Handling Dropouts by Copulas

Ene Käärik
Institute of Mathematical Statistics
University of Tartu
Estonia

ABSTRACT
The author shows how the copulas can be used to analyze repeated measurements with missing data. The main problem in practice is that there does not exist any theoretical joint distribution describing the repeated measurements, and then the copula function is a good tool to create the joint distribution. As the measurements are, in general, dependent, the conditional distribution of measurement in moment k, conditioned by the past measurements, might be highly informative for forecasting missing values. Imputations can be performed using conditional models. A simple method for imputation is presented using normal copula. The explicit formula to impute dropouts has been derived in the case of compound symmetry. The result can be easily applied to non-normal marginals using normalizing transformation.

KEY WORDS
Repeated measurements, Dropouts, Imputation, Normal copula, Conditional distribution

1 Introduction

The paper is focused on te repeated measurements study with missing data. Let us have a process with outcome variable measured at time points $1, \ldots, m$. Denote the outcome measured at time point j by X_j and suppose it has distribution F_j from the family $\boldsymbol{P}, j = 1, \ldots, m$. We do not put any restrictions on the family $\boldsymbol{P}$ of marginal distributions. Usually, the exponential family is used.

The measurements form an m-variate random vector $X = (X_1, \ldots, X_m)$. In general the distribution of the vector X is not known. Often it happens that it is possible to determine a model family $\boldsymbol{P}$ for marginal distributions X_j, but there does not exist any known family of multivariate distributions suitable to describe the joint distribution of the vector X. In practice, in most cases the multivariate normal distribution is used because of its easy definition via correlation structure. Nevertheless, often the marginal distributions do not fit with assumption of normality. As a novelty, Lindsey [1] suggested to use the skew Laplace and the multivariate t-distribution, which are not members of the exponential family.

Assume we have a sample of size n of measurements X_j; then the data form a matrix $\mathbf{X} = \{x_{ij}\}, i = 1, \ldots, n; j = 1, \ldots, m$. In practice, it often happens that not all the measurements are known, there are some missing values or dropouts in the matrix $\mathbf{X}$.

Common approach to treating missing data in repeated measurements study is to consider dropouts, where sequences of measurements on some units terminate prematurely. In the case of dropouts, the missingness pattern is said to be monotone if, whenever an observation x_{ik} is missing, x_{is} is also missing for all $s > k$. All observations on a subject are obtained until a certain time point, after which all measurements are missing. Let n_j denote the number of objects for which X_j is observed. If the pattern is monotone, then $n_j \geq n_{j+1},\ j = 1, \ldots, m-1$.

Let k be the time point at which the dropping out process starts. Without restrictions we can assume that until the time point $k-1$ we have complete data, $n_i = n,\ i = 1, \ldots, k-1$.

Usually, dropouts are to be distinguished from intermittent missing values, where an observed sequence has some gaps, i.e. the set of intended times of measurements is not common to all units (unbalanced data).

The classification of dropout processes is given by Rubin (see for example, [2], [3]), which includes a hierarchy of missing values mechanism:

- Completely random dropout (CRD) – dropout and measurement processes are independent, so dropouts are simply random missing values;
- Random dropout (RD) – dropout process depends on observed measurements;
- Informative dropout (ID) – dropout process depends additionally on unobserved measurements, i.e. those measurements that would have been observed if the subject had not dropped out.

Lindsey [4] proposed another typology of randomness for dropouts that relies on using a survival model for the dropout process. In terms of a stochastic process, dropping out is a change of state. In this case, the repeated measurements data and dropout process can be modelled simultaneously, each conditional on the complete previous history. According to Lindsey, types of dropout mechanisms are based on the risk of dropout.

If a dropout process is random, then a valid analysis can be performed using a likelihood method that ignores

the dropout mechanism: the parameters describing the measurement process are functionally independent of the parameters describing the dropout process. However, it may be difficult a priori to justify the assumption of random dropout.

We are concerned with missing outcome variable, i.e. measurements that potentially could be obtained. In many theoretical and practical tasks it is necessary to know the values of missing measurements, and there exists a long list of single and multiple imputation methods such as conditional and unconditional means, hot deck, linear prediction etc. Most imputation methods require the CRD assumption and *ad hoc* adjustments to yield satisfactory point estimates [5].

We use the idea of imputing a missing value based on conditional distributions. The problem is that the joint distribution may be unknown, but using the copula function it is possible to find approximate joint and conditional distributions.

The aim of this work is, using the fact that all imputation models are principally multivariate predictive models, to implement the concept of copula for developing a methodology for solving of the imputation problem.

2 Basic definitions

We consider a sequence of measurements up to k observations $X_1, X_2, \ldots, X_k$, where k is the dropout point and vector $H = (X_1, X_2, \ldots, X_{k-1})$ is called *history* (or *past*) of measurements.

In the case of repeated measurements often the variables X_j are rather strongly dependent (due to individual specialities of objects in the sample) and hence the vector H contains a big amount of information about the distribution of the variable X_k that can be used for forecasting the value of the missing observation. If the joint distribution of vector X is known, then the conditional distribution of X_k conditioned by the realized past H can be used to define the estimate of missing value, its limits and measures of variability.

We will generate the joint distribution using some copula. Using known marginal distributions $F_1(x_1), \ldots, F_k(x_k)$ and a copula C, the function

$$C(F_1(x_1), \ldots, F_k(x_k)) = F(x_1, \ldots, x_k)$$

defines a joint distribution function [6].

If marginal distributions are continuous, then the copula C is unique for every fixed F and equals

$$C(u_1, \ldots, u_k) = F(F_1^{-1}(u_1), \ldots, F_k^{-1}(u_k)),$$

where $F_1^{-1}, \ldots, F_k^{-1}$ are the quantiles functions of given marginals and $u_1, \ldots, u_k$ are uniform $[0, 1]$ variables.

If the random variables $X_1, \ldots, X_k$ are *independent*, then the copula function that links their marginals is the *product copula* $C(F_1, F_2, \ldots, F_k) = F_1 \cdot F_2 \cdot \ldots \cdot F_k$. Hence, in the case of completely random dropouts (CRD) the dropout is independent of the measurements and it is possible to use the product copula.

If C and $F_1, \ldots, F_k$ are differentiable, then the joint density $f(x_1, \ldots, x_k)$ corresponding to joint distribution $F(x_1, \ldots, x_k)$ can be written as a product of the marginal densities and the copula density ([6], [7]):

$$f(x_1, \ldots, x_k) = f_1(x_1) \times \ldots \times f_k(x_k) \times c(F_1, \ldots, F_k), \quad (1)$$

where $f_i(x_i)$ is the density corresponding to F_i and the copula density is defined in the following way

$$c = \frac{\partial^k C}{\partial F_1 \cdots \partial F_k}. \quad (2)$$

3 Normal copula approach

In the following we will use the multivariate normal copula ([1], [7], [8], [9], [10], [11]). Multivariate normal copula represents the dependencies between univariate marginals, allowing any positive-definite correlation matrix. By definition, the k-variate Gaussian copula with k Gaussian marginals corresponds to the k-variate Gaussian distribution.

For instance, the bivariate normal copula C_N is defined as

$$C_N(u, v, \theta) = \Phi_2[\Phi_1^{-1}(u), \Phi_1^{-1}(v), \theta], \quad (3)$$

where Φ_2 is the standardized bivariate normal distribution with correlation coefficient θ and Φ_1 is the univariate standard normal distribution. This notation can be easily extended to multivariate cases with θ replaced by a correlation matrix R.

One possibility is to use the normal copula that creates a multivariate distribution, generalizing normal distribution in the following sense: the dependencies are defined by some nonparametric measure of dependence - e.g., Spearman's rho or Kendall's tau that measure monotonic dependencies and are invariant to all scale transformations. Also the marginal distributions that are assumed to be continuous can substantially differ from normal ones.

The only condition for creating the normal copula is that the dependence matrix must be positively defined.

But, as the number of different dependence parameters is in the case of k-variate distribution is $k(k-1)/2$, usually some simple model structure describing the dependence matrix are used. The most simple of them are (see for example, [5]):

- *The compound symmetrical structure*: the correlation between any two time points is the same;
- *The autoregressive structure*: the dependence between observations decreases by geometric law as the measurements get further in time.

In the case of repeated measurements both structures have meaning, as the common dependence can be created by individual specialities; if there exists a time-dependent trend,

then the second model can fit. Probably, in many cases the mixture of two models works.

Let us use now on the whole history H of repeated measurements. There is (at least one) subject i such that the measurements $X_1, X_2, \ldots, X_{k-1}$ are observed and X_k drops out. Corresponding marginal distribution functions are $F_j(x)$ with density functions $f_j(x)$ and univariate standard normal distributions correspondingly are $\Phi_1(x)$ and $\phi_1(x)$, where $j = 1, \ldots, k$.

Let $\Phi_N(x_1, \ldots, x_k|\mu, \Sigma)$ represent the joint multivariate normal distribution with mean μ and covariance matrix Σ, and let $\phi_N(x_1, \ldots, x_k|\mu, \Sigma)$ represent the multivariate normal density. Assuming that marginals are standardized we can write correlation matrix R instead of covariance matrix Σ.

Then, according to (1), we get

$$\phi_N(x_1, \ldots, x_k|R) = \phi_1(x_1) \times \ldots \times \phi_1(x_k)$$
$$\times c_N[\Phi_1(x_1), \ldots, \Phi_1(x_k); R^*], \quad (4)$$

where c_N is the multivariate normal copula density (defined by (2)) we are interested in, R^* is matrix of dependence measures, usually Spearman's ρ or Kendall's τ.

If the multivariate normal distribution is parameterized by Pearson product-moment correlation r, the formula $\rho = \frac{6}{\pi} \arcsin(\frac{r}{2})$ can be used to transform the Pearson correlations to the Spearman ones.

To solve the equation (4) for the copula density, we use the normal inverse transformation Φ_1^{-1}. Defining $Y_i = \Phi_1^{-1}[F_i(X_i)]$, $i = 1, \ldots, k$, we get following formula (see [7], [11]):

$$c_N[\Phi_1(y_1), \ldots, \Phi_1(y_k); R^*] = \frac{\exp\{-y^T(R^{-1} - I)y/2\}}{|R|^{1/2}}, \quad (5)$$

where $y = (y_1, \ldots, y_k)$ and I is the $k \times k$ identity matrix.

For constructing a multivariate density we now use the marginals $F_1(x_1), \ldots, F_k(x_k)$ and copula density c_N from (5) as a dependence function. Thus, we obtain the joint density, as follows.

$$\phi_N(x_1, \ldots, x_k|R) = \phi_1(x_1) \times \ldots \times \phi_1(x_k) \times$$
$$\times \frac{\exp\{-Q_k^T(R^{-1} - I)Q_k/2\}}{|R|^{1/2}}. \quad (6)$$

where $Q_k = (\Phi_1^{-1}[F_1(x_1)], \ldots, \Phi_1^{-1}[F_k(x_k)])$.

3.1 Conditional density

To find the formula for imputation we should calculate the conditional density of the variable X_k (having the dropout) using the history H.

Taking into consideration the history, the whole correlation matrix can be partitioned as

$$R = \begin{pmatrix} R_{k-1} & r \\ r^T & 1 \end{pmatrix},$$

where R_{k-1} is the correlation matrix of the history $H = (X_1, \ldots, X_{k-1})$, and $r = (r_{1k}, \ldots, r_{(k-1)k})^T$ is the vector of correlations between the history and the time point k. If we suppose, for example, compound symmetry, then the correlations between all time points are equal, and $r = (\rho, \ldots, \rho)^T$, where ρ can be estimated by R_{k-1}.

Taking into consideration the definition of conditional density and expression (6), we get the conditional density as follows:

$$f(x_k|H; R^*) = f_k(x_k)$$
$$\times \frac{\phi_k(Q_k; R)}{\phi_1(\Phi_1^{-1}[F_k(x_k)]) \times \phi_{k-1}(Q_{k-1}; R_{k-1})}.$$

where $Q_{k-1} = (\Phi_1^{-1}[F_1(x_1)], \ldots, \Phi_1^{-1}[F_{k-1}(x_{k-1})])$. Substitution into the expression above normal densities and considering the notation $Y_i = \Phi^{-1}[F_i(X_i)]$ results in

$$f(y_k|H; R^*) = \phi_1(y_k)$$
$$\times \exp\{-\frac{1}{2}[\frac{(y_k - r^T R_{k-1}^{-1}(y_1, \ldots, y_{k-1})^T)^2}{(1 - r^T R_{k-1}^{-1} r)} - y_k^2]\}$$
$$\times (1 - r^T R_{k-1}^{-1} r)^{-1/2}. \quad (7)$$

3.2 Imputation

To impute the dropouts we should find the maximum likelihood estimate, i.e. maximize equation (7) with respect to y_k.

Taking into consideration the whole history consisting of $(k-1)$ measurements, the following *formula for imputation* y_k is valid:

$$\hat{y}_k = \frac{\rho}{1 + (k-2)\rho} \sum_{i=1}^{k-1} y_i. \quad (8)$$

Formula (8) is usable for any marginals in the case of compound symmetry and gives transformed imputing value y_k. To impute the initial value we have to use the inverse transformation $x_k = F^{-1}[\Phi(y_k)]$.

Detailed derivation of the formula (8) is given in Appendix A.

If we have correlated data $X_1, \ldots, X_k$ from a distribution with continuous margins $F_1, \ldots, F_k$, then above derived normal-copula-based approach to point estimation of missing value requires the following steps.

1. Estimate marginal distributions either using parametric or non-parametric procedure.

2. Estimate the correlation structure of data. If we can accept the hypothesis about the compound symmetry structure, then go on, if not, then do not use this approach.

3. If the marginals are not normal, use the normalizing transformation $Y_i = \Phi^{-1}(F_i(X_i))$.

4. Estimate the missing value using the formula (8).

5. Use inverse transformation to get imputed value $X_k = F^{-1}[\Phi(Y_k)]$.

4 Illustration

To illustrate the normal-copula approach in the case of strongly skewed marginals and CRD and ID dropouts the following data are generated: the three random variables X_1, X_2, X_3 with constant correlations $r = 0.7$ (compound symmetry).

That means $X = (X_1, X_2, X_3)$ has 3-variate normal distribution. To get a skewed non-normal marginals the variables are transformed using following rules:

1. For maximum value $y_i = \max\{X_i\}$, $y_i = C_1 y_i$.

2. For every positive value $y_i = C_2 y_i$.

The constants are chosen $C_1 = 10, C_2 = 5$.

The transformed data are extended to positive direction. Assume the data represent repeated measurements, when the sample size $n = 25$. Due to small sample size every value is important, hence we have to impute the missing values.

The dropouts occur at last time point (in random variable X_3) and we examine 2 cases: 1) the maximal value is dropped out (case of ID), 2) the random value is dropped out (case of CRD).

Two methods for imputation are used: 1) imputation by linear prediction $X_3 = \beta_0 X_1 + \beta_2 X_2$ and 2) imputation by formula (8). To compare the results the average bias is calculated as average difference between observed values and imputed values: $Bias1$ is the bias using linear prediction and $Bias2$ is the bias when using formula (8) for imputation. Also the standard deviations of estimates have been calculated (denoted by $StdDev_1$ and $StdDev_2$ accordingly). Results are given in units of standard deviation of given marginals.

The results of 50 simulations are in following table.

Table 1.

Dropout	$Bias1$	$StdDev_1$	$Bias2$	$StdDev_2$
Maximal	2.4958	1.5424	2.0671	1.2437
Random	-0.4343	0.7005	-0.2322	0.3832

From here it follows that the imputation of the maximal value causes rather large bias, the result that is not surprising. The other conclusion is that using formula (8) for imputation gives somewhat better results, in both cases the imputation via copula gives more stabile solution with less bias.

5 Remarks and conclusions

The *normal copula* is useful because of easy implementation in practice and simple simulation method. However, there is increasing evidence indicating that normal assumptions are inappropriate in the real world. In general a multivariate normal distribution is not ideal and is valid when only measurement error is present, ignoring or poorly modelling the dependence between repeated measurements (see for example [1]).

To resolve this problem some other copulas for the joint distribution can be applied. For example, Lindsey and Lindsey [1] suggested Student's t-distribution, power-exponential or skew Laplace distribution for modelling repeated responses instead of normal distribution. Vandenhende and Lambert [12] tested several marginal distributions, Cauchy, Gamma, log-normal, for dropout model.

An important class of parametric copulas to model non-normal data is *the Archimedean copula* ([6], [13]). Vandenhende and Lambert [12] used *Frank's copula* from this family to model the dependence between dropout and responses.

Members of Archimedean copula class are constructed by a continuous, strictly decreasing, and convex function ϕ such that for example in bivariate case

$$C(u, v) = \phi^{-1}\{\phi(u) + \phi(v)\}$$

and ϕ is called the *generator* of the copula. Different choices of generator yield several important families of copulas. If we assume an Archimedean form of the copula, then the conditional distribution of X_k given past H is (see for example, [14])

$$F(x_k|H) = \frac{\phi^{-1}\{c_{k-1} + \phi(F_k(x_k))\}}{\phi^{-1}(c_{k-1})},$$

where $c_k = \phi[F_1(x_1)] + \ldots + \phi[F_k(x_k)]$.

Genest [15] gave the conditional mean function in the case of Frank's copula (with generator function $\phi(t) = \ln \frac{\exp -\alpha x - 1}{\exp \alpha - 1}$) as follows

$$E(x_k|H) = \frac{(1 - e^{-\alpha})xe^{-\alpha x} + e^{-\alpha}(e^{-\alpha x} - 1)}{(e^{-\alpha x} - 1)(e^{-\alpha} - e^{-\alpha x})}.$$

A general copula-based approach to point estimation of missing value requires following steps.

1. Modelling univariate marginal distributions either parametrically or non-parametrically. The copula method works with any given marginal distributions i.e. it does not restrict the choice of margins.

2. Specifying a copula. Because copulas are parametric families, the standard technique such as maximum likelihood can be used for estimation. For selecting the best-fitting copula some Goodness-of-Fit statistics can be used (see for example [9])

3. Find the conditional density.

4. Estimate the conditional mean (or corresponding median etc.), take it for imputed value.

References

[1] Lindsey, J.K, Lindsey, P.J. Multivariate distributions with correlation matrices for nonlinear repeated measurements, 2002 (www.luc.ac.be/˜jlindsey : Nov, 2004).

[2] Little, R.J.A. . Modeling the dropout Mechanism in Repeated-Measures Studies. *JASA*, 90, 431, 1995, 1112–1121.

[3] Diggle, P. J., Kenward, M. G. Informative dropout in longitudinal data analysis. *Applied Statistics*, 43 (1),1994, 49–93.

[4] Lindsey, J.K. Dropouts in longitudinal studies: definition and models. *Journal of Biopharmaceutical Statistics*, 10 (4), 2000, 503–525.

[5] Verbeke, G., Molenberghs, G. Linear mixed models for longitudinal data. *Springer Series in Statistics*, (New York: Springer Verlag, 2001).

[6] Nelsen R.B. (1998). An Introduction to Copulas. *Lecture Notes in Statistics*, 139, (New York: Springer Verlag, 1998).

[7] Clemen, R.T., Reilly, T. Correlations and Copulas for Decision and Risk Analysis. Fuqua School of Business, Duke University. *Management Science*, 45 (2), 1999, 208–224.

[8] Reilly, T. Modelling correlated data using the multivariate normal copula. *Proceedings of the Workshop on Correlated Data*, Trieste, Italy, 1999.

[9] Kouros, O., Pranab, K.S. Copulas: concepts and novel applications. *METRON-International Journal of Statistics*, LXI (3), 2003, 323–353.

[10] Lambert, P., Vandenhende, F. A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine*, 21, 2002, 3197–3217.

[11] Song, P.X.K. Multivariate dispersion models generated from Gaussian Copula. *Scandinavian Journal of Statistics*, 27, 2000, 305–320.

[12] Vandenhende, F., Lambert, P. On the joint analysis of longitudinal responses and early discontinuation in randomized trials. *Journal of Biopharmaceutical Statistics*, 12 (4), 2002, 425–440.

[13] Genest, C., Rivest, L.P. Statistical Inference Procedure for Bivariate Archimedean Copulas. *JASA*, 88 (423) 1993, 1034–1043.

[14] Frees, E.W., Valdez, E.A. Understanding Relationships Using Copulas. *North American Actuarial Journal*, 2 (1), 1997, 1–25.

[15] Genest, C. Frank's Family of Bivariate Distributions. *Biometrika*, 74, 1987, 549–555.

Appendix A – Formula for imputation

In this appendix we derive formula (8) for imputation. Consider the conditional density given by (7):

$$f(y_k|H;R^*) =$$
$$= \phi_1(y_k) \times \exp\{-\tfrac{1}{2}[\tfrac{(y_k - r^T R_{k-1}^{-1}(y_1,\ldots,y_{k-1})^T)^2}{(1-r^T R_{k-1}^{-1} r)} - y_k^2]\}$$
$$\times (1 - r^T R_{k-1}^{-1} r)^{-1/2}.$$

We should find the maximum likelihood estimate for y_k. It is easy to see, that when maximizing the last equation with respect to y_k, we get

$$y_k = r^T \cdot R_{k-1}^{-1} \cdot Y_{k-1}^*, \tag{9}$$

where $Y_{k-1}^* = (y_1, \ldots, y_{k-1})$.

Suppose now compound symmetry, then $r = (\rho, \ldots, \rho)^T$ and

$$R_{k-1} = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & & \ddots & \\ \rho & \rho & \ldots & 1 \end{pmatrix}$$

is the $(k-1) \times (k-1)$ correlation matrix for history, and its inverse is R_{k-1}^{-1}.

To simplify formula (9), we should first find the elements of R_{k-1}^{-1}. Due to special form of R_{k-1}, the inverse matrix is symmetrical with equal elements a on the main diagonal and equal off-diagonal elements b, the following forms are well-known.

$$a = 1 + \frac{(k-2)\rho^2}{1-(k-2)\rho^2+(k-3)\rho},$$
$$b = -\frac{\rho}{1-(k-2)\rho^2+(k-3)\rho}$$

and, hence,

$$a + (k-2)b = \frac{1}{1+(k-2)\rho}.$$

On the other hand, formula (9) can be rewritten as

$$y_k = \rho \times (1,\ldots,1) \times \begin{pmatrix} a & b & \ldots & b \\ b & a & \ldots & b \\ \vdots & & \ddots & \\ b & b & \ldots & a \end{pmatrix} \times (y_1, \ldots, y_{k-1})^T$$
$$= \rho \times [a + (k-2)b] \times \textstyle\sum_{i=1}^{k-1} y_i.$$

Taking into account the last result we get formula (8):

$$\hat{y}_k = \frac{\rho}{1+(k-2)\rho} \sum_{i=1}^{k-1} y_i.$$

□

Acknowledgements

This work is supported by Estonian Science Foundation grants No 5521 and No 5203.

Author also thanks Ene-Margit Tiit for substantial comments from an early draft of this paper and Meelis Käärik for helpful suggestions in matrix calculus.