Journal of Al Rafidain University College Issue No. 46/ 2020
The (15th & the 2nd International) Conference of Statistical Applications ISSN (1681- 6870)

A Comparison between (ECM) and (KNN) Methods for the Multivariate skew-normal model with incomplete data

Lina Nidhal Shawkat[1] **, Prof. Dr. Qutaiba Nabeel Nayef**[2]

[1]Department of Statistics, University of Baghdad, lola1990@yahoo.com
[2]Department of Statistics, University of Baghdad, dr.qutaiba.n@gmail.com

Abstract:

Statistical parameter estimation for multivariate data leads to waste of information, if the missing data are omitted, and in return will lead to inaccurate estimates, that is why incomplete data should be estimated using one of the statistical estimation methods to get accurate results and in return good estimates for the parameters. This paper aims to estimate the missing values for a multivariate skew normal model using the Expectation Conditional Maximization (ECM) algorithm and the K-Nearest Neighbour (KNN) method. After estimating the missing values, the model parameters will be estimated using the Maximum Likelihood Estimation (MLE) method and Newton-Raphson algorithm. A comparison between the (ECM) algorithm and the (KNN) method was conducted using the Mean Squared Error (MSE) for the model through simulation to find the best estimation method.

Keywords: Expectation conditional maximization (ECM), K-nearest neighbour method, Maximum likelihood estimators (MLE), Newton-Raphson algorithm.

مقارنة بين طريقتي (ECM) و (KNN) لأنموذج متعدد المتغيرات الطبيعي الملتوي بوجود مشكلة البيانات المفقودة

لينا نضال شوكت **أ.د.قتيبة نبيل نايف**

قسم الإحصاء، كلية الإدارة والاقتصاد ، جامعة بغداد، العراق

المستخلص: ان تقدير المعلمات الإحصائية لبيانات متعددة المتغيرات تؤدي إلى إهدار المعلومات إذا تم اهمال القيم المفقودة ، وبالتالي تؤدي الى تقديرات غير دقيقة ، لذا يجب تقدير البيانات غير التامة بإحدى طرق التقدير الإحصائية ، للحصول على نتائج دقيقة وبالتالي الحصول على تقديرات معالم جيدة. يهدف البحث الى تقدير القيم لمفقودة لأنموذج متعدد المتغيرات الطبيعي الملتوي باستعمال بعض الطرق والخوارزميات ، منها استعمال خوارزمية توقع التعظيم الشرطي (ECM) ، و طريقة تعويض k-اقرب مجاور(KNN). ومن ثم يتم ايجاد مقدرات المعالم للبيانات بعد تقدير القيم المفقودة عن طريق إيجاد مقدرات الإمكان الأعظم (MLE) باستعمال خوارزمية نيوتن رافسون. وقد تمت المقارنة بين الطريقتين (ECM) و (KNN) باستعمال أسلوب المحاكاة من خلال إيجاد متوسط مربعات الخطأ (MSE) للأنموذج لمعرفة أفضل طريقة للتقدير.

الكلمات المفتاحية: توقع التعظيم الشرطي، طريقة تعويض k-اقرب مجاور، مقدرات الإمكان الأعظم، خوارزمية نيوتن رافسون

ISSN (1681- 6870)

1. Introduction:

The skew normal distribution is considered a member of the family of distributions that includes the normal distribution, except that the skew normal distribution contains an additional parameter to regulate the skewness. The skewness is the asymmetry of the statistical distribution where the curve appears distorted or skewed either to the left or the right, and this deviation can be determined to know how much the distribution differs from the normal distribution. The normal distribution appears in graphs as a classical bell-shaped curve that is symmetrical around the mean, median, and mode.

Missing data in the multivariate skew normal (MSN) distribution is one of the most common problems when collecting and analysing the data, which means losing part of the sample data, for example, in a manufacturing experiment missing data might happen because of mechanical malfunctions that has nothing to do with the experimental procedure, or in a survey, a person might not be able to express his preference of one candidate over another, or in a family survey some participants might refuse to report their income or refuse to answer some of the questions they were asked, or simply because some parts of the data are damaged or lost. Missing data is one of the major problems that researchers encounter, and the statistical methods that are used to analyse the data assumes a complete information on all the variables that are used in the analysis, and not solving this issue properly could cause some problems for the researcher like not estimating the variance appropriately or acquiring biased results, therefore it is necessary to estimate the missing values using some of the statistical method to overcome this problem, and in this paper the (ECM) algorithm and the (KNN) method will be used to estimate the missing values.

2. Patterns of the Missing Data: [1][6][12]

2.1 Pattern of Univariate Missing Data:

It is considered to be the simplest type of missing data and it is one of the special patterns in which all the variables are observed completely except for one variable that contains the missing values.

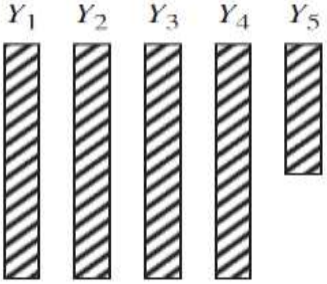


Figure (1): Pattern of Univariate Missing Data.

2.2 Multivariate with Two Patterns:

This pattern is considered one of the special patterns in multivariate data in which some variables are completely observed while others contain missing values in equal amounts.

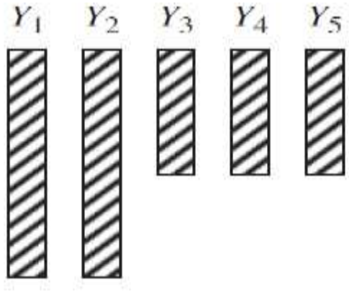


Figure (2): Multivariate with Two Patterns.

2.3 Monotone or Nested Missing Data:

This pattern is considered one of the special patterns in which data are arranged in an ascending or a descending order according to the number of missing values.

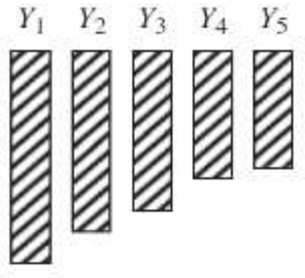


Figure (3): Monotone or Nested Missing Data.

2.4 Missing Data with Unidentified Parameters:

This pattern is also called the pattern of missing data in the case where the parameters are unidentified. This pattern is the last of the special patterns, and it emerges when two variables (X_i) and (X_k) have their observations not recorded in the same observations, which means any observation in (X_i) has a corresponding missing observation in the variable (X_k).

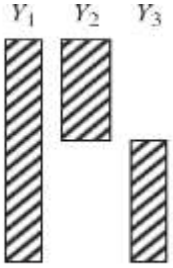


Figure (4): Missing Data with Unidentified Parameters.

2.5 General Pattern:

In this pattern the data is missing at random for any value from the values of the variables under study.

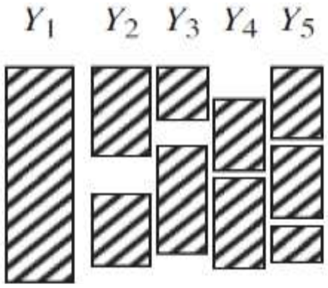


Figure (5): General Patterns.

3. Mechanisms that Lead to Missing Data:[1][2][3]

The mechanism that will be used in this paper is the Missing at Random (MAR) mechanism meaning that the distribution counted on the observed values of the variable and didn't count on the missing values which means the missing value is independent of any other values in the data according to the following formula:

$$P(M|\,X, \varphi) = P(M|\,X^o, \varphi) \qquad for\ all\ X^m$$

Where:

φ: The unknown parameters.

X: Data matrix of order $(n \times p)$.

M: Binary matrix taking the values of (0, 1) and is called the Missing Data Indicator Matrix.

And:

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

To generate a missing rate of (0.12):

$$\pi_i = \frac{1}{1 + e^{(-\,ln(3) - 0.3(y_i - \bar{y}) - 0.2(x_i - \bar{x})}}$$

And to generate a missing rate of (0.20):

$$\pi_i = \frac{1}{1 + e^{(-\ln(5) - 0.2(y_i - \bar{y}) - 0.2(x_i - \bar{x})}}$$

Here (y) was used as a contributing factor correlated to the variable in which we want to generate the missing values in. For example if we the data of three variables (y), (x_1), and (x_2) and there is a reasonable correlation between (y) and the two other variables and we wanted to consider some data are missing in variable (x_1) according to the mechanism above, we use (x_2) instead of (x_i) in the above equations and (y) stays the same, however if we want the missing data to be in the variable (x_2) we use (x_1) instead of (x_i), after that we generate binary values (0, 1) using the Bernoulli distribution ($Ber(\pi_i)$) and after that we consider the values corresponding to the (0's) as missing values.

4. The Multivariate Skew Normal Distribution:[13][16]

Let the random vector (X) follows the (MSN) distribution for (p) variables [13] with the location vector ($\xi \in \mathbb{R}^p$) (a Euclidian vector space consisting of the real numbers for every (p) of rows), (Σ) is the ($p \times p$) covariance matrix, ($\Lambda = Diag(\lambda)$) is the skewness matrix, and ($\boldsymbol{\lambda} = (\boldsymbol{\lambda_1}, \dots, \boldsymbol{\lambda_p})'$), then its p. d. f. is given by:

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

Where:

ξ: is the location vector consisting of (p) variables.

$\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}^2$

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

And ($\phi_p(\boldsymbol{x} \mid \boldsymbol{\xi}, \boldsymbol{\Omega})$) is the probability density function for the normal distribution ($N_p(\mu, \Sigma)$), that means ($\phi_p(\boldsymbol{x} \mid \boldsymbol{\xi}, \boldsymbol{\Omega}) \equiv \phi\text{p}(\,.\, \mid \mu, \boldsymbol{\Sigma})$), and ($\Phi_p(\boldsymbol{\Lambda}\, \boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi}) \mid \boldsymbol{\Delta})$) is the cumulative distribution function (c. d. f.) for the normal distribution ($N_p(0, \boldsymbol{\Sigma})$), which means ($\Phi_p(\boldsymbol{\Lambda}\, \boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi}) \mid \boldsymbol{\Delta}) \equiv \Phi\text{p}(\,.\, \mid \boldsymbol{\Sigma})$).

And ($f(\boldsymbol{x} \mid \boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$) can be expressed as ($SN_p(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$), and if ($\boldsymbol{\Lambda} = 0$): $\boldsymbol{X} \sim SN_p(\boldsymbol{\xi}, \boldsymbol{\Sigma})$

This distribution can be expressed in a form which is more appropriate and that is used for generating random numbers [12] :

$$\boldsymbol{X} = \boldsymbol{\xi} + \boldsymbol{\Lambda}\, |\zeta_1| + \boldsymbol{\Sigma}^{\boldsymbol{1/2}}\, \zeta_2 \qquad (2)$$

Where ($\boldsymbol{\zeta_1}$) and ($\boldsymbol{\zeta_2}$) are two independent random vectors that follows the distribution ($N_p(0, \boldsymbol{I_p})$), and if we assume that ($\boldsymbol{\tau} = |\zeta_1|$), then ($\boldsymbol{\Sigma}$) follows the Half-Normal distribution ($HN_p(0, I_p)$).

And by that, (X) can be expressed as follows:

$$\boldsymbol{X} \mid \boldsymbol{\tau} \sim N_p(\boldsymbol{\xi} + \boldsymbol{\Lambda}\, \boldsymbol{\tau}, \boldsymbol{\Sigma}) \qquad (3)$$

$\boldsymbol{\tau} \sim HN_p(0, \boldsymbol{I_p})$

5. Expectation Conditional Maximization Algorithm: [3][13]

The (EM) algorithm [3] is one of the most common iterative tools that is used in maximum likelihood estimation for models with missing data or hidden variables, in addition to having some desirable properties like its consistent rate or pattern of convergence and its simple implementation. However, the (EM) algorithm loses some of its features when the (M) step becomes analytically intractable, or in other word its analysis becomes very difficult.

The (ECM) algorithm that was suggested by (Rubin) and (Meng) is a simple modification to the (EM) algorithm in which the maximization (M) steps are replaced by a series of conditional maximization (CM) steps with simple calculations. And by treating the values of (X_i^m) and (τ_i) as eigenvalues, the (ECM) algorithm is used to find the (ML) estimators for the parameters.

For simplicity of notation, let ($\boldsymbol{X^o} = (\boldsymbol{X_1^o}, \dots, \boldsymbol{X_n^o})$) and ($\boldsymbol{X^m} = (\boldsymbol{X_1^m}, \dots, \boldsymbol{X_n^m})$) be the observed and missing parts of the experimental data, respectively, and let ($\boldsymbol{\tau} = (\boldsymbol{\tau_1}, \dots, \boldsymbol{\tau_n})$) be all the latent variables. The Log-likelihood function for all the data for the (θ) of the first model, after excluding the additional constant terms, is:

$$\ell_o(\theta\,|\boldsymbol{X^o},\boldsymbol{X^m},\boldsymbol{\tau}) = -\frac{1}{2}\sum_{j=1}^{n}\{log\,|\boldsymbol{\Sigma}| + (\boldsymbol{X_j} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\,\boldsymbol{\tau_j})'\,\boldsymbol{\Sigma^{-1}}(\boldsymbol{X_j} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\,\boldsymbol{\tau_j} + \boldsymbol{\tau_j}'\boldsymbol{\tau_j})\} \quad (4)$$

Where (θ) represents all the unknown parameters which means:

$\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$

In the (E) step (expectation step) the value of the function (Q) is calculated as follows:

$$Q\big(\theta\big|\hat{\theta}^{(k)}\big) = \frac{1}{2}\sum_{j=1}^{n}\Big\{log\,|\boldsymbol{\Sigma^{-1}}| - tr(\boldsymbol{\Sigma^{-1}}R_j^{(k)}(\boldsymbol{\xi},\boldsymbol{\Lambda}))\Big\} \quad (5)$$

Where:

$$R_j^{(k)}(\xi,\Lambda) = E\left(\big(\boldsymbol{X_j} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\,\boldsymbol{\tau_j}\big)\big(\boldsymbol{X_j} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\boldsymbol{\tau_j}\big)'|\boldsymbol{X_j^o}, \boldsymbol{\hat{\theta}}^{(k)}\right)$$

$$\begin{aligned} = &\left(\left(\boldsymbol{I_p} - \boldsymbol{\hat{\Sigma}}^{(k)}\boldsymbol{\hat{S}_j^{oo(k)}}\right)\boldsymbol{\hat{\Lambda}}^{(k)} - \boldsymbol{\Lambda}\right)\left(\boldsymbol{\hat{\Psi}_j^{(k)}} - \boldsymbol{\hat{\eta}_j^{(k)}}\boldsymbol{\hat{\eta}_j^{(k)'}}\right)\left(\left(\boldsymbol{I_p} - \boldsymbol{\hat{\Sigma}}^{(k)}\boldsymbol{\hat{S}_j^{oo(k)}}\right)\boldsymbol{\hat{\Lambda}}^{(k)} - \boldsymbol{\Lambda}\right)' \\ &+ \left(I_p - \boldsymbol{\hat{\Sigma}}^{(k)}\boldsymbol{\hat{S}_j^{oo(k)}}\right)\boldsymbol{\hat{\Sigma}}^{(k)} \\ &+ \left(\boldsymbol{\hat{X}_j^{(k)}} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\boldsymbol{\hat{\eta}_j^{(k)}}\right)\left(\boldsymbol{\hat{X}_j^{(k)}} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\boldsymbol{\hat{\eta}_j^{(k)}}\right)' \end{aligned} \quad (6)$$

And:

$$\boldsymbol{\hat{S}}_j^{oo(k)} = \boldsymbol{o}_j'(\boldsymbol{o}_j\boldsymbol{\hat{\Sigma}}^{(k)}\boldsymbol{o}_j')^{-1}\boldsymbol{o}_j$$

Where ($\hat{\eta}_j^{(k)}$) and ($\hat{\Psi}_j^{(k)}$) are defined as:

$$\widehat{\boldsymbol{\eta}}_j^{(k)} = E\left(\boldsymbol{\tau_j} \mid \boldsymbol{X}_j^o, \widehat{\boldsymbol{\theta}}^{(k)}\right) \quad , \quad \widehat{\boldsymbol{\Psi}}_j^{(k)} = E\left(\boldsymbol{\tau_j}\boldsymbol{\tau_j'} \mid \boldsymbol{X}_j^o, \widehat{\boldsymbol{\theta}}^{(k)}\right) \qquad (7)$$

$$\widehat{\boldsymbol{X}}_j^{(k)} = E\left(\boldsymbol{X}_j \mid \boldsymbol{X}_j^o, \widehat{\boldsymbol{\theta}}^{(k)}\right) = \widehat{\boldsymbol{\Sigma}}^{(k)}\widehat{\boldsymbol{S}}_{\boldsymbol{j}}^{\boldsymbol{oo(k)}}\boldsymbol{x_j} + \left(\boldsymbol{I_p} - \widehat{\boldsymbol{\Sigma}}^{(k)}\hat{S}_j^{oo(k)}\right)(\hat{\boldsymbol{\xi}}^{(k)} + \widehat{\boldsymbol{\Lambda}}^{(k)}\widehat{\boldsymbol{\eta}}_j^{(k)}) \qquad (8)$$

The (E) Step:

Computing ($\widehat{\boldsymbol{\eta}}_j^{(k)}$), ($\widehat{\boldsymbol{\Psi}}_j^{(k)}$), and ($\widehat{\boldsymbol{X}}_j^{(k)}$), where ($j = 1, \dots, n$), using equations (7) and (8).

The (CM) Steps:

- **The first (CM) step:**
  Updating the value of ($\hat{\boldsymbol{\xi}}^{(k)}$) by maximizing (5) with respect to ($\boldsymbol{\xi}$) which will result in:
  $$\hat{\boldsymbol{\xi}}^{(k+1)} = \frac{1}{2}\left(\sum_{j=1}^{n}\widehat{\boldsymbol{X}}_j^{(k)} - \widehat{\boldsymbol{\Lambda}}^{(k)}\sum_{j=1}^{n}\widehat{\boldsymbol{\eta}}_j^{(k)}\right)$$

- **The second (CM) step:**
  Updating the value of ($\widehat{\Sigma}^{(k)}$) by maximizing (5) with respect to ($\boldsymbol{\Sigma}$) which will result in:
  $$\widehat{\boldsymbol{\Sigma}}^{(k+1)} = \frac{1}{2}\sum_{j=1}^{n}\widehat{\boldsymbol{R}}_j^{(k)}$$
  Where $\widehat{\boldsymbol{R}}_j^{(k)}$is $\boldsymbol{R}_j^{(k)}(\boldsymbol{\xi}, \boldsymbol{\Lambda})$ in (6) with $\boldsymbol{\xi}$ and $\boldsymbol{\Lambda}$ replaced by $\hat{\boldsymbol{\xi}}^{(k+1)}$ and $\widehat{\boldsymbol{\Lambda}}^{(k)}$, respectively .
- **The third (CM) step:**
  Updating the value of ($\widehat{\Lambda}^{(k)}$) by maximizing (5) with respect to (Λ) which will result in:
  $$\widehat{\Lambda}^{(k+1)} = Diag\left\{\left(\widehat{\boldsymbol{\Sigma}}^{(k+1)^{-1}} \odot \sum_{j=1}^{n}\widehat{\boldsymbol{\Psi}}_j^{(k)}\right)^{-1}\left(\widehat{\boldsymbol{\Sigma}}^{(k+1)^{-1}} \odot \sum_{j=1}^{n}\widehat{\boldsymbol{\Upsilon}}_j^{(k+1)}\right)\boldsymbol{I}_p\right\}$$

  $$\widehat{\boldsymbol{\Upsilon}}_j^{(k+1)} = \left(\widehat{\boldsymbol{\Psi}}_j^{(k)} - \widehat{\boldsymbol{\eta}}_j^{(k)}\boldsymbol{\eta}_j^{(k)'}\right)\widehat{\boldsymbol{\Lambda}}^{(k)}\left(\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}^{(k)}\widehat{\boldsymbol{S}}_j^{00(k)}\right) + \widehat{\boldsymbol{\eta}}_j^{(k)}(\widehat{\boldsymbol{X}}_j^{(k)} - \hat{\boldsymbol{\xi}}^{(k+1)})'$$

$\odot$: represents the Hadamard product [7] of two matrices with the same dimensions.

Implementation of the (ECM) algorithm:

The initial values for (X) and ($Q(\xi, \Sigma, \Lambda)$) where ($\underline{Q}$) is a vector for the model parameters.

1- Replace every missing value in (X) by the mean for the values of (X).
2- Compute the mean and variance for the values of (X), ($\bar{x}$) and (S) respectively.
3- Generate random numbers from the Uniform (0, 1) distribution that are used in selecting initial values for the parameters:

$$\widehat{\boldsymbol{\theta}}^{(\mathbf{0})} = \left(\hat{\boldsymbol{\xi}}^{(\mathbf{0})}, \widehat{\boldsymbol{\Sigma}}^{(\mathbf{0})}, \widehat{\boldsymbol{\Lambda}}^{(\mathbf{0})}\right)$$

$$\widehat{\boldsymbol{\Sigma}}^{(0)} = S + (u-1)\,Diag(S)$$
$$\hat{\boldsymbol{\lambda}}_{\boldsymbol{i}}^{(\mathbf{0})} = (\pm)\sqrt{(1-\mathrm{u})s_{ii}/\,(1-2/\pi)} \qquad , i = 1, \dots, p$$
$$\hat{\boldsymbol{\xi}}^{(\mathbf{0})} = \overline{\boldsymbol{y}} - \sqrt{2/\pi}\,\hat{\boldsymbol{\lambda}}^{(0)}$$

ISSN (1681- 6870)

The missing values are replaced by new values according to the following equation:

$$\hat{\boldsymbol{X}}_j^m = \boldsymbol{M}_j(\hat{\boldsymbol{\xi}} + \hat{\boldsymbol{\Lambda}}\,\hat{\boldsymbol{\eta}}_{\mathrm{j}} + \hat{\boldsymbol{\Sigma}}\,\hat{\boldsymbol{S}}_{\mathrm{j}}^{\mathrm{oo}}\,(\boldsymbol{X}_j - \hat{\boldsymbol{\xi}} - \hat{\boldsymbol{\Lambda}}\,\hat{\boldsymbol{\eta}}_{\mathrm{j}})\,)$$

$$\hat{\boldsymbol{S}}_{\mathrm{j}}^{\mathrm{oo}} = \boldsymbol{O}_j'\,(\boldsymbol{O}_j\hat{\boldsymbol{\Sigma}}\,\boldsymbol{O}_j')^{-1}\,\boldsymbol{O}_j$$

6. K-Nearest Neighbours Imputation Method:[5][8][9]

The (KNN) method is a nonparametric classification method, it is simple but effective in many situations .The (KNN) is considered to be a successful method for multivariate standard data [9] and the idea is to use the distance measure to find the (K) most similar observations for a compound with missing values, and replace these missing values using the available variable information on its neighbours. The (KNN) imputation method can be summarized by filling the missing value with the mean value for the corresponding column of the nearest neighbour to the corresponding row that has no missing values. The nearest neighbours can be determined using the Euclidian Distance.

The distance or the variance between the samples can be measured by using the Euclidian Distance [5], and it can be said that the nearest neighbour uses the K-nearest points to do the classification.

The (KNN) algorithm is as follows:

1- The dataset (D) is divided into two parts, the first is (D_m) which represents the part of the data that has the missing values, and it has at least one missing value or instance, the second part of the data is the part that has no missing values or instances and it is denoted by (D_c).
2- For each vector (x) in (D_m):
    - Each vector is divided into two parts, observed and missing, $(x = [x_o;\ x_m])$.
    - The distance between (x_o) and the vectors for the set (D_c) is calculated. These features are only used in vectors in the case of complete dataset (D_c), and in the vector (x).
    - Use the K-nearest neighbours and estimate the majority of the missing values for the categorical instances. For continuous instances, replace the missing value using the mean value for the instance in the nearest neighbour. The median can be used instead of the mean.

Advantages of this method:

1- It can predict both the qualitative instances (the value that is repeated the most for the K-nearest neighbours) and the quantitative instances (the mean for the K-nearest neighbours).
2- It doesn’t need to create a prediction model for each instance containing missing data, in fact the (KNN) algorithm doesn’t create any particular models.
3- The situations with multiple missing values can be easily treated.

4- It takes into account the structure of the data correlation.

Disadvantages of the method:

1- Choosing the distance function could be Euclidian, Manhattan, Mahalanobis, Pearson, … etc. In this paper the Euclidian distance for two variables will be used and it can be expressed as follows:

$$G(x_1\,,x_2) = \sqrt{\sum_{i=1}^{n}(x_{1i}, x_{2i})^2}$$

2- The (KNN) algorithm searches through the dataset for the most similar cases, and this procedure takes a lot of time and it can be very important to extract the data where the large databases are analysed.
3- Choosing the K-number of neighbours, we substitute by it according to the accuracy of the classifier after the substitution , when choosing a small (K) we notice a reduction in the performance of the classifier after it is computed due to the extensive intensity in some common cases in estimating the missing values. In other hand, choosing a large (K) includes the cases which differs greatly from the analogous that has missing values weakens its estimation procedure, and in return the lowers the performance of the classifier. For small datasets, a (K) less than (10) can be used.

7. <u>Maximum Likelihood Estimation Method :</u> [2][15][16]

The (ML) method is considered as one of the most important methods to find the estimated parameters for complete data after estimating the missing values as previously explained, and now we will explain the procedure of estimating the parameters of a multivariate skew normal model (MSN). By taking the logarithm for equation (7):

$$\ell(\theta \mid \mathrm{X}) = \sum_{j=1}^{n} \log f(\boldsymbol{x_j} \mid \boldsymbol{\xi}\,, \boldsymbol{\Sigma}\,, \boldsymbol{\Lambda}\,)$$

Which means:

$$\ell_o(\boldsymbol{\theta} \mid \boldsymbol{X}) = -\frac{1}{2}\sum_{j=1}^{n}\left\{\log|\boldsymbol{\Sigma}| + (\boldsymbol{X_j} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\,)'\,\boldsymbol{\Sigma^{-1}}(\boldsymbol{X_j} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\,)\right\} \qquad (7)$$

And considering how difficult it is to solve the equations resulting from the partial derivatives of the likelihood function we will resort to the numerical methods by using the Newton-Raphson algorithm to obtain the shape parameter ($\boldsymbol{\xi}$), the covariance parameter ($\boldsymbol{\Sigma}$), and the skewness parameter ($\boldsymbol{\Lambda}$).

ISSN (1681- 6870)

8. The Experimental Side:

8.1. Generating the data:

Three sample sizes were used ($n = 400, 614, 800$) and (p=2) to run the simulations with (100) iterations for each experiment as follows:

1- The data was generated according to the multivariate skew normal distribution (MSN) using hypothetical values for the parameters ($\xi_1 = 12, \xi_2 = 4, \Sigma_1 = 9, \Sigma_2 = 5, \Lambda_1 = 6, \Lambda_2 = 3$) shown in Table (1), and a second hypothetical values for the parameters ($\xi_1 = 20, \xi_2 = 10, \Sigma_1 = 8, \Sigma_2 = 6, \Lambda_1 = 7, \Lambda_2 = 5$) shown in Table (2).

2- Two rates of missing were taken, (12 %) and (20 %), and the data are missing according to the missing at random (MAR) mechanism that was mentioned previously.

8.2. The Model Used in Simulation:

The mean squared error (MSE) for the model will be calculated after estimating the missing values using the (ECM) algorithm and the (KNN) method and estimating the parameters using (MLE), The R program was used for the statistical analysis.

8.3. Interpreting the Results of the Simulation:

The results of the simulation for Tables (1) and (2) will be interpreted as follows:

- **Sample size 400:**
  When the missing observations are at (12 %), and after the estimation of the missing values using (ECM) algorithm and (KNN) method we notice that the model estimated using (KNN) method was better than the model estimated using (ECM) algorithm.
  As for the case when the missing observations are at (20 %), and after the estimation of the missing values using (ECM) algorithm and (KNN) method we notice that the model estimated using (KNN) method was better than the model estimated using (ECM) algorithm.
- **Sample size 614:**
  When the missing observations are at (12 %), and after the estimation of the missing values using (ECM) algorithm and (KNN) method we notice that the model estimated using (ECM) algorithm was better than the model estimated using (KNN) method.
  As for the case when the missing observations are at (20 %), and after the estimation of the missing values using (ECM) algorithm and (KNN) method we notice that the model estimated using (ECM) algorithm was better than the model estimated using (KNN) method.
- **Sample size 800:**
  When the missing observations are at (12 %), and after the estimation of the missing values using (ECM) algorithm and (KNN) method we notice that the model estimated using (KNN) method was better than the model estimated using (ECM) algorithm.
  As for the case when the missing observations are at (20 %), and after the estimation of the missing values using (ECM) algorithm and (KNN) method we notice that the model estimated using (ECM) algorithm was better than the model estimated using (KNN) method.

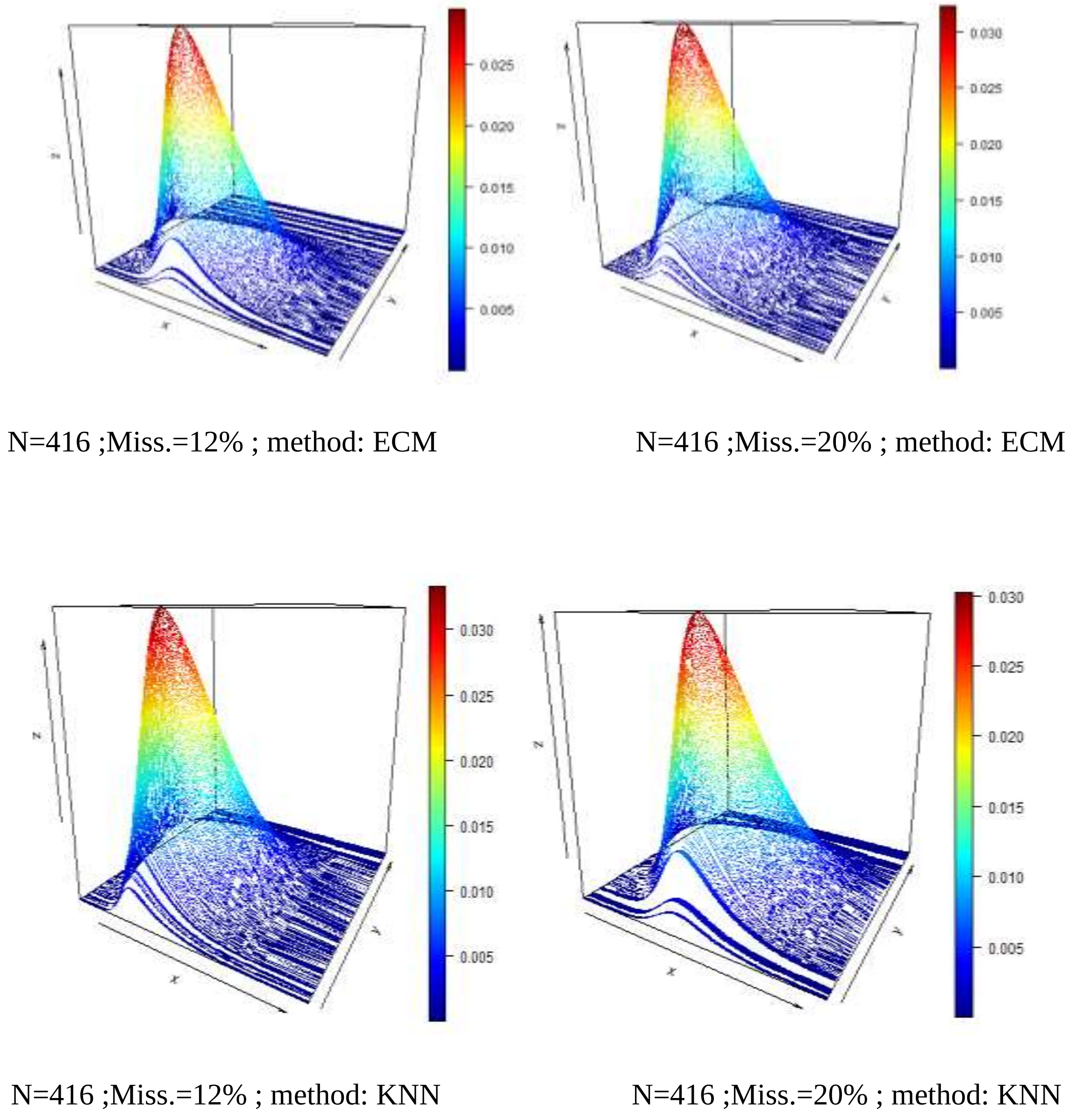
Table (1)

N=400	$\xi_1 = 12$		$\xi_2 = 4$		$\Sigma_1 = 9$		$\Sigma_2 = 5$		Λ_1 =6		Λ_2 =3		MSE	
Method Missing Rate	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	ECM	KNN
0.12	11.972190	11.992755	4.168256	4.174171	9.781955	9.529136	4.161180	4.129643	4.130137	4.060908	2.071783	2.005640	0.00005675219	0.00005539024
0.20	11.958057	11.987677	4.883160	4.899563	10.198317	9.857652	2.292834	2.291551	3.736854	3.682388	1.342513	1.301261	0.00008586726	0.0000762788
N=614	$\xi_1 = 12$		$\xi_2 = 4$		$\Sigma_1 = 9$		$\Sigma_2 = 5$		Λ_1 =6		Λ_2 =3		MSE	
Method Missing Rate	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	ECM	KNN
0.12	11.952236	11.944910	4.334198	4.307269	10.050067	9.949571	3.617136	3.725617	4.143404	4.186692	1.790107	1.806239	0.00004411781	0.0000464651
0.20	11.940702	11.956399	4.296350	4.265956	10.357712	10.040118	3.681946	3.694449	3.763790	3.862462	1.750797	1.795295	0.00006534279	0.00006747916
N=800	$\xi_1 = 12$		$\xi_2 = 4$		$\Sigma_1 = 9$		$\Sigma_2 = 5$		Λ_1 =6		Λ_2 =3		MSE	
Method Missing Rate	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	ECM	KNN
0.12	11.941069	11.986914	4.286915	4.260068	10.134210	9.900777	3.917675	3.902419	4.398699	4.263697	2.044866	2.038809	0.00003719075	0.00003313492
0.20	11.956640	12.023029	4.496946	4.512534	10.623968	10.373587	3.457218	3.479675	3.464408	3.364448	1.378545	1.357354	0.00005422449	0.00005817202

ISSN (1681- 6870)

Table (2)

N=400	ξ_1 =20		ξ_2 =10		Σ_1 =8		Σ_2 =6		Λ_1 = 7		Λ_2 =5		MSE	
Method Missing Rate	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	ECM	KNN
0.12	20.026137	20.087361	10.138293	10.173210	8.692531	8.427851	5.575072	5.391737	3.837823	3.672001	2.410354	2.331494	0.0001403665	0.0001193867
0.20	20.032916	20.130257	10.082907	10.165533	8.214829	7.967263	5.245145	5.006123	3.136178	2.964337	2.156868	2.050509	0.0001656719	0.0001549199
N=614	ξ_1 =20		ξ_2 =10		Σ_1 =8		Σ_2 =6		Λ_1 = 7		Λ_2 =5		MSE	
Method Missing Rate	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	ECM	KNN
0.12	20.077233	20.069538	10.085809	10.104498	8.230865	8.120328	5.641420	5.569634	3.384672	3.574532	2.296233	2.381271	0.0000136643	0.00009924098
0.20	20.068452	20.061563	10.186722	10.228881	8.402103	8.278845	5.272329	5.241268	3.076861	3.233440	1.771233	1.801482	0.0001177308	0.0001179611
N=800	ξ_1 =20		ξ_2 =10		Σ_1 =8		Σ_2 =6		Λ_1 = 7		Λ_2 =5		MSE	
Method Missing Rate	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	MLE FOR ECM	MLE FOR KNN	ECM	KNN
0.12	20.041265	20.118810	10.174469	10.185622	8.570184	8.431594	5.466199	5.369759	3.264229	3.114218	1.875900	1.892097	0.0000521167	0.0000186532
0.20	20.038145	20.009213	10.129646	10.261498	8.646109	8.545193	5.420922	5.270984	3.310216	3.514357	2.249315	2.119847	0.00001803284	0.00002160294



N=416 ;Miss.=12% ; method: ECM

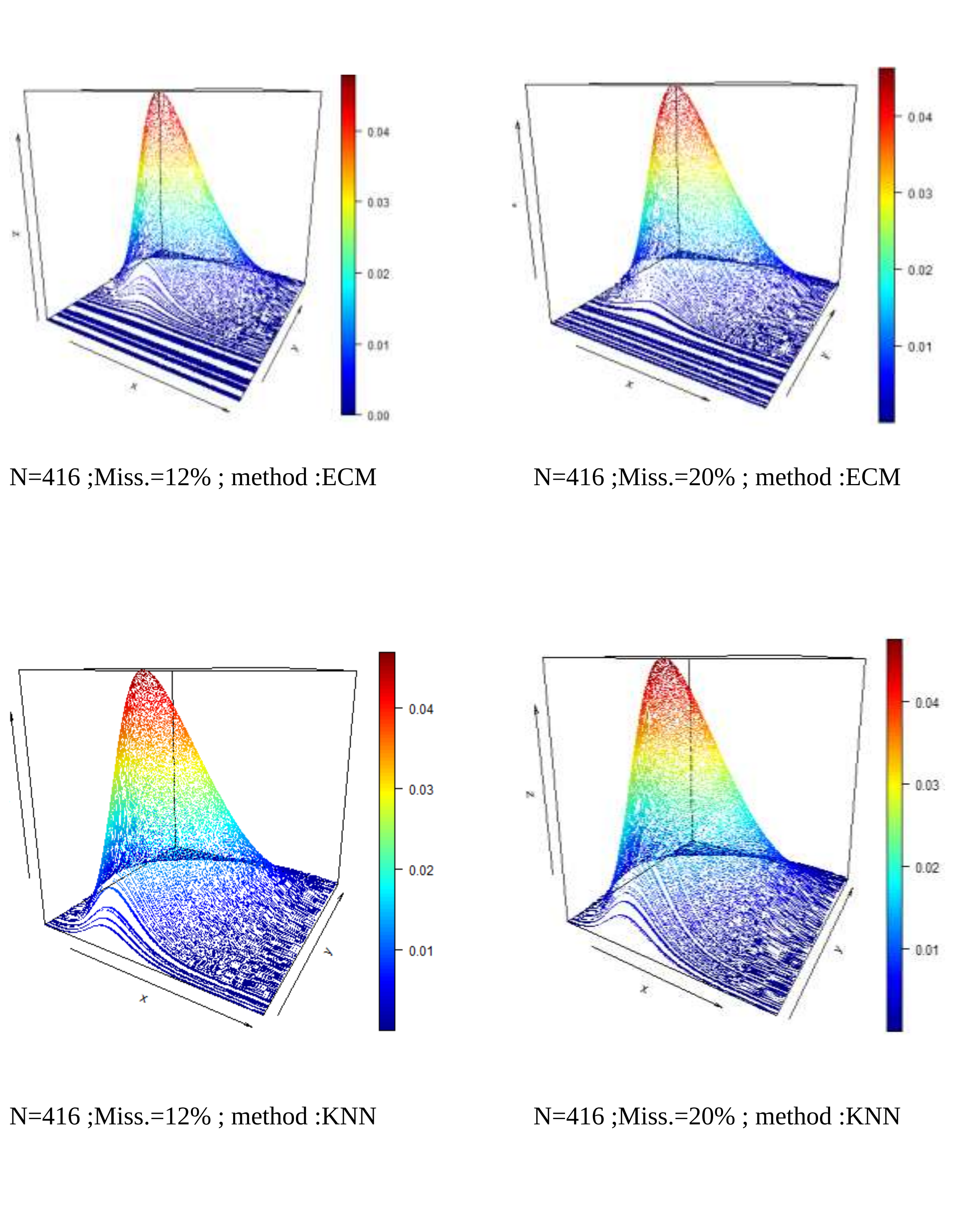
N=416 ;Miss.=20% ; method: ECM

N=416 ;Miss.=12% ; method: KNN

N=416 ;Miss.=20% ; method: KNN

Figure (6): hypothetical values for the parameters ($\xi_1 = 12, \xi_2 = 4, \Sigma_1 = 9, \Sigma_2 = 5, \Lambda_1 = 6, \Lambda_2 = 3$)

With Sample size 614



N=416 ;Miss.=12% ; method :ECM

N=416 ;Miss.=20% ; method :ECM

N=416 ;Miss.=12% ; method :KNN

N=416 ;Miss.=20% ; method :KNN

Figure (7): second hypothetical values for the parameters $(\xi_1 = 20, \xi_2 = 10, \Sigma_1 = 8, \Sigma_2 = 6, \Lambda_1 = 7, \Lambda_2 = 5)$ with Sample size 614

9. Conclusions:

In this paper, the problem of missing data for the multivariate skew normal model was studied, which is considered one of the major problems that many researchers encounter. The missing values were estimated using the (ECM) algorithm and the (KNN) method and a comparison was conducted between the two methods to figure out which one of them is better for the multivariate skew normal model and the following was attained:

1. In samples of size (400) or less, the (KNN) method was better than the (ECM) algorithm in estimating the missing values.
2. In samples of size between (500) and (700), the (ECM) algorithm was better than the (KNN) method in estimating the missing values.
3. In samples of size (800) or more, it was noticed that both of the methods were reasonably good, and the best method is determined according to the rate of missing values in the dataset and in this paper it was noticed that when the missing rate is at (12 %) the (KNN) method performs better than the (ECM) algorithm, while the (ECM) algorithm performs better than the (KNN) method when the missing rate is at (20 %).

المصادر العربية:

[1] القزار، قتيبة نبيل نايف (2007) ، **"مقارنة أساليب بيز الحصين مع طرائق اخرى لتقدير معالم انموذج الانحدار الخطي المتعدد في حالة البيانات غير التامة"**، أطروحة دكتوراه في الإحصاء ، كلية الإدارة والاقتصاد ، جامعة بغداد.

References:

[2] A. Azzalini, Capitanio A. , “ **Statistical applications of the multivariate skew-normal distribution**”, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 61 , Page 579_602 , (1999)

[3] Dempster A.P., Lard N.M , and Rubin, D.B. **“Maximum likelihood from incomplete data via the EM algorithm (with discussion)”** , Journal of the Royal Statistical Society: Series B (Statistical Methodology) , 39(1) , 1-38 ,(1977).

[4] Liu C. ,” **Efficient ML estimation of the multivariate normal distribution from incomplete data**”, J. Multivariate Anal. 69(2) , 206_217, (1999).

[5] Acuna E. , and Rodriguez C. , **“ The treatment of missing values and its effect in the classifier accuracy, Classification, Clustering and Data Mining"** In Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), pp.639-647, (2004).

[6] Beale E. ML, and Little R.JA., **“Missing values in multivariate analysis**”, Journal of the Royal Statistical Society Series B (Methodological): 37(1) , 129_146, (1975)

[7] Styan G.P.H. **, “Hadamard products and multivariate statistical analysis”** , Linear Algebra Appl. 6, Page 217_240, (1973).

[8] Malarvizhi R. , and Thanamani Antony Selvadoss **, ” K-Nearest Neighbor in Missing Data Imputation”** Volume 5, Issue 1 PP.05-07 , (2012).

[9] Troyanskaya Olga, Cantor Michael , Sherlock Gavin, Pat, Trevor Hastie, Robert Tibshirani , David Botstein and Russ B**.** Altman Brown,*“***Missing value estimation methods for DNA microarrays”,** Vol. 17 no.6 , (2001).

[10] R.R. Hocking, Wm.B. Smith ,” **Estimation of parameters in the multivariate normal distribution with missing observations”**, J. Amer. Statist. Assoc. 63 ,159_173 (1968).

[11] Arellano-Valle R.B. and Azzalini A., “**On the unification of families of skew-normal distributions**”, Scandinavian Journal of Statistics , 33(3) , 561_574. (2006).

[12] Arellano-Valle R.B., H. Bolfarine, and Lachos V.H. , “**Bayesian inference for skew-normal linear mixed models**” , Journal of Applied Statistics, 34(6) , 663_682, (2007).

[13] Little R. J., Rubin D. B., **“Statistical Analysis with Missing Data”**. thrid Edition, . John Wiley and Sons, New York, (2019).

[14] Sahu S.K., Dey D.K., and Branco M.D , " **A new class of multivariate skew distributions with applications to bayesian regression models** ", Canadian Journal of Statistics, 31(2), Page 129_150, (2003).

[15] Anderson T.W., **” Maximum likelihood estimates for a multivariate normal distribution when some observations are missing”**, Journal of the american Statistical Association, 52(278) , 200_203, (1957).

[16] Lin Tsung I., “**Maximum likelihood estimation for multivariate skew normal mixture models”** Journal of Multivariate Analysis 100(2), 257–265, (2009).

[17] Ghosh P., Branco M.D., and Chakraborty H.,” **Bivariate random effect model using skew-normal distribution with application to HIV-RNA** ”, Statistics in medicine, 26(6), Page 1255_1267, (2007).

[18] Styan G.P.H. , “**Hadamard products and multivariate statistical analysis**” , Linear Algebra Appl.6 , 217_240 ,(1973).