Paper 3605-2019

Multiple Imputation for Skewed Multivariate Data: A Marriage of the MI and COPULA Procedures

Zhixin Lun, Ravindra Khattree, Oakland University

ABSTRACT

Missing data is a common phenomenon in various data analyses. Imputation is a flexible method for handling missing-data problems since it efficiently uses all the available information in the data. Apart from regression imputation approach, the MI procedure in SAS® also provides the multiple imputation options which create multiple data sets based on Markov chain Monte Carlo (MCMC) and fully conditional specification (FCS) methods. However, these methods may not work very effectively for skewed multivariate data since they require the assumption of multivariate normal distribution. To deal with such data, we introduce an approach based on copula transformation. We combine imputation using PROC MI and copula theory using PROC COPULA to arrive at an approach to solve the missing data problem for skewed multivariate data. We implement and demonstrate the use of this method through simulated examples under the assumption that data are missing completely at random (MCAR).

INTRODUCTION

Most of the methodology available for missing data imputation assumes data distributed as multivariate normal (see Little and Rubin 2002, Rao et al. 2007). Applying normality-based imputation in skewed data may cause practical issues for the simple reason of violation of distributional assumptions. One common way to deal with non-normal data is to apply normalizing transformation prior to the imputation phase and then back-transform to original scale at the analysis phase. However, transformation of each variable individually may alter the association structure among variables and hence may impact the accuracy of imputations.

As Bahuguna and Khattree (2019) illustrated, based on copula transformation, multivariate skewed data can be transformed to any other multivariate distribution without losing dependence information among random variables. This property provides an approach to normalize multivariate skewed data and more importantly, ensures that existing normality-based imputation methods are applicable for the analysis of multivariate skewed data. Our work here builds on this crucial and important observation.

The objective of this work is to illustrate the implementation of above ideas by applying the copula transformation using PROC COPULA and to combine PROC MI for multiple imputation for the missing data in case of skewed multivariate data. In the following section, we revisit the basic concept of copula and the Sklar's theorem (Sklar, 1959), which is the foundation of copula transformation, and then we show the details of copula transformation algorithm and its implementation in SAS.

COPULAS AND COPULA TRANSFORMATION

THE COPULA TRANSFORMATION

In copula theory, copula is a multivariate probability distribution where the marginal probability distribution of each variable is uniform. In other words, a function C is a d-

dimensional copula if there is a random vector $U = (U_1, U_2, \dots, U_d)'$, such that for $i = 1,2,\dots,d$, $U_i \sim$ Uniform (0,1), and

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

The most important theorem in copula theory is the Sklar's theorem (Sklar, 1959), which states that a function $\mathrm{F}: \mathrm{R}^{\mathrm{d}} \rightarrow [0,1]$ is the distribution function of a random vector $X = (X_1, X_2, \dots, X_d)'$ if and only if there is a copula C from $[0,1]^d$ to $[0,1]$ and d univariate distribution functions $F_1, F_2, \dots, F_d$ such that

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

This theorem indirectly implies that two different continuous multivariate distributions can be transformed to each other via the same copula. Specifically, consider two different continuous multivariate cumulative distributions denoted by $F(\cdot)$ and $G(\cdot)$ and assume that they have a common copula. Then the transformation is shown as follows from Sklar's theorem,

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_d}) &= \boldsymbol{C}\big(\boldsymbol{F_1}(\boldsymbol{x_1}), \boldsymbol{F_2}(\boldsymbol{x_2}), \dots, \boldsymbol{F_d}(\boldsymbol{x_d})\big) \\ &= \boldsymbol{C}(\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_d}) \\ &= \boldsymbol{C}\big(\boldsymbol{G_1}(\boldsymbol{y_1}), \boldsymbol{G_2}(\boldsymbol{y_2}), \dots, \boldsymbol{G_d}(\boldsymbol{y_d})\big) \\ &= \boldsymbol{G}(\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_d}), \end{aligned} \qquad (\mathbf{1})$$

where $F_i(\cdot)$ and $G_i(\cdot)$ are the corresponding marginal cumulative distribution functions arising out of $F(\cdot)$ and $G(\cdot)$, respectively. Thus, a set of data on $(x_1, \dots, x_d)$ can be transformed as $(y_1, \dots, y_d)$ and vice versa via dependent uniform data $(u_1, u_2, \dots, u_d)'$ created in between.

In this study, since our purpose is to normalize multivariate variables, we assume that the common copula is a Gaussian copula $\Phi_{\mu,\Sigma}(\cdot)$, that is,

$$C_\Sigma(u_1, \dots, u_d) = \Phi_{\mu,\Sigma}\big(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\big),$$

where $\Phi_{\mu,\Sigma}(\cdot)$ is the cumulative distribution of multivariate normal distribution 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.

THE ALGORITHM

We start with the missing data problem, for which missingness occurs in one variable denoted by $\mathbf{Y}$ while the other variables $\mathbf{X}_{\mathrm{i}}$'s are fully observed. Then the missing data structure can be divided into two blocks: (i) the complete cases denoted by $(\mathbf{Y}_{\mathrm{obs}}, \mathbf{X}_{\mathrm{cc}})$ and (ii) incomplete cases denoted by $(\mathbf{Y}_{\mathrm{mis}}, \mathbf{X}_{\mathrm{ic}})$. Let

$$(\mathbf{Y}, \mathbf{X}) = \begin{bmatrix} \mathbf{Y_{obs}} & \mathbf{X_{cc}} \\ \mathbf{Y_{mis}} & \mathbf{X_{ic}} \end{bmatrix}.$$

According to the above process of copula transformation as stated in Equation (1), we implement the following algorithm,

1. Transform the complete cases $(\mathbf{Y_{obs}}, \mathbf{X_{cc}})$ to uniformly distributed data $\mathbf{U_{cc}} = \big(U_Y, U_{X_1}, U_{X_2}, \dots, U_{X_k}\big)$ using the empirical cumulative distribution function estimated from the data.
2. For the incomplete case, transform $\mathbf{X_{ic}}$ to uniformly distributed data $\mathbf{U_{ic}} = \big(U_{X_1}, U_{X_2}, \dots, U_{X_k}\big)$ using the empirical cumulative distribution function estimated from the data. There is no

U_Y data due to missingness.

3. Combine $\mathbf{U_{cc}}$ and $\mathbf{U_{ic}}$ into $\mathbf{U}$, that is

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_{cc} \\ \mathbf{U}_{ic} \end{bmatrix},$$

   and convert $\mathbf{U}$ to a new dataset $(\mathbf{Y}^*, \mathbf{X}^*)$ using inverse multivariate normal cumulative distribution, corresponding to the correlation matrix from the original data. Thus,

$$(\mathbf{Y}^*, \mathbf{X}^*) = \begin{bmatrix} \mathbf{Y^*_{obs}} & \mathbf{X^*_{cc}} \\ \mathbf{Y^*_{mis}} & \mathbf{X^*_{ic}} \end{bmatrix}.$$

   At this stage, after transforming from $\mathbf{U}$ to $(\mathbf{Y}^*, \mathbf{X}^*)$, one of the imputation methods can be applied on multivariate normally distributed $(\mathbf{Y}^*, \mathbf{X}^*)$ as in Step 4 below.

4. Use one of the imputation procedures (e.g. regression, MCMC, FCS) as desired to impute all missing values of $\mathbf{Y^*_{mis}}$. Multivariate normality of $(\mathbf{Y}^*, \mathbf{X}^*)$ makes this step easily implementable using PROC MI.

5. Back-transform the filled-in data to original scale via $\mathbf{U}$ according to the chosen copula function.

It is assumed that the missingness scheme is independent of any such transformation and hence will remain the same all through the transformation.

IMPLEMENTATION

We illustrate the implementation scheme step by step following the above algorithm on a sample dataset `misData` with four variables (Y, X_1, X_2, X_3), which contains missing values in Y and fully observed values in X_1, X_2, and X_3. We add an indicator column `Flag` into the dataset `misData` such that `Flag='X'` and `'.'` are for complete and incomplete cases, respectively.

Step 1 & 2: Transform complete cases $(\mathbf{Y}_{obs}, \boldsymbol{X}_{cc})$ and incomplete cases $\mathbf{X}_{ic}$ to uniform random variables using `PROC COPULA`, respectively. We specify parameter `normal` in `FIT` statement since we use Gaussian copula. The setting `marginals=empirical` indicates that we use the empirical cumulative distribution function estimated from the data. The resulting dataset `unif_cc_star` is the data set on the transformed uniform variables from complete cases $(\mathbf{Y}_{obs}, \mathbf{X}_{cc})$, while `unif_ic_star` is the data set on the transformed uniform variables from incomplete cases $\mathbf{X}_{ic}$.

```
%let misVar = y;
%let ccVarList = x1 x2 x3;

proc copula data=misData(where=(Flag='X'));
  var &misVar &ccVarList;
  fit normal / marginals=empirical outpseudo=unif_cc noprint;
run;

proc copula data=misData;
  var &ccVarList;
  fit normal / marginals=empirical outpseudo=unif_ic noprint;
run;

data unif_cc_star;
  set unif_cc;
  Flag = 'X';
run;
```

```
data unif_ic_star(where=(Flag='.'));
  merge unif_ic misData(keep=Flag);
run;
```

Step 3: Combine two datasets `unif_cc_star` and `unif_ic_star` and transform each of uniformly distributed column data to standard normal by using `quantile` function.

```
data unif_u;
  set unif_cc_star unif_ic_star;
run;

data std_norm;
  set unif_u;
  if Flag='X' then y = quantile("Normal", y);
  x1 = quantile("Normal", x1);
  x2 = quantile("Normal", x2);
  x3 = quantile("Normal", x3);
run;
```

Step 4: Apply the desired multiple imputation method on the dataset `std_norm`. MCMC method is selected as an example in the code given below.

```
proc mi data=std_norm nimpute=5 out=mi_std_norm seed=1234 noprint;
  mcmc;
  var &misVar &ccVarList;
run;
```

Step 5: Note that the imputed values in above dataset `mi_std_norm` are still in standard normal scale. The last step is to back-transform the filled-in data to original scale according to the copula. This process involves two steps:

(a) Simulate a large number (e.g., NSIM=10,000) of observations from multivariate uniform distribution corresponding to our copula and convert those simulated observations to the data on variables in original data scale and to the data on variables with standard normal distribution, respectively. This can be readily simulated by using `FIT` and `SIMULATE` statements in `PROC COPULA`. The `FIT` statement setting must be the same as we set in **Step 1 & 2** since we back-transform the data according to the same copula. The output dataset `sim_org` contains the simulated observations in original scale and `sim_unif` consists of the simulated observations distributed as multivariate uniform distribution. The dataset `sim_std_norm` is the converted data where each variable is distributed as standard normal.

```
%let NSIM=10000;

proc copula data=misdata(where=(Flag='X'));
  var &misVar &ccVarList;
  fit normal / marginals=empirical noprint;
  simulate /ndraws = &NSIM seed = 1234567
  out = sim_org outuniform=sim_unif;
run;

data sim_std_norm;
  set sim_unif;
  sy = quantile("Normal", y);
  sx1 = quantile("Normal", x1);
```

```
    sx2 = quantile("Normal", x2);
    sx3 = quantile("Normal", x3);
    keep sy sx1 sx2 sx3;
  run;
```

(b) Obtain the imputed values in original data scale by interpolation from above simulated observations in data sets `sim_org` and `sim_std_norm`. Denote the imputed value in **Step 4** by $\widehat{y_k}$, which is in standard normal scale. If $\widehat{y_k}$ is sandwiched between two values sy_t and sy_{t+1} in dataset `sim_std_norm`, then we predict y_k in its original scale by averaging, in general, by interpolating values corresponding to sy_t and sy_{t+1} in dataset `sim_org`. In the resulting dataset `impt_org_scale`, the variable `ry` with `MIS='Y'` are the imputed values in original scale.

```
  data sim_org;
    set sim_org;
    keep y;
    rename y=ry;
  run;

  data sim_std_norm(keep=sy);
    set sim_std_norm;
  run;

  proc sort data=sim_std_norm; by sy; run;

  proc sort data=sim_org; by ry; run;

  data sim_org_std;
    merge sim_std_norm sim_org;
  run;

  /* filter the imputed values in variable y*/
  data impt_std_norm;
    set mi_std_norm(where=(Flag='.'));
    keep y;
    rename y=sy;
  run;

  data impt_sim_comb;
    set impt_std_norm sim_org_std;
  run;

  proc sort data=impt_sim_comb; by sy; run;

  data impt_org_scale;
    merge impt_sim_comb impt_sim_comb(keep=ry firstobs=2
  rename=(ry=lead_mis));
    lag_mis=lag(ry);
    if ry=. then do;
    ry=mean(lag_mis, lead_mis);
    MIS='Y';
    end;
  run;
```

AN ILLUSTRATION VIA SIMULATED DATA SETS

COMPLETE DATA GENERATION AND MISSINGNESS MECHANISM

Using the Iman-Conover method given by Wicklin (2013), we generate data sets from the following two multivariate distributions, where marginals of components are as specified in Table 1,

Group	X_1	X_2	X_3	X_4
1	Log-normal $(0, \sigma)$	Pareto (1,1)	Normal (0,1)	Uniform (0,1)
2	Log-normal $(0, \sigma)$	Normal (0,1)	Exponential (1)	Uniform (0,1)

Table 1. Marginal distributions of simulated data sets

where σ was set as 1.0, 2.0 and 3.0. In each case, the following correlation structure was used,

$$\text{Corr} = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

where ρ was set as 0.9.

Missing values are assumed to be missing completely at random (MCAR).

EVALUATION OF OUR IMPUTATION METHOD

We select X_1 as the variate with missing values. The sample size is taken as 100 and the number of missing cases as 5. To evaluate the quality of imputation, we simulate each scenario NSIM=1,000 times and use k imputation(s). Then we compute the mean of the sum of squared residuals by

$$\text{MSSR} = \frac{1}{\text{NSIM}} \sum^{\text{NSIM}} \sum_{m=1}^{k} \sum_{i=1}^{5} \left(X_{1i}^{\text{impt}(m)} - X_{1i}^{true} \right)^2,$$

where $X_{1i}^{\text{impt}(m)}$ is the m-th imputed value for the i-th missing value X_{1i} and X_{1i}^{true} is the true observed value of X_{1i}.

MULTIPLE IMPUTATION METHODS

We select FCS regression (Van Buuren 2007) and MCMC (Schafer 1997) multiple imputation methods for multiple imputation. The general idea of FCS regression is to generate k sets of predicted values based on regression model, which involves filled-in phase and imputation phrase. MCMC method is used to generate k sets of predicted values according to the posterior distributions in Bayesian inference. Both methods require the assumption that the data are from a multivariate normal distribution. In our illustration, we choose $k = 5$.

SIMULATION RESULT

Table 2 and Table 3 give sample summaries using FCS and MCMC methods for original and copula-transformed data. The column **Ratio (O/C)** is the ratio of the MSSR values of above to respective data. The larger **Ratio (O/C)** indicates the better performance of copula transformation. The column **%SSR (O>C)** is the percent of times our approach results in smaller sum of squared residuals. Accordingly, the larger percentage value indicates superior performance of our transformation approach. The following results show that our approach performs substantially better than the case when multivariate normality of the data was blindly assumed. A more detailed extensive simulation work, not reported here due to lack of space, confirms to the above observations.

Method	σ	**MSSR**			**%SSR (O>C)**
		Original (Assumed multi-normality)	**Copula-transformed**	**Ratio (O/C)**	
FCS Regression	1.0	1,373.60	51.55	26.65	87.9%
	2.0	260,022.58	28,639.43	9.08	89.2%
	3.0	113,277,135.61	37,978,413.33	2.98	90.0%
MCMC	1.0	450.31	47.32	9.51	80.5%
	2.0	131,115.16	21,115.43	6.21	83.9%
	3.0	67,349,495.85	13,494,377.60	4.99	85.5%

Table 2 Comparison between original data and copula-transformed data using multiple imputation $(\boldsymbol{k} = 5)$ **for Group 1 with correlation choosing** $\boldsymbol{\rho} = \mathbf{0.9}$

Method	σ	**MSSR**			**%SSR (O>C)**
		Original (Assumed multi-normality)	**Copula-transformed**	**Ratio (O/C)**	
FCS Regression	1.0	84.70	53.40	1.59	86.4%
	2.0	50,264.93	30,896.51	1.63	90.3%
	3.0	75,184,163.76	42,523,936.26	1.77	90.8%
MCMC	1.0	65.26	48.47	1.34	81.2%
	2.0	35,414.91	25,431.77	1.39	84.3%
	3.0	44,235,290.84	29,549,312.57	1.50	85.5%

Table 3 Comparison between original data and copula-transformed data using multiple imputation $(\boldsymbol{k} = 5)$ **for Group 2 with correlation choosing** $\boldsymbol{\rho} = \mathbf{0.9}$

CONCLUSION

We have introduced a very powerful approach based on copula transformation to impute the missing values for general skewed multivariate data. We provide the algorithm and its implementation for one-variate missing pattern. Algorithm can be readily modified for k-variate $(k > 1)$ missing patterns. In view of ready accessibility of MI and COPULA procedures, this approach has a very wide scope for practical applications. A complete program combining all the procedures pieces is given as the supplementary material. An execution of our program results in imputed values (column rx1 in bold) shown in Table 4 for five missing values.

Obs	_Imputation_	sx1	sx2	sx3	sx4	Flag	Sr	rx1	MIS
1	1	-0.71624	-0.88485	-0.62092	-0.39469	.	1	**0.61905**	Y
2	1	-1.15206	-1.28721	-1.48097	-2.33008	.	2	**0.42109**	Y
3	1	-0.41595	-0.71397	-0.11191	-0.47658	.	3	**0.92597**	Y
4	1	0.37430	0.56179	0.65130	-0.11191	.	4	**1.78136**	Y
5	1	0.20480	0.21254	0.26361	0.16202	.	5	**1.38008**	Y

Table 4 The first five output (the imputed values are in column rx1 in bold) of execution of sample code in supplementary material

REFERENCES

Bahuguna, M. and Khattree, R. 2019. “A Generic All Purpose Transformation for Multivariate Modeling through Copulas.” To appear in *International Journal of Data Science and Analytics*.

Little, R.J.A. and Rubin, D.B. 2002. *Statistical Analysis with Missing Data*. 2nd Edition. Hoboken, NJ: John Wiley & Sons.

Rao, C.R., Toutenburg, H., Shalabh, Heumann, C. 2007. *Linear Models and Generalizations, Least Squares and Alternatives*. 3rd Extended Edition. New York, NY: Springer.

SAS Institute 2014. SAS/ETS 13.2 User's Guide The COPULA Procedure. Cary, NC: SAS Institute Inc.

SAS Institute 2017. SAS/STAT 14.3 User's Guide The MI Procedure. Cary, NC: SAS Institute Inc.

Sklar, A. 1959. "Distribution Functions of n Dimensions and Margins," *Publications of the Institute of Statistics of the University of Paris*, 8: 229-231.

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

Van Buuren, S. 2007. “Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification.” *Statistical Methods in Medical Research*, 16:219-242.

Wicklin, R. 2013. *Simulating Data with SAS*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Zhixin Lun, PhD Student
Department of Mathematics and Statistics, Oakland University
zlun@oakland.edu;zxlun9@gmail.com

Ravindra Khattree, Professor/Co-director
Department of Mathematics and Statistics/Center for Data Science and Big Data Analytics, Oakland University
khattree@oakland.edu