Missing Value Imputation for Non-Normal Data

Ravi Khattree
Distinguished University Professor of Applied Statistics
Co-Director, Center for Data Sciences and Big Data Analytics
Participating Member, Center for Biomedical Research
Oakland University

(Joint work with Zhixin Lun and Manoj Bahuguna)

email: khattree@oakland.edu

Marquette U., Oct. 25, 2024

Overview

Introduction

Basic Result (Casella and Berger, Theorem 2.1.10, p. 54):

For a continuous random variable W, the cumulative distribution function $U = F_W(\cdot)$ is uniformly distributed on [0,1].

Thus, any continuous random variable W can be transformed to a uniform random variable and conversely, any uniform random variable can be transformed to a random variable V with any desired continuous distribution.

Introduction

This fact, thus, provides an approach for the transformation of data to any desirable distribution.

In particular, this can be used as an all purpose approach to introduce normality so that analyses, such as linear model theory based modeling, can be performed under the usual normality assumptions.

NORTA (NORMAL to ANYTHING) algorithms use this fact backwards to simulate random numbers from any desired distribution after generating standard normal variates.

Introduction

However, one fact that should not be lost in this approach is that the information about the location and scale is essentially lost, in much the same way as the information about the shape and the skewness of the original distribution.

While symmetrized (specifically normally distributed) data are desirable for data analysis, the approach will therefore, prevent us from making any inference about the location in most situations.

A common approach to deal with skewed data is through the use of transformation (to normality). The logarithmic, square-root, arcsine and more generally, Box-Cox transformations have been the common tools to artificially induce, among other desirable features, symmetry for the asymmetric data and have been used extensively in a variety of statistical problems.

Introduction

In the multivariate context, when the interest is in studying the dependence structures and possibly prediction, a generalization of the approach described above can be extremely useful

and
the purpose of this talk is to introduce the usefulness of our suggested approach in various multivariate situations.

The objective here is to come up with an approach to transform the data where classical techniques of multivariate analyses [**Speci cally, here for missing data imputation**] can be readily adopted for the transformed data.

Yet, the method should be such, so that inference and especially the predictions for the transformed data can be brought back to the original context.

Multivariate Copula-Transformation

We will first define the concept of copula.

Definition

A function C from a d-dimensional rectangle $[0,1]^d$ to $[0,1]$ is called a copula if there is a random vector $\boldsymbol{U} = (U_1, U_2, \ldots, U_d)'$, such that for $i = 1, \ldots, d$, $U_i \sim U(0,1)$, the uniform distribution on interval [0,1] and $C(u_1, u_2, \ldots, u_d)$
$= P[U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d]$ where $U_1, U_2, \ldots, U_d \in [0,1].$

Thus, $C(\cdot)$ is essentially a d-dimensional multivariate cumulative distribution function of d random variables, each distributed uniformly in the interval [0,1].

The dependence structure is not stated in the definition and cannot be, in general, specified. It depends on the nature and the joint behavior of the particular set of the random variables.

Multivariate Copula-Transformation

Theorem

(Sklar's Theorem)
A function $F : R^d \to [0, 1]$ *is the distribution function of some continuous random vector* $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)'$ *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*

$$C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)) = F(x_1, x_2, \ldots, x_d) \tag{1}$$

for $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)' \in R^d.$

The functions $F_i(\cdot)$ are clearly the (marginal) cumulative distribution functions of corresponding random variables $X_i, i = 1, 2, \ldots, d.$ Thus, copula expresses the dependence among $X_1, X_2, \ldots, X_d$ through their marginal cumulative distribution functions.

Multivariate Copula-Transformation

It provides a way to express and obtain the joint cumulative distribution functions through an appropriate copula. Since $F(\cdot)$ is continuous and hence admits an inverse function $F^{-1}(\cdot)$, it follows from above that

$$(X_1; X_2; :::; X_d) \overset{dist}{=} F^{-1}C(F_1(x_1); F_2(x_2); :::; F_d(x_d)); \qquad (2)$$

where *dist* indidq8552 T

Multivariate Copula-Transformation

Let us concentrate on (**??**) namely,

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

Let $F(\cdot)$ and $G(\cdot)$ be two d-dimensional multivariate continuous CDFs, with corresponding marginal CDFs $F_1(\cdot); F_2(\cdot); \ldots; F_d(\cdot)$ and $G_1(\cdot); G_2(\cdot); \ldots; G_d(\cdot)$ respectively. Also assume that $F(\cdot)$ and $G(\cdot)$ both correspond to the same copula function $C(\cdot)$. Thus, with random vector $\boldsymbol{X}$ having the CDF $F(\cdot)$ and random vector $\boldsymbol{Y}$ having that as $G(\cdot)$, we have

$$F(x_1; x_2; \ldots; x_d) = C(F_1(x_1); F_2(x_2); \ldots; F_d(x_d)) \tag{4}$$

$$= C(u_1; u_2; \ldots; u_d) = C(G_1(y_1); G_2(y_2); \ldots; G_d(y_d))$$
$$= G(y_1; y_2; \ldots; y_d),$$

for some $y_1; y_2; \ldots; y_d$, so that $G^{-1}(y_i) = u_i$, $i = 1; 2; \ldots; d$ where $G^{-1}(\cdot)$ is the inverse function of $G(\cdot)$.

Multivariate Copula-Transformation

Note that the only assumption here is that the $F(\cdot)$ and $G(\cdot)$ share the common copula. By Sklar's Theorem, it also follows that given $F(\cdot)$ (or $G(\cdot)$), the copula is unique.

Thus, if $G(\cdot)$ is desired to be a particular cumulative distribution function then it automatically determines the choice of $C(\cdot)$*:*

Multivariate Copula-Transformation

The above calculation succinctly provides, an approach to transform the data on multivariate random vector $\boldsymbol{X}$ having cumulative distribution function $F(\cdot)$ to another random vector $\boldsymbol{Y}$ having the cumulative distribution function $G(\cdot)$: More succinctly, it can be

Multivariate Copula-Transformation

This is pictorially depicted in Figure on next slide.

For our work, with an intention to enable us to do classical multivariate modeling, the function $G(\cdot)$ will usually be a multivariate normal cumulative distribution function. Consequently, the choice of $C(\cdot)$ must be a Gaussian copula.

Multivariate Copula-Transformation

As a graphical representation for two given distribution functions say, $\boldsymbol{F}(\cdot)$ and $\boldsymbol{G}(\cdot)$ with common copula say $\boldsymbol{C}(\cdot)$, our transformation works as,

$$\boldsymbol{F}(\cdot) \longrightarrow \boldsymbol{C}(\cdot) \longrightarrow \boldsymbol{G}(\cdot)$$

Figure 1

Multivariate Copula-Transformation

Accordingly, implicit assumption on the distribution function $F(\cdot)$ of our raw data is that even though $F(\cdot)$ itself may not be multivariate normal distribution function, its copula function is Gaussian.

Such an assumption is very reasonable.

Of course, for the data analysis, we must resort to the empirical version of $F(\cdot)$*:* computed from data.

Multivariate Copula-Transformation

Thus, in essence, we make the assumption that the common copula is a Gaussian copula $\Phi_{(\boldsymbol{\mu};\quad)}(\cdot)$;

In principle, the choices of mean vector and covariance matrix are arbitrary.

Since our interest is in doing the multivariate analyses of dependence, we will choose cautiously to retain the essential dependence features of data.

On the other hand, since the choice of is often unimportant in such situations, we will take it's value to be the zero vector.

Multivariate Copula-Transformation

As a graphical representation for a given distribution of data, say $\boldsymbol{D}$, our transformation works as

$$\boldsymbol{D} \longrightarrow \boldsymbol{U} \longrightarrow \boldsymbol{N}$$

and in a reverse direction as

$$\boldsymbol{N} \longrightarrow \boldsymbol{U} \longrightarrow \boldsymbol{D};$$

where $\boldsymbol{U}$ U

Copula Transformed Multiple Regression: Data Sets

i) Wicklin's Data (2013):Taken from Wicklin's book, where we have four random variables, jointly exhibiting dependence, but each with marginal distributions which are functionally very di erent.

Speci cally, we have the response variable y distributed as standard lognormal (= 0, = 1) and explanatory variables x_1, x_2 and x_3 respectively, distributed as standard normal, uniform on [0,1] and standard exponential (= 1).

Clearly, considerable skewness is present in y and x_3: Also, the conditional distribution of y given x_1, x_2 and x_3 is clearly not normal. A total of 100 observations are available.

Copula Transformed Multiple Regression: Data Sets

Other Data sets (not discussed):

ii) Financial Indexes Data
iii) SENIC Data
iv) Prostate Cancer Data
v) Real Estate Sales Data
vi) Used Car Data
vii) University Admissions Data

Copula Transformed Multiple Regression: Analyses

For all the above data set we will t the multiple linear regression model regressing y on k explanatory variables $x_1; x_2; :::; x_k:$ Clearly, the value of k is di erent for various data sets. No cross product or higher degree polynomial terms are assumed.

The objective is to compare the regression models tted on the original data with those obtained by tting the equivalent model on the corresponding Gaussian-copula transformed data.

Copula Transformed Multiple Regression: Analyses

Speci cally, since the number of observations and the functional forms of the models will be the same in the two situations, the coe cient of determination R^2 values can be compared, along with the statistical signi cance of the models.

However, from a practical point of view, quality of prediction is also important and thus, we will also compare the prediction errors as well as the prediction intervals.

[**After all we are going to predict the missing values so quality of prediction better be superior.**]

For a fair comparison, these two will be obtained for the original response variable.

Copula Transformed Multiple Regression: Analyses

Here we naturally replace $C = C_{\boldsymbol{R}}$ (Gaussian Copula), $G = \Phi_{\boldsymbol{R}}$ (as we are using Gaussian-copula transformation) and $F(\cdot)$ is the empirical distribution function of the bivariate data on $\begin{matrix} y \\ x \end{matrix}$:

$\boldsymbol{R}$ is the correlation matrix.

Copula Transformed Multiple Regression: Analyses

To make these predictions independent of the model tting process, in all cases, divide the data into training and test sets by respectively assigning the odd (even) numbered observations to the training (test) sets.

Since the Gaussian copula is used, data on all the transformed variables have zero mean and unit standard deviation.

That is however a non-issue, since, R^2 as well as F-test for the model are invariant of such a transformation.

Again, since predictions are obtained in both the cases, for the data on original scale, such a location shift and scaling change do not gure in the comparison.

Copula Transformed Multiple Regression: Algorithm

Algorithmically, the following steps are adopted in the sequence.

1. Transform the training raw data on random vector $\begin{matrix} Y \\ \boldsymbol{X} \end{matrix}$ to data on uniform random variables $U_Y; U_{X_1}; \ldots; U_{X_k}$ using the empirical cumulative distribution function estimated from the data. From the estimated covariance matrix, a correlation matrix for $\begin{matrix} Y \\ \boldsymbol{X} \end{matrix}$ and the corresponding empirical correlation matrix of $\boldsymbol{U} = (U_Y; U_{X_1}; \ldots; U_{X_k})'$ are obtained. These provide the estimates of the copula parameters.

Copula Transformed Multiple Regression: Algorithm

2. This is our target correlation matrix and we want, more or less, the same correlation among the multivariate uniform variables and among the multivariate normal transformed variables. Using the inverse multivariate normal cumulative distribution function on $\boldsymbol{U}$; we obtain the transformed data which are jointly distributed as the multivariate normal. We denote this by $\overset{Y}{\boldsymbol{X}}$.

Copula Transformed Multiple Regression: Algorithm

4. For predictions, comparison is appropriate only in the original scale. That is readily available for the rst model.

   a78o 7 rg i rg ever,6(the)-326(o)27second7(014rst)-32onlyrea4rstfo

Copula Transformed Multiple Regression: Algorithm

Let the corresponding two predicted values of y be $\hat{y}$ and $\hat{y}_c$ respectively. Then for the test data set, the two sum of squared prediction errors (SSPE) are given by

$$SSPE_{Raw} = \sum_{test\ data} (y_i - \hat{y}_i)^2 \tag{8}$$

and

$$SSPE_{Copula} = \sum_{test\ data} (y_i - \hat{y}_{c;i})^2. \tag{9}$$

Copula Transformed Multiple Regression: Algorithm

To obtain the prediction interval (say for a future observation) on the original scale and using model in (**??**), a little more care is needed. For a given $\boldsymbol{X} = \boldsymbol{x}_f$ (and hence $\boldsymbol{X} = \boldsymbol{x}_f$), denote the predicted value of y using model (**??**) by $\hat{y}_f$ and let the corresponding prediction interval be $(\hat{y}_{f;L}, \hat{y}_{f;U})$. Since $\hat{y}_f$, $\hat{y}_{f;L}$ and $\hat{y}_{f;U}$ are the quantities about a future incoming observation, correspondence may not be readily available within the data set.

Copula Transformed Multiple Regression: Wicklin's Data

We circumvent this problem by simulating a large number of observations from the $k-$ dimensional multivariate uniform distribution corresponding to our copula, and compute the corresponding values of y and y . Let these simulated quantities be denoted by placing a tilde ( ) above the corresponding variable.

If $\hat{y}_f$ is sandwiched between two such (closest) simulated values, say $\tilde{y}_t$ and $\tilde{y}_{t+1}$, then the predicted value of y, say $\hat{y}_{c;f}$ can be obtained by interpolation from $\tilde{y}_t$ and $\tilde{y}_{t+1}$, each of which has one to one correspondence with $\tilde{y}_t$ and $\tilde{y}_{t+1}$ via $\tilde{u}_{y;t}$ and $\tilde{u}_{y;t+1}$.

The same procedure is followed to interpolate the two prediction limits corresponding to $\hat{y}_{f;L}$ and $\hat{y}_{f;U}$. Accordingly, a prediction interval $(\hat{y}_{c;f;L}\ ,\ \hat{y}_{c;f;U})$ is obtained.

Copula Transformed Multiple Regression: Wicklin's Data

We have done this for our data sets for all the observations and plotted them against the serial number, which represents the increasing order of the (raw) data, on the response variables.

Note that this approach will also be applicable and should be followed, in the real situations when the prediction is an important objective.

Copula Transformed Multiple Regression: Wicklin's Data

I will describe the analysis of Wicklin's data in detail so as to fully appreciate the steps of the modeling and interpretation.

The data set, consisting of 100 observations, is first arranged in the increasing order of the response variable.

We have divided the data into training data and test data, each consisting of fifty observations. Increasing order of values on response variable and taking alternative values in the training data and test data, respectively, ensure, that the two data sets are largely similar and represent the same underlying population.

R^2, Adjusted R^2, p-values and F-tests corresponding to model are obtained for the training data. For prediction, test data will be used.

Copula Transformed Multiple Regression: Wicklin's Data

Residual Plots:

Figures 2 and 3 respectively represent the residual plots for the training data for the traditional multiple regression analysis of raw data and that for the Gaussian copula transformed data. The patterns in Figure 2 clearly indicate non-randomness and a poor t. On the contrary, the residual plot is near-ideal in Figure 3.

Copula Transformed Multiple Regression: Wicklin's Data

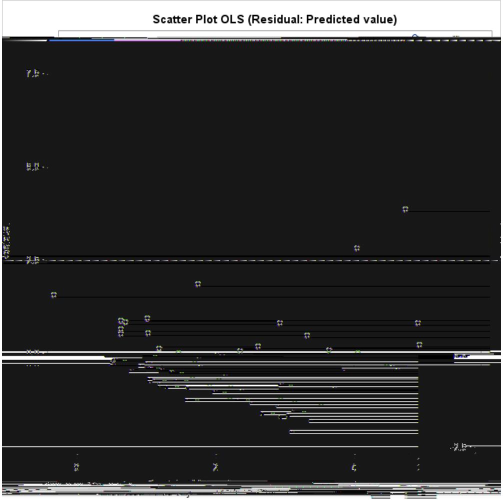


Figure: Figure 2: Wicklin's data: Scatter plot of residuals for raw training data

Copula Transformed Multiple Regression: Wicklin's Data

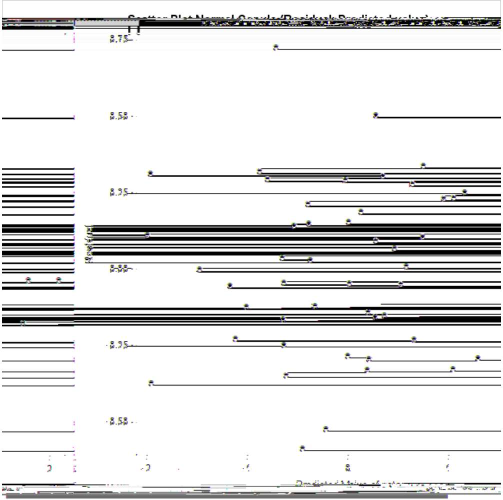


Figure: Figure 3: Wicklin's data: Scatter plot of residuals for copula

Copula Transformed Multiple Regression: Wicklin's Data

QQ Plots:
The same contrast between the two approaches is found between the two $Q - Q$ plots, given in Figures 4 and 5 for the corresponding residuals.

Copula Transformed Multiple Regression: Wicklin's Data

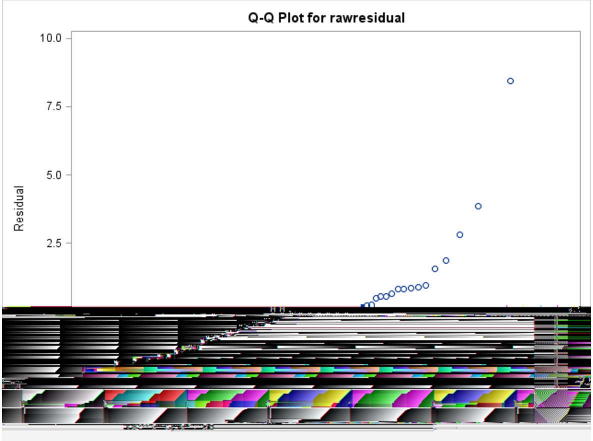


Figure: Figure 4: Wicklin's data: Residual $Q-Q$

Copula Transformed Multiple Regression: Wicklin's Data

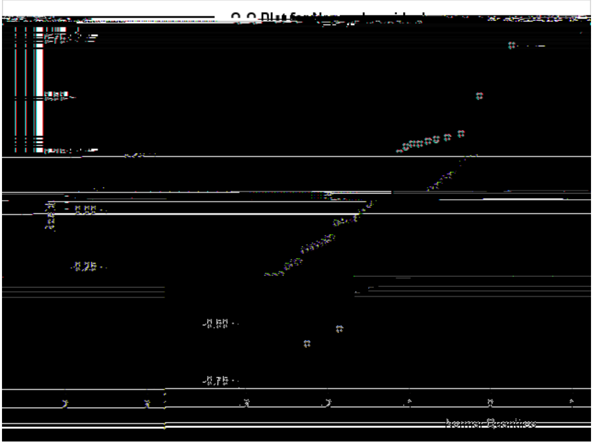


Figure: Figure 5: Wicklin's data: Residual $Q-Q$

Copula Transformed Multiple Regression: Wicklin's Data

Table 1 gives the values corresponding to model t and the statistical signi cance of the model. Drastic improvement in R^2 (38.86% vs. 92.47%) and Adjusted R^2 values is established. The same can be said about model F-statistics and corresponding p-values.

Copula Transformed Multiple Regression: Wicklin's Data

Table 2, shows point predictions for the data along with corresponding 95% prediction intervals.

For the sake of brevity, only the rst ten, last ten and middle ten observations of test data are presented. The superiority of Gaussian copula based approach is readily seen. All the point predictions using this approach are closer to the true observed responses. For the observation number 50 of the table, the observed value of the response variable is relatively very large.

Copula Transformed Multiple Regression: Wicklin's Data

Table: Table 2A: Wicklin's data: Comparison between raw and copula regression models for test data

Obs.	y	$\hat{y}$	$\hat{y}$	95% Raw Pred. Int.	95% Copula Pred. In
1	0.140	-1.265	0.144	(-5.050 , 2.521)	(0.132 , 0.271)
2	0.248	-0.359	0.248	(-4.076 , 3.359)	(0.133 , 0.356)
3	0.249	0.085	0.284	(-3.496 , 3.666)	(0.168 , 0.396)
4	0.283	0.063	0.334	(-3.565 , 3.691)	(0.247 , 0.492)
5	0.293	0.148	0.284	(-3.440 , 3.736)	(0.168 , 0.396)
6	0.316	0.539	0.374	(-2.998 , 4.076)	(0.254 , 0.615)
7	0.353	1.015	0.360	(-2.553 , 4.583)	(0.249 , 0.570)
8	0.369	-0.913	0.306	(-4.703 , 2.876)	(0.222 , 0.484)
9	0.385	1.536	0.485	(-2.015 , 5.088)	(0.321 , 0.913)
10	0.396	1.268	0.415	(-2.268 , 4.803)	(0.291 , 0.761)

Copula Transformed Multiple Regression: Wicklin's Data

Table: Table 2B: Wicklin's data: Comparison between raw and copula regression models for test data

Obs.	y	$\hat{y}$	$\hat{y}$	95% Raw Pred. Int.	95% Copula Pred. In
21	0.751	1.781	0.734	(-1.762 , 5.324)	(0.396 , 1.217)
22	0.770	-0.284	0.853	(-4.034 , 3.467)	(0.437 , 1.567)
23	0.822	0.455	0.923	(-3.194 , 4.105)	(0.487 , 1.570)
24	0.854	1.967	0.749	(-1.547 , 5.480)	(0.398 , 1.350)
25	0.922	1.731	1.079	(-1.807 , 5.269)	(0.669 , 2.306)
26	0.933	2.989	0.965	(-0.583 , 6.560)	(0.540 , 1.771)
27	0.981	2.526	1.414	(-1.030 , 6.082)	(0.801 , 2.758)
28	1.002	2.852	0.849	(-0.745 , 6.448)	(0.451 , 1.498)
29	1.078	1.230	0.918	(-2.321 , 4.780)	(0.487 , 1.564)
30	1.133	2.267	1.207	(-1.338 , 5.872)	(0.701 , 2.686)

Copula Transformed Multiple Regression: Wicklin's Data

Table: Table 2C: Wicklin's data: Comparison between raw and copula regression models for test data

Obs.	y	$\hat{y}$	$\hat{y}$	95% Raw Pred. Int.	95% Copula Pred. In
41	2.689	1.925	2.685	(-1.627 , 5.478)	(1.212 , 4.758)
42	2.751	3.061	3.191	(-0.489 , 6.612)	(1.563 , 7.035)
43	2.763	3.589	2.752	(-0.007 , 7.186)	(1.397 , 5.562)
44	3.456	1.830	4.112	(-1.726 , 5.387)	(1.865 , 8.987)
45	3.951	4.698	4.625	(0.914 , 8.481)	(2.559 , 11.435)
46	4.290	3.753	2.758	(0.103 , 7.402)	(1.400 , 5.999)
47	4.883	3.236	4.860	(-0.383 , 6.855)	(2.690 , 12.911)
48	6.952	3.790	5.731	(0.068 , 7.513)	(2.747 , 16.067)
49	9.133	3.052	9.577	(-0.499 , 6.604)	(4.236 , 20.029)
50	20.029	4.557	7.294	(0.809 , 8.305)	(3.426 , 18.878)

Copula Transformed Multiple Regression: Wicklin's Data

For Obs. 50: Observed response is very large; Both approaches underpredict the true response. Yet, the Gaussian-copula based prediction is still closer.

The prediction intervals as given in Table 2 and also graphed in Figures 6 and 7, further show that the prediction intervals are usually (and considerably) narrower when the approach is based on Gaussian copula, as compared to the raw data based regression.

The only few exceptions occur for the later few observations, but as seen in Table 2 for the last two observations, the prediction intervals based on usual regression analysis of data do not even contain the true observed value, while those based on the copula-regression approach do so for all observations except the last one.

Copula Transformed Multiple Regression: Other Data Sets

The residuals from the regression for the raw data for many of the datasets indicated earlier exhibit the violation of multivariate normality and linearity of regression.

The copula transformation to multivariate normality all together circumvents these issues rather than diagnosing and correcting each of them one by one. As Cherubini, Gobbi, Mulinacci and Romagnoli in their book (p. 30), explicitly point out,

Copula Transformed Multiple Regression: Wicklin's Data

\

Copula Transformed Multiple Regression: Wicklin's Data

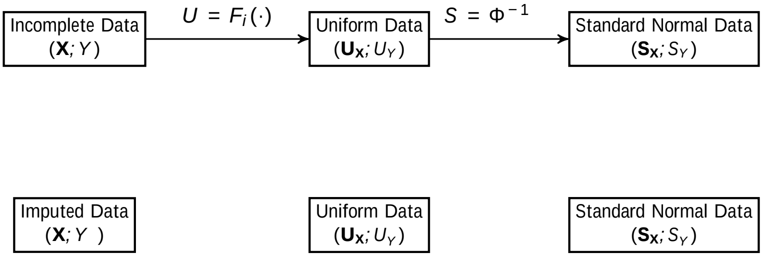
Table: Table 3: A comparison of the two regression models for various data sets (R = Raw data; G-C = Gaussian Copula Transformed)

Sr. No.	Data	Model based on[a]	Est. Skewness: Mardia's Skewness ($\hat{\beta}$)[b]	Est. Skewness: PC Skewness ($\hat{\eta}$)[b]	R^2	Adj. R^2	p-Val Model	Ave Sqrd Pred Err (Test Data)
1.	Wicklin's							
	$n_{training} = 50$	R	35.724	0.412	0.389	0.349	< 0:0001	6.997
	$n_{test} = 50$	G-C	0.439	0.024	0.923	0.920	< 0:0001	3.383
2.	Financial Indexes							
	$n_{training}$							

Imputing the missing values through copula transformation

Denote the fully observed variables (complete covariates) by $\mathbf{X} = (\mathbf{X}_{\text{obs}}; \mathbf{X}_{\text{mis}})^{005}$

Imputing the missing values through copula transformation



Imputing the missing values through copula transformation

Algorithm - Univariate Missing Data Pattern:

Imputing the missing values through copula transformation

- Use one of the imputation procedures (e.g. regression, MCMC, FCS) as desired, to impute all missing values and obtain dataset $(\mathbf{S_X}; S_Y)$ with imputed data.
- Back-transform the lled-in data to original scale via $U_Y = \Phi(S_Y)$ according to the inverse of empirical marginal distribution of Y, i.e., $Y = F_Y^{-1}(U_Y)$.

Imputing the missing values through copula transformation

We apply the Iman-Conover method to generate skewed multivariate datasets.

The reason we chose this method is that we can specify the marginal distribution of each variable and also the correlation structure.

We design two groups for multivariate data setting with marginals of components as follows.

Table: Table 7: Marginal distributions of simulated data sets using Iman-Conover method

Group	X_1	X_2	X_3	X_4
1	Log-normal (0,)	Pareto (1,1)	Normal (0, 1)	Uniform (0, 1)
2	Log-normal (0,)	Normal (0,1)	Exp (1)	Uniform (0, 1)

Imputing the missing values through copula transformation

The sample size is taken as 100 and the number of missing cases as 5.

To evaluate the quality of imputation, simulate each scenario NSIM=1,000 times and k imputation(s) and compute the mean of the sum of squared residuals by

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

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}. Here $k = 1$ for single imputation and $k > 1$ for multiple imputations.

Imputing the missing values through copula transformation

Table:

Imputing the missing values through copula transformation

Table: Table 9: Comparison between original data and copula-transformed data using single imputation for Group 2

σ	Correlation Structure	Orig.(nor.)	MSSR Cop-tran.	Ratio (O=C)	% SSR ($O > C$)
	$Corr_1$	13.64	13.27	1.03	58.3
	$Corr_2$	12.38	6.20	2.00	92.0
	$Corr_3(\rho = 0{:}5)$	16.29	16.24	1.00	55.4
1.0	$Corr_3(\rho = 0{:}9)$	8.70	6.48	1.34	72.2
	$Corr_1$	4,141.31	3,614.82	1.15	73.6
	$Corr_2$	4,220.18	3,273.75	1.29	91.2
	$Corr_3(\rho = 0{:}5)$	4,264.53	3,851.76	1.11	70.2
2.0	$Corr_3(\rho = 0{:}9)$	4,197.88	2,793.05	1.50	84.1
	$Corr_1$	2,773,953.52	1,760,982.56	1.58	81.7
	$Corr_2$	2,873,573.64	1,885,597.04	1.52	91.0
	$Corr_3(\rho = 0{:}5)$	2,603,019.87	1,790,836.43	1.45	80.1
3.0	$Corr_3(\rho = 0{:}9)$	3,436,021.98	1,604,055.17	2.14	88.4

Imputing the missing values through copula transformation

Table: Table 10: Comparison between original data and copula-transformed data using FCS Regression multiple imputation for **Group 1**

σ	Correlation	

Imputing the missing values through copula transformation

Table: Table 11: Comparison between original data and copula-transformed data using FCS Regression multiple imputation for **Group 2**

σ	Correlation Structure	MSSR Orig.(nor)	MSSR Cop.-tran.	Ratio $(O=C)$	% SSR $(O > C)$
1.0	$Corr_3(\rho = 0{:}5)$	164.68	123.08	1.34	85.00%
	$Corr_3(\rho = 0{:}6)$	149.42	109.95	1.36	83.00%
	$Corr_3(\rho = 0{:}7)$	132.64	95.48	1.39	82.20%
	$Corr_3(\rho = 0{:}8)$	112.58	79.57	1.41	83.10%
	$Corr_3(\rho = 0{:}9)$	84.70	53.40	1.59	86.40%
2.0	$Corr_3(\rho = 0{:}5)$	56,391.21	39,703.39	1.42	86.80%
	$Corr_3(\rho = 0{:}6)$	56,417.01	37,341.96	1.51	86.30%
	$Corr_3(\rho = 0{:}7)$	55,458.96	35,888.29	1.55	86.60%
	$Corr_3(\rho = 0{:}8)$	53,648.33	39,250.03	1.37	86.50%
	$Corr_3(\rho = 0{:}9)$	50,264.93	30,896.51	1.63	90.30%
3.0	$Corr_3(\rho = 0{:}5)$	74,206,031.68	44,532,451.31	1.67	86.90%
	$Corr_3(\rho = 0{:}6)$	77,116,532.29	42,631,176.04	1.81	87.50%
	$Corr_3(\rho = 0{:}7)$	76,957,607.66	41,943,270.76	1.83	88.00%
	$Corr_3(\rho = 0{:}8)$	75,842,964.49	55,855,584.64	1.36	87.70%
	$Corr_3(\rho = 0{:}9)$	75,184,163.76	42,523,936.26	1.77	90.80%

Imputing the missing values through copula transformation

Table: Table 12: Comparison between original data and copula-transformed data using MCMC multiple imputation for **Group 1**

σ	Correlation Structure	MSSR Orig.(nor)	Cop.-tran.	Ratio (O=C)	% SSR ($O > C$)
1.0	$Corr_3(\rho = 0{:}5)$	300.95	132.50	2.27	60.00%
	$Corr_3(\rho = 0{:}6)$	264.60	113.77	2.33	64.70%
	$Corr_3(\rho = 0{:}7)$	276.81	95.13	2.91	68.40%
	$Corr_3(\rho = 0{:}8)$	211.99	72.71	2.92	74.10%
	$Corr_3(\rho = 0{:}9)$	450.31	47.32	9.52	80.50%
2.0	$Corr_3(\rho = 0{:}5)$	71,610.82	36,851.77	1.94	66.10%
	$Corr_3(\rho = 0{:}6)$	156,457.61	34,240.23	4.57	70.20%
	$Corr_3(\rho = 0{:}7)$	171,850.60	32,943.64	5.22	75.00%
	$Corr_3(\rho = 0{:}8)$	60,038.31	26,745.73	2.24	79.40%
	$Corr_3(\rho = 0{:}9)$	131,115.17	21,115.43	6.21	83.90%
3.0	$Corr_3(\rho = 0{:}5)$	51,837,663.26	23,896,573.91	2.17	70.10%
	$Corr_3(\rho = 0{:}6)$	165,243,871.10	23,148,399.65	7.14	73.80%
	$Corr_3(\rho = 0{:}7)$	183,802,007.11	23,891,866.47	7.69	77.10%
	$Corr_3(\rho = 0{:}8)$	48,189,008.89	16,717,322.02	2.88	80.60%
	$Corr_3(\rho = 0{:}9)$	67,349,495.86	13,494,377.61	4.99	85.50%

Imputing the missing values through copula transformation

Table: Table 13: Comparison between original data and copula-transformed data using MCMC multiple imputation for **Group 2**

σ	Correlation Structure	MSSR Orig.(nor)	MSSR Cop.-tran.	Ratio $(O{=}C)$	% SSR $(O > C)$
1.0	$\mathrm{Corr}_3(\rho = 0{:}5)$	127.54	132.10	0.97	57.60%
	$\mathrm{Corr}_3(\rho = 0{:}6)$	115.44	113.55	1.02	61.10%
	$\mathrm{Corr}_3(\rho = 0{:}7)$	102.34	94.08	1.09	65.30%
	$\mathrm{Corr}_3(\rho = 0{:}8)$	86.99	74.35	1.17	72.20%
	$\mathrm{Corr}_3(\rho = 0{:}9)$	65.26	48.47	1.35	81.20%
2.0	$\mathrm{Corr}_3(\rho = 0{:}5)$	39,757.02	35,230.30	1.13	67.10%
	$\mathrm{Corr}_3(\rho = 0{:}6)$	39,631.01	30,936.11	1.28	69.00%
	$\mathrm{Corr}_3(\rho = 0{:}7)$	38,944.15	27,789.97	1.40	73.10%
	$\mathrm{Corr}_3(\rho = 0{:}8)$	38,106.04	29,270.04	1.30	79.20%
	$\mathrm{Corr}_3(\rho = 0{:}9)$	35,414.91	25,431.77	1.39	84.30%
3.0	$\mathrm{Corr}_3(\rho = 0{:}5)$	43,374,578.45	22,551,396.98	1.92	71.60%
	$\mathrm{Corr}_3(\rho = 0{:}6)$	44,902,158.06	19,198,351.45	2.34	72.20%
	$\mathrm{Corr}_3(\rho = 0{:}7)$	44,977,065.21	17,619,423.92	2.55	77.10%
	$\mathrm{Corr}_3(\rho = 0{:}8)$	45,876,052.72	30,256,223.22	1.52	80.60%
	$\mathrm{Corr}_3(\rho = 0{:}9)$	44,235,290.84	29,549,312.57	1.50	85.50%

Conclusions

- Much of the dependence based multivariate analyses for nonnormal data can be done using the copula transformation. However information about marginals is lost.
- Similar work has been done for principal component analyses, factor analyses, structural equation modeling.
- For missing data imputation for nonnormal situations, this approach is very handy. Further extensive studies showed that (for multivariate Lomax distribution) results under copula transformation are as good as those obtained by imputation by conditional expectations (assuming MCAR).
- Comparison of imputation done by our transformation to normality and that done by Box-Cox transformation showed our approach is much superior.