Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 2333–2350
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5932

Using some methods to estimate the parameters of the Multivariate Skew Normal (MSN) distribution function with missing data

Qutaiba Nabeel Nayef Al-Qazaz[a,*], Lina Nidhal Shawkat[a]

[a]University of Baghdad ,College of Administration & Economics, Statistics Department, Iraq

(Communicated by Madjid Eshaghi Gordji)

Abstract

The estimation of statistical parameters for multivariate data can lead to wasted information if the missing values are neglected, which in return will lead to inaccurate estimates, therefore the incomplete data must be estimated using one of the statistical estimation methods to obtain accurate results and thus obtaining good estimates for the parameters.

Missing values is considered one of the most important problems that researchers encounter and the most common, and in the case of the multivariate skew normal distribution (MSN) the presence of this problem will lead to weak and misleading conclusions for the research, which calls for treating this problem and in return obtaining efficient and convincing results. The aim of this paper is to estimate the missing values for the multivariate skew normal distribution function using the K-nearest neighbors Imputation (KNN). After estimating the missing values, the parameters are estimated using Genetic Algorithm (GA), and the Bayesian Approach was also used to estimate the missing values and find the estimates for the parameters. Using simulation, the Mean Squared Error (MSE) was calculated to find out which method is the best for estimation by comparing the two methods using different sample sizes (400, 600, and 800). The (GA) that is based on the (KNN) algorithm to estimate the missing values proved to be better and more efficient than the Bayesian Approach in terms of the results.

Keywords: Multivariate Skew Normal distribution (MSN), K-Nearest Neighbors Imputation (KNN), Genetic Algorithm (GA), Bayesian Approach, Mean Squared Error (MSE).

*Corresponding author
Email addresses: dr.qutaiba@coadec.uobaghdad.edu.iq (Qutaiba Nabeel Nayef Al-Qazaz), lola1990@yahoo.com (Lina Nidhal Shawkat)

Received: May 2021 *Accepted:* October 2021

1. Introduction

The skew normal distribution is a member of the family of distributions that contains the normal distribution, but the skew normal distribution has an additional parameter to regulate the skewness. The skewness is known as the irregularity in the statistical distribution, where the curve appears to be deformed or leaning towards the left or the right. The deviation can be determined to know the extent of the difference between this distribution and the normal distribution. In plot, the normal distribution appears to be a classic symmetrical bell-shaped curve and the mean, or the average, and the mode, or the upper limit for the point on the curve, are equal.

The occurrence of the missing values problem in the Multivariate Skew Normal distribution (MSN) is one of the most common problems when collecting the data and analysing it, and it means losing part of the sample data, like in an industrial experiment where some of the results might be missing because of mechanical malfunctions that are not related to the experimental operation, for example, participants in a family survey might refuse to report their income or refuse to answer some of the questions they're asked, or some of the data might be damaged or missing. Missing data is one of the major problems that researchers encounter, and the statistical mechanisms that are used to analyze the data assumes complete information about all the variables that are used in the analysis, and not addressing this problem in the proper manor might cause some problems for the researcher like not estimating the variance correctly, or getting biased results, therefore the missing values must be estimated using some statistical methods. From here the missing data will be estimated using K-Nearest Neighbor (KNN) imputation and Bayesian approach.

There are many researchers touched on this subject, In (2008), (Hmoud, Abbas and Nayef) [4] studied some methods to estimate multivariate (p. d. f.). The focus was on a bivariate normal distribution and also a polluted bivariate normal distribution. Three estimators were compared (one parametric and two nonparametric) using simulation. In (2016), (Nayef and Ayub) [3] worked on data mining by using (K-Means) algorithm to classify the data, and it was noticed that changing the clustering as well as the number of clusters required can change the condition of the experiment to reach the optimum clustering that allows for a new dataset that answers all the questions. In (2020), (Shawkat and Nayef) [13] estimated the missing values for a multivariate skew normal model using (ECM) algorithm and (KNN) method as well as using (MLE) method and (N-R) algorithm to estimate the parameters of the model. Simulation was used to compare between (ECM) and (KNN) and it was found that with samples of size (400) or less, the (KNN) was better, while with larger sample sizes, the (ECM) was better, and finally with samples of size (800) or more, both methods were reasonably good.

In this paper, missing data mechanisms will be discussed, depending on the Multivariate Skew Normal (MSN) distribution function, which was introduced for the first time by [6] (*Azzalini and Dalla Valle*) in (1996), which is considered one of the important distributions with three parameters, and they are: position parameter (ξ), scale parameter (Σ), skewness parameter (shape parameter) (Λ).

2. Missing Data Mechanisms

Methods that specialize in analyzing data with missing values are different in their hypotheses around the mechanisms that lead to missing data. Understanding the mechanism and determining its nature is very important for choosing the proper method for analysis, the missing mechanisms can be divided into the following [2]:

1. Missing Completely At Random (MCAR):

Missing Completely At Random (MCAR) data happens when the reason for missing is independent from the missing value itself and independent from the values of other variables, and this case is very rare.

2. Missing At Random (MAR):

Missing At Random (MAR) data happens when the reason for missing is independent from the missing value itself and could be correlated to the values of another variable, and this case is common and easy to deal with.

3. Missing Not At Random (MNAR):

Missing Not At Random (MNAR) data happens when the reason for missing is caused by the missing value itself, (which means the missing value is correlated to the other values for the same variable), and this case is difficult to deal with.

The missing mechanism can be expressed in a mathematical formula using its own distribution that was suggested in (1976) by (*Rubin*) which is represented by the conditional distribution for $(X|R)$ with unknown parameters $(\underline{\theta})$ [1]:

$$P\left(R|\ X, \underline{\theta}\right)$$

Where:

X: Real data matrix of order $(n\ x\ p)$.

R: Binary matrix that takes the values (0,1), and it is called (Missing data indicator matrix).

And:

$$r_{ij} = \begin{cases} 1 & if\ X_{ij}\ is\ obs \\ 0 & if X_{ij}\ is\ mis \end{cases}$$

And let us assume that:

$$P\left(R|X, \underline{\theta}\right) = P\left(R|\ \underline{\theta}\right) \qquad for\ all\ X^m \tag{2.1}$$

Then the data is missing completely at random $(MCAR)$.

From (2.1) it appears that the distribution does not depend on the observed values (X^O) or on the missing values (X^m).

But if:

$$P\left(R|X, \underline{\theta}\right) = P\left(R|X^O, \underline{\theta}\right) \qquad for\ all\ X^m \tag{2.2}$$

Then the data is missing at random (MAR).

From the above it appears that the distribution depends on the observed values (X^O) but does not depend on the missing values (X^m).

And if:

$$P\left(R|X, \underline{\theta}\right) = P\left(R|X^m, \underline{\theta}\right) \qquad for\ all\ X^m \tag{2.3}$$

Then the data is missing not at random (MNAR).

From (2.3) it can be noticed that the distribution depends on the missing values (X^m).

The distribution of the missing mechanism must be taken into account when analyzing this type of data, but it can also be neglected in the cases of (MAR) and (MCAR).

3. The Multivariate Skew Normal Distribution (MSN)

Let the random vector ($\underline{X}$) follows the distribution (MSN) for (p) variables [11] with position vector ($\underline{\xi} \in \mathbb{R}^p$) (a Euclidean space vector that consists of the real numbers for every (p) rows), (Σ) is the variance-covariance matrix of order (pxp), and ($\Lambda = Diag(\lambda)$) is the skewness matrix, and ($\underline{\lambda}= (\lambda_1, \ldots, \lambda_p)'$), if its probability density function (pdf) is [9]:

$$f(x \mid \underline{\xi}, \Sigma, \Lambda) = 2^p \phi_p(x \mid \underline{\xi}, \Omega)\, \Phi_p(\Lambda\, \Omega^{-1}(x-\underline{\xi}) \mid \Delta) \tag{3.1}$$

ξ: is the position vector consisting of (p) variables which means ($\underline{\xi}=(\xi_1, \ldots, \xi_p)'$).
And the variance-covariance matrix (Σ) can be expressed as follows:

$$\Sigma= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}$$

$\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp}$: represents the variances.
$\sigma_{12}, \sigma_{13}, \ldots, \sigma_{1p}, \sigma_{2p}, \ldots, \sigma_{pp}$: represents the covariances.

$$\Omega=\Sigma+\Lambda^2$$

Ω: is a matrix that represents the sum of the variance-covariance matrix (Σ) and the skewness matrix (Λ^2) of order ($p\times p$).

$$\Delta=(I_p+\Lambda\, \Sigma^{-1}\Lambda)^{-1}=I_p-\Lambda\, \Omega^{-1}\Lambda$$

I_p: is the identity matrix for (p) dimensions.
And ($\phi_p(.| \underline{\mu}, \Sigma)$) is the (pdf) for the normal distribution ($N_p(\underline{\mu}, \Sigma)$), and ($\Phi_p(.| \Sigma)$) is the cumulative distribution function (cdf) for the normal distribution ($N_p(0, \Sigma)$). And ($\underline{X} \sim SN_p(\underline{\xi}, \Sigma, \Lambda)$) can also be written for the function as if ($\underline{X}$) has a density function, equation (3.1).

4. Models of Multivariate Skew Normal Distribution with Missing Information

Let ($\underline{X}= (X_1, \ldots, X_n)$) be a random sample of size (n) where each ($\underline{X}_j$) is taken from ($SN_p(\underline{\xi}, \Sigma, \Lambda)$) where ($j = 1, \ldots, n$) and ($n > p$) and to analyze the (MSN) model for a group of data that have general missing patterns, meaning the observations for each (X_i) might not be completely observed, ($i = 1, \ldots, p$).
And to make the estimation equations for multivariate data which allows for missing data, and on that base we can divide ($\underline{X}_{j(p\times 1)}$) into two components $\left(\underline{X}_j^o, \underline{X}_j^m\right)$, where ($\underline{X}^o_{j\left(p_j^o\times 1\right)}$) is the observed component, and ($\underline{X}^m_{j((p-p_j^o)\times 1)}$) is the missing component, in addition to that we have two matrices in the paper (M_j, O_j), that corresponds to ($\underline{X}_j$) where ($\underline{X}_j^o=O_j\underline{X}_j$), and ($\underline{X}_j^m=M_j\underline{X}_j$), respectively, where ($O_j$) is a matrix that shows the observed values in ($\underline{X}$), and (M_j) shows the missing values in ($\underline{X}$).
More specifically, ($O_{j(p_j^o\times p)}$) and ($M_{j((p-p_j^o)\times p)}$) are partial matrices extracted from the rows of the matrix (I_p) that corresponds to the row positions of ($\underline{X}_j^o$) and ($\underline{X}_j^m$) in ($\underline{X}_j$), respectively. When ($\underline{X}_j = \underline{X}_j^o$), ($O_j=I_p$), and ($M_j$) is an idempotent matrix, the properties of these two matrices are [9]:
a) $\underline{X}_j=O_j'\, \underline{X}_j^o+M_j'\underline{X}_j^m$
b) $O_j'\, O_j + M_j'M_j = I_p$

5. K-Nearest Neighbors Imputation Method (KNN)

The (KNN) is a nonparametric classification method, and it is a simple but very effective method in many cases [8]. The (KNN) is considered a very successful method for multivariate standard data, and the idea is to use a distance scale to find the (K) most similar observations for a compound that has missing values, and replacing the missing values using the available information from the neighboring variables. The (KNN) method can be summarized into filling the missing value with the mean value of the column that corresponds to the nearest neighbor to the corresponding row that has no missing values, and the nearest neighbor can be determined using the Euclidian Distance [18]. The distance or the variances between the samples can be measured using the Euclidian distance, and it can be said that the nearest neighbor using the nearest (K) points to do the classification.
The following will illustrate Mean Imputation (MI) [7]:
This method is considered one of the most commonly used methods, it consists of substituting the missing data for a specific attribute (trait) with the mean for all the known values for this (attribute) in the class that contains the case with the missing traits. Considering the value (x_{ij}) in class (d), (C_d), is missing and will be replaced with:

$$\hat{x}_{ij} = \sum_{i:x_{ij}\in C_d} \frac{x_{ij}}{n_d} \tag{5.1}$$

Where (n_d) is the number of values that are not missing in attribute (j) in class (d).
The steps for (KNN) imputation algorithm are as follows [5]:

1. The data set (X) is divided into two parts. Let (X^m) be the data set that contains the missing values or at least one missing value, as for the second part, it is the part that doesn't have missing values in the attributes or observations, denoted by (X^o).

2. For every vector $(\underline{x_j})$ in (X^m):

3. The vector is divided into two parts, observed and missing

$$(\underline{x_j} = \left[\underline{x}_j^o;\ \underline{x}_j^m\right]).$$

4. The distance is measured between $(\underline{x}_j^o)$ and the other vectors in (X^o). These features are only used in vectors in the case of complete data set (X^o), that are noticed in vector $(\underline{x_j})$.

5. The closest data vectors to the vectors that has the missing values are used (K-Nearest Neighbors), and then conduct a majority estimation for the missing values of the categorical variables. For continuous variables, the missing values are replaced with the mean value of the variable for the K-nearest neighbor. The median can be used instead of the mean.

For more illustration on how to select the nearest neighbors, let us assume that there are three categories and see how the (KNN) will determine the categories for the experiment data [18].

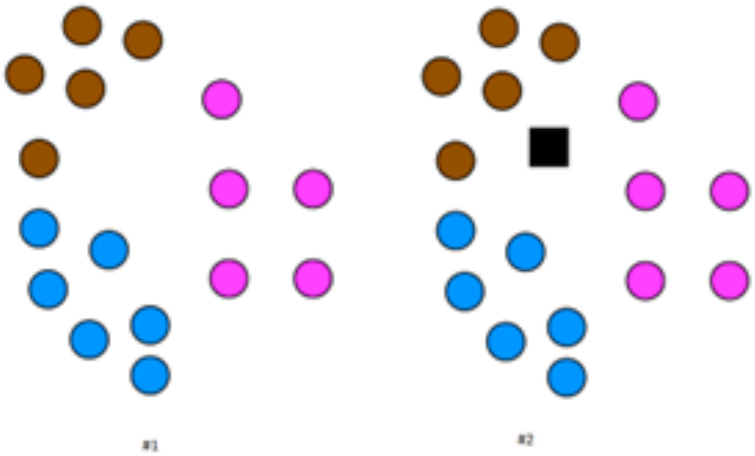

Figure 1: shows hypothetical data to illustrate the work of (KNN).

The black square in Figure 1 is the Data point, and to find the category that this data point belongs to, and with the help of (KNN), we determine the value of (K) which represents the number of the nearest neighbors to the data point, as shown below in Figure 2:

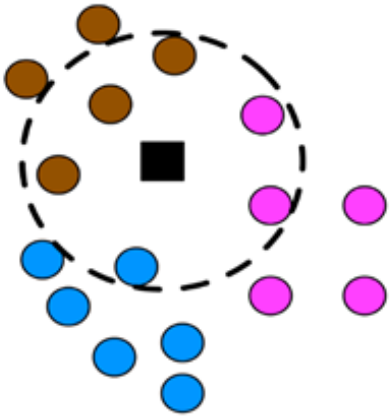

Figure 2: shows the way in which (KNN) chooses the nearest neighbors.

Finding the nearest neighbors is by using the distance scale which measures the distance between the data point and the set of its neighboring points, where the nearest neighbors for the data sets is called (K), then check the number of times the neighbors appear for each category, and in return choose the best according to the nearest sequence of neighbors as shown below in Figure 3.

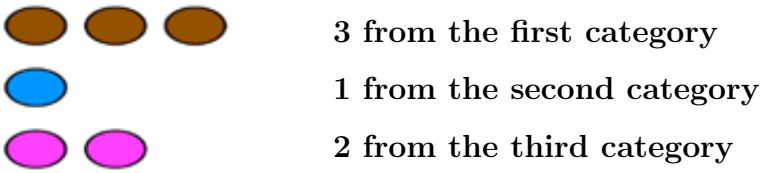

Figure 3: shows the number of times similar neighbors appear for each category.

From Figure 3, it is clear that the best category among the three for the nearest neighbors was the first category, and that is because the number of similarities for the same category appeared more than the other categories.

As for choosing the distance function, it is possible to use the Euclidean distance, or Manhattan distance, or it can be Mahalanobis, Pearson, ... etc. and depending on the type of the used data, we'll use the Euclidean distance in this paper for two variables. The Euclidean distance is one of the most used distance scales, and it is calculated as follows:

$$G\left(x_1\ ,\ x_2\right) = \sqrt{\sum_{i=1}^{n}\left(x_{1i} -\ x_{2i}\right)^2} \tag{5.2}$$

6. Genetic Algorithm (GA)

The Genetic Algorithm (GA) is basically one of the evolutionary ideas for natural selection and genetics, and it is a useful and effective random (stochastic) search method, inspired by Darwin's theory for survival of the fittest, and it is common in nature that in a contest where individuals search for resources, the most skilled individuals have the upper hand over the weak ones. Evolutionary computation today considers the genetic algorithm as one of the important methods of random search that are used to solve optimization problems, where the (GA) is a smart and easy to apply structure.

For any specific problem, the (GA) tries to solve it by simulations and using natural processes, like selection, crossover, and mutation to evolve a good solution for this problem. The (GA) starts with a randomly chosen population that consists of possible solutions and ending up with an optimal solution by iterative updates that copies the biological evolution mechanisms. In the (GA), the fittest individuals (solutions) dominate the weakest using these mechanisms [12].

- Operators of the Genetic Algorithms:

  The (GA) uses genetic operators to maintain genetic diversity. It is important to maintain genetic diversity or variety for the evolution process. The genetic operators are themselves inspired from the natural genetic structure. The following is the operators used in the (GA) [12]:

a) Reproduction \ Selection: usually it is the first operator to work on the population. We select chromosomes from the population to be the parents for the reproduction step and produce the offspring, according to Darwin's theory survival of the fittest, meaning only the best will stay alive and create a new dynasty. The reproduction operator is also called the selection operator because it basically works on extracting a subset of genes from the current population based on some criteria for quality. The Fitness Function is the measure of quality that can determine the best subset of genes, and each gene has a specific meaning, and we get it by transforming the Objective Function to a proper function for solving in the algorithm.

b) Crossover \ Reconstruction: this operator is called crossover because it combines two chromosomes to produce a new dynasty. Each chromosome consists of a set of genes which gets encoded using one of the encoding methods like the Binary Encoding which is the best solution, since it represents the chromosomes in the (GA) as a chain of pieces partitioned according to the length of the chromosome containing ones and zeros, or by using the Value Encoding where the chromosomes gets encoded as a form of letters or other symbols like real numbers, or by using Permutation Encoding which can be represented as a series of normal numbers ... etc. In this paper the binary encoding was used, and the Roulette Wheel Selection was used to select the crossover in the algorithm.

c) Mutation: the mutation happens during the evolution stage where the user decides the mutation probability, usually this probability is set to a small value to some degree, like (0.01) is the first good choice. Mutation is the genetic operator that is used to maintain the genetic diversity from one generation of chromosomes to the next generation.

- Steps of the Genetic Algorithm:

The steps of the (GA) can be summarized into [16]:

Step 1: Define the convergence criteria, objective function, search space (possible solution intervals), and the initial values for the (GA) parameters (like the population size (N), Elite Number (EN), Mutation Probability (MP), Crossover Probability (CP), and Selection Ratio (SR)).

Step 2: Generate the initial population that consists of (N) chromosomes in the search space using preparation strategy. The initial population is written in the form $\left\{\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_n^{(0)}\right\}$ and $(\underline{\theta}=\left\{\underline{\xi}, \Sigma, \Lambda\right\})$ in this study. The (GA) deals with a population of possible solutions, each solution is represented by a chromosome and a fitness function, given that the fitness function used in this paper is the likelihood function like in the following form:

$$\ell_c\ (\underline{\theta}|\underline{X}, \underline{\tau}\ ) = -\frac{1}{2}\sum_{j=1}^{n}\left\{log\,|\Sigma| + \left(\underline{X}_j - \underline{\xi} - \Lambda\underline{\tau}_j\right)'\Sigma^{-1}\left(\underline{X}_j - \underline{\xi} - \Lambda\underline{\tau}_j\right) + \underline{\tau}_j'\underline{\tau}_j\ \right\} \tag{6.1}$$

The chromosomes in the population are represented by $(\underline{\theta}_j)$, $(j = 1, \ldots, n)$, there are many preparation strategies in the literature, but in this paper, random generation strategy was used.

Step 3: The fitness function is calculated to find the fitness value for each chromosome in the population at any iteration (k), $(ln\ell_c(\underline{\theta}_j^{(k)}))$.

Step 4: At a predetermined selection rate, the solutions (individuals) that has the worst fitness values, and amid evaluation of the individuals at the previous step, are replaced with new individuals randomly generated. In addition to that, a certain number of individuals (EN), which have the best fitness values, are accepted as elite individuals and are transferred to the next generation without any update.

Step 5: By applying the roulette wheel method (based on the concept that there is a bigger chance for selection if there is a better fitness) as a proportional selection method, two candidate individuals are selected as parents, except for the elite individuals.

Step 6: Crossover and mutation are applied, as disruption mechanisms, to nominate individuals according to (CP) and (MP). Parents crossover is applied to acquire new offspring and mutate new individuals. Thus generation $(k+1)$ is acquired and it is denoted by $\left\{\theta_1^{(k+1)}, \theta_2^{(k+1)}, \ldots, \theta_N^{(k+1)}\right\}$.

Step 7: Finally, make $(k = k + 1)$ and continue to iterate with the fitness evaluation step until achieving the convergence criteria. When the evolution stops, the solution with the best fitness value is the optimum solution. The values for the best solution are denoted by $\left\{\widehat{\underline{\xi}}, \widehat{\Sigma}, \widehat{\Lambda}\right\}$ and they are called the parameter estimates.

7. Concept of the Bayesian Theorem

Considering the importance of statistics and its use in different fields, in addition to the wide interest of researchers, had a big effect in its development, which led to the emerging of many statistical schools, one of these schools is the Bayesian school and the classical school. The Bayesian school is considered one of the most important schools in statistics, and it was named after its founder (Thomas Bayes) in (1763), and it was met with wide interest by science students and researchers. The Bayesian school states that the available information from the observations is adopted, but it is considered insufficient to complete all the statistical analyses, so the Prior Information about the parameters that needs to be estimated must be known, and these parameters are assumed to be random variables that have a probability distribution and not fixed values. The prior information can be acquired by personal beliefs or based on previous experiments about the studied phenomenon, and these information are represented as a Prior Probability Distribution (Prior Distribution), and thus we can acquire the Posterior Probability Distribution (Posterior Distribution) by merging the distribution of the current sample observations (under study) with the prior distribution, and the posterior distribution contains all the information (prior information and current sample information) about the parameters that needs to be estimated [1].

8. The Bayesian Approach

Bayesian estimation is an important statistical approach that is widely used in many papers and studies, and in this paper this approach will be adopted in the analysis of the (MSN) models with missing data. Regarding the (MSN) models, it is difficult to draw random samples that follows the posterior distribution $(\underline{\theta}|\underline{X}^o)$ directly, which is why certain strategies where used similar to Data Augmentation (DA) algorithm [14] which is considered an effective statistical tool in multiple assignments to treat the problem of missing data. to make things easier , the observed portion of the data will be denoted as $\underline{X}^o = (X_1^o, \ldots, X_n^o)$ the missing portion of the data will be denoted as $\underline{X}^m = (X_1^m, \ldots, X_n^m)$ and the hole latent variables will be denoted as $\underline{\tau} = (\tau_1, \ldots, \tau_n)$, the complete data log-likelihood function of $(\underline{\theta})$, without additive constant terms will be :

$$\ell_c\left(\underline{\theta} \mid \underline{X}^O, \underline{X}^m, \underline{\tau}\right) = -\frac{1}{2}\sum_{j=1}^{n}\left\{log\,|\Sigma| + \left(\underline{X}_j - \underline{\xi} - \Lambda\underline{\tau}_j\right)'\Sigma^{-1}\left(\underline{X}_j - \underline{\xi} - \Lambda\underline{\tau}_j\right) + \underline{\tau}_j'\underline{\tau}_j\right\} \tag{8.1}$$

In this paper the Gibbs Sampling was employed to merge the latent variables $(\underline{\tau})$ and the missing data $(\underline{X}^m)$ to create the integrated posterior likelihood function [10], at iteration (k) the procedure generates posterior variable by drawing consecutively from the posterior distribution $\left(p\left(\underline{\theta}\left|\underline{X}_j^o,\ \underline{X}^{m(k+1)},\ \underline{\tau}^{(k+1)}\right.\right)\right)$, $\left(p\left(\underline{X}_j^m\left|\underline{X}_j^o,\ \underline{\tau}_j^{(k+1)},\ \underline{\theta}^{(k)}\right.\right)\right)$, *and* $(p(\underline{\tau}_j\left|\underline{X}_j^o, \underline{\theta}^{(k)}\right.)$just as it is shown in (*Gelfand & Smith*) [7]. In simulation, $(\underline{\tau}_j^{(k)})$, $(\underline{X}_j^{m(k)})$, and $(\underline{\theta}^{(k)})$ will converge to its desired distribution after a sufficiently long interval. In terms of Bayesian modeling, we need to select a prior distribution for the (MSN) parameters, $(\underline{\theta} = (\underline{\xi}, \Sigma, \Lambda))$, let (W) be a random matrix of order $(p \times p)$, we say that (W) follows the Wishart distribution, that means $(W \sim w_p(\nu, V))$ if its (pdf) is proportional to [9]:

$$|W|^{\frac{1}{2}(\nu-p-1)}\exp\left\{-\frac{1}{2}\ tr(V^{-1}W)\right\} \tag{8.2}$$

Where (V) is a parameter matrix of order $(p \times p)$.

When prior information are unavailable, one of the proper strategies to avoid an improper posterior distribution is to use common proper prior distributions, and the used prior distributions are:

$$\underline{\xi} \sim N_p(a, \kappa^{-1})$$

$$\Sigma^{-1} \left| B \sim w_p 2\alpha, (2B)^{-1} \right.)$$

$$B \sim w_p(2\gamma, (2H)^{-1})$$

$$\lambda \sim \ N_p(0, \ \Gamma)$$

And (B) is the matrix of the parameters of the prior distribution (Hyperparameters) which follows the Wishart distribution, and $(a, \kappa, \ \alpha \ , \gamma, \ H \ , \Gamma)$ are fixed as proper quantities "depending on the data" which reflects the monotone of the prior distributions. For convenience, but this is not always optimal, we assume that the parameters are independent intuitively and the joint prior distribution for $(\underline{\theta})$ and (B) is:

$$\pi\left(\underline{\theta}, B\right) \propto |B|^{\alpha+(2\gamma-p-1)/2} \exp\left\{-\frac{1}{2}\left(\underline{\xi}-a\right)'\kappa\left(\underline{\xi}-a\right) - tr\left(\left(\Sigma^{-1}+H\right)B\right) - \frac{1}{2}\lambda' \Gamma^{-1}\lambda\right\} \tag{8.3}$$

And from (8.1) and (8.3) we get the joint posterior distribution:

$$p\left(\underline{\theta}, B, \underline{X}^m, \ \tau \left| X^o \right.\right) \propto \pi\left(\underline{\theta}, B\right) \prod_{j=1}^{n} f(\underline{X}_j^m \left| \underline{X}_j^o \right., \underline{\tau}_j, \underline{\theta}) f(\ \underline{\tau}_j \left| \underline{X}_j^o, \underline{\theta}\right.) f(\ \underline{X}_j^o \ \left| \underline{\theta}\right.) \tag{8.4}$$

And the conditional distribution for (8.4) can be found in theorem (2.3) in [9].

Theorem 8.1. *The complete conditional posterior distributions for each of $(\underline{\theta})$, (B), $(\underline{\tau}_j)$, and $(\underline{X}_j^m)$ are as follows [9]:*
("$\left|\ldots\right.$" denotes the condition over all the other variables).

1. $p\ (\underline{\tau}_j \left| \underline{X}_j^o \right., \underline{\theta}) \propto \exp\left\{-\frac{1}{2}\ \left(\underline{\tau}_j - \underline{\nu}_j\right)'\ V_j^{-1}\left(\underline{\tau}_j - \underline{\nu}_j\right)\right\} I_{\mathbb{R}_+^p}(\underline{\tau}_j)$ *Where:*

$$\begin{aligned} \underline{\nu}_j &= \Lambda C_j^{oo}\left(\underline{Y}_j - \underline{\xi}\right), \\ V_j &= I_p - \Lambda C_j^{oo}\Lambda, \\ C_j^{oo} &= O_j'\ \Omega_j^{oo-1} O_j\ , \\ \Omega_j^{oo} &= O_j \Omega O_j' \end{aligned} \tag{8.5}$$

2. $p(\ \underline{X}_j^m ? \underline{X}_j^o, \ \underline{\tau}_j, \underline{\theta}) \ \propto \ \exp\left\{-\frac{1}{2}\left(\underline{X}_j^m - \underline{\zeta}_j^{m.o}\right)' \Sigma_j^{mm.o^{-1}} (\underline{X}_j^m - \underline{\zeta}_j^{m.o})\right\}$

$$\begin{aligned} \underline{\zeta}_j^{m.o} &= M_j(\underline{\xi} + \ \Lambda\underline{\tau}_j + \Sigma S_j^{oo}(\underline{X}_j - \underline{\xi} - \Lambda\underline{\tau}_j)) \\ \Sigma^{mm.o} &= M_j(I_p - \Sigma S_j^{oo})\Sigma M_j' \\ S_j^{oo} &= O_j'(O_j \Sigma O_j')^{-1} O_j \end{aligned}$$

3. $p(\underline{\xi}|\ldots) \propto \exp\left\{-\frac{1}{2}(\underline{\xi}-\underline{\mu}^*)'\Sigma^{*-1}(\underline{\xi}-\underline{\mu}^*)\right\}$

$$\Sigma^* = (n\Sigma^{-1}+\kappa)^{-1}, \tag{8.6}$$
$$\underline{\mu}^* = \Sigma^*\left(\Sigma^{-1}\sum_{j=1}^{n}(\underline{X}_j - \Lambda\underline{\tau}_j) + \kappa a\right)$$
$$\kappa = Diag\ (\frac{1}{R_1^2}, \frac{1}{R_2^2}, \ldots, \frac{1}{R_p^2})$$

R: is the range for the variable (i).

a:is the middle point for the ranges.

4. $p(B|\ldots) \propto |B|^{(2\gamma^*-p-1)/2}\exp\left\{-\frac{1}{2}tr(2H^*B)\right\}$

$$\gamma^* = \alpha+\gamma, \qquad H^* = \Sigma^{-1}+H, \qquad H = 10\ \kappa \tag{8.7}$$

5. $p(\ \Sigma^{-1}|\ \ldots) \propto |\Sigma^{-1}|^{(\alpha^*-p-1)/2}\exp\left\{-\frac{1}{2}\ tr(B^*\Sigma^{-1})\right\}$

$$\alpha^* = n+2\alpha \quad , \qquad B^* = 2B + \sum_{j=1}^{n}(\underline{X}_j - \underline{\xi} - \Lambda\underline{\tau}_j)(\underline{X}_j - \underline{\xi} - \Lambda\underline{\tau}_j)' \tag{8.8}$$

6. $p(\ \underline{\lambda}|\ \ldots) \propto \exp\left\{-\frac{1}{2}(\underline{\lambda}-\Gamma^*\underline{\delta}^*)'\Gamma^{*-1}(\underline{\lambda}-\Gamma^*\underline{\delta}^*)\right\}$

$$\Gamma^* = \left(\Gamma^{-1}+\Sigma^{-1}\,?\sum_{j=1}^{n}\underline{\tau}_j\underline{\tau}_j'\right)^{-1} \quad , \quad \underline{\delta}^* = \left(\Sigma^{-1}\ ?\sum_{j=1}^{n}\underline{\tau}_j(\underline{X}_j - \underline{\xi})'\right)\underline{1}_p \tag{8.9}$$

The proof follows from direct calculations.
In simulation, samples for each of ($\underline{\tau}_j$, $\underline{X}_j^m$, B, and $\underline{\theta}$) are generated in turns. And following the theorem above, Gibbs sampling continues as follows:

1. Generating values for ($\underline{\tau}_j$) from the distribution ($TN_p(\underline{\nu}_j\ ,\ V_j\ ,\ \mathbb{R}_+^p)$) (Truncated Normal) with positive values, and each of (V_j) and ($\underline{\nu}_j$) is given by (8.5).

2. Generate values for ($\underline{X}_j^m$) from the distribution ($N_p\left(\underline{\zeta}_j^{m.o}, \Sigma_j^{mm.o}\right)$) (Normal Distribution), and each of ($\underline{\zeta}_j^{m.o}$) and ($\Sigma_j^{mm.o}$) is given is theorem (3b) in [9].

3. Generate values for ($\underline{\xi}$) from the distribution ($N_p\left(\underline{\mu}^*, \Sigma^*\right)$) (Normal Distribution), and each of ($\underline{\mu}^*$) and (Σ^*) is given in equation (8.6).

4. Generate values for (B) from the distribution ($w_p\left(2\gamma^*, (2H^*)^{-1}\right)$) (Wishart Distribution), and each of (γ^*) and (H^*) is given in equation (8.7).

5. Generate values for (Σ^{-1}) from the distribution ($w_p(\alpha^*, B^{*-1})$) (Wishart Distribution) and each of (α^*) and (B^{*-1}) is given in equation (8.8).

6. Generate values for ($\underline{\lambda}$) from the distribution ($N_p(\Gamma^*\underline{\delta}^*, \Gamma^*)$) (Normal Distribution) and each of (Γ^*) and ($\underline{\delta}^*$) is given in equation (8.9).

To achieve "The concept of constant estimation" (*Edwards et al.*) [7] from an objective Bayesian perspective, we need to assign (a, κ, α , γ, H , Γ) so that they are not sensitive or are not affected by the change in the prior distributions. And as advised by (*Richardson & Green*) [10], we assume that [9]:

$$\kappa^{-1} = \text{Diag}(R_1^2, R_2^2, ..., R_p^2)$$

This assignment gives weak prior information for ($\underline{\xi}$) for the parameters of the previous Wishart distributions, we choose:

$$\alpha = \ p + 1$$

Meaning $\alpha = number\ of\ variables + 1$, and:

$$\gamma = \frac{p+1}{10} = \frac{\alpha}{10}$$

In addition to that, (Γ) is selected as a diagonal matrix with relatively large variances, like:

$$\Gamma = 10^4 I_p$$

Γ: Is a diagonal matrix of order (pxp) and each diagonal element is multiplied by (10^4).

9. Experimental Side

9.1. Preface

In this part, simulations were used to compare between estimation methods to find which one of them is the best by using the Mean Squared Error (MSE), and the results of the simulation was analyzed using R programming language.

9.2. Concept of Simulation

Analysis using simulation is considered a natural extension for analytical techniques in a reasonable way, and it is considered a solid method because it is a method for testing before applying the experiment on the real data, and it is a method that researchers turn to for some situations that can't be represented mathematically for many reasons, like when it is too complicated to phrase the problem under study, or that the problem is of random nature, or because of the formulations necessary for describing the problem accurately, and for all the problems that are difficult to phrase mathematically. Simulation is known as the process of representing the behavior of the real phenomenon under study in a way that is close to reality, and it is considered one of the most important problem-solving techniques, it can also be said that it is the last and only way to solve any problem if it was difficult to solve it using numerical methods or analytical methods. Simulation also depends on resampling methods and generating random variables and numbers with specific characteristics.

9.3. Steps of the Simulation Experiment

The simulation experiment using R programming language will be expressed by the following steps:
The First Stage:
In this stage, the initial values for the function parameters were assumed, where these values are close to the estimated values for the real data, and this stage is one of the most important stages which will be relied on primarily in the following stages as shown in the following table:

Table 1: shows the initial values for the parameters of the Bivariate Skew Normal Distribution.

function	ξ_1	ξ_2	σ_{11}	σ_{21}	σ_{22}	λ_1	λ_2
1	8.0	5.0	3.0	0.0	1.0	2.0	1.0
2	9.0	6.0	4.0	0.0	2.0	3.0	2.0
3	10.0	7.0	5.0	0.0	3.0	4.0	3.0

Three different sample sizes were also selected to conduct the simulation (400, 600, 800), and the experiment was repeated (500) times.

The Second Stage:
In this stage, variables are generated according to the (MSN) distribution using predefined functions as follows:

$$(\underline{X}_1\ , \underline{X}_2)' \sim MSN(\underline{L},\ Sigma,\ Lambda)$$

Where:
$\underline{L}$: Is the position vector ($\underline{\xi}$).
$Sigma$: Is the variance-covariance matrix (Σ).
$Lambda$: Is the skewness matrix (Λ).

The Third Stage:
In this stage, the missing data mechanism will be explained below:

1. Calculated ($y = f(\underline{X}\,|\underline{\xi}\ , \Sigma, \Lambda)$) that is shown in equation (3.1) using the generated data.
2. Use ($\underline{X}_1$), ($\underline{X}_2$), and ($\underline{y}$) in the (MAR) mechanism to calculate (π_i) for each observation.
3. Generate random numbers (1,0) using the Bernoulli distribution ($Ber(\pi_i)$).
4. Lose the observations that corresponds to (0).

Thus, data with missing rates are generated. Two missing rates were generated and they are (12%) and (20%), and the used missing mechanism is (MAR) aforementioned in equation (2.2) for ($\underline{X}_2$), which was generated according to the suggestion of the researcher and based on (*Wang*, [pp:56, [15]]) as follows:

$$\begin{aligned} p\,(y, x_1) &= p\,(R = 1 \mid \underline{Y} = y, \underline{X}_1 = x_1) = \pi_i \\ &= 1/(1 + \exp(-ln\,(\varrho_1) - \varrho_2\,(y - \overline{y}) - \varrho_3\,(x_1 - \overline{x}_1) \end{aligned} \tag{9.1}$$

To generate the (12%) missing rate, let (ϱ_1=3, ϱ_2= 0.3, ϱ_3 = 0.3) in the above formula (9.1) which was found using simulation.
Using the same method to generate the (20%) missing rate (ϱ_1= 5, ϱ_2= 0.2, ϱ_3 = 0.2) in (9.1).
Which means the missing depended on the observed values of the variable and didn't depend on the missing value, meaning the missing value is independent of any other value in the data.

The Fourth Stage:
The missing observations are estimated using the following methods:

1. K-Nearest Neighbors (KNN).

2. Bayes Method.

The Fifth Stage:
The parameters of the bivariate skew normal distribution function are estimated using the genetic algorithm and Bayes method.

The Sixth Stage:
In this stage, the (MSE) criterion for the model is used to compare between the estimation methods and finding which one is the best, as shown in the following formula:

$$MSE\left[\underline{\hat{Y}}\right] = \frac{1}{k}\sum_{i=1}^{k}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\underline{\hat{Y}} - \underline{Y}\right]^2\right\} \tag{9.2}$$

k: The iteration of the experiment.
n: The sample size.

10. Analyzing the Results of the Simulation

The results of the analysis of the simulation according to the initial values of the parameters of the bivariate skew normal distribution function will be shown after estimating the missing values for missing rates (0.12%) and (0.20%) as shown below:Table 2 shows the estimates of the parameters for the first function with missing rate (0.12%) for all sample sizes.

Table 2: represents the values of the measure ARL For control charts at different significant levels

	Missing Rate	Parameters / Method	ξ_1= 8	ξ_2= 5	σ_1= 3	σ_{12} = 0	σ_2= 1	λ_1= 2	λ_2= 1
N=400	0.12	GA for KNN	8.8099614	5.0214461	2.3845751	-0.5133339	1.3747442	0.4210459	0.8553179
		Bayes	9.1279540	5.4012179	2.0765837	-0.1319928	1.8271391	2.0104509	1.0002036
	Missing Rate	Parameters / Method	ξ_1= 8	ξ_2= 5	σ_1= 3	σ_{12} = 0	σ_2= 1	λ_1= 2	λ_2= 1
N=600	0.12	GA for KNN	8.1629024	5.2798586	2.3777766	-0.5228044	0.9877376	1.9194407	0.7037679
		Bayes	9.1027602	5.3759444	2.0817608	-0.1570675	1.7324595	1.9985579	0.9961794
	Missing Rate	Parameters / Method	ξ_1= 8	ξ_2= 5	σ_1= 3	σ_{12} = 0	σ_2= 1	λ_1= 2	λ_2= 1
N=800	0.12	GA for KNN	8.2500446	4.9887114	2.2530473	-0.2241111	1.0414799	1.5816371	0.9011015
		Bayes	9.1743515	5.3265541	2.0383445	-0.1088979	1.7038134	1.9993176	0.9986298

Table 3: shows the estimates of the parameters for the first function with missing rate (0.20%) for all sample sizes.

	Missing Rate	Parameters / Method	ξ_1= 8	ξ_2= 5	σ_1= 3	σ_{12} = 0	σ_2= 1	λ_1= 2	λ_2= 1
N=400	0.20	GA for KNN	9.1204851	6.1918406	3.0782047	-0.3508509	1.7818439	1.8452623	1.5504247
		Bayes	10.1607716	6.7201135	3.0473978	-0.1760932	5.2108559	2.9978775	2.0016053
	Missing Rate	Parameters / Method	ξ_1= 8	ξ_2= 5	σ_1= 3	σ_{12} = 0	σ_2= 1	λ_1= 2	λ_2= 1
N=600	0.20	GA for KNN	9.0663555	6.4759207	3.1982838	-0.8621723	2.1283933	2.2282410	1.1704172
		Bayes	10.2280575	6.6216008	3.0138835	-0.1949773	5.2790555	3.0049681	2.0064444
	Missing Rate	Parameters / Method	ξ_1= 8	ξ_2= 5	σ_1= 3	σ_{12} = 0	σ_2= 1	λ_1= 2	λ_2= 1
N=800	0.20	GA for KNN	8.9551510	6.2250839	3.1597356	-0.4985113	1.8275938	2.7730661	1.5799084
		Bayes	10.2300881	6.5832515	3.0104555	-0.1557614	5.1685245	3.0071409	1.9961121

Table 4: shows the estimates of the parameters for the second function with missing rate (0.12%) for all sample sizes.

	Missing Rate	Parameters / Method	ξ_1= 9	ξ_2= 6	σ_1= 4	σ_{12} = 0	σ_2= 2	λ_1= 3	λ_2= 2
N=400	0.12	GA for KNN	9.4660538	5.4191096	3.0066047	-0.4671901	2.9493001	1.2957205	1.9413553
		Bayes	10.3549689	6.5646579	3.0449073	-0.1809513	5.2033314	2.9956296	1.9934686
	Missing Rate	Parameters / Method	ξ_1= 9	ξ_2= 6	σ_1= 4	σ_{12} = 0	σ_2= 2	λ_1= 3	λ_2= 2
N=600	0.12	GA for KNN	9.6350910	5.2822114	3.2231167	-0.7968911	3.0931638	1.2303037	2.2957748
		Bayes	10.4147833	6.4701936	3.0505336	-0.2116168	5.3138965	3.0045497	1.9984123
	Missing Rate	Parameters / Method	ξ_1= 9	ξ_2= 6	σ_1= 4	σ_{12} = 0	σ_2= 2	λ_1= 3	λ_2= 2
N=800	0.12	GA for KNN	9.1446100	6.2382298	3.2356584	-0.5082414	2.7026968	3.0724806	2.1131503
		Bayes	10.1720752	6.6120382	3.0556513	-0.2125329	5.3512420	2.9985188	1.9995881

Table 5: shows the estimates of the parameters for the second function with missing rate (0.20%) for all sample sizes.

	Missing Rate	Parameters / Method	ξ_1= 9	ξ_2= 6	σ_1= 4	σ_{12} = 0	σ_2= 2	λ_1= 3	λ_2= 2
N=400	0.20	GA for KNN	9.0530555	5.9516929	3.1463821	-0.3777404	2.0836474	2.9307700	2.0170883
		Bayes	10.2956659	6.4890443	3.0667846	-0.2509221	5.4121143	2.9993344	2.0013349
	Missing Rate	Parameters / Method	ξ_1= 9	ξ_2= 6	σ_1= 4	σ_{12} = 0	σ_2= 2	λ_1= 3	λ_2= 2
N=600	0.20	GA for KNN	8.8864657	6.4400817	3.4724829	-0.9290718	2.3326312	3.1332305	1.4633680
		Bayes	10.212146	6.612245	3.060047	-0.214642	5.405647	2.988821	1.998194
	Missing Rate	Parameters / Method	ξ_1= 9	ξ_2= 6	σ_1= 4	σ_{12} = 0	σ_2= 2	λ_1= 3	λ_2= 2
N=800	0.20	GA for KNN	8.9405466	6.1005540	3.0890493	-0.4173828	2.0378583	1.8627951	0.6255540
		Bayes	10.2123174	6.5734467	3.0247528	-0.1983989	5.3305819	3.0052261	1.9983973

Table 6: shows the estimates of the parameters for the third function with missing rate (0.12%) for all sample sizes.

	Missing Rate	Parameters / Method	ξ_1= 10	ξ_2= 7	σ_1= 5	σ_{12} = 0	σ_2=3	λ_1=4	λ_2=3
N=400	0.12	GA for KNN	10.285287	6.605522	3.685891	-0.569276	4.724055	1.508633	2.389329
		Bayes	11.5045279	7.7490308	4.0771460	-0.1975196	10.7244408	3.9898691	2.9912015
	Missing Rate	Parameters / Method	ξ_1= 10	ξ_2= 7	σ_1= 5	σ_{12} = 0	σ_2=3	λ_1=4	λ_2=3
N=600	0.12	GA for KNN	10.1139101	7.1166405	4.1670083	-0.7989466	3.3631598	2.2429585	1.5377812
		Bayes	11.5372379	7.8672222	4.0140401	-0.1644978	10.7571008	4.0257169	2.9994155
	Missing Rate	Parameters / Method	ξ_1= 10	ξ_2= 7	σ_1= 5	σ_{12} = 0	σ_2=3	λ_1=4	λ_2=3
N=800	0.12	GA for KNN	10.0870696	7.0726298	4.0620440	-0.3847556	4.3763441	3.8966957	2.9894216
		Bayes	11.2962057	7.7465020	3.9900585	-0.2327017	11.0951170	4.0044200	3.0049644

Table 7: shows the estimates of the parameters for the third function with missing rate (0.20%) for all sample sizes.

	Missing Rate	Parameters / Method	ξ_1= 10	ξ_2= 7	σ_1= 5	σ_{12} = 0	σ_2=3	λ_1=4	λ_2=3
N=400	0.20	GA for KNN	10.1933037	6.6638231	4.1577300	-0.5382748	3.7339044	2.5951507	2.3170002
		Bayes	11.4222847	7.6259594	4.1126781	-0.2798102	11.5053717	3.9999447	3.0012919
	Missing Rate	Parameters / Method	ξ_1= 10	ξ_2= 7	σ_1= 5	σ_{12} = 0	σ_2=3	λ_1=4	λ_2=3
N=600	0.20	GA for KNN	10.2322819	6.5626969	3.8113400	-0.2164446	4.0580198	1.7569416	2.2917484
		Bayes	11.4428037	7.7345661	4.0459555	-0.2258907	10.9917370	4.0122840	3.0069837
	Missing Rate	Parameters / Method	ξ_1= 10	ξ_2= 7	σ_1= 5	σ_{12} = 0	σ_2=3	λ_1=4	λ_2=3
N=800	0.20	GA for KNN	10.2901773	6.7189520	4.0419350	-0.3775511	3.5226661	2.6927565	2.9769257
		Bayes	11.2938850	7.7583988	4.0407081	-0.2414053	11.2694677	4.0127042	3.0026480

11. Conclusions

The problem of missing values is one of the most common problems that researchers encounter when collecting the data and analyzing them, which means missing part of the sample data under study, and not treating this problem properly might cause a lot of problems for the researcher, like poorly estimating the variance, or obtaining biased results, that is why this paper aims to estimate the missing values for the Multivariate Skew Normal (MSN) model using some methods, like the K-Nearest Neighbors (KNN). After estimating the missing values, the parameters are estimating using the Genetic Algorithm (GA), and Bayes method, and form the experimental side, a number of conclusions where found, it was found form the simulation that the (GA) is the best method for estimation because it had the lowest (MSE) values for all missing rates and sample sizes, especially at size (800).

Table 8: shows the results of the (MSE) for the function.

Model	Missing Rate	N	GA_ KNN	Bayes	Best
1	0.12	400	0.0004192601	0.003826618	GA_ KNN
		600	0.0004732944	0.003817759	GA_ KNN
		800	0.0004168924	0.002288662	GA_ KNN
	Missing Rate	N	GA_ KNN	Bayes	Best
	0.20	400	0.0002268837	0.00199855	GA_ KNN
		600	0.00020432	0.002451352	GA_ KNN
		800	0.0002531902	0.003075343	GA_ KNN
Model	Missing Rate	N	GA_ KNN	Bayes	Best
2	0.12	400	0.0003352251	0.002984836	GA_ KNN
		600	0.0002680636	0.00212315	GA_ KNN
		800	0.0001357652	0.001933735	GA_ KNN
	Missing Rate	N	GA_ KNN	Bayes	Best
	0.20	400	0.0005446133	0.006961795	GA_ KNN
		600	0.0002141637	0.005884033	GA_ KNN
		800	0.0001807105	0.0032971325	GA_ KNN
Model	Missing Rate	N	GA_ KNN	Bayes	Best
3	0.12	400	0.000670485	0.005176874	GA_ KNN
		600	0.0004020423	0.004151064	GA _KNN
		800	0.0001819244	0.00126218	GA_ KNN
	Missing Rate	N	GA_ KNN	Bayes	Best
	0.20	400	0.00054468997	0.006114805	GA_ KNN
		600	0.00007872525	0.001151139	GA_ KNN
		800	0.00004812375	0.001127537	GA_ KNN

References

[1] Q. N. N. Al-Qazaz, *A Comparison of Robust Bayesian Approaches with other Methods for Estimating Parameters of Multiple Linear Regression Model with missing Data*, Thesis in Ph.D. from University of Baghdad; college of Administration and Economics / Department of Statistics, 2007.

[2] I. A. Hussein, *Analysis of incomplete data for multiple regression models using ECME, ECM, EM algorithms with practical application*, Master of Science in Statistics, College of Administration and Economics, University of Baghdad, 2010.

[3] Q. N. N. Al-Qazaz and M. K. Ayoub, , *User (K-Means) for clustering in Data Mining with application*, J. Econ. Administrative Sci. , 22(91)(2016) 389-406.

[4] Q. N. N. Al-Qazaz, M. Y. Hammoud and T. M. Abbas , *A nonparametric estimation of a multivariate probability density function* Al-Nahrain J. Sci., 11 (2)(2008) 55-63.

[5] E. Acuna and C. Rodriguez, *The treatment of missing values and its effect on classifier accuracy*, In: Classif. Clustering Data Min. Appl., Springer, Berlin, Heidelberg, 2004, pp. 639-647.

[6] A. Azzalini and A. D. Valle, *The multivariate skew-normal distribution* , Biometrika, 83 (4)(1996) 715-726. .

[7] A. E. Gelfand and A. F. Smith, *Sampling-based approaches to calculating marginal densities*, J. Am. Stat. Assoc., 85 (410)(1990) 398-409.

[8] G. Guo, H. Wang, D.Bell, Y. Bi and K. Greer ,*KNN model-based approach in classification.* In: OTM Confederated Int. Conf. "On the Move to Meaningful Internet Systems", Springer, Berlin, Heidelberg, 2003, pp. 986-996.

[9] T. I. Lin , H. J. Ho and C. L. Chen , *Analysis of multivariate skew normal models with incomplete data*, J. Multivar. Anal., 100 (10)(2009), 2337-2351.

[10] S. Richardson and P. J. Green, *On Bayesian analysis of mixtures with an unknown number of components (with discussion)*, J. R. Stat. Soc. Ser. B (Statistical Methodology), 59 (4)(1997) 731-792.

[11] S. K. Sahu, D. K. Dey and M. D. Branco, *A new class of multivariate skew distributions with applications to Bayesian regression models*, Can. J. Stat. , 31 (2)(2003) 129-150.
[12] W. Shahzad, Q. Rehman and E. Ahmed, *Missing data imputation using genetic algorithm for supervised learning*, Int. J. Adv. Comput. Sci. Appl. (IJACSA), 8(3)(2017) 438-445 .
[13] L. N. Shawkat and Q. N. Nayef, *A Comparison between (ECM) and (KNN) Methods for the Multivariate skew-normal model with incomplete data* , J. Al Rafidain Univ. Coll. , 46(2020) 378-393.
[14] M. A. Tanner and W. H. Wong, *The calculation of posterior distributions by data augmentation*, J. Am. Stat. Assoc. , 82 (398)(1987), 528-540.
[15] Q. H. Wang, *Statistical estimation in partial linear models with covariate data missing at random*, Ann. Inst. Stat. Math., 61 (1) (2009) 47-84.
[16] A. Yalçınkaya, B. Şenoğlu and U. Yolcu , *Maximum likelihood estimation for the parameters of skew normal distribution using genetic algorithm*, Swarm Evol. Comput. , 38(2018) 1-28.
[17] N. A. Zainuri, A. A. Jemain and N. Muda ,*A comparison of various imputation methods for missing values in air quality data*, Sains Malaysiana, 4 (3)(2015) 449-456.
[18] https://towardsdatascience.com/importance-of-distance-metrics-in-machine-learning-modelling-e51395ffe60d