The 8th Tartu Conference on MULTIVARIATE STATISTICS, The 6th Conference on MULTIVARIATE DISTRIBUTIONS with Fixed Marginals

Modelling Dropouts by Conditional Distribution, a Copula-Based Approach

Ene Käärik

Institute of Mathematical Statistics, University of Tartu

Tartu, 27th June 2007

Outline

1. Introduction: Missing data. Handling missing data

2. Copula. Gaussian copula

3. Imputation. Gaussian copula approach

4. Different structures of correlation matrix. Formulas for imputation

5. Simulation study. Results

6. Advantages of Gaussian copula approach

7. Remarks

8. References

Missing data. Dropouts

Let $X = (X_1, \ldots, X_m)$ be an outcome variable with repeated measurements at the time points $1, \ldots, m$.

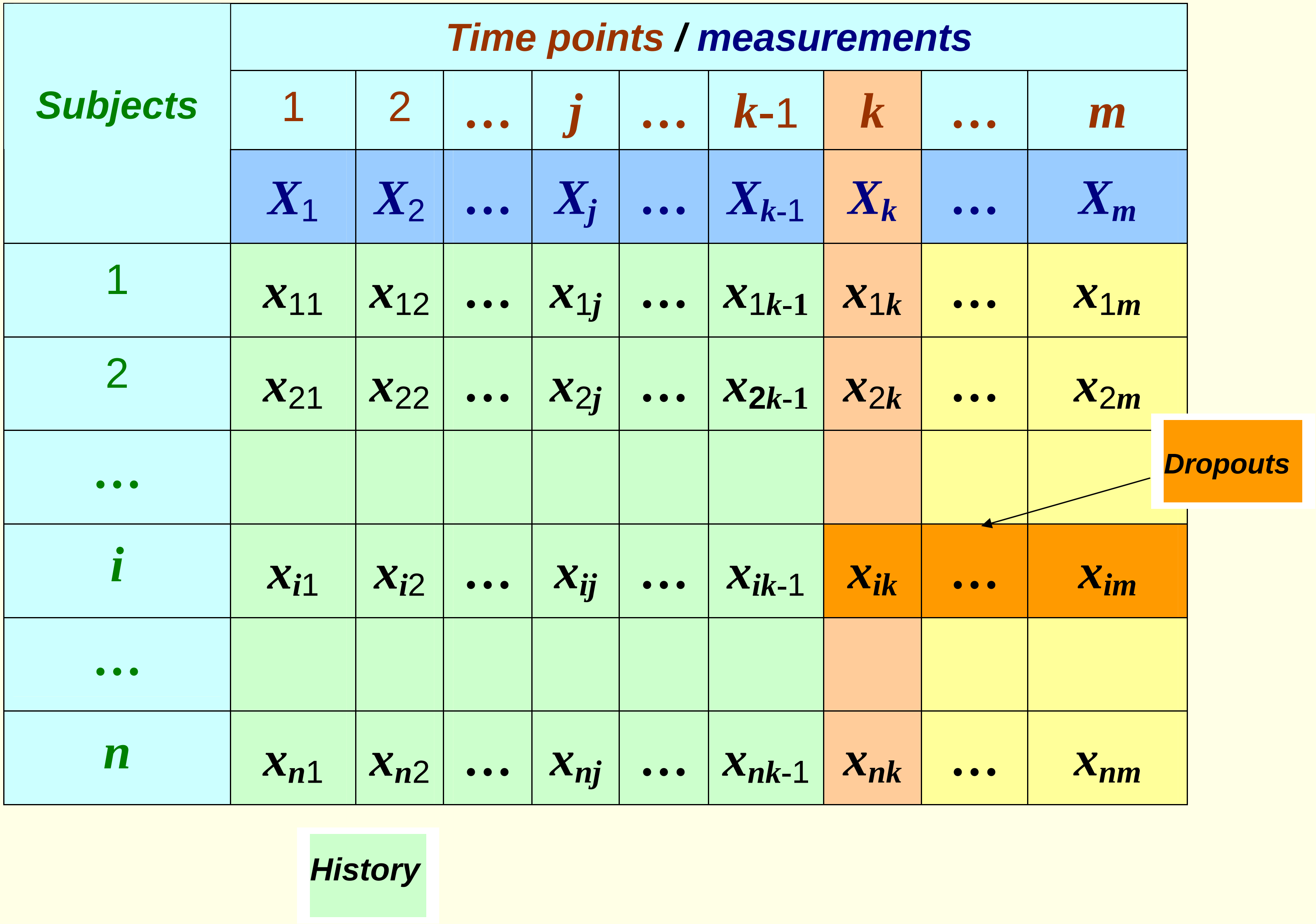
Suppose that $X_j \sim F_j$ $(j = 1, \ldots, m)$, $F_j \in \mathcal{P}$.

For n subjects the data are $\mathsf{X} = (x_{ij})$, $i = 1, \ldots, n$; $j = 1, \ldots, m$

Definition 1. *Dropout* or *attrition* is missingness in data which occurs when subject leaves the study prematurely and does not return.

Definition 2. Let k be the time point at which the dropout process starts. The vector $H = (X_1, X_2, \ldots, X_{k-1})$ is called *history* of measurements.

Data matrix



Subjects	Time points / measurements								
	1	2	…	j	…	k-1	k	…	m
	X_1	X_2	…	X_j	…	$X_{k\text{-}1}$	X_k	…	X_m
1	x_{11}	x_{12}	…	x_{1j}	…	$x_{1k\text{-}1}$	x_{1k}	…	x_{1m}
2	x_{21}	x_{22}	…	x_{2j}	…	$x_{2k\text{-}1}$	x_{2k}	…	x_{2m}
…									
i	x_{i1}	x_{i2}	…	x_{ij}	…	$x_{ik\text{-}1}$	x_{ik}	…	x_{im}
…									
n	x_{n1}	x_{n2}	…	x_{nj}	…	$x_{nk\text{-}1}$	x_{nk}	…	x_{nm}

Dropouts

History

Handling missing data. Classification

The classification of dropout processes

- *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 additionally depends on unobserved measurements, ie those that would have been observed if the subject had not dropped out.

Rubin (1976), Little & Rubin (1987)

We work with the data we don't have

Imputation

Definition 3. *Imputation (filling in, substitution)* is a strategy for completing missing value in the data with plausible value which is an estimate of the true value of the unobserved observation.

Methods for handling missing data

- *Single imputation methods.* Missing value is replaced with a single value.
- *Multiple imputation methods* (MI).
- *Model based analysis* (selection model, pattern-mixture model).

Imputation – drawing the value from a predictive (conditional) distribution of the missing values
⇒ requires the method of creating a predictive (conditional) distribution for the imputed value based on the observed values.

Little & Rubin (1987):

Imputation is especially important in case of small sample sizes.
It makes sense to consider imputation of dropouts separately from modelling data.

Imputation by conditional distribution

We use the idea of imputing a missing value *based on conditional distribution* of dropout at time point k conditioned to the history up to time point $k-1$.

Steps:

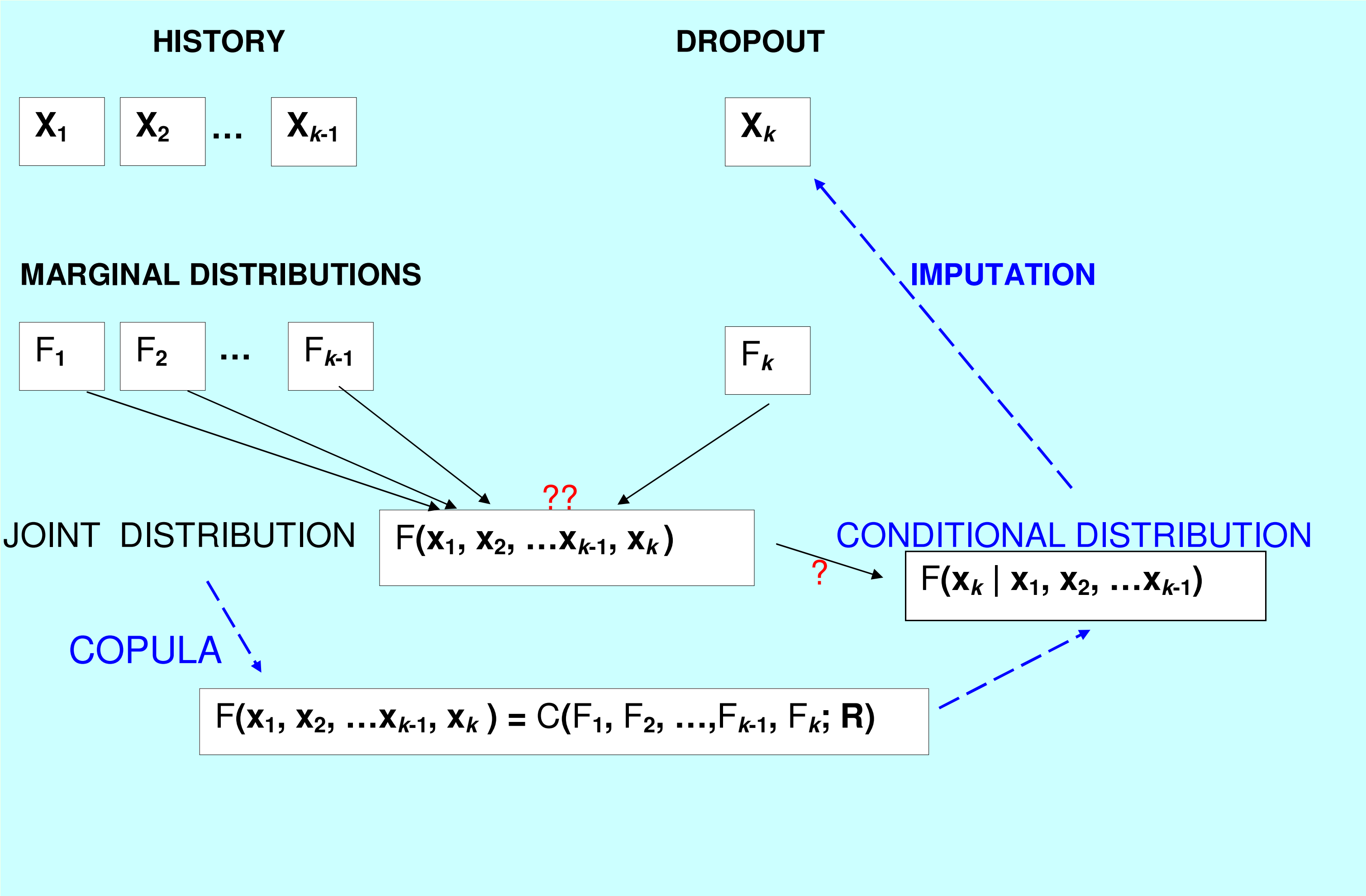
1. Estimate univariate marginal distributions $F_1, \ldots, F_k$
2. Use marginal distributions to construct the multivariate joint distribution $F_1, \ldots, F_k \quad \Rightarrow \quad F$
3. Find the conditional distribution $F(x_k|H)$ for missing data conditioned to history as a predictive distribution
4. Predict the missing value from the conditional distribution

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

Sklar (1959): copula links joint distribution function to their one-dimensional marginals

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*"

Handling missing data. Situation



HISTORY

DROPOUT

$\mathbf{X}_1$ $\mathbf{X}_2$ … $\mathbf{X}_{k\text{-}1}$

$\mathbf{X}_k$

MARGINAL DISTRIBUTIONS

IMPUTATION

F_1 F_2 … $F_{k\text{-}1}$

F_k

??

JOINT DISTRIBUTION

$F(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_{k\text{-}1}, \mathbf{x}_k)$

CONDITIONAL DISTRIBUTION

?

$F(\mathbf{x}_k \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_{k\text{-}1})$

COPULA

$F(\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_{k\text{-}1}, \mathbf{x}_k) = C(F_1, F_2, \ldots, F_{k\text{-}1}, F_k; \mathbf{R})$

Copula. Repeated measurements

Repeated measurements $X_1, \dots, X_k$, $\quad X_j \sim F_j$

Joint distribution

Using marginal distributions $F_1(x_1), \dots, F_k(x_k)$ and a **copula** C, the function $C(F_1(x_1), \dots, F_k(x_k))$ defines *a joint distribution function*

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

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

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

$F_1^{-1}, \dots, F_k^{-1}$ – the quantile functions of given marginals
$u_1, \dots, u_k$ – uniform $[0, 1]$ variables

Sklar (1959), Nelsen (1998)

Copula. Joint and conditional densities of repeated measurements

Joint density

If C and $F_1, \dots, F_k$ are differentiable, then joint density $f(x_1, \dots, x_k)$ corresponding to the joint distribution $F(x_1, \dots, x_k)$ can be written by canonical representation as a product of the marginal densities and the copula density

$$f(x_1, \dots, x_k) = f_1(x_1) \cdot \dots \cdot f_k(x_k) \cdot c(F_1, \dots, F_k),$$

where $f_i(x_i)$ is the density corresponding to F_i and the copula density c is defined as derivative of the copula (*dependence function*)

$$c = \frac{\partial^k C}{\partial F_1 \cdots \partial F_k}.$$

Conditional density

$$f(x_k | x_1, \dots, x_{k-1}) = f_k(x_k) \frac{c(F_1, \dots, F_k)}{c(F_1, \dots, F_{k-1})},$$

where $c(F_1, \dots, F_k)$, $c(F_1, \dots, F_{k-1})$ – corresponding copula densities

Gaussian copula. Basic definitions

Definition 4. Let R be a symmetric, positive definite matrix with $diag(R) = (1, 1, \dots, 1)^T$ and Φ_k the standardized k-variate normal distribution function with correlation matrix R. Then the *multivariate Gaussian copula* is following:

$$C_k(u_1, \dots, u_k; R) = \Phi_k(\Phi_1^{-1}(u_1), \dots, \Phi_1^{-1}(u_k)).$$

Joint distribution

$$F(x_1, \dots, x_k; R) = C_k(u_1, \dots, u_k; R) = \Phi_k[\Phi_1^{-1}(u_1), \dots, \Phi_1^{-1}(u_k); R],$$

$u_i \in (0, 1), i = 1, \dots, k$, Φ_k – the standard multivariate normal distribution function with correlation matrix R
Φ_1^{-1} – the inverse of the standard univariate normal distribution function

Joint density

$$f_k(x_1, \dots, x_k|R) = \phi_1(x_1) \cdot \dots \cdot \phi_1(x_k) \cdot c_k[\Phi_1(x_1), \dots, \Phi_1(x_k); R^*],$$

Φ_1 and ϕ_1 – the univariate standard normal distribution function and density, respectively,

c_k – dependence function (copula density), R^* – matrix of dependence measures.

Gaussian copula. Copula density

$Y_i = \Phi_1^{-1}[F_i(X_i)]$, $i = 1, \ldots, k$ Clemen and Reilly (1999); Song (2000)

Density of normal copula

$$c_k[\Phi_1(y_1), \ldots, \Phi_1(y_k); R^*] = \frac{\exp[(-1/2)Y^T R^{-1} Y + (1/2)Y^T Y]}{|R|^{1/2}}$$

$$= \frac{\exp\{-Y^T(R^{-1} - I)Y/2\}}{|R|^{1/2}}$$

$Y = (Y_1, \ldots, Y_k)$ and I is the $k \times k$ identity matrix.

Conditional density

$$f(x_k|x_1, \ldots, x_{k-1}) = f_k(x_k)\frac{c(F_1, \ldots, F_k)}{c(F_1, \ldots, F_{k-1})}$$

Partition of correlation matrix

$R = (r_{ij}), \quad r_{ij} = corr(X_i, X_j), \quad i, j = 1, \ldots, k$

Partition:

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

R_{k-1} is the correlation matrix of the history $H = (X_1, \ldots, X_{k-1})$

$r = (r_{1k}, \ldots, r_{(k-1)k})^T$ is the vector of correlations between the history and the time point k

Derivation of general formula for imputation

$Y_i = \Phi^{-1}[F_i(X_i)]$

Conditional density

$$f(y_k|H;R^*) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}[\frac{(y_k - r^T R_{k-1}^{-1}(y_1,\ldots,y_{k-1})^T)^2}{(1-r^T R_{k-1}^{-1} r)}]\} \cdot (1 - r^T R_{k-1}^{-1} r)^{-1/2}$$

$\arg\max_{y_k}[\ln(f(y_k|H;R^*)]]$

$$\frac{\partial \ln f(y_k|H;R^*)}{\partial y_k} = \frac{-y_k + r^T R_{k-1}^{-1}(y_1,\ldots,y_{k-1})^T}{(1-r^T R_{k-1}^{-1} r)} = 0$$

$\Rightarrow$

$$\boxed{\widehat{y}_k = r^T \cdot R_{k-1}^{-1} \cdot Y_{k-1}^*} \qquad (2)$$

r – the vector of correlations between the history and the time point k
R_{k-1}^{-1} – the inverse of correlation matrix of the history
$Y_{k-1}^* = (y_1,\ldots,y_{k-1})^T$ – the vector of observations for the subject who drops out at the time point k.

Results

Proposition 1

Let $Y_1, \ldots, Y_k,\ \ Y_j = (y_{1j}, \ldots, y_{nj})^T,\ \ j = 1, \ldots, k,$ be repeated measurements with standard normal marginals. Let the correlation matrix of the data be partitioned as in (1), and let the dropout process start at the time point k, so that the history has complete observations. Then the imputation of dropout at the time point k is given by formula:

$$\boxed{\widehat{y}_k = r^T \cdot R_{k-1}^{-1} \cdot Y_{k-1}^*} \qquad (2)$$

.

Corollary

Let $X_1, \ldots, X_k,\ X_j = (x_{1j}, \ldots, x_{nj})^T,\ j = 1, \ldots, k,$ be repeated measurements with arbitrary marginals $F_1, \ldots, F_k$. Then the imputation of dropout at the time point k is given by formula (2).

We apply the following three-steps procedure to Proposition 1.

1. Use the normalizing transformation $Y_j = \Phi_1^{-1}(F_j(X_j))$, $j = 1, \ldots, k$.

2. Impute the dropout using formula (2).

3. Use the inverse transformation $X_k = F_k^{-1}[\Phi_1(Y_k)]$ for imputing the dropout in initial measurements.

Structures of correlation matrix

- *Compound symmetry (CS)*, where the correlations between all time points are equal, $r_{ij} = \rho, \quad i, j = 1, \ldots, k, i \neq j$

- *First order autoregressive (AR)*, where the dependence between observations decreases as the measurements get further in time
$r_{ij} = \rho^{|j-i|}, \quad i, j = 1, \ldots, k, \; i \neq j.$

- 1-*banded Toeplitz (BT)*, where only two sequential measurements are dependent, $r_{ij} = \rho, \quad j = i + 1, \quad i = 1, \ldots, k - 2$

Verbeke & Molenberghs (2001)

Examples

Let $k = 4$

1. Compound symmetry structure:

$$\begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}.$$

2. First order autoregressive structure:

$$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

3. 1-banded Toeplitz structure:

$$\begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}.$$

Special cases: 1. Compound symmetry

$r = (\rho, \ldots, \rho)^T$

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

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

(2) ⇒

$$\boxed{\widehat{y}_k^{CS} = \frac{\rho}{1+(k-2)\rho}\sum_{i=1}^{k-1} y_i} \qquad (3)$$

$y_1, \ldots, y_{k-1}$ – the observed values for the subject.

Special cases: 2. Autoregressive dependencies

$r = (\rho^{k-1}, \rho^{k-2} \ldots, \rho)^T$

$$R_{k-1} = \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{k-2} \\ \rho & 1 & \rho & \ldots & \rho^{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{k-2} & \rho^{k-3} & \rho^{k-4} & \ldots & 1 \end{pmatrix}$$

$$R_{k-1}^{-1} = \frac{1}{\rho^2 - 1} \begin{pmatrix} -1 & \rho & 0 & \ldots & 0 & 0 \\ \rho & -(1+\rho^2) & \rho & \ldots & 0 & 0 \\ 0 & \rho & -(1+\rho^2) & \ldots & 0 & 0 \\ 0 & 0 & \rho & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -(1+\rho^2) & \rho \\ 0 & 0 & 0 & \ldots & \rho & -1 \end{pmatrix}.$$

(2) $\Rightarrow$

$$\boxed{\widehat{y}_k^{AR} = \rho \frac{S_k}{S_{k-1}} (y_{k-1} - \bar{Y}_{k-1}) + \bar{Y}_k} \qquad (4)$$

y_{k-1} – the last observed value for the subject
$\bar{Y}_{k-1}$, $\bar{Y}_k$ – the mean values of kth and $(k-1)$th time points
S_k, S_{k-1} – the corresponding standard deviations

Special cases: 3. 1-banded Toeplitz structure

$r = (0, \ldots, 0, \rho)$

$$R_{k-1} = \begin{pmatrix} 1 & \rho & 0 & \ldots & 0 & 0 \\ \rho & 1 & \rho & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & \rho \\ 0 & 0 & 0 & \ldots & \rho & 1 \end{pmatrix}.$$

$\Rightarrow$ considering $\widehat{y}_k = r^T R_{k-1}^{-1} Y_{k-1}^*$ we are interested only in *last row of inverse matrix*.

$\Rightarrow$

$$\boxed{\widehat{y}_k^{1BT} = \frac{1}{|R_{k-1}|} \sum_{j=1}^{k-1} (-1)^{k-j+1} |R_{j-1}| \rho^{k-j} y_j} \qquad (5)$$

$y_1, \ldots, y_{k-1}$ – the observed values for the subject
$|R_j|, j = 1, \ldots, k-1,$ – the determinant of correlation matrix of history
$|R_0| = 1$, $|R_1| = 1$.

Example: dropping out starts at $k = 3$ and $k = 4$

$k = 3$: $\widehat{y}_3 = \frac{1}{1-\rho^2}(-\rho^2 y_1 + \rho y_2) = \frac{1}{|R_2|}(-\rho^2 y_1 + \rho y_2)$;

$k = 4$: $\widehat{y}_4 = \frac{1}{1-2\rho^2}(\rho^3 y_1 - \rho^2 y_2 + \rho(1-\rho^2) y_3 = \frac{1}{|R_3|}(\rho^3 y_1 - \rho^2 y_2 + \rho |R_2| y_3)$;

Simulation study (1)

The goal of the simulation study is

to test the *effectiveness of the new imputation formulas*

by comparison with some well-known imputation methods with different missing data mechanisms and sample sizes.

As quality measures we use the *standardized difference between observed value and imputed value*.

Experimental design:

I − normal distribution, II− skewed distribution

$3 \times 2 \times 2 \times 3 = 36$ different data sets (*CS*, *AR*)
$3 \times 2 \times 3 = 18$ different data sets (1-BT)

$k = 3, 6, 12$ (data from 3-, 6-, 12-dimensional normal distribution)

$n = 10, 20$ (small sample sizes)

$\rho = 0.5, \rho = 0.7$

3 missingness mechanisms (CRD, RD and ID).

For each combination 1000 runs were performed.

Simulation study (2)

Algorithms for imputation were compared:

- in the case of *CS*: imputation by formula (3) vs imputation by linear prediction $Y_k = \beta_0 Y_1 + \ldots + \beta_{k-1} Y_{k-1}$;
- in the case of *AR* and *BT*: imputation by formulas (4), (5) vs imputation using *LOCF* (*Last Observation Carried Forward*).

Results

- Bias is smaller in the case of CRD and RD.
- Standard deviations are more stable.
- The formula (3) could be used for small data sets with several repeated measurements $(k > n)$, when linear prediction does not work.
- The formulas (4) and (5) contain more information about data than the *LOCF*-method.

⇒ In all simulation studies the results showed that *the imputation algorithms based on the copula approach are quite appropriate for modelling dropouts.*

Advantages of Gaussian copula approach

The *Gaussian copula* is useful for its easy simulation method and is perhaps easiest to employ in practice.

1. Normality of marginals is not necessary. Furthermore, the marginals may have different distributions. The normalizing transformation will be used.

2. For a simple dependence structures simple formulas can be found for calculating conditional mean as imputed value.

3. Effectiveness, especially in the case of small sample size n relative to the number of measurements (time points) k.

$\Rightarrow$ *Gaussian copula represents a good approach for modelling dropouts in repeated measurements study.*

Remarks

The Gaussian copula is not the only possibility to use in this approach.

- Lindsey and Lindsey (2002) suggested Student's t-distribution, power-exponential or skew Laplace distribution for modelling repeated responses instead of normal distribution.

- Vandenhende and Lambert (2002) 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* (Genest, Rivest, 1993)
  – Vandenhende and Lambert (2002) used *Frank's copula* to model the dependence between dropout and responses.

The family of copulas is sufficiently large and allows a wide range of multivariate distributions as models.

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

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

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

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

6. Song, P.X.K. (2000). Multivariate dispersion models generated from Gaussian Copula. *Scandinavian Journal of Statistics*, **27**, 305–320.

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

8. Verbeke, G., Molenberghs, G. (2001). Linear mixed models for longitudinal data. *Springer Series in Statistics*, Springer-Verlag, NY Inc.

Thank you for your attention !