Imputation for Non-Normal Multivariate Continuous Data using Copula Transformation

Zhixin Lun* Ravindra Khattree †

Abstract

Dealing with missing data problems for skewed data is one of the most difficult tasks in imputation since most of data augmentation methodologies assume multivariate normality. The performance of imputation and the accuracy of parameters inference become questionable when the violation of above assumption occurs. One approach to solve the normality violation is to apply normalizing transformation prior to the imputation phase. However, this approach may introduce new problems such as altering dependence structure among random variables. This article describes the multiple imputation approach based on the Copula transformation, which we use to effectively transform multivariate non-normal data into normal. We compare the performance of the Copula transformation method with traditional normality-based multiple imputation approaches through real non-normal multivariate datasets. We demonstrate that our approach significantly mitigates the impact of blind assumption of multivariate normality for the non-normal multivariate data under the scenario when the data are missing completely at random (MCAR).

Key Words: Copula transformation, Gaussian Copula, Missing data, Multiple Imputation, Skewed data

1. Introduction

Imputation has been treated as a flexible and effective method to handle missing data problems since it utilizes all the available information in the data. Instead of deleting incomplete cases, we fill in some plausible values for the missing data so that the standard complete-data analysis methods are still largely applicable. The first suggestion in the very early years was to replace a missing value by corresponding average, but it was soon realized that doing so results in an underestimation of variability. Among various imputation methods, multiple imputation (MI), first proposed by Rubin (1977), has been widely used for the last few decades as this method enables us to provide valid inference through repeated imputation. However, most of multiple imputation methods, such as Markov chain Monte Carlo (MCMC) (Schafer, 1997) and fully conditional specification (FCS) (van Buuren et al., 2006; van Buuren, 2007) methods using linear regressions, assume that the data are from a multivariate normal distribution. The multivariate normality (MVN) assumption facilitates us to conveniently impute plausible values since the conditional distribution of the missing data given the observed data is also multivariate normal. However, the violation of MVN assumption may result in grossly inaccurate imputations and hence provide inference about parameters which may be utterly invalid. Case in point is the imputation of nonnegative data based on MVN assumption where negative imputed values, may inevitably be produced.

In order to deal with the missing value problem for the non-normal data, a common approach is to individually apply transformation on each variable so that each of the corresponding marginal distribution is approximately normal. Such a method may not work effectively, especially the highly skewed data. Enders (2010) specifically points out two

*Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309

†Department of Mathematics and Statistics and Center for Data Science and Big Data Analytics, Oakland University, Rochester, MI 48309

main potential problems resulting from such data transformations namely, (i) difficulty in choosing an appropriate transformation and (ii) the drastic change in the correlation or more generally the dependence structure after the transformation. An ideal solution must remain, as much as possible, the original association structure among all the random variables.

Bahuguna and Khattree (2020) provide and illustrate via a number of examples a generic all purpose multivariate transformation based on copula. The advantage of the approach is that, under mild assumptions, by using Gaussian copula, any skewed multivariate data can be transformed to data having multivariate normal distribution without losing the dependence structure among the random variables. The objective of our work is to explore multiple imputation possibilities based on copula-transformation methods, formalize it and then evaluate its performance in terms of imputation discrepancy and parameter estimation.

This article is organized as follows. First, we revisit in Section 2 the basic concept of copula and the Sklar's theorem (Sklar, 1959) and illustrate the implementation of copula transformation in missing data imputation. A large simulated data and a real data set is considered next in Sections 3 and 4 respectively to illustrate how the copula-transformed approach performs superiorly under the missing completely at random (MCAR) missing mechanism. Few general remarks are made at the conclusion of the article.

2. Copulas and copula transformation

We here introduce an approach for imputation by using the copula transformation which, using the Gaussian copula effectively normalizes the data and retains its dependence structure. To set the stage with a context, we first revisit the definition of copula and an important result – Sklar's theorem, which is the basis for the copula transformation, we reply on.

2.1 Copula function and Sklar's theorem

Copula is a multivariate probability distribution function where each marginal probability distribution is uniform. Formally speaking, a function C is a d-dimensional copula if there is a random vector $\mathbf{U} = (U_1, U_2, \ldots, U_d)'$, such that for $i = 1, \ldots, d$, $U_i \sim \text{Uniform}(0, 1)$, and

$$C(u_1, u_2, \ldots, u_d) = P[U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d].$$

The most central theorem in copula theory is the Sklar's theorem (Sklar, 1959), which is stated as follows.

Theorem 3.1 (Sklar's Theorem): A function $F : \mathbf{R}^d \to [0, 1]$ is the distribution function of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)'$ if and only if there is a copula C from $[0, 1]^d$ to $[0, 1]$ and d univariate distribution functions $F_1, F_2, \ldots, F_d$ such that

$$C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) = F(x_1, x_2, \ldots, x_d).$$

If the marginals F_i are continuous, then $C(\cdot)$ is unique.

This theorem indirectly implies that vectors from two different continuous multivariate distributions can be transformed to each other provide they share the same copula. Specifically, consider two different continuous multivariate cumulative distributions denoted by $F(\cdot)$ and $G(\cdot)$ and assume that they share a common copula. Then,

$$\begin{aligned} F(x_1, x_2, \ldots, x_d) &= C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) \\ &= C(u_1, u_2, \ldots, u_d) \\ &= G(G_1^{-1}(u_1), G_2^{-1}(u_2), \ldots, G_d^{-1}(u_d)) \\ &= G(s_1, s_2, \ldots, s_d), \end{aligned} \tag{1}$$

where $F_i(\cdot)$ and $G_i(\cdot)$ are the corresponding marginal cumulative distribution functions to $F(\cdot)$ and $G(\cdot)$, respectively. Thus, a set of data on $(x_1, \ldots, x_d)$ can be transformed as $(s_1, \ldots, s_d)$ and vice versa, via the multivariate uniform variable $(u_1, \ldots, u_d)$. The process of copula transformation is pictorially depicted in Figure 1.

It must be noted that there are two important components to be estimated in the process of multivariate variables transformation according to common copula in (1): (i) the marginal distribution functions F_i's and G_i's and (ii) copula function $C(\cdot)$. Let $\mathbf{x}_i' = (x_{1i}, x_{2i}, ..., x_{ni})$ be the observed values for a random sample for the i-th variate X_i. We nonparametrically estimate the marginal cumulative distribution function $F(t)$ by using empirical distribution function as

$$\hat{F}_i(t) = \frac{1}{n+1} \sum_{k=1}^{n} \mathcal{I}[x_{ki} \leq t],$$

where $\mathcal{I}[\cdot]$ is an zero-one indicator function. Accordingly, the corresponding uniformly distributed sample (often called pseudo-observations) $\mathbf{u}_i' = (u_{1i}, u_{2i}, ..., u_{ni})$ is obtained by

$$u_{ji} = \hat{F}_i(x_{ji}) = \frac{1}{n+1} \sum_{k=1}^{n} \mathcal{I}[x_{ki} \leq x_{ji}] \quad \text{for } j = 1, \cdots, n. \tag{2}$$

Since our purpose is to *normalize* the multivariate variables, we assume that the common copula between original multivariate random vector and the transformed one is a Gaussian copula $\Phi_{\mu,\Sigma}(\cdot)$. Accordingly, we specify each marginal distribution G_i as standard normal, $i = 1, \cdots, d$. That is,

$$\begin{aligned} F(x_1, \ldots, x_d) &= C_\Sigma(u_1, \ldots, u_d) \\ &= \Phi_{\mu,\Sigma}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)) \\ &= \Phi_{\mu,\Sigma}(s_1, \ldots, s_d) \end{aligned}$$

where $\Phi_{\mu,\Sigma}(\cdot)$ is the cumulative distribution function of a multivariate normal vector with mean vector μ and covariance matrix Σ. $\Phi(\cdot)$ is the cumulative distribution function of the standard univariate normal and $\Phi^{-1}(\cdot)$ is its inverse function.

According to the Sklar's theorem the copula $C(\cdot)$ is uniquely determined by the multivariate distribution function if all its marginal distribution functions are continuous. The underlying copula of a multivariate data may not be the Gaussian copula; however, we can use Gaussian copula to approximately construct data $(S_1, \ldots, S_d)$ with the same dependence structure as for the original raw data $(X_1, \cdots, X_d)$.

An important point must be made. Note that, $(X_1, \cdots, X_d)$, $(U_1, \cdots, U_d)$ and $(S_1, \cdots, S_d)$ all have the same rank-correlation matrix. Thus if τ_{ij} is the (rank) correlation between u_i

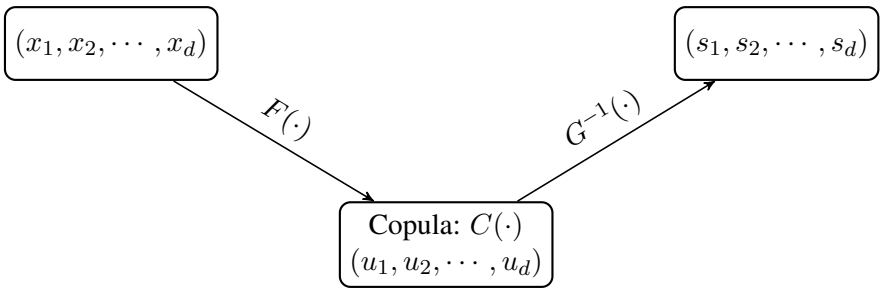


Figure 1: Directional diagram of copula transformation.

and u_j then the Pearson's correlation ρ_{ij} between Y_i and Y_j is given by

$$\rho_{ij} = 2\sin\left(\frac{\pi}{6}\tau_{ij}\right), \quad i,j = 1,\cdots,d, \quad i \neq j. \tag{3}$$

The derivation of the above equation can be found in Meyer (2013). Since we plan to transform data to $(S_1,\cdots,S_d)$, each having a marginal distribution as standard normal, we hence assume that $\mu = 0$ and the assumed variance covariance matrix is $R = (\rho_{ij})$.

3. An illustration using simulated data

To illustrate the performance of copula transformation for skewed data, we here use a sample data of size $n = 10,000$ from a bivariate Lomax (Pareto Type II) distribution (Lindley and Singpurwalla, 1986) for which the probability density function is given by,

$$f(x_1,x_2) = \frac{\theta_1\theta_2\beta(\beta+1)}{(1+\theta_1x_1+\theta_2x_2)^{\beta+2}}, \quad x_1,x_2 > 0, \quad \beta,\theta_1,\theta_2 > 0.$$

For our simulation, we assume $\beta = 3.1$, $\theta_1 = 0.5$ and $\theta_2 = 1.5$. The skewness of above distribution depends only on parameter β. With $\beta = 3.1$, data are highly skewed. Plot in Figure 2 (a) presents this highly skewed data.

We may also transform the data to symmetry by log-transformation or more generally by Box-Cox transformation (Box and Cox, 1964). The empirically obtained optimum choice for the latter is $\lambda_1 = 1.54 \times 10^{-5}$ and $\lambda_2 = 1.89 \times 10^{-5}$ (by using the maximum likelihood estimation). As these power parameters are very close to zero, the Box-Cox transformed data behave very similar to the logarithm transformed data. See the two scatter plots in Figure 2 (b) and (c). However, the effects of either of these two transformations are not very satisfactory since these corresponding scatters do not seem to be elliptic thereby suggesting that the joint distribution of transformed data may not be sufficiently close to bivariate normal. The univariate histograms, also shown in the same figures also indicate that the marginal distributions are somewhat skewed and are not quite close to normal distribution. In contrast, the scatter of copula transformed data in Figure 2 (d) exhibits an approximate ellipse-shape. The marginal distributions of both transformed variables also exhibit symmetry in their empirical histograms.

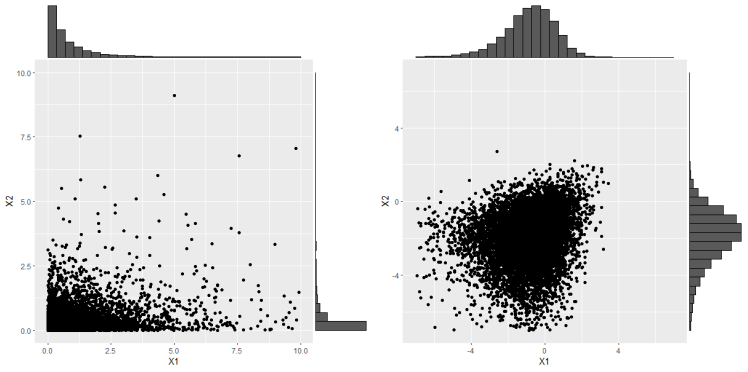
We now consider the missing data problem. We also change the notations somewhat. We assume that missingness occurs only in one variable, to be denoted by $Y = (Y'_{\text{obs}}, Y'_{\text{mis}})'$ while each variable X_i in $\mathbf{X}$ is fully observed. We also assume that missingness is of *missing completely at random* (MCAR) type. The algorithm for imputation through copula transformation is given below.

Algorithm - Univariate Missing Data Pattern:

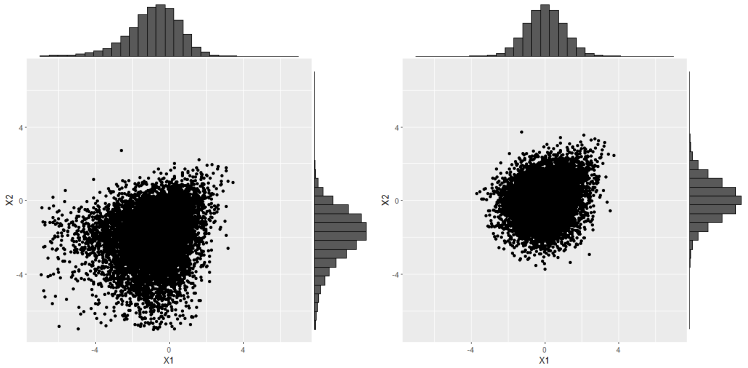
1. Transform the complete data on covariates $\mathbf{X}$ to uniformly distributed data $\mathbf{U_X}$ by empirical distributions defined in (2).

2. For the observed data Y_{obs}, transform variable Y to uniformly distributed variable U_Y by empirical distribution defined in (2).

3. Convert the data $(\mathbf{U_X}, U_Y)$ to standard normal data $(\mathbf{S_X}, S_Y)$ by using the standard inverse multivariate normal cumulative distribution. That is, each column vector is transformed by $S_i = \Phi^{-1}(U_i)$. At this stage, the data set $(\mathbf{S_X}, S_Y)$ are assumed to be distributed as multivariate normal distribution with zero mean and variance-covariance matrix $R = (\rho_{ij})$ where ρ_{ij} is defined in (3).



(a) Raw bivariate Lomax data (b) Logarithm-transformed data



(c) Box-Cox transformed data ($\lambda_1 = 1.54 \times 10^{-5}, \lambda_2 = 1.89 \times 10^{-5}$) (d) Gaussian-Copula transformed data

Figure 2: Raw data and transformed bivariate Lomax data with different transformation methods

4. Use a suitably chosen imputation procedure (e.g. regression, MCMC, FCS) to impute all missing values and obtain the dataset $(\mathbf{S_X}, S_Y^*)$ along with filled in imputed data.

5. Back-transform the filled-in data to original scale via $U_Y^* = \Phi(S_Y^*)$ according to the inverse of empirical marginal distribution function of Y, i.e., $Y^* = F_Y^{-1}(U_Y^*)$.

Algorithm is pictorially depicted in Figure 3. The implementation of algorithm is readily available in SAS as described in Lun and Khattree (2019).

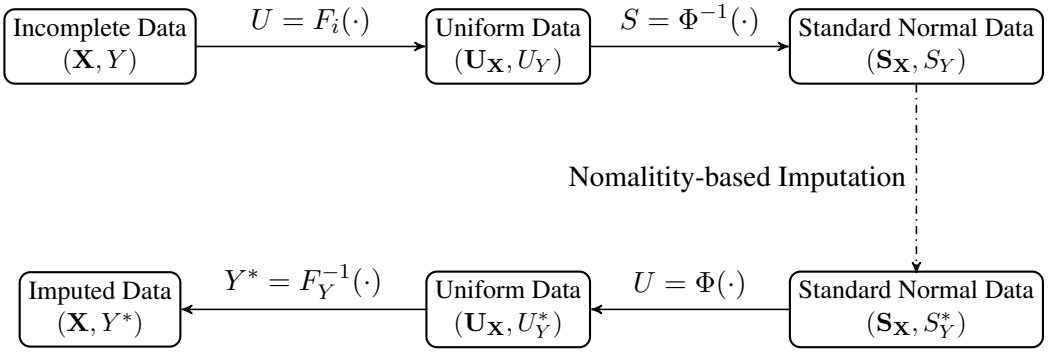


Figure 3: Procedure of imputation implementation using copula transformed data.

Table 1: Missingness pattern of **UCI air quality data set**: label $= 1$ indicates the observed variate and label $= 0$ indicates missing variate.

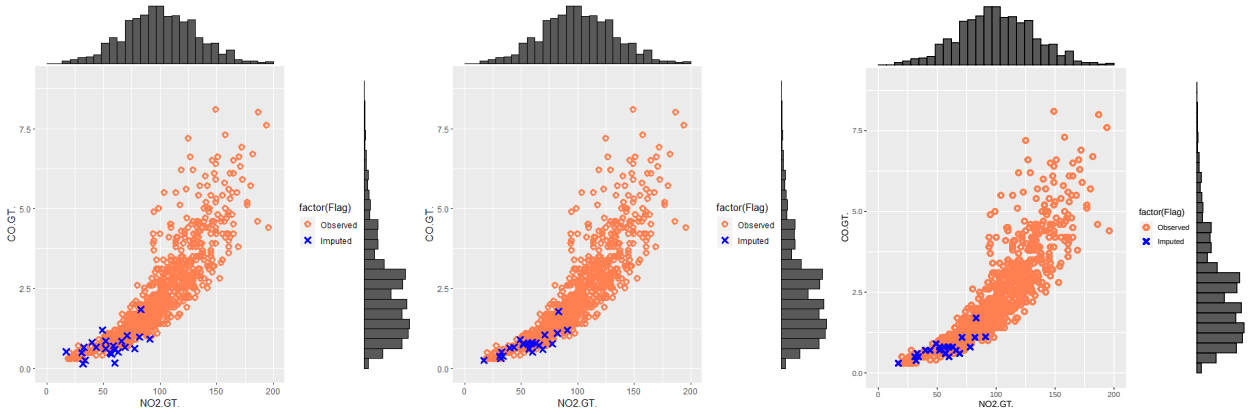
# Instances	C6H6.GT.	NOx.GT.	NO2.GT.	CO.GT.	NMHC.GT.	# Missing variables
827	1	1	1	1	1	0
6114	1	1	1	1	0	1
24	1	1	1	0	1	1
428	1	1	1	0	0	2
3	1	1	0	1	0	2
36	1	0	0	1	1	2
364	1	0	0	1	0	3
1195	1	0	0	0	0	4
26	0	1	1	1	1	1
291	0	1	1	1	0	2
5	0	1	1	0	0	3
1	0	0	0	1	1	3
12	0	0	0	1	0	4
31	0	0	0	0	0	5

4. Imputation on real data: skewed air quality data

We use a dataset of air quality (De Vito et al., 2008) from the University of California Irvine Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets/Air+quality`). This highly skewed dataset records 9358 instances of hourly averaged responses from an array of five metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. These are Ground truth hourly averaged concentrations for CO (CO.GT), Non Metanic Hydrocarbons (NMHC.GT), Benzene (C6H6.GT), Total Nitrogen Oxides (NOx.GT) and Nitrogen Dioxide (NO2.GT). There exist various missingness patterns including univariate and multivariate missingnesses and these are summarized in Table 1. As shown there, there are only 827 instances that are fully observed; 6114 instances are recorded with only NMHC missing; 24 instances are recorded with only CO missing; others instances have two or more variates missing in a variety of missingnesses. Due to the presence of large amount of missing data, data analyses are restricted to those fully observed cases, which are only small portion of the original data set. For example, Luo and Qi (2019) used only 355 observations of this data set in their study.

We impute only the univariate missing instances for CO and NMHC, respectively, and compare the performance of imputations for the raw data (RAW), upon applying logarithm-transformation data (LOG) and by using Gaussian copula transformation data (CPL). Imputation method is selected as the single imputation $k_{ipmt} = 1$ using linear regression model ignoring model error and under the assumption of MCAR. This is implemented in function `mice()` with `method = 'norm.nob'` in **mice** package (van Buuren, 2019) in R (R Core Team, 2019). We will compare the relative placements of missing values with in the patterns of 827 fully observed instances. Any substantial departure from the patterns indicates the inadequacy of the approach.

As shown in Figure 4 (a) for the imputation of missing variate CO.GT by using raw data, most of the imputed values fall within the scatterplots of observed instances but some of them somewhat depart away from the trend of the scatterplot between CO and NO_2. On the other hand, both imputations through logarithm-transformed and copula-transformed data show imputed values well within or around the other observed data values corresponding to fully observed cases. See Figures 4 (b) and (c). Logarithm transformation indeed does equally well for this particular data set. However, such a transformation for every variable is neither always possible nor will it be always superior.



(a) Imputation on raw data (b) Imputation via log-transformation (c) Imputation via copula-transformation

Figure 4: Scatterplots of NO_2 vs. observed and imputed CO obtained by linear regression imputation through the (a) raw data, (b) log-transformation and (c) copula transformation for **UCI air quality data**

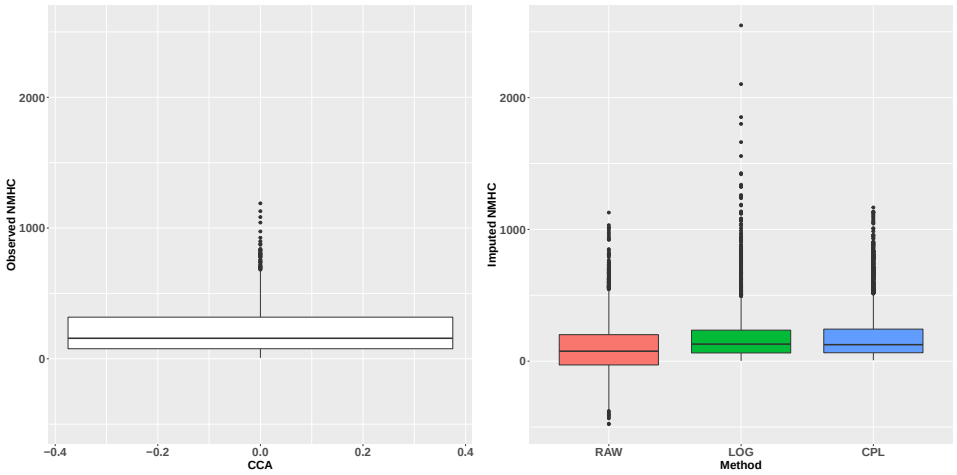
We now consider imputations for the missing variate NMHC.GT, which has missing values in 6114 instances out of a total of 9358 observations. Thus, the number of missing instances is over seven times the fully observed cases (=827). Thus, imputation for such high missing rate (6114/(6114+827) = 88%) may not be quite appropriate in real context as the imputed data may overly distort the underlying true distribution due to the preponderance of missing data. Nonetheless, we will proceed by using the same single imputation through linear regression method out of curiosity about the distributions of the resulting imputed values obtained by previous three approaches and compare these with those for the fully observed cases.

A box-plot of 827 complete cases (CCA) is given in Figure 5 (a). As shown in Figure 5 (b), the imputation assuming multivariate normality of raw data (RAW), produces many negative values of NMHC, which could not be viewed as valid since it is a measure of concentrations and nonnegative. Consequently, the median of imputed data (=75.64) turns out to be much lower than that of data with fully observed cases (=157). The imputation via logarithmically transformed data (LOG) results in all non-negative imputations with the median of imputed values (=129.38), which is much closer to that of fully observed cases compared with the imputation through raw data. However, there are a few very large imputed values, some even over 2000. This indicates the possible drawback of logarithm transformation method where some imputed values may go out of valid range.

However, all imputed values obtained by applying Gaussian copula-transformation (CPL) are confined within the valid range in Figure 5 (b). No negative value or excessively large imputation are observed. The median of imputed data is 124.98, which is also much closer to that of fully observed cases when compared with the median obtained from imputed values assuming multivariate normality of raw data. This illustrates the possible superiority of the Gaussian copula transformation approach. To establish that conviction more firmly, we have taken upon extensive simulation studies, which we present else where as a separate communication.

5. Concluding Remarks

We have introduced an imputation method through copula-transformation for univariate missing pattern under MCAR missing mechanism. Technique is general in that no assumption on raw data is made except that its copula is Gaussian.



(a) 827 complete cases (b) 6114 imputed cases

Figure 5: Boxplots of complete cases and imputed NMHC data for 6114 missing instances through raw (RAW), logarithm-transformed (LOG) and copula-transformed (CPL) data for **UCI air quality data set**

To illustrate the usefulness of our approach we have taken two data sets. The first one is a large simulated data exhibiting strong skewness. The second data set is a real data which not only exhibits skewness but also has a large proportion of a variety of missingness patterns. We have demonstrated via these two examples, the utility of our approach and shown, how our method is superior to other traditional approaches. We strongly believe that our general purpose imputation approach holds much promise of real applications. A sample SAS code is also available for the implementation of our approach.

It was pointed out to us that Robbins, Ghosh and Habiger (2013) have also used imputation through a Gaussian copula for skewed data in an application. Their suggestion is more specialized in that they assumed the skew-normal distribution for the observed data while we have chosen to be the less restrictive by relying on the empirical CDF to obtain the uniform intermediate random variables.

Simulation studies, to be presented else where, indicate that this approach does have considerably much superior performance with respect to several criteria. They reinforce our confidence further on the suggested approach.

REFERENCES

Bahuguna, M. and Khattree, R. (2020), "A generic all purpose transformation for multivariate modeling through copulas," *International Journal of Data Science and Analytics*, 10:1–23.

Box, G. E. P. and Cox, D. R. (1964), "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

De Vito, S., Massera, E., Piga, M., Martinotto, L., and Di Francia, G. (2008), "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sensors and Actuators B: Chemical*, 129:750–757.

Enders, C. K. (2010), *Applied Missing Data Analysis*, New York: The Guilford Press.

Lindley, D. V. and Singpurwalla, N. D. (1986), "Multivariate distributions for the life lengths of a system sharing a common environment," *Journal of Applied Probability*, 23:418–431.

Lun, Z. and Khattree, R. (2019), "Multiple imputation for skewed multivariate data: A marriage of the MI and COPULA procedures," *Proceedings of the SAS Global Forum 2019*.

Luo, R. and Qi, X. (2019), "Interaction model and model selection for function-on-function regression," *Journal of Computational and Graphical Statistics*, 28:309–322.

Meyer, C. (2013), "The bivariate normal copula," *Communications in Statistics - Theory and Methods*, 42(13):2402–

2422.

R Core Team. (2019), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Robbins, M. W., Ghosh, S. K., and Habiger, J. D. (2013), "Imputation in high-dimensional economic data as applied to the agricultural resource management survey," *Journal of the American Statistical Association*, 108(501):81–95.

Rubin, D. B. (1977), "Formalizing subjective notions about the effect of nonrespondents in sample surveys," *Journal of the American Statistical Association*, 72(359):538–543.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Sklar, A. (1959), "Distribution functions of n dimensions and margins," *Publications of the Institute of Statistics at the University of Paris* , 8:229–231.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006), "Fully conditional specification in multivariate imputation," *Statistical Methods in Medical Research*, 76(12):1049–1064.

Van Buuren, S. (2007), "Multiple imputation of discrete and continuous data by fully conditional specification," *Statistical Methods in Medical Research*, 16:219–242.

Van Buuren, S. (2019), mice: Multivariate Imputation by Chained Equations, R package version 3.7.0.