(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{k} \alpha_j \frac{{x_{1i}}^{a_j-1} {x_{2i}}^{b_j-1} (1-x_{1i})^{b_j+c_j-1} (1-x_{2i})^{a_j+c_j-1}}{B(a_j, b_j, c_j)(1-x_{1i}x_{2i})^{a_j+b_j+c_j}} \right\}$$

The log-possibility function above is difficult to maximize because it contains a logarithm of the sum. One way to overcome the above problem is to use the EM algorithm. For example $\boldsymbol{Z} = \left(Z_{ij}; i = 1,2, \dots, n_i, j = 1,2, \dots, k\right)$, which is the latent variable that determines the group with observations originating,

$$Z_{ij} = \begin{cases} 1\ ; \text{observations } (x_{1i}, x_{2i}) \text{ derived from distribution } f_j \\ 0\ ; \text{others} \end{cases}$$

The possibility function is:

$$L_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} \ln \left( \alpha_j \frac{{x_{1i}}^{a_j-1} {x_{2i}}^{b_j-1} (1-x_{1i})^{b_j+c_j-1} (1-x_{2i})^{a_j+c_j-1}}{B(a_j, b_j, c_j)(1-x_{1i}x_{2i})^{a_j+b_j+c_j}} \right)^{Z_{ij}}$$

The log-possibility function is:

$$l_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} \ln \left( \alpha_j \frac{{x_{1i}}^{a_j-1} {x_{2i}}^{b_j-1} (1-x_{1i})^{b_j+c_j-1} (1-x_{2i})^{a_j+c_j-1}}{B(a_j, b_j, c_j)(1-x_{1i}x_{2i})^{a_j+b_j+c_j}} \right). \quad (3)$$

**Stage E (expectation stage)**

Substitute $Z_j$ in Equation (3) to be $E\left(Z_{ij}\right) = T_{ij}$, i.e.

$$E[l_2(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\alpha})] = \sum_{i=1}^{n} \sum_{j=1}^{k} T_{ij} \ln \left( \alpha_j \frac{{x_{1i}}^{a_j-1} {x_{2i}}^{b_j-1} (1-x_{1i})^{b_j+c_j-1} (1-x_{2i})^{a_j+c_j-1}}{B(a_j, b_j, c_j)(1-x_{1i}x_{2i})^{a_j+b_j+c_j}} \right). \quad (4)$$

with $T_{ij}$ is

$$T_{ij} = P\left(Z_{ij} = 1 \middle| (X_i, Y_j) = (x_i, y_j)\ ; a, b, c, \alpha\right)$$

$$= \frac{\alpha_j \dfrac{{x_{1i}}^{a_j-1} {x_{2i}}^{b_j-1} (1-x_{1i})^{b_j+c_j-1} (1-x_{2i})^{a_j+c_j-1}}{B\left(a_j, b_j, c_j\right)(1-x_{1i}x_{2i})^{a_j+b_j+c_j}}}{\sum_{l=1}^{k} \alpha_l \dfrac{{x_{1i}}^{a_l-1} {x_{2i}}^{b_l-1} (1-x_{1i})^{b_l+c_l-1} (1-x_{2i})^{a_l+c_l-1}}{B(a_l, b_l, c_l)(1-x_{1i}x_{2i})^{a_l+b_l+c_l}}}$$

**Stage M (maximization stage)**

Maximize Equation (4) to estimate $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\alpha}$

$$\frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial \alpha_j} = 0,$$

will get an estimate for the parameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_k)$, i.e.

$$\widehat{w}_j = \frac{1}{n}\sum_{\Gamma=1}^{n} T_{ij} \quad ; \quad j = 1, 2, \dots k.$$

Whereas the estimated parameters $\boldsymbol{a} = (a_1, a_2, \dots, a_k)$, $\boldsymbol{b} = (b_1, b_2, \dots, b_k)$, dan $\boldsymbol{c} = (c_1, c_2, \dots, c_k)$ is the solution of the following equations

$$\frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial a_j} = 0; j = 1,2, \dots k,$$

$$\frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial b_j} = 0; j = 1,2, \dots k,$$

$$\frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial c_j} = 0; j = 1,2, \dots k,$$

with

$$\frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial a_j} = \sum_{\Gamma=1}^{n} T_{ij}\left[\ln x_{1i} + \ln(1 - x_{2i}) - \ln(1 - x_{1i}x_{2i}) + \frac{\Gamma'(a_j + b_j + c_j)}{\Gamma(\mathrm{a}_j + b_j + c_j)} - \frac{\Gamma'(b_j)}{\Gamma(b_j)}\right]$$

$$\frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial b_j} = \sum_{\Gamma=1}^{n} T_{ij}\left[\ln x_{2i} + \ln(1 - x_{1i}) - \ln(1 - x_{1i}x_{2i}) + \frac{\Gamma'(a_j + b_j + c_j)}{\Gamma(\mathrm{a}_j + b_j + c_j)} - \frac{\Gamma'(a_j)}{\Gamma(a_j)}\right],$$

$$\frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial c_j} = \sum_{\Gamma=1}^{n} T_{ij}\left[\ln(1 - x_{1i}) + \ln(1 - x_{2i}) - \ln(1 - x_{1i}x_{2i}) + \frac{\Gamma'(a_j + b_j + c_j)}{\Gamma(\mathrm{a}_j + b_j + c_j)} - \frac{\Gamma'(b_j)}{\Gamma(b_j)}\right].$$

There is no analytical solution for the alleged parameters $\boldsymbol{a} = (a_1, a_2, \dots, a_k)$, $\boldsymbol{b} = (b_1, b_2, \dots, b_k)$, and $\boldsymbol{c} = (c_1, c_2, \dots, c_k)$. Numerical solutions using the Newton-Raphson iteration method can be used to obtain the expected parameters $\boldsymbol{a} = (a_1, a_2, \dots, a_k)$, $\boldsymbol{b} = (b_1, b_2, \dots, b_k)$, and $\boldsymbol{c} = (c_1, c_2, \dots, c_k)$. Iteration equation to get the estimated parameters $\boldsymbol{a} = (a_1, a_2, \dots, a_k)$, $\boldsymbol{b} = (b_1, b_2, \dots, b_k)$, and $\boldsymbol{c} = (c_1, c_2, \dots, c_k)$ is

$$\begin{pmatrix} \boldsymbol{a}^{(k+1)} \\ \boldsymbol{b}^{(k+1)} \\ \boldsymbol{c}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}^{(k)} \\ \boldsymbol{b}^{(k)} \\ \boldsymbol{c}^{(k)} \end{pmatrix} - \begin{pmatrix} \frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial a_j} \\ \frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial b_j} \\ \frac{\partial E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial c_j} \end{pmatrix} \times \begin{pmatrix} \frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{(\partial a_j)^2} & \frac{\partial^2 E[\mathrm{l}_2(a,b,c,\alpha)]}{\partial a_j \partial b_j} & \frac{\partial^2 E[\mathrm{l}_2(a,b,c,\alpha)]}{\partial a_j \partial c_j} \\ \frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial b_j \partial a_j} & \frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{(\partial b_j)^2} & \frac{\partial^2 E[\mathrm{l}_2(a,b,c,w)]}{\partial b_j \partial c_j} \\ \frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial c_j \partial a_j} & \frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial c_j \partial b_j} & \frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{(\partial c_j)^2} \end{pmatrix}^{-1} \quad (5)$$

with

$$\frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{(\partial a_j)^2} = \sum_{i=1}^{n} T_{ij}\left[\Psi'\left(a_j,b_j,c_j\right) - \Psi'(a_j)\right],$$

$$\frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{(\partial b_j)^2} = \sum_{i=1}^{n} T_{ij}\left[\Psi'\left(a_j,b_j,c_j\right) - \Psi'(b_j)\right],$$

$$\frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{(\partial c_j)^2} = \sum_{i=1}^{n} T_{ij}\left[\Psi'\left(a_j,b_j,c_j\right) - \Psi'(c_j)\right],$$

$$\frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial b_j \partial a_j} = \frac{\partial^2 E[\mathrm{l}_2(a,b,c,\alpha)]}{\partial a_j \partial b_j} = \sum_{i=1}^{n} T_{ij}\left[\Psi'\left(a_j+b_j+c_j\right)\right],$$

$$\frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial c_j \partial a_j} = \frac{\partial^2 E[\mathrm{l}_2(a,b,c,\alpha)]}{\partial a_j \partial c_j} = \sum_{i=1}^{n} T_{ij}\left[\Psi'\left(a_j+b_j+c_j\right)\right],$$

$$\frac{\partial^2 E[l_2(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{\alpha})]}{\partial c_j \partial b_j} = \frac{\partial^2 E[\mathrm{l}_2(a,b,c,\alpha)]}{\partial b_j \partial c_j} = \sum_{i=1}^{n} T_{ij}\left[\Psi'\left(a_j+b_j+c_j\right)\right].$$

The initial values for the Newton-Raphson iteration process above use the moment estimation results for the parameters of the beta distribution, i.e.

$$\hat{a}_j^{(0)} = \bar{x}_j\left(\frac{\bar{x}_j\left(1-\bar{x}_j\right)}{S_{Xj}^2} - 1\right)$$

$$\hat{b}_j^{(0)} = \bar{y}_j\left(\frac{\bar{y}_j\left(1-\bar{y}_j\right)}{S_{Yj}^2} - 1\right)$$

$$\hat{c}_j^{(0)} = \frac{\left(1-\bar{x}_j\right)\left(\frac{\bar{x}_j\left(1-\bar{x}_j\right)}{S_{Xj}^2} - 1\right) + \left(1-\bar{y}_j\right)\left(\frac{\bar{y}_j\left(1-\bar{y}_j\right)}{S_{Yj}^2} - 1\right)}{2}. \quad (6)$$

where $\bar{x}_j$ and $S_{Xj}^2$ each state the mean and variety of examples for the $X$ variable of the $j$ cluster. Whereas $\bar{y}_j$ and $S_{Yj}^2$ each state the mean and variety of examples for the $Y$ variable of the $j$ cluster.

## 4. SIMULATION STUDY

In this section an evaluation of the performance of the mixed beta two variables using the Monte Carlo simulation will be discussed. The data that will be used to evaluate the distribution of beta mix two variables is simulation data generated from MATLAB software from two distribution cases. Case 1, the simulation data are generated from the beta mixture distribution of two variables proposed in this paper which involve correlations between variables. Case 2, simulation data generated from the distribution of beta mixture developed by Sahu et al. (2016) which does not involve correlation between variables.

**Case Simulation Data 1**

**Case 1** simulation data is obtained from the generation of data through MATLAB software from the beta mixture distribution of two variables with the number of groups 2 and sample sizes 100, 300 and 500. Pearson correlation between the two variables tried is 6, namely 2 low Pearson correlation values (correlation value between 0.1 to 0.2), 2 moderate Pearson correlation values (correlation values between 0.5 to 0.6) and 2 high Pearson correlation values (correlation values between 0.8 to 0.9). There are 6 combination parameters for the beta mixed model two variables that will be tried (Table 1).

Table 1. Combinations of Beta Mixed Model Parameters Two variables Case 1 for Monte Carlo Simulation

| No. | $a_1$ | $a_2$ | $b_1$ | $b_2$ | $c_1$ | $c_2$ | $w_1$ | $w_2$ | Pearson Correlation |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 35 | 25 | 25 | 35 | 30 | 35 | 0,35 | 0,65 | Low |
| 2 | 35 | 25 | 27 | 35 | 30 | 35 | 0,35 | 0,65 | Low |
| 3 | 35 | 15 | 25 | 35 | 20 | 35 | 0,35 | 0,65 | Moderate |
| 4 | 25 | 25 | 20 | 40 | 45 | 35 | 0,35 | 0,65 | Moderate |
| 5 | 15 | 30 | 25 | 30 | 25 | 16 | 0,35 | 0,65 | High |
| 6 | 15 | 30 | 25 | 30 | 35 | 16 | 0,35 | 0,65 | High |

Thus there are 18 possible data scenarios to be simulated. Each data scenario is generated 1,000 times with a certain sample size and a certain Pearson correlation coefficient value. All data scenarios are presented in Table 2.

Table 2. Case Simulation Data Scenarios 1

| Scenario Number | Parameter Combination Number | $n$ | Pearson Correlation |
|---|---|---|---|
| 1 | 1 | 100 | Low |
| 2 | 2 | 100 | Low |
| 3 | 3 | 100 | Moderate |
| 4 | 4 | 100 | Moderate |
| 5 | 5 | 100 | High |
| 6 | 6 | 100 | High |
| 7 | 1 | 300 | Low |
| 8 | 2 | 300 | Low |
| 9 | 3 | 300 | Moderate |
| 10 | 4 | 300 | Moderate |
| 11 | 5 | 300 | High |
| 12 | 6 | 300 | High |
| 13 | 1 | 500 | Low |
| 14 | 2 | 500 | Low |
| 15 | 3 | 500 | Moderate |
| 16 | 4 | 500 | Moderate |
| 17 | 5 | 500 | High |
| 18 | 6 | 500 | High |

**Case Simulation Data 2**

**Case 2** simulation data were obtained from the generation of data through MATLAB software from the beta mixture distribution of two variables with the number of groups 2 and sample sizes of 100, 300 and 500. Table 3 presents 12 combinations of parameters tested. The combination of parameters is made in such a way that clearly visible distance between the center of the group. Distances between center groups are categorized as close, medium and far. Because the sample size was tested there were 3, and the combination of parameters there were 12, overall, 36 data case scenarios were generated (see Table 4).

Table 3. Parameters of the Beta Variable Mixed Model Combination Case 2 for Monte Carlo Simulation

| No. | $a_{11}$ | $a_{12}$ | $a_{21}$ | $a_{22}$ | $b_{11}$ | $b_{12}$ | $b_{21}$ | $b_{22}$ | $\alpha_1$ | $\alpha_2$ | Distance Between Cluster Center |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 30 | 40 | 25 | 35 | 30 | 20 | 25 | 20 | 0,35 | 0,65 | Close |
| 2 | 25 | 40 | 25 | 35 | 40 | 30 | 35 | 20 | 0,35 | 0,65 | Medium |
| 3 | 20 | 40 | 20 | 35 | 40 | 20 | 35 | 20 | 0,35 | 0,65 | Long |
| 4 | 40 | 20 | 20 | 25 | 25 | 25 | 40 | 20 | 0,35 | 0,65 | Close |
| 5 | 40 | 20 | 20 | 30 | 25 | 25 | 40 | 20 | 0,35 | 0,65 | Medium |
| 6 | 40 | 20 | 20 | 35 | 20 | 35 | 40 | 20 | 0,35 | 0,65 | Long |
| 7 | 40 | 15 | 40 | 25 | 25 | 25 | 30 | 20 | 0,35 | 0,65 | Close |
| 8 | 40 | 15 | 40 | 25 | 20 | 25 | 30 | 20 | 0,35 | 0,65 | Medium |
| 9 | 40 | 15 | 40 | 25 | 15 | 25 | 30 | 20 | 0,35 | 0,65 | Long |
| 10 | 35 | 25 | 20 | 40 | 35 | 25 | 45 | 35 | 0,35 | 0,65 | Close |
| 11 | 35 | 25 | 20 | 40 | 35 | 25 | 50 | 30 | 0,35 | 0,65 | Medium |
| 12 | 35 | 25 | 20 | 40 | 35 | 25 | 50 | 20 | 0,35 | 0,65 | Long |

Table 4. Case 2 Simulation Data Scenario

| Scenario Number | Parameter Combination Number | $n$ | Pearson Correlation |
|---|---|---|---|
| 1 | 1 | 100 | Close |
| 2 | 2 | 100 | Medium |
| 3 | 3 | 100 | Long |
| 4 | 1 | 300 | Close |
| 5 | 2 | 300 | Medium |
| 6 | 3 | 300 | Long |
| 7 | 1 | 500 | Close |
| 8 | 2 | 500 | Medium |
| 9 | 3 | 500 | Long |
| 10 | 4 | 100 | Close |
| 11 | 5 | 100 | Medium |
| 12 | 6 | 100 | Long |
| 13 | 4 | 300 | Close |
| 14 | 5 | 300 | Medium |
| 15 | 6 | 300 | Long |
| 16 | 4 | 500 | Close |
| 17 | 5 | 500 | Medium |
| 18 | 6 | 500 | Long |
| 19 | 7 | 100 | Close |

| | | | |
|---|---|---|---|
| 20 | 8 | 100 | Medium |
| 21 | 9 | 100 | Long |
| 22 | 7 | 300 | Close |
| 23 | 8 | 300 | Medium |
| 24 | 9 | 300 | Long |
| 25 | 7 | 500 | Close |
| 26 | 8 | 500 | Medium |
| 27 | 9 | 500 | Long |
| 28 | 10 | 100 | Close |
| 29 | 11 | 100 | Medium |
| 30 | 12 | 100 | Long |
| 31 | 10 | 300 | Close |
| 32 | 11 | 300 | Medium |
| 33 | 12 | 300 | Long |
| 34 | 10 | 500 | Close |
| 35 | 11 | 500 | Medium |
| 36 | 12 | 500 | Long |

In this section, we will discuss the performance comparison results of the two-variable beta mixture model discussed in this paper with the beta mixture model Sahu et al. (2016) for data containing correlations. The comparison is done using a Monte Carlo simulation. The size of the comparison used is the percentage of accuracy of the number of groups of the results of each model. The results of the comparison are presented in Table 5.

Table 5. Comparison of Proposed Model Performance with Sahu et al Model, for Case 1

| Scenario Number | $n$ | Pearson Correlation | Percentage of Accuracy Number of Cluster | |
|---|---|---|---|---|
| | | | Proposed Model | Sahu *et al.* Model |
| 1 | 100 | Low | 91,0 | 0 |
| 2 | 100 | Low | 93,8 | 0 |
| 3 | 100 | Moderate | 100 | 98,5 |
| 4 | 100 | Moderate | 100 | 53,1 |
| 5 | 100 | High | 100 | 98,4 |
| 6 | 100 | High | 100 | 100 |
| 7 | 300 | Low | 98,3 | 0 |
| 8 | 300 | Low | 99,3 | 0 |
| 9 | 300 | Moderate | 100 | 100 |
| 10 | 300 | Moderate | 100 | 69,5 |
| 11 | 300 | High | 100 | 100 |
| 12 | 300 | High | 100 | 100 |
| 13 | 500 | Low | 99,7 | 0 |
| 14 | 500 | Low | 100 | 0 |
| 15 | 500 | Moderate | 100 | 100 |
| 16 | 500 | Moderate | 100 | 77,8 |
| 17 | 500 | High | 100 | 100 |
| 18 | 500 | High | 100 | 100 |

Based on Table 5 it can be seen that the percentage of accuracy of the number of groups for the proposed model in this paper is above 90%. While the percentage accuracy of the number of groups for the model of Sahu et al. (2016), at least 0%. This happens for data with low

Pearson correlation. Both the proposed model and the Sahu et al. (2016), the greater the Pearson correlation value, the greater the accuracy of the number of groups. Both the proposed model and the Sahu et al. (2016), the greater the sample size, the greater the accuracy of the number of groups. Based on the results of the comparison of the performance of the proposed model with the model of Sahu et al. (2016), which is in Table 5 shows that the percentage of the number of groups for the proposed model is greater than the percentage of the number of groups for the model of Sahu et al. (2016), except for high correlation measures. This shows that for correlated data, the proposed model is better to be used than the Sahu et al. (2016).

The performance of the method for estimating the parameters of a beta mix two-variable model involving correlations proposed in this paper is presented in Tables 6 and 7. Based on Table 4.4, it appears that the greater the sample size, the EM estimator values for the proposed model parameters in this paper are closer to the actual parameter values for all measures of Pearson correlation. This means that the larger the sample size, the more accurate the EM estimation method in estimating the proposed model parameters in this paper. Based on Table 4.5, it appears that the greater the sample size, the deviation of the EM estimator values for the proposed model parameters from the actual parameters is smaller for all measures of Pearson correlation. This means that the larger the sample size, the more precise the EM estimation method in estimating the parameters of the proposed model in this paper.

Table 6. Accuracy of EM Estimation Methods for Proposed Models

| Scenario Number | $n$ | Pearson Correlation | Accuracy for estimators | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\hat{\alpha}_1$ | $\hat{\alpha}_2$ | $\hat{b}_1$ | $\hat{b}_2$ | $\hat{c}_1$ | $\hat{c}_2$ | $\hat{w}_1$ | $\hat{w}_2$ |
| 1 | 100 | Low | 39.082 | 26.470 | 28.403 | 36.564 | 33.722 | 36.809 | 0.352 | 0.648 |
| 2 | 100 | Low | 38.849 | 26.522 | 31.124 | 35.947 | 34.103 | 36.244 | 0.348 | 0.652 |
| 3 | 100 | Moderate | 37.292 | 15.544 | 26.592 | 36.306 | 21.272 | 36.334 | 0.349 | 0.651 |
| 4 | 100 | Moderate | 26.653 | 26.134 | 21.292 | 41.832 | 48.094 | 36.536 | 0.350 | 0.650 |
| 5 | 100 | High | 16.311 | 31.350 | 27.226 | 31.354 | 27.322 | 16.721 | 0.351 | 0.649 |
| 6 | 100 | High | 15.931 | 31.111 | 26.592 | 31.145 | 37.224 | 16.604 | 0.351 | 0.649 |
| 7 | 300 | Low | 36.761 | 25.428 | 26.126 | 35.578 | 31.346 | 35.628 | 0.349 | 0.651 |
| 8 | 300 | Low | 36.734 | 25.476 | 28.322 | 35.634 | 31.465 | 35.659 | 0.350 | 0.650 |
| 9 | 300 | Moderate | 35.625 | 15.137 | 25.459 | 35.319 | 20.347 | 35.325 | 0.349 | 0.651 |
| 10 | 300 | Moderate | 25.550 | 25.438 | 20.408 | 40.740 | 46.001 | 35.633 | 0.351 | 0.649 |
| 11 | 300 | High | 15.311 | 30.442 | 25.534 | 30.446 | 25.565 | 16.216 | 0.350 | 0.650 |
| 12 | 300 | High | 15.341 | 30.198 | 25.554 | 30.191 | 35.830 | 16.110 | 0.350 | 0.650 |
| 13 | 500 | Low | 35.957 | 25.225 | 25.575 | 35.361 | 30.683 | 35.371 | 0.348 | 0.652 |
| 14 | 500 | Low | 35.914 | 25.268 | 27.600 | 35.443 | 30.688 | 35.423 | 0.351 | 0.649 |
| 15 | 500 | Moderate | 35.491 | 15.060 | 25.339 | 35.115 | 20.277 | 35.124 | 0.350 | 0.650 |
| 16 | 500 | Moderate | 25.342 | 25.176 | 20.259 | 40.312 | 45.569 | 35.252 | 0.349 | 0.651 |
| 17 | 500 | High | 15.249 | 30.361 | 25.445 | 30.344 | 25.457 | 16.183 | 0.349 | 0.651 |
| 18 | 500 | High | 15.160 | 30.144 | 25.266 | 30.163 | 35.385 | 16.071 | 0.350 | 0.650 |

Table 7. Precision of the EM Estimation Method for the Proposed Model

| Scenario Number | $n$ | Pearson Correlation | Precision for estimators | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\hat{\alpha}_1$ | $\hat{\alpha}_2$ | $\hat{b}_1$ | $\hat{b}_2$ | $\hat{c}_1$ | $\hat{c}_2$ | $\hat{w}_1$ | $\hat{w}_2$ |
| 1 | 100 | Low | 106.425 | 19.083 | 59.845 | 52.892 | 71.947 | 47.957 | 0.006 | 0.006 |
| 2 | 100 | Low | 113.029 | 22.460 | 78.953 | 56.716 | 81.761 | 53.402 | 0.007 | 0.007 |
| 3 | 100 | Moderate | 54.274 | 4.663 | 26.030 | 25.768 | 16.953 | 26.042 | 0.002 | 0.002 |
| 4 | 100 | Moderate | 32.393 | 14.291 | 18.977 | 38.962 | 108.509 | 27.236 | 0.003 | 0.003 |
| 5 | 100 | High | 14.143 | 27.699 | 40.320 | 26.263 | 45.234 | 6.775 | 0.003 | 0.003 |
| 6 | 100 | High | 10.602 | 18.467 | 29.657 | 18.486 | 60.142 | 5.128 | 0.002 | 0.002 |
| 7 | 300 | Low | 40.929 | 5.938 | 15.822 | 15.530 | 24.273 | 14.697 | 0.002 | 0.002 |
| 8 | 300 | Low | 42.123 | 6.020 | 20.584 | 16.918 | 25.198 | 15.927 | 0.003 | 0.003 |
| 9 | 300 | Moderate | 15.139 | 1.181 | 7.398 | 6.807 | 4.742 | 6.766 | 0.001 | 0.001 |
| 10 | 300 | Moderate | 9.619 | 4.468 | 5.465 | 12.585 | 32.143 | 8.935 | 0.001 | 0.001 |
| 11 | 300 | High | 3.301 | 7.233 | 10.096 | 6.866 | 11.263 | 1.838 | 0.001 | 0.001 |
| 12 | 300 | High | 2.802 | 5.493 | 7.767 | 5.371 | 16.294 | 1.553 | 0.001 | 0.001 |
| 13 | 500 | Low | 19.391 | 3.170 | 7.240 | 8.236 | 11.346 | 7.644 | 0.002 | 0.002 |
| 14 | 500 | Low | 21.752 | 3.185 | 8.973 | 8.530 | 11.952 | 8.140 | 0.002 | 0.002 |
| 15 | 500 | Moderate | 8.117 | 0.756 | 4.003 | 4.277 | 2.558 | 4.333 | 0.000 | 0.000 |
| 16 | 500 | Moderate | 5.541 | 2.723 | 3.243 | 7.511 | 18.305 | 5.196 | 0.000 | 0.000 |
| 17 | 500 | High | 1.918 | 4.057 | 5.870 | 3.868 | 6.444 | 0.990 | 0.001 | 0.001 |
| 18 | 500 | High | 1.574 | 3.313 | 4.572 | 3.255 | 9.205 | 0.886 | 0.001 | 0.001 |

Table 8. Comparison of Proposed Model Performance with Sahu et al. Model for Case 2

| Scenario Number | Parameter Combination Number | $n$ | Distance Between Cluster Center | Percentage of Accuracy Number of Cluster | |
|---|---|---|---|---|---|
| | | | | Model Usulan | Sahu *et al.* Model |
| 1 | 1 | 100 | Close | 12.1 | 1.6 |
| 2 | 2 | 100 | Medium | 86.4 | 100 |
| 3 | 3 | 100 | Long | 100 | 100 |
| 4 | 1 | 300 | Close | 25.3 | 40.3 |
| 5 | 2 | 300 | Medium | 97.9 | 100 |
| 6 | 3 | 300 | Long | 100 | 100 |
| 7 | 1 | 500 | Close | 17 | 49.7 |
| 8 | 2 | 500 | Medium | 99.7 | 100 |
| 9 | 3 | 500 | Long | 100 | 100 |
| 10 | 4 | 100 | Close | 100 | 96.8 |
| 11 | 5 | 100 | Medium | 100 | 99.9 |
| 12 | 6 | 100 | Long | 100 | 100 |
| 13 | 4 | 300 | Close | 100 | 100 |
| 14 | 5 | 300 | Medium | 100 | 100 |
| 15 | 6 | 300 | Long | 100 | 100 |
| 16 | 4 | 500 | Close | 100 | 100 |
| 17 | 5 | 500 | Medium | 100 | 100 |
| 18 | 6 | 500 | Long | 100 | 100 |
| 19 | 7 | 100 | Close | 100 | 1.6 |
| 20 | 8 | 100 | Medium | 100 | 37.8 |
| 21 | 9 | 100 | Long | 100 | 93.4 |
| 22 | 7 | 300 | Close | 100 | 1.1 |
| 23 | 8 | 300 | Medium | 100 | 70.9 |
| 24 | 9 | 300 | Long | 100 | 100 |
| 25 | 7 | 500 | Close | 100 | 0.2 |
| 26 | 8 | 500 | Medium | 100 | 90.2 |
| 27 | 9 | 500 | Long | 100 | 100 |
| 28 | 10 | 100 | Close | 100 | 20.8 |

| | | | | | |
|---|---|---|---|---|---|
| 29 | 11 | 100 | Medium | 100 | 94.2 |
| 30 | 12 | 100 | Long | 100 | 100 |
| 31 | 10 | 300 | Close | 100 | 46.8 |
| 32 | 11 | 300 | Medium | 100 | 100 |
| 33 | 12 | 300 | Long | 100 | 100 |
| 34 | 10 | 500 | Close | 100 | 55.8 |
| 35 | 11 | 500 | Medium | 100 | 100 |
| 36 | 12 | 500 | Long | 100 | 100 |

In this section, we will discuss the performance comparison results of the two-variable beta mixture model discussed in this paper with the beta mixture model Sahu et al. (2016) for data generated from the model of Sahu et al. (2016) which does not contain correlation. The size of the comparison used is the percentage of accuracy of the number of groups of the results of each model. The results of the comparison are presented in Table 8. It appears that both the proposed model and the Sahu et al. (2016), the farther the distance between the centers of the cluster, the greater the percentage of accuracy of the number of groups. In general, the larger the sample size, the greater the percentage of accuracy of the number of groups for each model. The results of the comparison show that overall it can be concluded that the mixed beta model of the two variables discussed in this paper is better than the beta mixture model Sahu et al. (2016) for the case of uncorrelated data except for the case of a combination of parameters 1, 2 and 3 for the distance between close and medium cluster centers.

Based on the results of comparisons for data case 1 and data case 2 it can be concluded that in general the mixed beta model of the two variables discussed in this paper is better than the mixed model beta Sahu et al. (2016) for the case of correlated data or non-correlated data.

## 5. CONCLUSIONS

In this article the mixed beta model of two variables has been discussed which involves the correlation between the variables by utilizing the results of previous studies. Estimation of parameters for the distribution of a beta mixture of multiple variables involving correlations between the variables discussed in this my paper uses the EM algorithm. Whereas the selection of the best distribution or model or in other words the determination of the optimal number of clusters uses the ICL-BIC criteria. Simulation results show that in general the two variable beta mixed model proposed in this article is better than the Sahu et al beta mixed model. (2016) for the case of correlated data or non-correlated data. The simulation results also show that the larger the sample size, the more accurate and more precise the EM estimation method in estimating the parameters of the proposed model in this article.

## CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

## REFERENCES

[1] C.C. Aggarwal, Data Classification: Algorithms and Applications, CRC Press, Boca Raton, London, New York, 2014.

[2] K. Blekas, A. Likas, N.P. Galatsanos, I.E. Lagaris, A Spatially Constrained Mixture Model for Image Segmentation, IEEE Trans. Neural Netw. 16 (2005), 494–498.

[3] J.F. Hair, et al. Multivariate Data Analysis: A Global Perspective. 7th ed. Upper Saddle River: Prentice Hall, 2009.

[4] T. Hofmann, Unsupervised Learning by Probabilistic Latent Semantic Analysis. Mach. Learn. 42 (2001), 177–196.

[5] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis. 6th Ed. Prentice-Hall, London, 2007.

[6] T.P. Sahu, N.K. Nagwani, S. Verma, Multivariate Beta Mixture Model for Automatic Identification of Topical Authoritative Users in Community Question Answering Sites, IEEE Access. 4 (2016), 5343–5355.

[7] C. Stauffer, 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), Fort Collins, CO, USA, 1999: pp. 246–252.

[8] I. Olkin, R. Liu, A bivariate beta distribution, Stat. Probab. Lett. 62 (2003), 407–412.

[9] M. Revow, C.K.I. Williams, G.E. Hinton, Using generative models for handwritten digit recognition, IEEE Trans. Pattern Anal. Machine Intell. 18 (1996), 592–606.

[10] W.G. Zikmund, Business research methods., South-Western Cengage Learning, Mason, OH, 2010.