Imputation by Gaussian Copula Model with an Application to Incomplete Customer Satisfaction Data

Meelis Käärik, Ene Käärik

Institute of Mathematical Statistics, University of Tartu, Estonia

OVERVIEW

1. Motivating example
2. Imputation. Basic definitions
3. Framework
4. Problem setting
5. Copula. Gaussian copula approach
6. Imputation algorithm
7. Application to Incomplete Customer Satisfaction Data
8. Summary. Remarks

MOTIVATING EXAMPLE

Customer satisfaction survey

Questionnaire – respondents (customers) give scores from least to most satisfied

Blocks of similar questions (correlated variables)

Each customer represents a company

Individual scores are important!

MOTIVATING EXAMPLE

Customer satisfaction survey

Questionnaire – respondents (customers) give scores from least to most satisfied

Blocks of similar questions (correlated variables)

Each customer represents a company

Individual scores are important!

$\Rightarrow$ Finding reasonable substitutes for missing values is of high interest

INCOMPLETE DATA

Consider correlated incomplete data

DEF. **Imputation** (filling in, substitution) is a strategy for completing missing values in data with plausible estimates.

Little & Rubin (1987)

- Imputation might seem like an unimportant distinction.
- There are many situations where the non-response mechanism needs to be considered explicitly, since it is of scientific interest itself.
- It makes sense to consider imputation of missing values separately from modelling data.

FRAMEWORK

Let $Y = (Y_1, \ldots, Y_v)$ be the random vector with correlated components Y_j

Consider data with n subjects

$$\mathbf{Y} = (\mathrm{Y}_1, ..., \mathrm{Y}_v),\ \mathrm{Y}_j = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix},\quad j = 1, \ldots, v$$

Ordered missingness: the columns of data matrix are sorted starting from the column with least missing values to the column with most missing values

Assume that first k $(k \geq 2)$ components are complete, then

$$\mathbf{Y} = (\mathbf{Y}^c, \mathbf{Y}^m)$$

$\mathbf{Y}^c = (\mathrm{Y}_1, \ldots, \mathrm{Y}_k)$ – complete data,
$\mathbf{Y}^m = (\mathrm{Y}_{k+1}, \ldots, \mathrm{Y}_v)$ – incomplete data.

DEPENDENCE between variables

$Y^c = (Y_1, \ldots, Y_k)$, Y_{k+1}

Correlation matrix: $\mathbf{R} = (r_{ij})$, $r_{ij} = corr(Y_i, Y_j)$, $i, j = 1, \ldots, k+1$

Partition of correlation matrix

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_k & r \\ r^T & 1 \end{pmatrix}$$

$\mathbf{R}_k$ – the correlation matrix of complete part $Y^c = (Y_1, ..., Y_k)$

$r = \begin{pmatrix} r_{1,k+1} \\ \vdots \\ r_{k,k+1} \end{pmatrix}$ – the vector of correlations between Y^c and Y_{k+1}.

PROBLEM SETTING

We use the *idea of imputing a missing value based on conditional distribution* of missing value conditioned to the observed values.

The joint distribution may be unknown, but *using the copula function* it is possible to find approximate joint and conditional distributions.

H. Joe (2001): "... if there is no natural multivariate family with a given parametric family for the univariate margins, *a common approach has been through copulas*"

COPULA

In 1959 Sklar introduced a new class of functions which he called *copulas.*

Sklar: if Q is a bivariate distribution function with margins $F(x)$, $G(y)$, then there exist a *copula* C such that

$$Q(x,y) = C(F(x), G(y)).$$

$\Rightarrow$ copula links joint distribution function to their one-dimensional marginals.

DEF. A *copula* is a function $C : [0,1]^2 \to [0,1]$ which satisfies:

- for every u, v in $[0,1]$, $C(u,0) = 0 = C(0,v)$, and $C(u,1) = u, \quad C(1,v) = v$;
- for every u_1, u_2, v_1, v_2 in $[0,1]$ such that $u_1 \leq u_2, \ v_1 \leq v_2$,

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$$

Example: *product copula* $\Pi(u,v) = uv$ characterizes independent random variables when the distribution functions are continuous.

GAUSSIAN COPULA APPROACH (1)

DEFINITION:

Let $\mathbf{R}$ be a symmetric, positive definite matrix with $diag(\mathbf{R}) = (1, 1, \ldots, 1)^T$ and Φ_{k+1} be the $k+1$-variate normal distribution function with correlation matrix $\mathbf{R}$, then the multivariate **GAUSSIAN COPULA** is defined as

$$C(u_1, \ldots, u_{k+1}; \mathbf{R}) = \Phi_{k+1}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_{k+1}); \mathbf{R})$$

$u_j \in (0, 1), j = 1, \ldots, k+1$

Joint multivariate distribution function:

$F_Y(y_1, \ldots, y_{k+1}; \mathbf{R}) =$

$$= [C[F_1(y_1), \ldots, F_{k+1}(y_{k+1}); \mathbf{R}] = \Phi_{(k+1)}[\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F_{k+1}(y_{k+1}))]$$

GAUSSIAN COPULA APPROACH (2)

Conditional probability density function (see Käärik and Käärik (2009))

$$f_{Z_{k+1}|Z_1,\ldots,Z_k}(z_{k+1}|z_1,\ldots,z_k;\mathbf{R}) = \frac{\exp\{-\frac{(z_{k+1}-\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{z}_k)^2}{2(1-\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{r})}\}}{\sqrt{2\pi(1-\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{r})}} \qquad (1)$$

$Z_j = \Phi^{-1}[F_j(Y_j)], \quad j = 1,\ldots,k+1$ – standard normal r.v.-s from Y_j

$\mathbf{z}_k = (z_1,\ldots,z_k)^T$

GAUSSIAN COPULA APPROACH (2)

Conditional probability density function (see Käärik and Käärik (2009))

$$f_{Z_{k+1}|Z_1,\ldots,Z_k}(z_{k+1}|z_1,\ldots,z_k;\mathbf{R}) = \frac{\exp\{-\frac{(z_{k+1}-\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{z}_k)^2}{2(1-\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{r})}\}}{\sqrt{2\pi(1-\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{r})}} \quad (1)$$

$Z_j = \Phi^{-1}[F_j(Y_j)], \quad j = 1,\ldots,k+1$ – standard normal r.v.-s from Y_j

$\mathbf{z}_k = (z_1,\ldots,z_k)^T$

As a result we have the (conditional) probability density function of a normal random variable with expectation $\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{z}_k$ and variance $1-\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{r}$:

$$E(Z_{k+1}|Z_1 = z_1,\ldots,Z_k = z_k) = \mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{z}_k, \quad (2)$$

$$Var(Z_{k+1}|Z_1 = z_1,\ldots,Z_k = z_k) = 1-\mathbf{r}^T\mathbf{R}_k^{-1}\mathbf{r}. \quad (3)$$

IMPUTATION FORMULA

The formula (2) leads us to the general formula of replacing the missing value z_{k+1} by the estimate $\hat{z}_{k+1}$ using the conditional mean imputation

$$\boxed{\hat{z}_{k+1} = \mathbf{r}^T \mathbf{R}_k^{-1} \mathbf{z}_k} \tag{4}$$

$\mathbf{r}$ – the vector of correlations between $(Z_1, \ldots, Z_k)$ and Z_{k+1}
$\mathbf{R}_k^{-1}$ – the inverse of the correlation matrix of $(Z_1, \ldots, Z_k)$
$\mathbf{z}_k = (z_1, \ldots, z_k)^T$ – the vector of complete observations for the subject which has missing value z_{k+1}.

IMPUTATION FORMULA

The formula (2) leads us to the general formula of replacing the missing value z_{k+1} by the estimate $\hat{z}_{k+1}$ using the conditional mean imputation

$$\boxed{\hat{z}_{k+1} = \mathbf{r}^T \mathbf{R}_k^{-1} \mathbf{z}_k} \tag{4}$$

$\mathbf{r}$ – the vector of correlations between $(Z_1, \ldots, Z_k)$ and Z_{k+1}
$\mathbf{R}_k^{-1}$ – the inverse of the correlation matrix of $(Z_1, \ldots, Z_k)$
$\mathbf{z}_k = (z_1, \ldots, z_k)^T$ – the vector of complete observations for the subject which has missing value z_{k+1}.

From expression (3) we obtain the (conditional) variance of imputed value as follows

$$\boxed{(\hat{\sigma}_{k+1})^2 = 1 - \mathbf{r}^T \mathbf{R}_k^{-1} \mathbf{r}} \tag{5}$$

These results for dropouts are proved by Käärik and Käärik (2009)

DEPENDENCE STRUCTURES

Start from a simple correlation structure, depending on one parameter only.

(1) The *compound symmetry (CS)* or the constant correlation structure, when the correlations between all measurements are equal, $r_{ij} = \rho,\ i, j = 1, \ldots, m, i \neq j$.

(2) The *first order autoregressive* correlation structure (*AR*), when the observations on the same subject that are closer are more highly correlated than measurements that are further apart, $r_{ij} = \rho^{|j-i|},\quad i, j = 1, \ldots, m, i \neq j$.

DEPENDENCE STRUCTURES

Start from a simple correlation structure, depending on one parameter only.

(1) The *compound symmetry (CS)* or the constant correlation structure, when the correlations between all measurements are equal, $r_{ij} = \rho,\ i, j = 1, \ldots, m, i \neq j$.

(2) The *first order autoregressive* correlation structure (*AR*), when the observations on the same subject that are closer are more highly correlated than measurements that are further apart, $r_{ij} = \rho^{|j-i|},\quad i, j = 1, \ldots, m, i \neq j$.

Imputation strategy in the case of an existing *CS* correlation structure is studied in detail in Käärik and Käärik (2009).

For the ordered missing data with CS correlation structure, we had the following imputation formula

$$\hat{z}_{k+1}^{CS} = \frac{\rho}{1 + (k-1)\rho} \sum_{j=1}^{k} z_j, \tag{6}$$

$z_1, \ldots, z_k$ – the observed values for the subject with missing value z_{k+1}.

AR STRUCTURE

Lemma 1. Let $\mathbf{Z} = (Z_1, \ldots, Z_{k+1})$ be a random vector with standard normal components and let the corresponding correlation matrix have *AR* correlation structure with correlation coefficient ρ. Then the following assertions hold:

$$E(Z_{k+1}|Z_1 = z_1, \ldots, Z_k = z_k) = E(Z_{k+1}|Z_k = z_k) = \rho z_k, \tag{7}$$

$$Var(Z_{k+1}|Z_1 = z_1, \ldots, Z_k = z_k) = 1 - \rho^2. \tag{8}$$

By Lemma 1, the conditional mean imputation formula for standardized measurements with an *AR* structure has the simple form

$$\hat{z}_{k+1}^{AR} = \rho z_k, \tag{9}$$

z_k – the last observed value for the subject

The corresponding variance is

$$(\hat{\sigma}_{k+1}^{AR})^2 = 1 - \rho^2 \tag{10}$$

IMPUTATION ALGORITHM

IMPUTATION ALGORITHM

Step 1. Sort the columns of the data matrix to get ordered missing data, fix Y_{k+1} (column with the least number of missing values) as the starting point for imputation.

IMPUTATION ALGORITHM

Step 1. Sort the columns of the data matrix to get ordered missing data, fix Y_{k+1} (column with the least number of missing values) as the starting point for imputation.

Step 2. Estimate the marginal distribution functions of $Y_1, \ldots, Y_k, Y_{k+1}$.

IMPUTATION ALGORITHM

Step 1. Sort the columns of the data matrix to get ordered missing data, fix Y_{k+1} (column with the least number of missing values) as the starting point for imputation.

Step 2. Estimate the marginal distribution functions of $Y_1, \ldots, Y_k, Y_{k+1}$.

Step 3. Estimate the correlation structure between variables $Y_1, \ldots, Y_k, Y_{k+1}$. If we can accept the hypothesis of compound symmetry or autoregressive structure, estimate the Spearman's correlation coefficient ρ. If there is no simple correlation structure, estimate $\mathbf{R}$ by an empirical correlation matrix.

IMPUTATION ALGORITHM

Step 1. Sort the columns of the data matrix to get ordered missing data, fix Y_{k+1} (column with the least number of missing values) as the starting point for imputation.

Step 2. Estimate the marginal distribution functions of $Y_1, \ldots, Y_k, Y_{k+1}$.

Step 3. Estimate the correlation structure between variables $Y_1, \ldots, Y_k, Y_{k+1}$. If we can accept the hypothesis of compound symmetry or autoregressive structure, estimate the Spearman's correlation coefficient ρ. If there is no simple correlation structure, estimate $\mathbf{R}$ by an empirical correlation matrix.

Step 4. In the case of *CS* correlation structure, use imputation formula (6). In the case of *AR* correlation structure, use imputation formula (9) and estimate the variance of the imputed value using formula (10). If there is no simple correlation structure, then use general formulas (4) and (5).

IMPUTATION ALGORITHM

Step 1. Sort the columns of the data matrix to get ordered missing data, fix Y_{k+1} (column with the least number of missing values) as the starting point for imputation.

Step 2. Estimate the marginal distribution functions of $Y_1, \ldots, Y_k, Y_{k+1}$.

Step 3. Estimate the correlation structure between variables $Y_1, \ldots, Y_k, Y_{k+1}$. If we can accept the hypothesis of compound symmetry or autoregressive structure, estimate the Spearman's correlation coefficient ρ. If there is no simple correlation structure, estimate $\mathbf{R}$ by an empirical correlation matrix.

Step 4. In the case of *CS* correlation structure, use imputation formula (6). In the case of *AR* correlation structure, use imputation formula (9) and estimate the variance of the imputed value using formula (10). If there is no simple correlation structure, then use general formulas (4) and (5).

Step 5. Repeat step 4 until all missing values in column Y_{k+1} are imputed. If $k < m-1$, then take $k = k+1$, take a new Y_{k+1}, estimate the marginal distribution of Y_{k+1} and go to step 3. In the following steps the imputed values are treated as if they were observed.

CASE STUDY: INCOMPLETE CS DATA

Questionnaire where the respondents (customers) are requested to give scores (in our example on a scale from 0 to 10, from least to most satisfied)

We are focusing on a group of five questions (from 20 customers) directly related to customer satisfaction.

We have complete data and we will delete the values from one variable step by step and analyze the reliability of the proposed method.

IMPUTATION (1)

The imputation study has the following general steps:

1. *Estimation of marginal distributions.*

Kolmogorov-Smirnov and Anderson-Darling tests for normality did not reject the normality assumption.

IMPUTATION (1)

The imputation study has the following general steps:

1. *Estimation of marginal distributions.*

Kolmogorov-Smirnov and Anderson-Darling tests for normality did not reject the normality assumption.

2. *Estimation of the correlation structure.*

Calculation of the 'working' correlation matrix gave us Spearman's $\hat{\rho} = 0.784$ as an estimate of the parameter of the *AR*-structure.

IMPUTATION (2)

3. *Estimation of the missing values.*

To validate the imputation algorithm we repeat the imputation procedure for every value in the data column Y_5.

Modified formulas (for nonstandard normal variables instead of (9) and (10)):

$$\hat{z}_{k+1}^{AR} = \rho \frac{s_{k+1}}{s_k}(z_k - \bar{Z}_k) + \bar{Z}_{k+1}, \tag{11}$$

$\bar{Z}_k$, $\bar{Z}_{k+1}$ – the mean values of data columns Z_{k+1} and Z_k respectively
s_{k+1} and s_k – the corresponding standard deviations

$$(\hat{\sigma}_{k+1}^{AR})^2 = s_{k+1}^2(1-\rho^2). \tag{12}$$

QUALITY OF IMPUTATION

L_1 error (absolute distance between the observed and imputed value)

$$e_1 = 0.641$$

L_2 error (root mean square distance)

$$e_2 = 0.744$$

RESULTS

4. *Estimation of the variance of imputed values.*

No	y_5	$\hat{z}_5^{AR}$	0.95 CI	No	y_5	$\hat{z}_5^{AR}$	0.95 CI
1	6	6.77	(5.12; 8.41)	11	8	7.57	(5.89; 9.24)
2	8	8.52	(6.84; 10.19)	12	4	5.43	(3.89; 6.97)
3	9	8.46	(6.79; 10.13)	13	7	6.68	(5.01; 8.35)
4	6	5.85	(4.21; 7.50)	14	5	6.89	(5.28; 8.49)
5	9	8.46	(6.79; 10.13)	15	10	9.30	(7.66; 10.95)
6	10	9.30	(7.66; 10.95)	16	8	8.52	(6.84; 10.19)
7	10	9.30	(7.66; 10.95)	17	7	6.68	(5.01; 8.35)
8	10	9.30	(7.66; 10.95)	18	8	8.52	(6.84; 10.19)
9	9	8.46	(6.79; 10.13)	19	7	7.62	(5.95; 9.29)
10	9	8.46	(6.79; 10.13)	20	9	9.40	(7.73; 11.07)

y_5 – the observed value, $\hat{z}_5^{AR}$ – the corresponding imputed value
0.95 CI – 0.95-level confidence interval based on the normal approximation

SUMMARY

It is important to remember that the imputation methodology does not give us qualitatively new information but enables us to use all available information about the data with maximal efficiency.

In general, most of the missing data handling methods deal with incomplete data primarily from the perspective of estimation of parameters and computation of test statistics rather than predicting the values for specific cases. We, on the other hand, are interested in *small sample sizes* where every value is essential and *imputation results are of scientific interest itself*.

The results of this study indicate that in the empirical context of the current study the algorithm performs well for modeling missing values in correlated data.

As importantly, the following advantages can be pointed out.

(1) The marginals of variables do not have to be normal, they can even be different.

(2) The simplicity of formulas (9)–(12).

REMARKS

The class of copulas is wide and growing, the copula approach used here can be extended to the case of other copulas.

Choosing a copula to fit the given data is an important but difficult problem.

In some cases analytical solutions are not available (copula density might not exist).

These relevant problems obviously merit further research.

Acknowledgements

This work is supported by Estonian Science Foundation
grants No 7313 and No 8294.

References

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

2. Käärik, E. (2007). Handling dropouts in repeated measurements using copulas. *Diss. Math. Universitas Tartuensis*, 51, Tartu, UT Press.

3. Käärik, E., Käärik, M. (2009). Modelling Dropouts by Conditional Distribution, a Copula-Based Approach. *Journal of Statistical Planning and Inference*, 139(11), 3830 - 3835.

4. Little, J. A., Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley.

5. Nelsen R.B. (2006). An Introduction to Copulas. 2nd edition. Springer Verlag, New York.

6. Song, P.X.K. (2007): Correlated data analysis. Modeling, analytics, and applications. Springer, New York.

7. Song, P.X-K., Li, M., Yuan, Y. (2009). Joint Regression Analysis of Correlated Data Using Gaussian Copulas. *Biometrics,* **64** (2), 60–68.

THANK YOU!