

Missing Data Imputation in a t -Distribution with Known Degrees of Freedom Via Expectation Maximization Algorithm and Its Stochastic Variants

P.K. Kinyanjui¹, C.L. Tamba^{1*} and J.O. Okenye¹

¹Mathematics Department, Egerton University, Njoro,
536-20115, Egerton (Kenya).

Emails: kinyanjui.access@gmail.com, cox.tamba@egerton.ac.ke, justinoke69@gmail.com

ABSTRACT

During the data collection process, the field personnel encounters some inevitable challenges, making it difficult to gather all the required information. Since most of the statistical tools assume complete data, a data analyst is prompted to pre-treat the data before analyzing it. Model-based imputation techniques are widely used because they utilize all the collected information and preserve the original distribution of the data. In this paper, missing data imputation in multivariate t -distribution with known degrees of freedom using expectation maximization (EM), Stochastic EM (SEM) and Monte Carlo EM (MCEM) algorithms is considered. The method of mean square error (MSE) is used to compare the imputation efficiencies of the three procedures. In this study, the SEM technique yields the most efficient imputations when compared to EM and MCEM methods. Throughout all the iterations, SEM explores a wide log-likelihood region, thereby enabling it to avoid local maximal traps and converge at the most optimal points possible. Further, the efficiency of the three methods improves with higher correlation coefficients between variables.

Keywords: expectation maximization (EM), stochastic EM, Monte Carlo EM, missing values, imputation

Mathematics Subject Classification: 62H12

1. INTRODUCTION

The missing data problem is a common phenomenon in research studies. In meteorological studies for instance, instruments may fail as a result of extreme weather conditions. Such a situation may inhibit the collection of all information regarding the measurement of interest. In a survey, missing values may be caused by recording errors, deliberate or accidental omission of some questions on the part of respondents, and the dropping out of some respondents before completion of the study.

Missing values present significant hurdles in the field of research particularly because incomplete data sets cannot be analyzed directly using most of the statistical approaches. It is salient to find an appropriate way of dealing with the problem before proceeding with the actual data analysis. Salgado *et al.* (2016), Little and Rubin (2014), McLachlan and Krishnan (2007), and Allison (2002) have comprehensively discussed the techniques that are widely used to handle incomplete data and or find

the estimates for the missing values. Some of the techniques discussed include deletion methods, substitution techniques, and model-based imputation approaches.

The efficiency of the various techniques for handling missing data problems significantly depends on the missing data mechanism (MDM). The main MDMs are missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR) (Rubin, 1976). In practice, deletion methods require that the data is MCAR (Little, 1992; Garg, 2013). Model-based techniques work well when the data is MAR.

The model-based approaches are considered most powerful particularly because they make use of all the available data and also estimate the errors due to imputation of the missing values (Little and Rubin, 2014). Nevertheless, these methods may prove inefficient in situations where the observed and the missing values of a particular dataset are independent. One of the commonly employed model-based imputation techniques is the expectation maximization (EM) algorithm. It is a maximum likelihood (ML) method that facilitates the imputation of missing values and estimation of parameters from the resulting complete dataset.

The multivariate t -distribution may be required for robust inferences in cases where incomplete datasets are characterized by heavy tails and or outliers (Liu, 1995). This paper examines the efficiency of the EM technique and two of its stochastic versions, the stochastic EM (SEM) and Monte Carlo EM (MCEM), in missing data imputation for datasets assuming the multivariate t -distribution with known degrees of freedom. The data used are assumed to follow an MAR mechanism. The mean square error (MSE) metric is used to assess the efficiencies of the three imputation techniques.

The subsequent sections of the paper are as follows: the methods used in the study are presented in section 2. Results are presented in section 3. Section 4 discusses the obtained results. Section 5 consists of a conclusion for the study.

2. METHODS

2.1. Missing Data Imputation Using EM and Its Stochastic Variants

The EM technique was first formulated by Dempster *et al.* (1977). It is a deterministic approach that iteratively imputes missing values and estimates parameters of the distribution of interest in two phases namely the expectation (E) step and maximization (M) step. The method uses all the available information and does not alter the original distribution of the data. During the E-step, the missing values for every observation are imputed given the observed variables (McLachlan and Krishnan, 2007). The M-step then updates the parameter values of the data model. The EM technique iterates between the two steps until convergence has been reached.

In the multivariate t -distribution context, at any t^{th} iteration, the E-step is implemented as follows:

$$Y_{imp}^{(t)} = E(Y_{mis}^{(t)} | Y_{obs}; \theta^{(t)}) = E(Y_{mis}^{(t)} | Y_{obs}; \mu^{(t)}, \Sigma^{(t)}, \nu) = \mu_{mis}^{(t)} + \Sigma_{mis, obs}^{(t)} (\Sigma_{obs}^{(t)})^{-1} (Y_{obs} - \mu_{obs}^{(t)}) \quad (1)$$

where Y_{mis} are the missing values, Y_{imp} are the corresponding imputed values, and Y_{obs} are the observed values. $\theta = (\mu, \Sigma, \nu)$ in which case μ is the location vector, Σ is the scatter matrix (scale parameter), and ν denotes the degrees of freedom (Liu, 1995). The incomplete dataset can be partitioned as $Y = (Y_{mis}, Y_{obs})$ with corresponding location vector and scatter matrix being partitioned as $\mu = \begin{bmatrix} \mu_{mis} \\ \mu_{obs} \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{mis} & \Sigma_{mis,obs} \\ \Sigma_{obs,mis} & \Sigma_{obs} \end{bmatrix}$ respectively. The degrees of freedom for the two partitions of the data remain unchanged.

The algorithm's M-step updates the parameter estimates given the resulting augmented data set as follows:

$$\hat{\mu}^{(t+1)} = \frac{\sum_{i=1}^n W_{i,obs}^{(t+1)} Y_i^{(t)}}{\sum_{i=1}^n W_{i,obs}^{(t+1)}}; i = 1, \dots, n \quad (2)$$

$$\hat{\Sigma}^{(t+1)} = \frac{1}{n} \sum_{i=1}^n W_{i,obs}^{(t+1)} (Y_i^{(t)} - \hat{\mu}^{(t)})(Y_i^{(t)} - \hat{\mu}^{(t)})^T + \frac{1}{n} \psi_i^{(t)} \quad (3)$$

where

For every entry in the i^{th} unit and j^{th} variable;

$$Y_{ij}^{(t)} = \begin{cases} Y_{ij}, & \text{if } Y_{ij} \text{ is observed} \\ E(Y_{ij} | Y_{obs,i}; \theta^{(t)}), & \text{if } Y_{ij} \text{ is missing} \end{cases}$$

$$\psi_i^{(t)} = \begin{cases} 0, & \text{if } Y_{ij} \text{ or } Y_{ik} \text{ is observed} \\ \Sigma_{i,mis}^{(t)} - \Sigma_{i,mis,obs}^{(t)} (\Sigma_{i,obs}^{(t)})^{-1} \Sigma_{i,obs,mis}^{(t)}, & \text{if } Y_{ij} \text{ and } Y_{ik} \text{ are missing} \end{cases}$$

$W_{i,obs}^{(t+1)} = E(\tau_i | \theta^{(t)}, Y_{i,obs}) = \frac{\nu + p_i}{\nu + \delta_{i,obs}^{(t)}}$, that is, the weights associated with each i^{th} unit.

$\delta_{i,obs}^{(t)} = \delta_{i,obs} = (Y_{i,obs} - \mu_{i,obs})^T \Sigma_{i,obs}^{-1} (Y_{i,obs} - \mu_{i,obs})$, that is, the Mahalanobis distance.

Initially, the parameter values are fixed. The EM technique then updates these values in the subsequent iterations. The algorithm iterates between the E-step and the M-step until it achieves convergence of the imputed values and the parameter estimates (Meng, 1994.; Wu, 1983). Convergence takes place once the difference $L(\theta^{i+1}) - L(\theta^i) \leq \delta$, where δ is a set maximum tolerance level.

In SEM, the missing values are not replaced by expected values of Y_{mis} given the current parameter estimates and the observed data. Instead, the algorithm fills in each missing position with a single draw from the posterior distribution of the unobserved values given the observed data values and the current parameter estimates, that is, $Y_{imp}^{(t)} \sim g(Y_{mis}/Y_{obs}; \theta^{(t)})$ (Nielsen, 2000). The set of parameter values that maximize the augmented data log likelihood function are then computed in the M-step. Convergence is expected to occur as the number of iterations increase provided the algorithm is ergodic.

The MCEM method is implemented in a similar manner as the SEM. However, instead of a single simulation during the E-step, multiple random samples, say N , are drawn from the conditional distribution of the missing values given the observed data values and the updated parameter estimates

(Levine and Casella, 2001; Jank, 2005). Therefore, $Y_{imp}^{*(t)} = \frac{1}{N} \sum_{i=1}^N U_i^{(t)}$, where $U_1^{(t)}, \dots, U_N^{(t)}$ constitute samples drawn from the conditional distribution of the unobserved data given the current parameter estimates and the observed data (i.e. $U_i^{(t)} \sim g(Y_{mis}/Y_{obs}; \theta^{(t)})$). The multiple samples are independently and identically distributed (IID). The average values of the N random samples are then used to fill in the missing positions. The M-step then computes $\theta^{(i+1)}$ that maximizes the average of the N generated complete data log likelihood functions.

2.2. Simulation Study

Data of dimension $n \times p = 25 \times 3$ is simulated from a multivariate t -distribution with fixed degrees of freedom using the *mvtnorm* package in *R* statistical software. 20% of the data is then removed at random, resulting in 6 missing values. The EM, SEM, and MCEM techniques are then used to perform data imputation. In order to determine the effect of correlation between variables on the imputation efficiency of the three techniques, different values of correlation coefficients ($\rho = 0.3$, $\rho = 0.5$, and $\rho = 0.9$) are used. The scatter matrices used for simulation are constructed using the relation $\Sigma = D^{\frac{1}{2}} R_h D^{\frac{1}{2}}$; where D is a diagonal matrix containing the diagonal elements of the scatter matrix and R_h is the correlation coefficient matrix. In all cases, the diagonal elements for the scatter matrix used were (435 500 398). Throughout the study, the location vector $\mu = (93 \ 139 \ 188)$ and degrees of freedom $\nu = 3$ are arbitrarily used for simulation. The three imputation procedures are run over 3,000 iterations. The starting values for the three imputation techniques are arbitrarily set as $\mu_{stat} = (10, \ 20, \ 30)$, $\Sigma_{stat} = \begin{bmatrix} 14 & 10 & 12 \\ 10 & 13 & 9 \\ 12 & 9 & 18 \end{bmatrix}$, and $\nu = 3$. In the EM technique, the tolerance level for convergence is set at $L(\theta^{(t+1)}) - L(\theta^{(t)}) \leq 0.00001$. In the stochastic methods, SEM and MCEM, the first 100 iterations are burnt-in in order to eliminate the effects of the initial values. During MCEM's E-step, we draw $N = 1,000$ values from the conditional trivariate t distribution for each of the missing values. In each iteration, the resulting average values for the drawn samples fill in the respective missing positions.

In this study, chains of imputed values for the three imputation procedures are plotted. The points of convergence are then compared with the true data points. The denotations ASW72YChain[i, j], ASW71YChain[i, j], and ASW711YChain[i, j] depict the chains for the imputed values at row i , column j in the case of EM, SEM, and MCEM respectively. The 'Index' notation indicates the number of iterations.

3. RESULTS

Figures (1), (2), and (3) depict the convergence for the imputed values using the EM algorithm when the correlation coefficients for the simulated data are 0.3, 0.5, and 0.9 respectively. The red lines indicate where the true values for the missing positions lie.

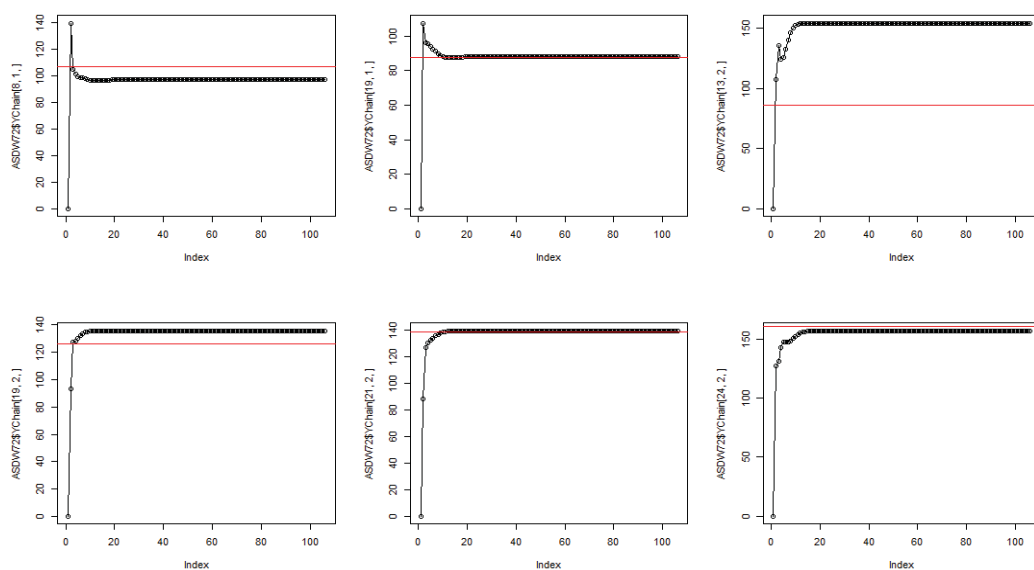


Figure 1: Convergence of imputed values when the correlation is 0.3 (EM)

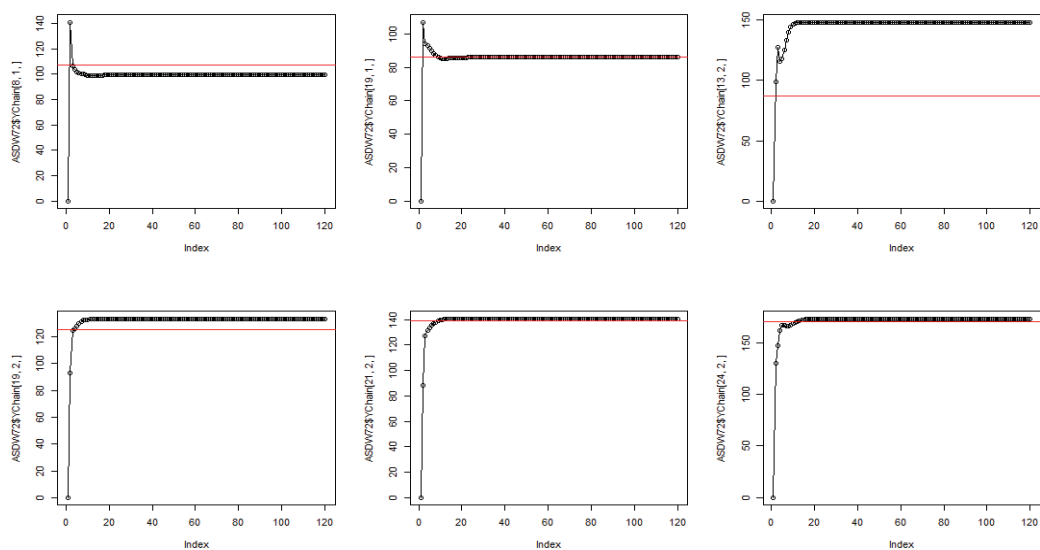


Figure 2: Convergence of imputed values when the correlation is 0.5 (EM)

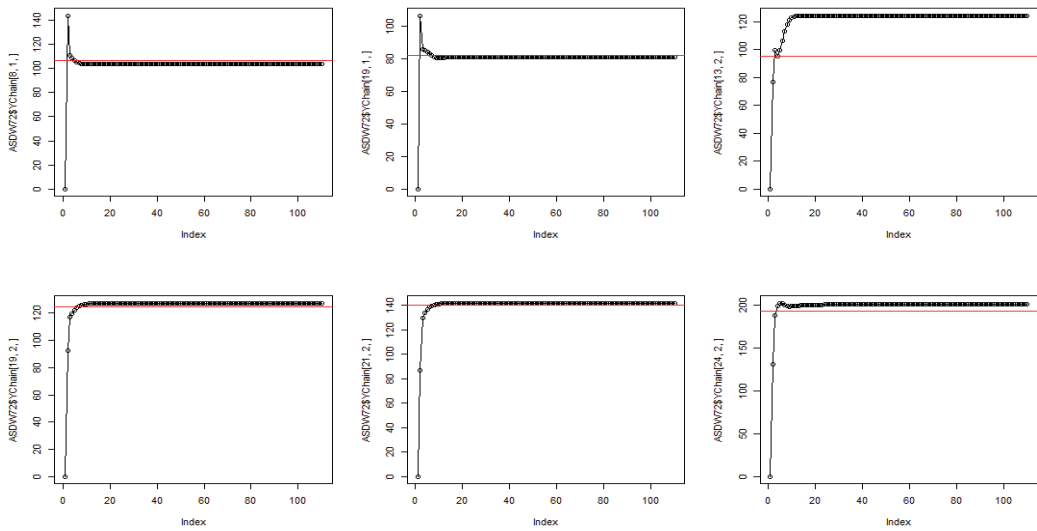


Figure 3: Convergence of imputed values when the correlation is 0.9 (EM)

The figures show that most of the imputed values lie close to the true values. This is indicated by the close proximities between the red lines and the chains for the imputed values. In all cases, only the imputation for the 13th row, 2nd column is relatively far from the true value.

Figures (4), (5), and (6) show the convergence for the imputed values using the SEM algorithm when the correlation coefficients are 0.3, 0.5, and 0.9 respectively.

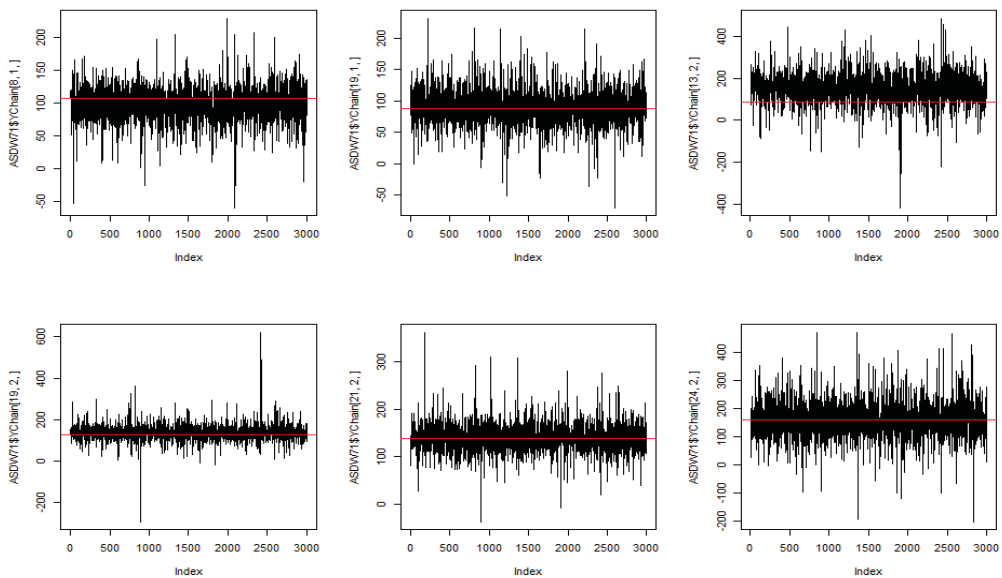


Figure 4: Convergence of imputed values when the correlation is 0.3 (SEM)

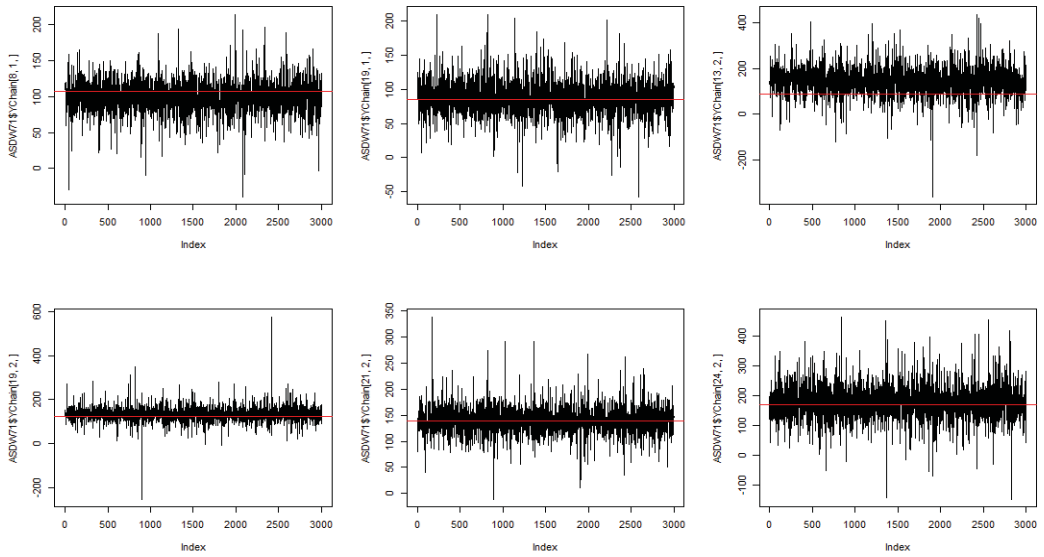


Figure 5: Convergence of imputed values when the correlation is 0.5 (SEM)

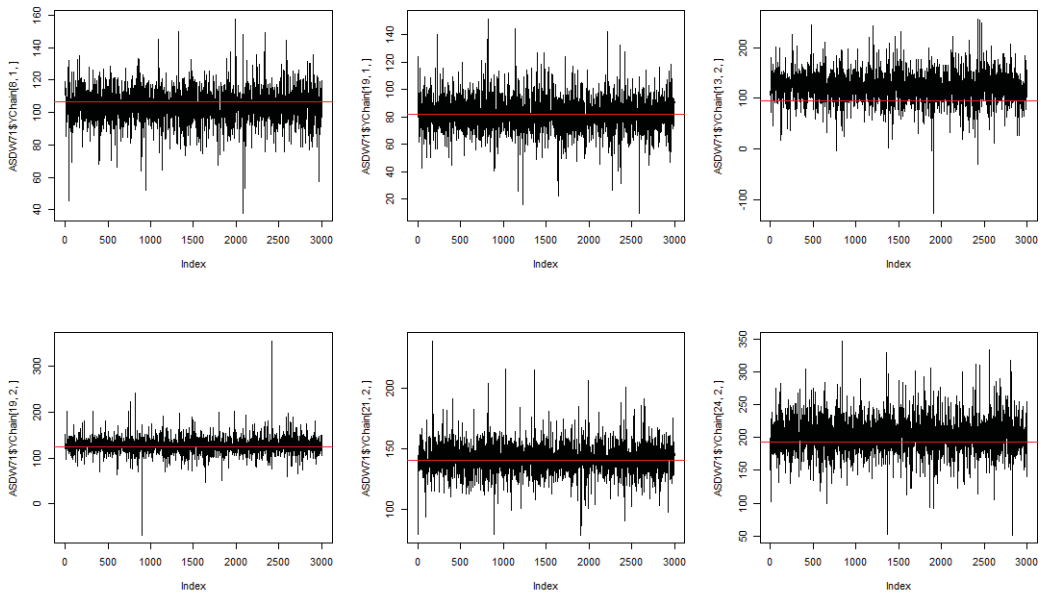


Figure 6: Convergence of imputed values when the correlation is 0.9 (SEM)

Figures (4), (5), and (6) show that the SEM iterative procedure has converged in distribution. In all the trace plots for the imputed values, it can be seen that the true values lie within the regions created by the chains of imputed values over all the iterations.

Figures (7), (8), and (9) depict the convergence for the imputed values using the MCEM algorithm when the correlation coefficients are 0.3, 0.5, and 0.9 respectively.

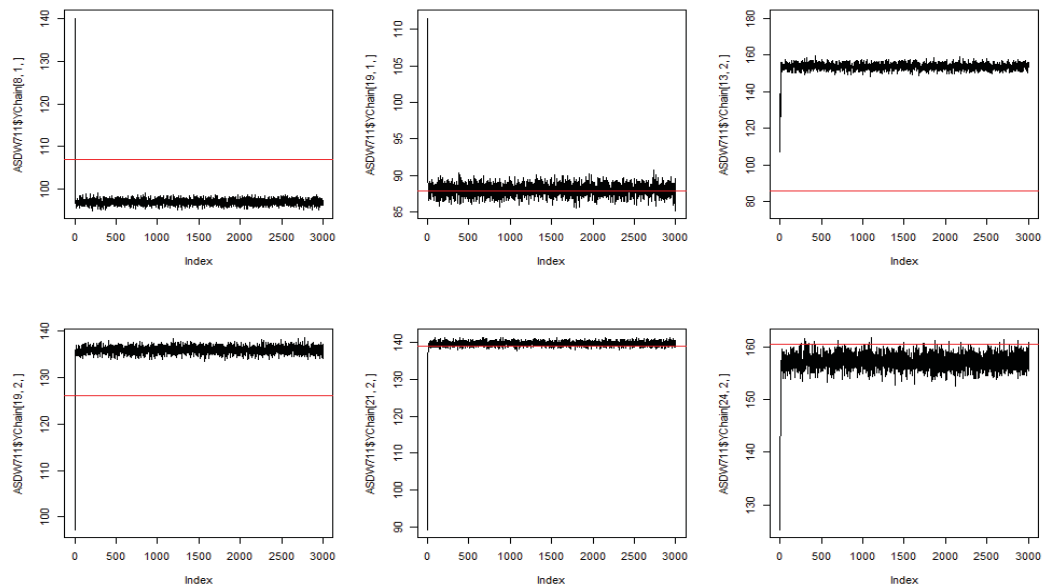


Figure 7: Convergence of imputed values when the correlation is 0.3 (MCEM)

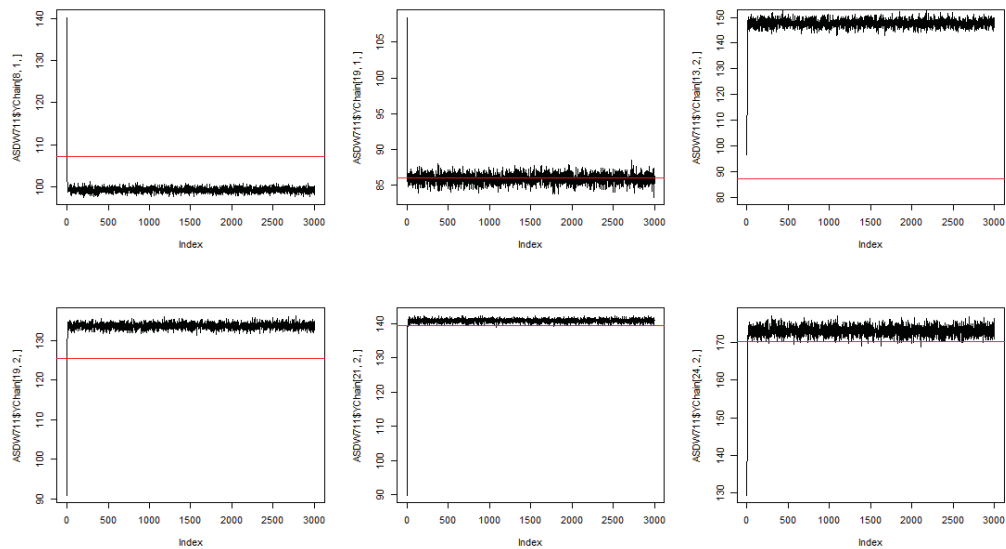


Figure 8: Convergence of imputed values when the correlation is 0.5 (MCEM)

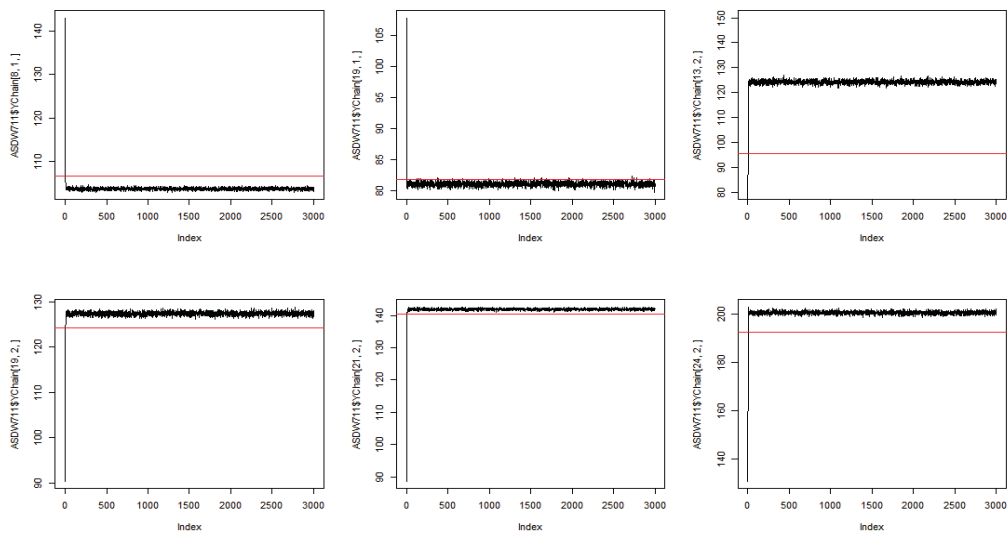


Figure 5: Convergence of imputed values when the correlation is 0.9 (MCEM)

Based on figures (7), (8), and (9) above, the imputed values have converged in distribution over the various iterations. Most of the estimates are also relatively close to the true values. However, the chains are shrunk, implying that the region of exploration across the iterations is relatively narrower as compared to the case of the SEM method.

Table 1 below shows the MSE values for the three imputation procedures when the correlation coefficients are 0.3, 0.5, and 0.9 respectively.

Table 1: MSE Values for EM, SEM, and MCEM with $\rho = 0.3$, $\rho = 0.5$, and $\rho = 0.9$ respectively.

Method	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.9$
EM	804.0973	633.7757	151.3958
SEM	733.5204	571.3648	137.3529
MCEM	803.7219	633.4792	151.343

From the table above, it can be seen that in all the imputation procedures, the MSE values decrease with increase in the correlation coefficients between variables. The SEM method yields the lowest MSE values in all the correlation levels considered.

4. DISCUSSION

The EM method is widely used in parameter estimation when the data under consideration contains missing values. The technique is also used in missing data imputation. Throughout the iterations, EM ensures that the likelihood function does not decrease, which facilitates the monotonic convergence of parameter estimates as well as imputed values (Jank, 2006). Furthermore, because the technique works on the log-scale, it facilitates the simplification of many models, particularly those in the exponential family.

The EM method may not be feasible in cases where the log-likelihood function is intractable. Additionally, the technique may be trapped in local saddle points particularly in complex data models. Such a shortcoming may inhibit its efficiency in data imputation as well as parameter estimation. The problem can be addressed by using different starting values when running the algorithm (Gupta and Chen, 2011). However, this approach may not necessarily be sufficient. Stochastic variants to the EM have been proposed as alternatives. Two of the widely used stochastic techniques are the SEM and the MCEM methods. Unlike the EM method which is deterministic in nature, the stochastic variants randomly sample values from the posterior distribution of the missing data given the observed values. The ability of the iterative methods to search over a wide region of possible values in the “filling in” process ensures that the global maximal points are achieved.

In this study, the main focus is on the efficiency of the EM method and its stochastic variants in recovering missing information in data assuming the multivariate t -distribution. In this regard, simulation studies are employed. From the results presented in the previous section, it can be observed that the efficiency of the imputations in a particular dataset is partly dependent on the correlation coefficients between variables. If the scatter matrix yields low correlation coefficients, then the imputations are poorer and the MSEs absurdly huge. On the contrary, a scatter matrix that yields relatively high correlation coefficients leads to better imputations and considerably smaller MSEs in all the imputation procedures. This can be explained by the fact that the methods assume that the missing values are response variables; dependent on the observed values. For more accurate results, therefore, the missing variables should be highly correlated with the observed variables. These results are consistent with Richman *et al.* (2009) findings.

Notably, the SEM algorithm has many adorable characteristics. In most cases, the algorithm achieves highly efficient imputations, which can be justified by the lower MSE values as compared to the other imputation procedures under consideration; that is, the EM and the MCEM. The wide regions explored by the SEM's trace plots make it possible for the chains of imputed values to include the true values in most of the considered cases.

When the sample size being simulated in the MCEM's E-step is relatively large, the algorithm behaves more like the EM procedure ((Biscarat *et al.*, 1992; Nielsen, 2000). Therefore, the imputations for the two methods are almost similar; which is evidenced by the close MSE values displayed in table 1. A data analyst can therefore use either of the algorithms to achieve more or less the same results.

However, he or she must ensure that the EM algorithm will not be locked in local maximal traps before using it. In case there is no guarantee, it would be advisable to use the MCEM algorithm instead.

5. CONCLUSION

The paper has considered the imputation of missing values using EM, SEM, and MCEM in multivariate t context when the degrees of freedom are known. The three imputation procedures employed are considerably efficient in recovering lost information. SEM is the most efficient imputation technique among the three algorithms considered in this study. Large number of simulations in MCEM's E-step yields results similar to those obtained when the EM method is employed. In cases where the conditional expectation of the log likelihood function under consideration is not in closed-form, MCEM would come in handy given its ability to approximate the expectation through simulation. The SEM method would also be an ideal alternative. Additionally, the imputations of the EM, SEM, and MCEM algorithms are more accurate when the correlation coefficients between data variables are relatively high. It would be interesting to assess the efficiency of the three imputation procedures in the multivariate t context with unknown degrees of freedom. This will be handled in our next issue.

ACKNOWLEDGEMENT

C.-L.T. and J.-O.O. are hereby acknowledged for the conceptualization and supervision of the study. The Egerton University Council Scholarship is also recognized for the partial funding of this research work.

REFERENCES

- Allison, P. D., 2002, *Missing data*. Thousand Oaks, CA: Sage.
- Biscarat, J. C., Celeux, G., and Diebolt, J., 1992, *Stochastic versions of the EM algorithm* (No. TR-227). Washington University Seattle Department of Statistics.
- Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.
- Garg, P., 2013, Robustness of multiple imputation under missing at random (MAR) mechanism: A simulation study.
- Gupta, M. R., and Chen, Y., 2011, Theory and Use of the EM Algorithm. *Foundations and Trends® in Signal Processing*, 4(3), 223-296.
- Jank, W., 2005, Stochastic Variants of EM: Monte Carlo, Quasi-Monte Carlo and More. In *Proceedings of the American Statistical Association*.
- Jank, W., 2006, The EM algorithm, its randomized implementation and global optimization: Some challenges and opportunities for operations research. In *Perspectives in operations research* (pp. 367-392). Springer, Boston, MA.
- Levine, R. A. and Casella, G., 2001, Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3), 422-439.
- Little, R. J., 1992, Regression with missing X's: a review. *Journal of the American Statistical Association*, 87(420), 1227-1237.

- Little, R. J., and Rubin, D. B., 2014, *Statistical analysis with missing data* (Vol. 333). John Wiley & Sons.
- McLachlan, G., and Krishnan, T., 2007, *The EM algorithm and extensions* (Vol. 382). John Wiley & Sons.
- Liu, C., 1995, Missing data imputation using the multivariate t distribution. *Journal of multivariate analysis*, 53(1), 139-158.
- Meng, X. L., 1994, On the rate of convergence of the ECM algorithm. *The Annals of Statistics*, 22(1), 326-339.
- Nielsen, S.F., 2000, The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*, 6(3), 457.
- Richman, M. B., Trafalis, T. B., and Adrianto, I., 2009, Missing data imputation through machine learning algorithms. In *Artificial intelligence methods in the environmental sciences* (pp. 153-169). Springer, Dordrecht.
- Rubin, D. B., 1976, Inference and missing data. *Biometrika* 63 (3), 581–592.
- Salgado, C.M., Azevedo, C., Proença, H., and Vieira, S.M., 2016, Missing Data. In: *Secondary Analysis of Electronic Health Records*. Springer, Cham.
- Wu, C. J., 1983, On the convergence properties of the EM algorithm. *The Annals of statistics*, 95-103.