DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
51

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
51

# HANDLING DROPOUTS IN REPEATED MEASUREMENTS USING COPULAS

**ENE KÄÄRIK**

TARTU UNIVERSITY
PRESS

Faculty of Mathematics and Computer Science, University of Tartu, Tartu

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (Ph. D.) in mathematical statistics on January 26, 2007, by the Council of the Faculty of Mathematics and Computer Science, University of Tartu

Supervisor:
Professor Emeritus, Cand. Sc. Ene-Margit Tiit
University of Tartu, Tartu, Estonia

Opponents: Professor, Ph. D. Esa Läärä
University of Oulu, Oulu, Finland

Associate Professor, Cand. Sc. Ebu Tamm
Tallinn University of Technology Tallinn, Estonia

The public defence will take place on March 13, 2007.

ISSN 1024–4212
ISBN 978–9949–11–539–6 (trükis)
ISBN 978–9949–11–540–2 (PDF)

Autoriõigus Ene Käärik, 2007

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr. 40

# Contents

**Acknowledgements** **7**

**List of original publications** **8**

**Introduction** **9**

**1 Missing data** **14**
1.1 Basic assumptions and concepts . . . . . 14
1.2 Dropout . . . . . 17
1.3 Types of dropout . . . . . 20
1.3.1 The risk of dropout . . . . . 21
1.4 Objectives relating to dropouts . . . . . 22
1.5 Handling missing data . . . . . 23
1.5.1 Traditional approaches . . . . . 23
1.5.2 Single imputation methods . . . . . 25
1.5.3 Multiple imputation methods . . . . . 28
1.5.4 Model based analysis . . . . . 29
1.5.5 Imputation by conditional distribution . . . . . 31
1.5.6 Cautions to imputation . . . . . 33

**2 Copula** **35**
2.1 Basic definitions and theorems . . . . . 36
2.2 Joint and conditional density functions . . . . . 38
2.3 Gaussian copula . . . . . 40
2.4 Modelling joint density of repeated measurements by Gaussian copula . . . . . 41

2.5 Other copulas . . . . . 43
2.5.1 Elliptical copula . . . . . 43
2.5.2 Archimedean copula . . . . . 45
2.6 Modelling using copulas . . . . . 48

**3 Imputation using copula** **51**
3.1 Dependence between measurements . . . . . 51
3.2 Correlation structures . . . . . 53
3.3 Derivation of general formula for imputation . . . . . 56
3.4 Special cases . . . . . 59
3.4.1 Compound symmetry structure . . . . . 59
3.4.2 Autoregressive dependence structure . . . . . 62
3.4.3 Banded Toeplitz structure . . . . . 65
3.5 Illustration . . . . . 67
3.6 Implementation of copula approach . . . . . 68

**4 Simulation study** **71**
4.1 Generation of the complete data . . . . . 71
4.2 Generation and imputation of the dropouts . . . . . 72
4.3 Experimental design and calculations . . . . . 73
4.4 Results . . . . . 74
4.5 Analysis of dependence of experimental design . . . . . 76
4.6 Conclusions . . . . . 77

**Appendix** **79**

**Bibliography** **83**

**Summary in Estonian** **91**

**Curriculum Vitae** **95**

# Acknowledgements

I would like to express my sincere gratitude to my supervisor Professor Emeritus *Ene-Margit Tiit* for her guidance and encouragement during my studies. Her support and great enthusiasm were of the utmost importance for me.

My special thanks to my son Meelis for his helpful discussions, theoretical suggestions and practical hints.

I am very obliged to Tatjana Nahtman for her all-round help and professional advice.

I am grateful to Professor Tõnu Kollo for thorough reading of this thesis and for his valuable comments.

I am thankful to all my colleagues and friends who believed in the completion of the present thesis.

I am also grateful to my family for their enormous support and patience.

The work is partially supported by Estonian Science Foundation grants 5521, 5203 and 6702.

# List of original publications

1. Tiit, E.-M., Käärik, E. (1996). Generation and investigation of multivariate distribution having fixed discrete marginals. In: *Proceedings in Computational Statistics. Compstat' 96*, Ed. A. Prat. Physica-Verlag, Springer, 471–476.

2. Käärik, E., Sell, A. (2004). Estimating $ED_{50}$ using the up-and-down method. In: *Proceedings in Computational Statistics. Compstat'04*, Ed. J. Antoch. Physica-Verlag, Springer, 1279–1286.

3. Käärik, E. (2005). Handling dropouts by copulas. In: *WSEAS Transactions on Biology and Biomedicine*, Ed. N. Mastorakis. Vol **1** (2), 93–97.

4. Käärik, E. (2006a). Imputation algorithm using copulas. *Advances in Methodology and Statistics*, Ed. A. Ferligoj. Vol **3** (1), 109–120.

5. Käärik, E. (2006b). Imputation by conditional distribution using Gaussian copula. In: *Proceedings in Computational Statistics. Compstat'06*, Ed. A. Rizzi and M. Vichi. Physica-Verlag, Springer , 1447–1454.

# Introduction

Longitudinal and/or repeated measures studies have extensive implementation in medicine, epidemiology, biology and social sciences. Repeated measures studies contain data representing multiple measurements from a single subject for a given variable.
Repeated measurements are often taken on the same experimental over time, but they could be taken over space and/or under different conditions as well. For example, in longitudinal analysis the measurements on subjects are recorded over a certain time period. Since measurements made on the same subject for a given variable are not independent, in repeated measurements analysis one should model the dependencies between observations in an appropriate way.

In practice, the sequence of measurements could often be terminated due to reasons that are outside the control of the investigator, which yield incomplete data. Missing values may cause complicated problems in many statistical analyses, especially in case of small sample sizes. Common approach for treating missing data in repeated measurements studies is to consider dropouts, where sequences of measurements on some units terminate prematurely. It might be necessary to accommodate dropout in the modelling process, which itself could be of scientific interest. The problem of dropouts is extremely important for small samples where every value is substantial.

Usually, dropouts should 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). As a matter of fact, the only difference is that in order to handle intermittent missingness

we could use information before and after the missing value, in case of dropouts we do not have any information after dropouts.
Thus, the methodology suitable for handling dropouts can be used for handling intermittent missingness as well, but it might not be the most effective method in this case.

Though missing data cause the statistical analysis of available data to be subject to bias, there are no universally applicable methods for handling incomplete data. Imputation of the missing values is a widely used strategy to deal with missing measurements. The basic idea of imputation is to fill in the missing data by using existing values following certain model with given assumptions. In general, imputation is a process used to determine and assign replacement values for missing, invalid or inconsistent data. Usually, the goal of imputation is not to predict missing values or describe the data, but to preserve important relationships in data using the observed values in order to do statistical inference with maximal effectiveness.

The first attempt to identify a missing data structure and impute the missing data was done by McKendrick in 1926 (see Meng, 2000). McKendrick analyzed the data from an epidemic study of cholera in an Indian village. The existence of unexposed households complicated the analysis, and to avoid this problem McKendrick derived a zero-truncated Poisson model. His algorithm is similar to EM-algorithm to obtain estimates from a sample with missing values.

Extensive development of the missing data theory began in 1970's with the case deletion and single imputation methods. In 1980's the likelihood based imputation procedures (EM-algorithm, etc.) and in 1990's the multiple imputation method and joint models have been developed. There are currently available many approaches to handle missing data. A comprehensive overview and guidelines for handling missing data can be found, for example, on the website developed by J. Carpentier and M. Kenward[1].

One possible approach for imputing dropouts is to use conditional distributions. Therefore, we need to know the joint distribution of repeated measurements which is a multivariate distribution with a special depen-

[1] www.lshtm.ac.uk/msu/missingdata

dence structure. The main problem is that though there is a vast selection of flexible parametric univariate distributions, only a few suitable multivariate distributions are available beside the multivariate normal distribution. Lindsay (Lindsay, 2000a) proposed a method for generating a useful family of multivariate distributions by substituting one distribution into the other (outer) distribution. Lindsey suggested the Pareto distribution for outer distribution. The parameters of the outer distribution could then be used to create the dependence structure between observations. The procedure suggested by Lindsay is similar to that used in copula theory, but the multivariate models obtained by Lindsay are not copulas.

*Copula* function creates the joint distribution with given marginals. The dependence between successive repeated measurements and between dropout and response can be modeled using copula function. Copula is one of the most useful tools for handling multivariate distributions with dependent components and it provides a convenient way to express joint distribution of two or more random variables.

In particular, copulas are joint distribution functions of random variables with standard uniform marginal distributions. There are two principal ways of using the copula's theory. We can extract copulas from well-known multivariate distribution functions, but we can also create new multivariate distribution function by joining arbitrary univariate distributions together with copulas. These ideas are used in this work.

Copulas form a flexible tool for multivariate model construction because no restrictions are placed on the marginal distributions.
Working with copulas has several advantages compared with working with the given (classical) multivariate distribution. Firstly, it is more flexible in applications. Secondly, in many cases it is complicated to specify a joint distribution directly when distribution of the data does not fit to any known family.

Copula theory is related to the study of multivariate distributions with given marginals. A copula $C$ is a function that links univariate marginal distributions to the multivariate distribution. It is defined as a multivariate distribution function on $[0,1]^k$, where $k$ is the dimension of the distribution.

The review of methods for constructing discrete and continuous joint distributions from the component marginal distributions is given in Miller and Liu (2002, p. 263–264), who pointed out the paper Tiit and Käärik (1996) as one of the origins of copula-based approach to data analysis.

In recent years, copula models became an increasingly popular tool for modelling dependencies between random variables, especially in biostatistics, actuarial and financial mathematics.
One advantage of copula models is their relative mathematical simplicity. Another advantage is the possibility to build a variety of dependence structures based on existing parametric or nonparametric models of the marginal distributions.

Using the copula approach to multivariate data, we can first estimate the marginal distributions, and then construct a copula that captures the dependence between the random variables. This two-step approach gives the investigator many options in model specifications. Secondly, in a copula model approach, we obtain a dependence function explicitly. Besides linear correlation, there are several other measures of dependence, among which Spearman's $\rho$ and Kendall's $\tau$ are most popular in the copula model building. Rank correlations are useful because, unlike the Pearson's product-moment correlation, they are invariant under monotonic transformations of marginal distributions.
A copula is called normal when it is created using the dependence structure of multivariate normal distribution. The normal copula is useful as it is defined for arbitrary dimension $k$ and it is easy to simulate. This family arises naturally in the case when data is multivariate normal. However the model may also be used in many situations where the corresponding marginal distributions are not normal.

If we have a multivariate distribution (classical or created by copulas), we can find different conditional distributions. We will apply the idea of imputing a missing value based on conditional distributions conditioned to the history of measurements, which can be derived forthrightly as the joint distribution is known. This conditional distribution gives complete information about incomplete data and gives many possibilities to impute missing values. The problem is that the joint distribution may be unknown,

but using the copula approach it is possible to find joint and conditional distributions modelling the data.
The aim of this work is to implement the concept of copula into the methodology for solving the imputation problem. As example, we will use Gaussian copula to derive three simple imputation formulas according to the chosen correlation structures using conditional mean as the imputed value.

The thesis is organized in the following way.

Chapter 1 describes the missing data problem and presents basic definitions and hierarchy of missingness mechanisms. A brief overview of methods for handling missing data with the emphasis to repeated measurements and handling dropouts is given as well. In this chapter we consider the imputation problem when conditional distributions are used as a key problem of the thesis and point out some open questions.
Chapter 2 introduces necessary tools in the copula theory. In particular, Gaussian copula is considered as a tool for finding joint and conditional distributions.
In Chapter 3 the correlation structure of repeated measurements is handled, and three new imputation algorithms are derived using Gaussian copula. In this chapter the following original results are presented: general form of imputation formula (3.3) (Proposition 3.1, Corollary 3.1) and its applications, formulas (3.5), (3.8), (3.9), (3.11) (Propositions 3.2, 3.3, 3.4, Corollary 3.2–3.5). An example with real incomplete repeated measurements data is given to illustrate the work of the new proposed algorithm.
Chapter 4 consists of results obtained from simulation study carried out to estimate the bias and effectiveness of new imputation rules. Simulation study showed that the suggested new imputation techniques are appropriate for imputing dropouts in the case of small sample sizes.

Most of the results given in Chapter 3 and 4 are published in Käärik (2005), Käärik (2006a) and Käärik (2006b) and presented at international conferences (WSEAS *Mathematical Biology and Ecology* (MABE'05) Udine, Italy, January, 2005; *Applied Statistics*, Ribno, Slovenia, Sept, 2005; *Compstat* 2006, Rome, August, 2006).

# Chapter 1

# Missing data

"The topic of missing data is as old and as extensive as statistics itself – after all, statistics is about knowing the unknowns" (Meng, 2000, p. 1328).

## 1.1. Basic assumptions and concepts

Incomplete or missing data is a common problem in empirical research and occur in every study, including sample surveys, where nonresponse is often a big problem, even in well-controlled situations. Whatever the reason, thus missing data requires the analyst to consider additional issues.

Common notation is following. Let $Y$ be partially observed data, $Y_{obs}$ and $Y_{mis}$ be the observed part and the missing part of $Y$, respectively. Therefore we can write full data $Y$ as $Y = (Y_{obs}, Y_{mis})$.

Let $M$ be the associated missing value indicator, which elements take the values 1 and 0 indicating, whether the corresponding values of $Y$ are observed ($M = 1$) or missing ($M = 0$). $Y$ can be a vector or a matrix, and $M$ has always the same dimension and is completely observed.

Usually it is assumed that $M$ has a distribution which may be unknown. The distribution of $M$ is called the *response mechanism* or *missingness mechanism*, but, to avoid some misunderstanding, Schafer and Graham (2002) suggested for the distribution of $M$ to use term the *distribution of missingness* or the *probabilities of missingness*.

The joint distribution of full data is

$$P(Y, M|\theta, \psi),$$

where $\theta$ parameterizes the measurement distribution and $\psi$ the missingness distribution.
Missingness distribution, in general, depends on the full data $Y$, hence missingness distribution can be described by

$$P(M|Y, \psi) = P(M|Y_{obs}, Y_{mis}, \psi).$$

In particular, Rubin (1976) and Little and Rubin (1987) made important distinctions between different missing data processes. They introduced the hierarchy of missingness mechanisms and characterized the assumptions regarding the nature of the missing values (Rubin, 1976; Little and Rubin, 1987).

1. *Missing Completely at Random* (MCAR): missingness is independent of the measurements $P(M|Y, \psi) = P(M|\psi)$.

2. *Missing at Random* (MAR): missingness is independent of the missing measurements, but depends on the observed measurements $P(M|Y, \psi) = P(M|Y_{obs}, \psi)$.

3. *Missing Not at Random* (MNAR): missingness depends on the observed and missing values $P(M|Y, \psi) = P(M|Y_{obs}, Y_{mis}, \psi)$.

First two types of missingness are called also *noninformative* or *ignorable nonresponse*. MNAR is called *informative* or *nonignorable nonresponse*.

*Missing completely at random* exists when missing values are randomly distributed across all observations. A missing value does not depend on the variable itself or on the values of other variables in the database. It means that the probability of an item being missing is unrelated to any measured or unmeasured characteristic for that unit and this is a very strong assumption.

*Missing at random* is a condition which is fulfilled when missing values are not randomly distributed across all observations but are randomly distributed within one or more subsamples. The probability of missing data of

any variable is not related to its particular value. The pattern of missing data is traceable or predictable from other variables in the database, but there is no residual relationship, the missingness is completely described by the observed variable.
Under MNAR there exists some residual dependence between missingness and $Y$ after accounting for the observed variable (Schafer and Graham, 2002, p. 151).

In case of MAR the probability of an item being missing depends only on other items that have been measured for that unit. This is a weaker assumption underlying most imputation methods, since they use the observed data to predict what is missing. There are some misunderstandings and problems in definition of MAR which are explained in Kenward and Molenberhgs (1998), Shafer and Graham (2002).

*Missing not at random* is the most problematic form, existing when missing values are not randomly distributed across observations and depend on the values that are missing. Thus, missingness is related to the variables under study, which is the weakest assumption, but complicated. The pattern of data missingness is non-random and it is not predictable from other variables in the data. It implies that the missing observations if measured, would have a different distribution from that predicted from what is observed. It is not possible to correct data by a nonignorable mechanism, except by using outside information.
In fact, missing data are closely related with other concepts, such as *coarse* data[1], which includes missing data as special case, or latent variable concept, which handles with unobservable quantities and several models (see Schafer and Graham, 2002; Roy, 2003).

[1]Heitjan and Rubin proposed a general model of data incompleteness and defined data to be *coarse* when one observes not the exact value of the data but only some set that contains the exact value. That is, data are neither entirely missing nor perfectly present (Heitjan and Rubin, 1991).

## 1.2. Dropout

We focus on the longitudinal or repeated measurements study with missing data. A characteristic of repeated measurements design is that each subject (unit) is observed at several different time points or under different experimental conditions. Unfortunately, repeated measurement studies are rarely balanced and complete.

A convenient framework for longitudinal study is the following.
Let $X = (X_1, \ldots, X_m)$ be an outcome variable with repeated measurements at time points $t_1, \ldots, t_m$. In this work we consider discrete time points and instead of $t_1, \ldots, t_m$ we will write $1, \ldots, m$.
Suppose that $n$ units are sampled repeatedly over time. The aim is to measure each unit $m$ times (in general, at the same time points), but due to dropouts some of them are measured at $s \leq m$ time points.

**Definition 1.1.** *Dropout* or *attrition* is missingness in data which occurs when subject leaves the study prematurely and does not return.

In the subsequent, we consider a sample of $n$ measurements $X_j$ that form a data matrix $\mathbf{X} = \{x_{ij}\},\ i = 1, \ldots, n;\ j = 1, \ldots, m$, in which due to dropouts some values are missing. In general $x_{ij}$ can be a vector of several measurements on the $i$-th subject at the $j$-th time point (*unit* of measurements). Usually, $x_{ij}$ means one measurement on the $i$-th subject at the $j$-th time point (*item* of measurements). In sample surveys, the corresponding missing values are called *unit nonresponse* or *item nonresponse* (see, for example, Lundström and Särndal, 2001; Durrant, 2005).
In our framework we do not distinguish between response and covariates, that is, between missing values in dependent variables and missing values in independent variables, therefore all variables are denoted as $X$.

**Definition 1.2.** In the case of dropouts, the missingness matrix is said to be *monotone* if, whenever an observation $x_{ik}$ is missing, $x_{is}$ is also missing for all $s > k$.

Monotonicity of the missingness matrix follows from two natural assumptions:

1. Subject which drops out does not return (Definition 1.1).

2. Order of subjects in sample does not matter, important is, that one dimension of the data matrix (*time*) has the fixed ordering.

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 subjects for which $X_j$ is observed. If the pattern is monotone, then $n_j \geq n_{j+1},\ j = 1, \ldots, m-1$. Hence, there always exists a permutation of the measurements such that a measurement earlier in the permuted sequence is observed for at least those subjects that are observed at later measurements. That is, in general, we can order the measurements so, that $n$ corresponds to the subject which drops out first, $n-1$ corresponds to the subject which drops out secondly, etc. (see Figure 1, here $n - 1 = n_k$).

**Definition 1.3.** Let $k$ be the time point at which the dropping out process starts. The vector $H = (X_1, X_2, \ldots, X_{k-1})$ is called *history* of measurements.

Herewith, natural assumption is that a history always has complete data.

Let $k$ be the time point when the dropping out process starts. Without restrictions we can assume, that until the time point $k-1$ we have complete data and that the rows have been sorted as in Figure 1.

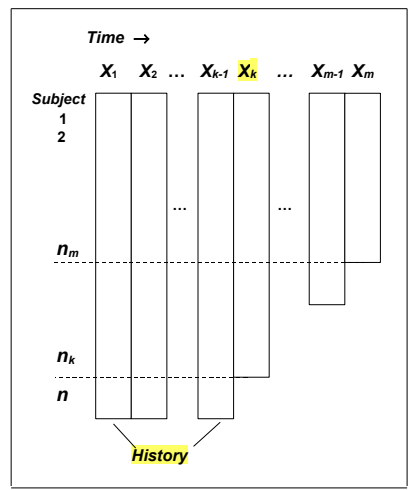


*Figure 1. Monotone missing data pattern with repeated measures and blocks representing data. Dropping out started at the time point $k$.*

Thus, in a longitudinal study, where the measurements are made over time, we can say that dropout in the sense of Definition (1.1) and monotone missingness (Definition 1.2) are equivalent. Generally, for any correlated measurements, the monotone missingness in the case of dropouts may be not obvious. Monotone missingness can appear also in the case of nonresponse in survey samples when the pattern of missingness is *nested*. By *nested* we mean that variables can be ordered in such a way that once a subject has a missing value at one observation, then it is subsequently missing everywhere else (see Little, 1992).

Analogously to general missing data approach, we can use here the missing data indicator matrix $M$, which elements are equal to 1 or 0 depending on whether the corresponding observation is taken or not. Diggle and Kenward (1994) used the concept of the dropout time $D$ which is a random variable, such that $2 \leq D \leq m$ identifies the dropout and $D = m + 1$ (or $D = 0$) identifies no dropout. Particularly, $D = k$ for some subject, if this subject drops out between the $(k-1)$th and $k$th timepoint, namely, if dropout process starts at timepoint $k$.

For longitudinal data we usually observe two types of missingness patterns: intermittent missing and dropout. Dropouts are distinguished from intermittent missing values, in which the set of intended times of measurements is not common for all units and which sometimes will be handled as unbalanced data.

For simplicity, hereafter we use notations without subscript for the subject's indicator $i$. Usually, the lowercase letter is used for subject which drops out and subscript denotes the time point.

## 1.3. Types of dropout

Consider the probability model for dropout time $D$ which depends on the history $H$ of a measurement process

$$P(D|X_1, \ldots, X_k) = P(D|H, X_k),$$

where $D = k$ is the dropout time and $H = (X_1, \ldots, X_{k-1})$. That means, in general, that the dropout probability depends on the observed measurements history $H$ and the unobserved variable $X_k$.

The classification of dropout processes is analogous to Rubin (see, for example, Little, 1990; Diggle and Kenward, 1994)

- *Completely random dropout* (CRD): dropout and measurement processes are independent

$$P(D|H, X_k) = P(D);$$

- *Random dropout* (RD): dropout process depends on observed measurements but not on unobserved measurements

$$P(D|H, X_k) = P(D|H);$$

- *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.

By Hogan *et al* (2004) there does not exist a unified terminology for describing dropout mechanism in longitudinal studies. They introduced the notion *Sequential Missing at Random*, which in their opinion, naturally fits to stochastic process formulation.

**Definition 1.4.** The dropout process is called *Sequential Missing at Random* (S-MAR), when conditionally on history the dropout process does not depend on current or future measurements.

S-MAR has definite meaning in general repeated measurements design where several covariates and responses are measured repeatedly. If we consider only one variable $X$, which is observed at $m$ time points $X_1, \ldots, X_m$

and there is monotone missingness, then we do not have any measurements of given object after dropout process has started. Thus, in our case there is no difference between MAR and S-MAR processes[2].

### 1.3.1. The risk of dropout

Lindsey (2000) proposed another typology of randomness for dropouts that relies on a survival model for the dropout process. In terms of a stochastic process, dropping out corresponds to a change of state of the subject. All subjects which do not drop out are censored in terms of dropout process. In this case, the repeated measurements data and dropout process can be modeled simultaneously, each conditional on the complete previous history. According to Lindsey (2000, p. 510), there are three types of missingness processes based on the *risk*[3] of dropout.

- The dropout is *random* if risk of dropout for all subjects can be described by the same homogeneous Poisson process, so that the risk of dropout for all subjects is not varying in time over the period in study.
- The dropout process is *ignorably nonrandom* if risk of dropout varies over time or depends on some factors in the same way for all subjects.
- The dropout process is *nonignorably nonrandom* if risk of dropout depends on any of the variables relevant to the process under study, including any specially collected as reasons for dropping out.

To model the dropout process given by these definitions, Lindsay proposed to implement some procedure for the survival data. He demonstrated how parametric proportional hazards model for failure time data can be fitted by Poisson regression.

Applying Lindsay's definitions there may be a good possibility to use survival copula and achieve good results in modelling dropouts, but it is not our task here. We considered the traditional approach to dropouts.

[2]See also Robins *et al* comment to his assumption 2*a* (Robins *et al*, 1995, p. 107).

[3]Risk or hazard function is the probability that a subject having not failed up to time $t$ will fail during the small interval $t + \triangle t$. Mathematically $h(x) = \frac{f(x)}{1-F(x)}$.

## 1.4. Objectives relating to dropouts

Dropping out is a difficult problem which often occurs in repeated measurements study. Depending on the missingness mechanism, different strategies can be used to analyze the data. Though a lot of research has been carried out, there does not exist the best approach valid for all situations. Non of the considered method dominates in the practical data analysis.

In the case of CRD the dropout does not dependent on data as subjects are randomly selected to dropout. This yields unbalanced data and one has adjust available statistical methods to this situation.

If the dropout process is RD, then the dropout is determined by the observed variables. In practice this means, that we usually know the reasons, why each subject has dropped out. Thus, 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.

In the case of ID, the dropout depends on an unobserved variable at the time of dropout. All analysis may be biased unless we do not have some additional information, sensitivity analysis may be reasonable in this case.

Currently we are interested in missing outcome variable, i.e. measurements that potentially could be obtained. Dropout may be an important outcome itself. 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. Next we will give a short overview about the most popular methods of handling missing data.

## 1.5. Handling missing data

In the literature there is a variety of methods proposed to deal with incomplete data (for example, monographs: Rubin, 1987; Little and Rubin, 1987; Schafer, 1997; Verbeke and Molenberghs, 2000). This area has been developed particularly in biostatistical and biomedical applications (see Schafer and Graham, 2002; Fitzmaurice, 2003; Hogan *et al*, 2004; Durrant, 2005; Hedeker and Gibbons, 2006). Among others, there are new techniques for imputation non-respondents in survey processes, developed by Laaksonen (see Laaksonen, 2002).
In fact, the proper method for handling incomplete data depends on the missingness mechanism.

### 1.5.1. Traditional approaches

Traditional approaches for handling missing data are well known. The simplest way to deal with missing data is to omit incomplete cases from analysis or *case deletion*:

**(i)** *Listwise (case wise) deletion* uses only complete cases,

**(ii)** *Pairwise deletion* uses all available cases.
The method is ordinary when analyzing two variables together and all cases observed in both variables have been analyzed. In general, the method is also useful for $3-, 4-$*wise* etc. deletion, that means in a statistical procedure the complete subsets of data are in use.

Listwise deletion omits cases which do not have data for all variables. This approach is implemented as the default method of handling incomplete data by many statistical procedures in commonly-used statistical software packages.
Pairwise deletion omits cases which do not have data on two variables used in the current calculation only. This means that different calculations (for example, different correlation coefficients) will use different cases and will have different samples. This effect is undesirable and may cause serious misinterpretations. As parameters are estimated from different sets, it is difficult to compute standard errors (Schafer and Graham, 2002).

Listwise deletion is preferred over pairwise deletion when sample size is large compared with the number of cases which have missing data. Already Little and Rubin (1987) have demonstrated the danger of simply deleting cases. Case deletion strategies assume that the deleted cases form a relatively small proportion of the entire dataset and they are representative. Rule of thumb: if a variable has less than 5% missing values with completely random missingness, then we can use case deletion.

Deletion of cases may cause two problems: (a) sampling-theoretical: the rest of sample may be not random and representative for the population; (b) loss of information, especially crucial in case of small samples and big amount of missing values.

In longitudinal study with dropouts, the listwise deletion means that we exclude all subjects which do not attend the study until the end. It means we lose a lot of information and in the case of small sample sizes we cannot allow this, the reduction in the number of subjects will lead to a reduction in statistical power which causes additional problems (Fitzmaurice, 2003). An alternative approach to case deletion is the correction of the missing values.

**Definition 1.5.** *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.

Imputation replaces a missing value for a variable with an imputed value, which has to be as correct as possible with regard to the true but unknown value. In general, the basic aim of imputation is to fill in the missing data by using values based on a specific model with certain assumptions.

There are methods based on a single imputation and methods based on multiple imputation, which, instead of filling in a single value for each missing value, one replaces each missing value with a set of plausible values. As result of imputation, missing data are filled-in (imputed) and all the statistical tools available for the complete data may be applied.
The parameter estimates could be obtained then from imputed data, the general aim is to get unbiased and efficient estimates by choosing an appro-

priate imputation method, which ideally has to be robust under misspecification of underlying assumptions.
As a result of imputation we get the point estimate of a missing value and sometimes it may be the aim in itself, but usually researchers are more interested in some statistics or models constructed from the completed data. Usually the imputation procedure starts by substituting the missing values for the variable with the fewest missing values from variables with complete data. Then the complete and imputed values are used to predict the missing values for the next variable, and so on until all the missing data are replaced. There could be a problem with this method, since variables that have their data replaced first using reduced model lack some important dependencies. Thus, it is important to know missing data mechanism.

The list of most popular methods for handling missing data is the following.

- *Single imputation methods.* Missing value is replaced with a single value.
  1. *Mean substitution.* Replace each missing value by the mean of observed values.
  2. *Regression methods.* Replace each missing value by the predicted value from a regression model estimated from the observed data.
  3. *Last observation carried forward* (LOCF) approach.
  4. *Hot Deck* approach, *nearest neighbor* imputation.
  5. *Expectation Maximization* (EM) approach.
- *Multiple imputation methods* (MI). A simulation-based approach to missing data.
- *Model based analysis.*

Next we will give a short overview of above mentioned methods accentuating to longitudinal data and introduce the method of imputation by *conditional distribution* which is of main interest in our work afterward.

### 1.5.2. Single imputation methods

**1. Mean substitution.** Replace a missing observation of the variable with its sample mean computed from available cases to fill in missing data values

on the remaining cases. When using longitudinal data, we can replace a missing value with the mean of the individual responses from earlier measurements for this individual. The essential drawback here is that the trend in the data is not considered. Mean substitution was once the most popular method for imputing missing values but is no longer preferred.
The problem is, when the data is MAR, this approach leads to biases in both, the standard errors and the parameters. The method shifts possible extreme values to the middle of the distribution, and it reduces variance in the variable being imputed; the correlations are inflated as well. Thus, mean substitution is no longer recommended.

**2. Regression-based imputation**[4]**.** In this approach a regression equation based on complete case data for a given variable is used to obtain predictions for missing values. When longitudinal data are used, an individual-specific regression can be used to predict the missing value.
This is probably one of the best simple approaches, but this underestimates standard errors by underestimating the variance. A simple remedy is to add some random error to the predicted value (called *stochastic substitution*) from the regression, but this rises another question concerning the distribution that should the error follow. The regression method assumes that missing values are MAR. The regression method also assumes that the same model explains the data for the non-missing cases as well as for the missing cases, which, of course, is not necessarily true.

**3. Last observation carried forward (LOCF).** This method is implemented specially in the case of repeated measurements, the last observed value is used to fill in the missing values at later points. That means we assume that the value at the time of dropout is the same as the previous one. Method can be accepted if measurements are expected to be relatively constant over time (the assumption of constant profile) or when the main interest is the outcome at the endpoint of the study, but typically using LOCF produces bias (Molenberghs *et al*, 2004).
Roy and Lin (2005) called this method a naive method as well as those

[4]Sometimes called also as *conditional mean* imputation (see, for example, Schafer and Graham, 2002)

using baseline measures[5] and ignoring missing data completely (Roy and Lin, 2005). This method assumes that an individual's missing value follows the same distribution as the previously measured values for that individual. Despite criticism by statisticians, the LOCF-method is still used to handle dropout in clinical trials because of its simplicity. The method may be useful for single use but certainly not for sequential imputation.

**4. Hot deck imputation.** Hot deck procedures contain the imputation methods in which missing values are replaced with values from another (most similar) subject in the current sample.

The hot deck procedures have some advantages (especially conceptual simplicity) and disadvantages. Hot deck can be superior to case deletion, and mean substitution approaches for handling missing data.

The methods are ordinarily used for the imputation of non-response in sample surveys and they are widely accepted as providing accurate samples of study population (see, for example, Fuller and Kim, 2005).

Using the hot deck imputation methods, the standard variance estimates are reduced because of the additional variability due to missing values and imputation is not taken into account. Hot deck imputation has a long history of use and there are many complementations made since Rao and Shao (1992), who suggested a jackknife method for estimation of variance in hot deck imputation.

Hot deck methods may be particularly difficult to implement in the case of continuous variables, they are simpler to use in practice with categorical data. The more variables are used to match the missing observation, the better, but also the less likely to find a match.

**4a. Nearest neighbor imputation** or *distance function matching* (see Chen and Shao, 2000; Durrant, 2005) is an approach where a random selection is made from several *closest* nearest neighbors. This imputation method is one of the hot deck methods used in sample surveys. The suitable distance measure is defined, the observed unit with the smallest distance (the nearest neighbor) to the missing observation is identified, and the

[5]Baseline approach considered that measurements are not changed since baseline and some baseline value is used to fill in the missing value

missing value is substituted by the value of the nearest neighbor.

**5. Expectation Maximization (EM)** approach. The EM algorithm (original from Dempster *et al*, 1977; comprehensive assay from Schafer, 1997) is a method that finds maximum likelihood estimates for incomplete data using an iterative procedure that proceeds in two steps. First, the expectation step (*E-step*) calculates the conditional expectation for missing data of the complete-data log likelihood, given the observed data and the parameter estimates.
The maximization step (*M-step*) substitutes the missing data by the expected values obtained from the E-step and then maximizes the likelihood function as if no data were missing to obtain new parameter estimates. The procedure iterates through these two steps until it converges.
EM-algorithm is simple to program and each iteration always increases the likelihood, but the convergence is often too slow. The algorithm is more used to obtain parameter estimates than to create imputation for individual missing data.

In general, single imputation methods have two general drawbacks. Firstly, the standard errors due to imputation are almost never calculated to account for the uncertainty behind imputed data, and secondly, they may cause systematic bias.

### 1.5.3. Multiple imputation methods

Multiple imputation (MI) avoids both problems associated with single imputation. Proper standard errors are estimated as a part of the process, thereby reflecting additional uncertainty that comes from using imputed data. In addition, MI produces unbiased estimates of the eventual statistics under reasonable assumptions.

Multiple imputation (Rubin 1987; Rubin, 1996; Schafer, 1997; Horton and Lipsitz, 2001; King *et al*, 2001) is a strategy of replacing each missing value with a set of plausible values that represent the uncertainty about the right value to impute. The multiple imputed data sets are then analyzed using the standard procedures for complete data and the results from these

analyzes are combined. Since each multiple imputation represent a random sample of missing values, this process yields valid statistical inference that properly reflects the uncertainty due to missing values. So, the multiple imputation inference involves the following three distinct phases:

1. Missing data are filled in $q$ times to generate $q$ complete data sets.
2. The $q$ complete data sets are analyzed using standard statistical methods.
3. The results from the analysis of $q$ complete data sets are combined to produce inferential results.

It has been shown that the efficiency of data imputation using MI is high even when the number of imputed datasets is low (in the range 3 to 10). The amount of calculations and the circumstance that we do not have single imputed value itself (which sometimes is important to know), may cause problems.

Depending on the patterns of missingness, various methods of multiple imputation can be implemented, the most well-known of them are the following:
(1) *parametric regression method* or *propensity scores* (non-parametric method) for the data sets with monotone missing patterns;
(2) *Markov Chain Monte Carlo* (MCMC) method for data sets with arbitrary missing patterns.

### 1.5.4. Model based analysis

Consider the joint distribution of the full data and dropouts with density function $f(Y, D|\psi, \theta)$ (see notation in subsections 1.1 and 1.3), where $\theta$ parameterizes the measurements distribution and $\psi$ the dropout distribution. Choosing the model implies specification for the density function. The joint distribution can be factored in different ways (Little and Rubin, 1987). According to the factorization there are two types of models: *selection models* and *pattern-mixture models*. Both of them do not require random missingness. Nice overview of the implementation of these models to informative dropouts in longitudinal data is given in Fitzmaurice (2003).

**1. Selection Model.** The measurement model and the dropout model can be fitted separately, provided that the parameters of the measurement process $\theta$ and the dropout process $\psi$ are statistically independent of each other. If the interest is only in the measurement model, the dropout model can be ignored.
In selection models we use the following factorization

$$f(Y, D|\theta, \psi) = f(D|Y, \psi)f(Y|\theta),$$

where $f(Y|\theta)$ is the density of $Y$ and $f(D|Y, \psi)$ is the conditional density of $D$ given $Y$.

**2. Pattern-mixture model.** The alternative factorization of the joint distribution is

$$f(Y, D|\theta, \psi) = f(Y|D, \theta)f(D|\psi),$$

which corresponds to pattern-mixture models. This model classifies subjects according to their missingness and describes the observed data within each missingness group (*pattern*).

There are some suggestions in literature how to deal with these models (Rubin, 1987; Little and Rubin 1987; Verbeke and Molenberghs, 2000). Pattern-mixture models are very sensitive to the assumptions made about the distributions of the variables with missing data and there is no standard way to test these assumptions. Hence the most important requirement is the good *a priori* knowledge of the mechanism of generating missing data.

**3. Latent dropout class model.** This approach is an alternative to the pattern-mixture models. Here the missingness pattern membership is unobserved itself, but the probability of being in a particular dropout pattern is determined by the dropout times. The correlation between the response and dropout time is modeled separately from serial correlation of the response (Roy, 2003).

In general, there is a relationship between pattern-mixture model and the structural equation modelling procedure (Schafer and Graham, 2002).

**4. Sensitivity analysis.** Sensitivity analysis (see, for example, Rotnitzky et al, 1998, 2001; Verbeke and Molenberghs, 2000; Daniels and Hogan, 2000;

Troxel et al, 2004) is a set of analyses showing the influence of different methods of handling missing data in the study. In sensitivity analysis we explore the results of the imputation method under a range of plausible assumptions about the dependence of dropout, etc. If informative dropout is assumed we have to collect more information on reasons of missingness to get better outcomes.
If the results of the sensitivity analysis are similar and consistent, then the robustness is guaranteed and missing values are acceptable. If the sensitivity analysis gives inconsistent results, then the validity of the chosen method may be questioned.
When substantial amounts of data are missing, the only analysis that matters is often the sensitivity analysis.

### 1.5.5. Imputation by conditional distribution

Beside other methods we are interested in imputation by conditional distribution. A full distribution model allows us to impute values from the distribution of missing observations conditional upon observed data. Using this approach we have to formulate the conditional distribution and draw a value from it.

In our repeated measurements framework complete data are presented by history $H = (X_1, \ldots, X_{k-1})$ until the time point $k-1$ and the measurement $X_k$ at time point $k$, which has at least one missing value $x_k$ for some subject. In general, we can assume that data have a $k$-variate distribution with joint density function $f_{H,X_k}$. Then the conditional density function of $X_k$, conditioned by the history, can be expressed as

$$f_{X_k|H}(x_k|x_1, \ldots, x_{k-1}) = \frac{f_{H,X_k}(x_1, \ldots, x_{k-1}, x_k)}{f_H(x_1, \ldots, x_{k-1})}.$$

Imputation from the conditional distribution usually means simulation or drawing value from $f_{X_k|H}$ (Schafer and Graham, 2002). We will use a somewhat different approach. Our goal is to find the imputed value that would be observed most likely, that is, we shall find the *argmax* of conditional density function in order to estimate the dropout $x_k$. The procedure is technically similar to the maximum likelihood estimation. That means, we

can find the conditional mean as the imputed value using the maximum likelihood method.

Thus, the conditional distribution approach consists of the following steps:

1. Construct the joint distribution function $F_{H,X_k}$ and the density function $f_{H,X_k}$ using marginals $F_1, \ldots, F_{k-1}, F_k$.

2. Find the conditional density function $f_{X_k|H}$.

3. Find the *argmax* of the conditional density function to estimate the dropout, and then use it as imputed value

$$\hat{x}_k = \arg\max_{x_k}[f_{X_k|H}(x_k|x_1, \ldots, x_{k-1})].$$

Hence, we will apply the idea of imputing a missing value based on conditional distributions conditionally to all observed variables up to dropout. The idea of using conditional distribution is distinguished and extensively used. Conditional distribution contains all the information about the history of measurements (as condition) and about marginal distributions (unconditioned information that is specified using this condition).

Using conditional distribution of the missing value conditioned to the history of measurements we can solve the following tasks.

1. Estimate all distribution parameters using any method for estimation of parameters. For example, we can use the maximum likelihood method to estimate conditional mean (the most likely value of dropout), median, etc.

2. Estimate the dropout using some other loss-function for estimation of missing value (for instance, when losses of overestimation and underestimation are different).

3. Estimate the precision of an estimate (confidence interval, standard deviation) using standard statistical methods.

4. Find possible extreme values (or quantiles) for dropout.

5. Generate one or several draws from conditional distribution. In the case of several generated draws we obtain the multiple imputation rule.

Further we will use the conditional mean or the expected value as an imputed value.
The main problem is that the joint distribution may be unknown and finding conditional distribution may be therefore impossible. By using the copula theory the approximate joint and conditional distributions can be still found, which motivates us to use the copulas later on.

### 1.5.6. Cautions to imputation

Missing data analysis procedures do not generate something out of nothing. Missing data analysis procedures do make the most out of the data available, maximizing precision of estimation and eliminating biases.

In many papers in this field we can read the following cautionary citation of Dempster and Rubin[6]:

*The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard errors applied to the real and imputed data have substantial biases.*

Of course, we have to take into account that even through imputation we have complete data, inference, in particular point estimation, is valid only if the additional underlying assumptions are satisfied. Most conventional methods are inefficient and produce biased estimates, except under strict assumptions.

There are many open questions here and many solutions for missing data problems have been available in the statistical literature for some time now, the best or most reasonable procedure for imputing is often complicated to

[6]See for example Verbeke and Molenberghs, 2000, p. 224.

choose. No universally best and generally accepted approach for handling missing data exists.
The researcher must assume that missing observations differ from observations where values are present. The problem with missing data is mostly the possibility that the remaining data are biased not so much that the sample size is reduced.

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 prediction of values for specific cases. Important is to remember that *imputation of the dropout does not give us qualitatively new additional information but enables with maximal effectiveness to use all available information about the data for achieving our purpose in the best way.*
As a matter of fact we are interested in small sample sizes where every value is important and imputation results are of scientific interest itself.

# Chapter 2

# Copula

A fundamental problem in mathematical statistics is to determine a relationship between a multivariate distribution function and its lower dimensional margins. In many situations we are interested in construction of multivariate distribution with given marginal distributions and dependence structure. The problem of existence multivariate distribution function with discrete marginals was introduced by Tiit and Käärik some years ago (Tiit and Käärik, 1996).

One of the most useful tools for handling multivariate distributions with dependent components is the copula. We give here a brief review of some important concepts of copula.

Copula is a function that allows to represent a joint distribution of random variables as a function of marginal distributions specifying the dependence structure. Copula links univariate marginal distribution functions to their joint multivariate distribution function.

In fact, copula function was introduced independently in 1940s by Hoeffding and Fréchet, whose research area consisted of the analysis of the relationship between a multidimensional probability distribution and its lower-dimensional marginals, especially in case of maximal and minimal distributions. Basic developments of the properties of the copula function can be found in three fundamental papers by Hoeffding (1940–1942) in German and these were long time unnoticed (see Fisher, 1997).

Sklar (1959) defined and provided some general properties of copulas. He established the copula function and showed that any joint distribution function can be considered as a copula function.

There exists rapidly growing literature in copula theory. The first principal books in this area were written by Joe (1997) and Nelsen (1999), for an exhaustive overview see Lindskog (2000) or Embrechts *et al* (2001).
Applications of copula theory appeared in econometrics, finance and actuarial science (see, for example, Frees and Valdez, 1998; Embrecht *et al*, 1999, 2001; Clemen and Reilly, 1999) and have been rapidly developing in recent years. Copulas have been applied to a wide range of problems in biostatistics (Lambert and Vandenhende, 2002; Vandenhende and Lambert, 2002, 2005; Lindsay and Lindsay, 2002) and recently to hydrology and environmental data as well (see, for example, Dupuis, 2006; De Michele and Salvadori, 2006; Zhang and Singh, 2006).

Recently, there have appeared some critical remarks about fast growing copula applications (Mikosch, 2005).

## 2.1. Basic definitions and theorems

**Definition 2.1.** A copula is a function $C : [0,1]^k \to [0,1]$ which has following properties

1. $C(u_1, \ldots, u_{j-1}, 0, u_{j+1}, \ldots, u_k) = 0$ (is grounded);
   $C(1, \ldots, 1, u_j, 1, \ldots, 1) = u_j$ for all $j \in \{1, \ldots, k\}$, $u_j \in [0,1]$;

2. $C(u_1, \ldots, u_k)$ is nondecreasing in each component $u_j$;

3. For all $(u_{11}, \ldots, u_{k1}), (u_{12}, \ldots, u_{k2}) \in [0,1]^k$ with $u_{i1} \leq u_{i2}$ we have the rectangle inequality:

$$\sum_{i_1=1}^{2} \ldots \sum_{i_k=1}^{2} (-1)^{i_1+\ldots+i_k} C(u_{1i_1}, \ldots, u_{ki_k}) \geq 0.$$

Because of these properties, a copula is the distribution function of a random vector in $\mathbb{R}^k$ with uniform (0,1) marginals. Property 1 is necessary

for the existence of the marginal uniform distributions. Properties 2 and 3 correspond to the properties of distribution function.
If $F_1, \ldots, F_k$ are univariate distribution functions, then $C(F_1(x_1), \ldots, F_k(x_k))$ is a multivariate distribution function with marginals $F_1, \ldots, F_k$ because $U_j = F_j(X_j), j = 1, \ldots, k$, are uniformly distributed random variables.
In definitions standard uniform marginals are used, but in general the marginals might be arbitrary.

**Theorem 2.1** (**Sklar**). *Suppose that $F$ is a distribution function on $\mathbb{R}^k$ with one dimensional marginal distribution functions $F_1(x_1), \ldots, F_k(x_k)$, then there exists a copula $C$ so that*

$$F(x_1, \ldots, x_k) = C(F_1(x_1), \ldots, F_k(x_k)). \tag{2.1}$$

*If $F$ is continuous, then $C$ is unique and is given by*

$$C(u_1, \ldots, u_k) = F(F_1^{-1}(u_1), \ldots, F_k^{-1}(u_k)) \tag{2.2}$$

*for $u = (u_1, \ldots, u_k) \in \mathbb{R}^k$, where $F_i^{-1} = \inf\{x : F_i(x) \geq u\}, i = 1, \ldots, k$, is the generalized inverse of $F_i$.*
*Conversely, if $C$ is a copula on $[0,1]^k$ and $F_1, \ldots, F_k$ are distribution functions in $\mathbb{R}$, then the function defined in* (2.1) *is a distribution function on $\mathbb{R}^k$ with one-dimensional marginal distribution functions $F_1, \ldots, F_k$.*

This theorem provides an easy way to form multivariate distributions from known marginals that need not to be necessarily from the same distribution, combining them with a copula function and getting a suitable joint distribution. There are two principal ways to use the copula idea. We can extract copulas from well-known multivariate distribution functions. We can also create a new multivariate distribution function by joining arbitrary marginal distributions together with a copula.

Hence, we have the random vector $X = (X_1, \ldots, X_k) \in \mathbb{R}^k$, marginal distribution functions $F_1, \ldots, F_k$, and the joint continuous distribution function $F$, so that $X_i \sim F_i$ and $X \sim F$. Suppose now that we transform the random vector component-wise to have standard uniform marginal distributions $U(0,1)$. This can be achieved using *probability integral transformation* $X_i \mapsto F_i(X_i) = U_i$. Thus from (2.1) and (2.2) we see that the copula is the multivariate distribution which links univariate uniform marginals.

The following theorem (see proof, for instance, Lindskog, 2000) shows one important feature of the copula representation, namely that copula is invariant under increasing and continuous transformations of the marginals.

**Theorem 2.2** (Invariance theorem)**.** *Consider $k$ random variables $X_1$, $\ldots, X_k$ with a copula $C$. Then, if $g_1(X_1), \ldots, g_n(X_n)$ are continuous strictly increasing on the ranges of $X_1, \ldots, X_k$, then the random variables $Y_1 = g(X_1), \ldots, Y_k = g(X_k)$ have exactly the same copula $C$.*

In the case where all marginal distributions are continuous it suffices that the transformations are increasing.

This theorem shows that the dependence between the random variables is completely captured by the copula, independently of the shape of the marginal distributions. This property is very useful as transformations are commonly used in statistical analysis. For example, no matter whether we are working with $X$ or $\log X$, we get the same copula.

Another fundamental property of copulas is that Frechet-Hoeffding bounds exist for copulas (Joe, 1997; Nelsen, 1999). For example, in two-dimesional case, for any copula $C$ and for all $(u, v) \in [0, 1]$

$$W(u, v) = max(u + v - 1, 0) \leq C(u, v) \leq min(u, v) = M(u, v),$$

where $W(u, v)$ are called the *minimum* copula and $M(u, v)$ the *maximum* copula which correspond to perfect negative and positive dependence, respectively.

## 2.2. Joint and conditional density functions

We focus to the case where each marginal distribution $F_i$ is continuous and differentiable. If $C$ and $F_1, \ldots, F_k$ are differentiable, then the *joint density* $f(x_1, \ldots, x_k)$ corresponding to the joint distribution function $F(x_1, \ldots, x_k)$ can be written by canonical representation as a product of the marginal densities and the copula density

$$f(x_1, \ldots, x_k) = f_1(x_1) \cdot \ldots \cdot f_k(x_k) \cdot c(F_1, \ldots, F_k), \tag{2.3}$$

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

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

Copulas which are not absolutely continuous do not have joint densities. Equation (2.3) is known as the density version of Sklar's theorem: the joint density can be decomposed into product of the marginal densities and the copula density. Underlying theory of the equation (2.3) is essence of the *copula density*, which is equal to the ratio of the joint density $f$ to the product of all marginal densities $f_i$.

For example, in bivariate case the copula density can be found as follows. Consider two random variables $X_1, X_2$, so that $X_1 \sim F_1$ and $X_2 \sim F_2$, and let the joint distribution function $F$ be defined by copula $C$, $F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$. The probability integral transformations of random variables are $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$, so we have $X_1 = F_1^{-1}(U_1)$ and $X_2 = F_2^{-1}(U_2)$. These transformations are strictly increasing and continuous and we get:

$$c(u_1, u_2) = \frac{\partial^2 C}{\partial u_1 \partial u_2} = f(F_1^{-1}(u_1), F_2^{-1}(u_2))|\mathcal{J}| = \frac{f(F_1^{-1}(u_1), F_2^{-1}(u_2))}{f_1(F_1^{-1}(u_1)) f_2(F_2^{-1}(u_2))},$$

where the Jacobian of the transformation is following:

$$\mathcal{J} = \begin{vmatrix} \frac{\partial X_1}{\partial U_1} & \frac{\partial X_1}{\partial U_2} \\ \frac{\partial X_2}{\partial U_1} & \frac{\partial X_2}{\partial U_2} \end{vmatrix},$$

whith $\frac{\partial X_i}{\partial U_i} = (\frac{\partial U_i}{\partial X_i})^{-1} = (\frac{\partial F_i(X_i)}{\partial X_i})^{-1} = f_i^{-1}(X_i)$, and $\frac{\partial X_i}{\partial U_j} = \frac{\partial X_j}{\partial U_i} = 0$, $i \neq j, \quad i, j = 1, 2.$

□

The next essential notion is conditional distribution. Taking into account the joint density defined by copula and univariate marginals (2.3) and basic definition of the conditional density we get the *conditional density* defined by copula as follows:

$$f(x_k | x_1, \ldots, x_{k-1}) = \frac{f(x_1, \ldots, x_k)}{f(x_1, \ldots, x_{k-1})}$$

$$= \frac{f_1(x_1) \cdot \ldots \cdot f_k(x_k) \cdot c(F_1, \ldots, F_k)}{f_1(x_1) \cdot \ldots \cdot f_{k-1}(x_{k-1}) \cdot c(F_1, \ldots, F_{k-1})}$$

$$= f_k(x_k) \frac{c(F_1, \ldots, F_k)}{c(F_1, \ldots, F_{k-1})} \ , \tag{2.4}$$

where $c(F_1, \ldots, F_k)$ and $c(F_1, \ldots, F_{k-1})$ are corresponding copula densities.

## 2.3. Gaussian copula

One of the most important examples of copulas is the normal or Gaussian copula (Clemen and Reilly, 1999; Reilly, 1999; Song, 2000; Lindsay and Lindsay, 2002; Lambert and Vandenhende, 2002). Gaussian copula belongs to the class of *implicit* copulas which are derived from distributions. Implicit copulas do not have a simple closed form, but are implied by well-known distribution functions (see, for example, Aas, 2005).

By definition, the $k$-variate Gaussian copula with $k$ Gaussian marginals corresponds to the $k$-variate Gaussian distribution. From copulas point of view, the multivariate normal distribution has normal marginal distributions and Gaussian copula dependence. Gaussian copula handles dependence in the same way as multivariate normal distribution, that means it uses only pairwise dependencies among the variables, but it does so for variables with arbitrary marginals.

Hence, the marginal distributions of $k$-variate normal copula are assumed to be continuous and can substantially differ from normal ones and can, in principle, be different. The advantage of using normal dependence structure arises from its simplicity, analytical manageability and the easy estimation of its only parameter, the matrix of pairwise dependencies.
For instance, the bivariate Gaussian copula is defined as

$$C_2(u, v, \theta) = \Phi_2[\Phi_1^{-1}(u), \Phi_1^{-1}(v), \theta], \ u, v \in (0, 1), \tag{2.5}$$

where $\Phi_2$ is the standardized bivariate normal distribution function with correlation coefficient $\theta$ and $\Phi_1$ is the univariate standard normal distribution function. This notation can be easily extended to multivariate case with replacing $\theta$ by a correlation matrix $R$.

**Definition 2.2.** Let $R$ be a symmetric, positive definite matrix with $diag(R) = (1, 1, \ldots, 1)^T$ and $\Phi_k$ be the standardized $k$-variate normal distribution function with correlation matrix $R$. Then the *multivariate Gaussian copula* is:

$$C_k(u_1, \ldots, u_k; R) = \Phi_k(\Phi_1^{-1}(u_1), \ldots, \Phi_1^{-1}(u_k)). \tag{2.6}$$

*Remark.* Under linear independence (when $R$ is the identity matrix) the copula $C_k(\cdot, R)$ reduces to independence or product copula $\Pi(\cdot)$. The product copula is defined as $\Pi(u_1, \ldots, u_k) = u_1 \cdot \ldots \cdot u_k$.

## 2.4. Modelling joint density of repeated measurements by Gaussian copula

Let us use the whole history $H$ of repeated measurements. Suppose there is a subject $i$ such that until the time point $k-1$ the measurements $X_1, X_2, \ldots, X_{k-1}$ are observed and at the time point $k$ the subject drops out, thus the measurement $X_k$ has a missing value.
Suppose the measurement $X_j$ has continuous distribution function $F_j$ ($j = 1, \ldots, k$), in general we can normalize this using normalizing transformation

$$Y_j = \Phi_1^{-1}[F_j(X_j)], \quad j = 1, \ldots, k, \tag{2.7}$$

where $\Phi_1^{-1}$ is the inverse of the standard univariate Gaussian distribution function.
Then by using the $k$-variate normal copula we get the following expression for the joint multivariate distribution function:

$$F(y_1, \ldots, y_k; R) = C_k(u_1, \ldots, u_k; R) = \Phi_k[\Phi_1^{-1}(u_1), \ldots, \Phi_1^{-1}(u_k); R],$$

where $u_j \in (0, 1), j = 1, \ldots, k$; $\Phi_k$ is the standard $k$-variate normal distribution function with the correlation matrix $R$.
For getting the normal joint density function we have to find the normal copula density $c_k$ as a derivative from normal copula $C_k$. According to (2.3)

the joint density function is the product of the marginal densities and the copula density

$$\phi_k(y_1, \ldots, y_k|R) = \phi_1(y_1) \cdot \ldots \cdot \phi_1(y_k) \cdot c_k[\Phi_1(y_1), \ldots, \Phi_1(y_k); R], \quad (2.8)$$

where $\Phi_1$ and $\phi_1$ are the univariate standard normal distribution function and density function, respectively.
From here we get the copula density

$$c_k[\Phi_1(y_1), \ldots, \Phi_1(y_k); R] = \frac{\phi_k(y_1, \ldots, y_k|R)}{\phi_1(y_1) \cdot \ldots \cdot \phi_1(y_k)}. \quad (2.9)$$

The copula density contains information about dependencies among marginals and is called also *dependence* function.
Using standard normal density expressions we get the normal copula density in the following form

$$\begin{aligned} c_k[\Phi_1(y_1), \ldots, \Phi_1(y_k); R] &= \frac{\exp[(-1/2)Y^T R^{-1} Y + (1/2)Y^T Y]}{|R|^{1/2}} \\ &= \frac{\exp\{-Y^T(R^{-1} - I)Y/2\}}{|R|^{1/2}}, \end{aligned} \quad (2.10)$$

where $Y = (Y_1, \ldots, Y_k)$ and $I$ is the $k \times k$ identity matrix.
To construct a multivariate density for arbitrary marginals we now use the marginal densities $f_1(x_1), \ldots, f_k(x_k)$ and copula density $c_k$ (2.10) as the dependence function. Thus, we obtain the joint density, as follows

$$f_k(x_1, \ldots, x_k|R) = f_1(x_1) \cdot \ldots \cdot f_1(x_k) \cdot \frac{\exp\{-Q_k^T(R^{-1} - I)Q_k/2\}}{|R|^{1/2}}, \quad (2.11)$$

where, following (2.7) we denoted $Q_k = (\Phi_1^{-1}[F_1(x_1)], \ldots, \Phi_1^{-1}[F_k(x_k)])$ to stress on arbitrary marginals.

Eventually, considering (2.4) and (2.10) we can find the conditional density of Gaussian copula. We will derive the conditional density after examining structure of the correlation matrix and its partition in Chapter 4.

## 2.5. Other copulas

The classical approach to describe dependence is based on the multivariate normal distribution. The *normal copula* is useful because of easy implementation in practice and simple simulation rule.

Problem for normal copula is that it takes into account only second order moments and all higher order moments are uniquely determined by them. It follows that dependence structure of Gaussian copula considers only pairwise dependencies and does not account higher order dependencies. On the other hand it should be pointed out that for estimating higher order dependencies a lot of parameters are needed and this may be a complication for practical usage.

However, there exist increasing evidences indicating that normal assumptions are inappropriate in many situations in the real world. In general, a multivariate normal distribution is not an ideal model and is valid when only measurement error is present, at the same time ignoring or poorly modeling the dependencies between repeated measurements.
To solve this problem some other distributions and/or some other copulas for joint distributions can be applied. For example, Lindsey and Lindsey (2004) suggested Student's t-distribution, power-exponential or skew Laplace distribution for modeling repeated responses instead of normal distribution (Lindsay and Lindsay, 2004). Lambert and Vandenhende (2002) used normal copula when marginals were gamma, Weibull, inverse Gaussian, normal, log-normal, Student and log-Student distributions (Lambert and Vandenhende, 2002).

From the wide variety of copulas probably the elliptical and Archimedean copulas are the most useful in applications.

### 2.5.1. Elliptical copula

A natural extension of the multivariate normal distribution is the class of elliptical distributions. An elliptical distribution is the multivariate generalization of the family of univariate symmetric distributions. Classical examples of elliptical distributions are the multivariate normal and the mul-

tivariate $t$-distributions. The class of elliptical distributions shares many tractable properties of the multivariate normal distribution and enables modelling multivariate extremes and other forms of non-normal dependencies. In the family of elliptical distributions, additionally to the correlation matrix, the dependence structure takes into account the forth moments as well.

Elliptical copulas are generally defined as copulas of elliptical distributions (Bouyé, 2000; Lindskog, 2000; Embrecht *et al*, 2001; Demarta and McNeil, 2004). So they are useful when observed data are not normally distributed and tend to have marginal distributions with heavier tails. Elliptical copulas are able to support tail dependencies. Tail dependence is a concept that is relevant to dependence in extreme values. Joe (1997) introduced the tail dependence to describe the tail behavior of copulas. The tail dependence between the random variables exists when the probability of joint extreme events is higher than what could be predicted from the marginal distributions.

For example, in bivariate case the upper tail dependence is defined as follows.

**Definition 2.3.** Let $X_1$ and $X_2$ be random variables with continuous marginal distribution functions $F_1, F_2$ and copula $C$ as the joint distribution. The coefficient of *upper tail dependence* of two random variables is

$$\lambda_U = \lim_{u \to 1} P(X_1 > F_1^{-1}(u) | X_2 > F_2^{-1}(u)) = \lim_{u \to 1} \frac{1 - 2uC(u,u)}{1-u}.$$

For elliptical copulas the coefficient of upper tail dependence is equal to the coefficient of lower tail dependence, since elliptical distributions are radially symmetric.

One example of elliptical copulas is the Student $t$-copula defined as

$$C_t(u_1, \ldots, u_k; \nu, R) = \Psi_t(\Psi_1^{-1}(u_1, \nu), \ldots, \Psi_1^{-1}(u_k, \nu); \nu, R),$$

where $\Psi_t$ denotes the $k$-variate Students $t$ distribution function, $\Psi_1^{-1}$ denotes the inverse of the univariate Students $t$ distribution function, $\nu$ is the number of degrees of freedom and $R$ is the correlation matrix.

The main difference between Student's and Gaussian copulas lies in the probability of extreme events. A Gaussian copula has zero tail dependence, that means that the probability that variables are in their extremes is asymptotically zero unless linear correlation coefficient is equal to one, while the Student $t$ has symmetric, but nonzero tail dependence. Of course for moderate values of the correlation coefficient, the Student copula with large number of degrees of freedom may be close to the Gaussian copula.

There are, however, drawbacks of elliptical copulas: they do not have closed form expressions and their applicability is restricted, asymmetries cannot be modeled with elliptical copulas.

### 2.5.2. Archimedean copula

An important class of parametric copulas to model non-normal data is the class of *Archimedian* copulas, which is often used in applications because of easy construction and estimation (see Genest and MacKay, 1986; Genest and Rivest, 1993; Frees and Valdez, 1998; Nelsen, 1999; Lindskog, 2000).

The term *Archimedean*[1] copula first appeared in the statistical literature in the paper by Genest and Mackay (1986).

In particular, Archimedean copulas belong to *explicit* copulas. They have closed form expressions and are defined explicitly, not derived from multivariate distributions using Sklar's theorem. A disadvantage of this model is that extensions of bivariate Archimedean copulas to multivariate ones need some technical assumptions about their parameters, so the choice of free parameters is restricted.

Archimedean copulas are widely used in applications due to their simple form and a variety of dependence structures.
The main advantages of Archimedean copulas are the following.

- Archimedean copulas can be easily constructed. In general, when comparing with elliptical copulas, they have simpler, closed form ex-

[1] Nelsen explains the utilization of term *Archimedean* copula by *Archimedean* property (see Nelsen, 2005, p. 2).

pressions. For instance, the expression of the Gaussian copula involves the inverse standard Gaussian distribution function, i.e. the inverse of a function defined by an integral.

- Class of Archimedean copula allows to use a variety of dependence structures. In particular, Archimedean copulas can have asymmetric tail dependence.

Archimedean copulas are constructed using a continuous, strictly decreasing convex function $\varphi$.

**Definition 2.4.** A copula function $C$ is called *Archimedean* if it can be written in the following form:

$$C(u_1, \ldots, u_k) = \varphi^{-1}[\varphi(u_1) + \cdots + \varphi(u_k)]$$

for all $0 \leq u_1, \ldots, u_k \leq 1$ and for some continuous function $\varphi$ (often called the *generator*) satisfying three conditions

- $\varphi(1) = 0$;
- $\varphi$ is strictly decreasing and convex.
  That is, for all $x \in (0, 1)$, $\varphi'(x) < 0$, $\varphi''(x) \geq 0$;
- $\varphi^{-1}$ is completely monotonic on $[0, \infty)$.

A collection of twenty-two families of Archimedean copulas can be found in Nelsen, 1999, pp 94–97.

As already mentioned, copulas do not always have densities. Copulas which are absolutely continuous have densities, and for Archimedean copulas with generator $\varphi$, the density is given by

$$f_k(x_1, \ldots, x_k) = \varphi^{-1(k)}\{\varphi[F_1(x_1)] + \cdots + \varphi[F_k(x_k)]\} \prod_{i=1}^{k} \varphi^{(1)}[F_i(x_i)] F_i^{(1)}(x_i),$$

where $\varphi^{-1(k)} = \frac{\partial^k}{\partial x_1 \ldots \partial x_k} \varphi^{-1}$, that means the superscript notation $(k)$ is used for the $k$-th mixed partial derivative.

The conditional density of $X_k$ with given past $H = (X_1, \ldots, X_{k-1})$ is accordingly (Frees and Valdez, 1998)

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

where $c_{k-1} = \varphi[F_1(x_1)] + \ldots + \varphi[F_{k-1}(x_{k-1})]$.

For Archimedean copulas the dependence measure can be expressed in terms of the generator as shown in bivariate case (Genest and McKay, 1986).

**Theorem 2.3** (Kendall's tau)**.** *Let $X_1$ and $X_2$ be random variables with an Archimedean copula $C$ generated by $\varphi$, then Kendall's tau of $X_1$ and $X_2$ is given by*

$$\tau_C = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt \tag{2.12}$$

**Example: Frank's copula**

In repeated measurements study the Frank's copula from class of Archimedean copulas was implemented by Vandenhende and Lambert (2000, 2002) to describe dependence between dropout and response. They used Frank's copula in the case of ordinal responses for the dropout model and tested several marginal distributions (Cauchy, Gamma, log-normal). In Vandenhende and Lambert (2005) Archimedean copulas (among others the Frank's copula) are used for lifetime study of Danish twins.

The generator function of the Frank's copula (Genest, 1987; Nelsen, 1999) is

$$\varphi(t) = -\ln\frac{e^{\{-\alpha t\}} - 1}{e^{\{-\alpha\}} - 1},$$

with one parameter $\alpha$, which measures strength of dependence between marginals.

Hence, the $k$-variate Frank's copula has the form

$$C(u_1, \ldots, u_k) = -\frac{1}{\alpha}\left(1 + \frac{\prod_{i=1}^{k}(e^{-\alpha u_i} - 1)}{(e^{-\alpha} - 1)^{k-1}}\right).$$

Genest (1987) gave the conditional mean function in the case of bivariate Frank's copula:

$$E(X_2|X_1 = x) = \frac{(1 - e^{-\alpha})xe^{-\alpha x} + e^{-\alpha}(e^{-\alpha x} - 1)}{(e^{-\alpha x} - 1)(e^{-\alpha} - e^{-\alpha x})}, \tag{2.13}$$

where relationship between copula parameter $\alpha$ and Kendall's tau is the following

$$\tau = 1 - \frac{4(1 - G_1(\alpha))}{\alpha}$$

and

$$G_1(\alpha) = \frac{1}{\alpha}\int_0^{\alpha} \frac{t}{e^t - 1}dt.$$

The equation (2.13) can be used for imputation in the case of two repeated measurements.

## 2.6. Modelling using copulas

The canonical representation for the joint density function (2.3) permits in general a statistical modelling of copula to be decomposed into the following steps:

- determination of the marginal distributions $F_1, \ldots, F_k$ and estimation their parameters (in our case marginals are distributions of repeated measurements at the time points $1, \ldots, k$, where any univariate distribution can be used as a marginal),
- determination of the appropriate copula function which completely describes the dependence structure of random variables (in our case the dependencies between repeated measurements),
- determination of the joint and conditional distributions (in our case the conditional distribution of the missing value conditioned to the history of measurements).

This decomposition of modelling into steps is the main advantage of the copula approach. Instead of estimating all the parameters of the distribution simultaneously, we can estimate parameters of marginal distributions

separately from the joint distribution. Given the estimated marginal distributions we use appropriate copula to construct the joint and conditional distribution.

Compared to using the joint distribution directly, working with the copula model has several advantages.

*Firstly*, in many cases, it may be complicated to specify a joint distribution directly within any well-known families. Besides, traditional representation of multivariate distribution requires that all random variables come from the same family of marginals. Using the copula approach we can first estimate arbitrary marginal distributions. By changing the types of marginal distributions and their parameters we can select the best model for data, and then estimate the dependence structure as the copula parameter.

*Secondly*, in the copula model approach we obtain a dependence function explicitly, which enables us to provide a more specific description of dependence. In repeated measurements study the assessment of dependence structure is extremely important. We can vary the dependence structure by choosing different copulas or the same copula with different parameter values.
Furthermore, the family of copulas is sufficiently large and allows a wide range of multivariate distributions as models.

Now, we can summarize and present the following important features of copulas:

- A copula describes how the marginals are connected together in the joint distribution. Every joint distribution may be written as a copula which entirely assigns the dependence between random variables and the type of dependence is not limited only to correlation.

- The marginal distribution functions and the copula can be estimated separately. Copula separates dependence structure and marginal behavior.

- Given a copula we can obtain many multivariate distributions by selecting different marginals. Given marginal distributions we can

vary different copulas and obtain different multivariate distributions having different dependence structure.

Comparing copulas available for connecting arbitrary marginals to a multivariate distribution the Gaussian copula seems to be the best for practical use because

1. It has good opportunities for describing the dependence structure: it is possible to estimate all $\frac{k(k-1)}{2}$ dependence coefficients of a $k$-variate random variable.

2. It is possible to find a dependence structure to describe models with small number of parameters.

3. The conditional distribution of the missing value can be found for every existing dependence structure.

4. For a simple dependence structure simple formulas can be found for calculating conditional mean (as imputed value) or standard deviation of conditional distribution.

# Chapter 4

# Simulation study

The goal of the simulation study is to test the effectiveness of the original imputation methods given by formulas (3.5), (3.9) and (3.11) by comparison with some well-known imputation methods in the case of different missing data mechanisms, dependence structure and sample sizes.
We start of comparisons with normally distributed data and then check the robustness of the imputation methods by moving away from the normal distribution. We have performed two simulation experiments:
(1) using standard normal distribution;
(2) using a skewed distribution.
At first we generated the complete dataset and then datasets with dropout are formed using three missingness mechanisms from the complete set.
As a quality measure the standardized absolute difference between the observed value and the imputed value was used.

## 4.1. Generation of the complete data

In the first simulation experiment we generated the complete data matrix from a multivariate normal distribution using

(a) constant correlation structure;

(b) autoregressive correlation structure with the correlation coefficients $\rho = 0.5$ and $\rho = 0.7$;

(c) banded Toeplitz correlation structure with the correlation coefficient $\rho = 0.5$ (here correlations $\rho = 0.7$ are not possible in Toeplitz structure, see Remarks in page 67).

We generated data from 3-, 6- and 12-dimensional normal distribution with sample sizes $n = 10$ and $n = 20$, assuming that the data represent repeated measurements. Small sample sizes are typical for studies with repeated measurements.

The second simulation study we performed in the case of skewed marginals. Suppose $X = (X_1, \ldots, X_k)$ has the $k$-variate normal distribution. To get a skewed marginals $Z_1, \ldots, Z_k$ the data were transformed using the following rules

$$z_{ij} = \begin{cases} C_1 v_i & \text{for maximum value } v_j = \max_i x_{ij}, \\ C_2 x_{ij}, & \text{for every other positive value } x_{ij}, \\ x_{ij}, & \text{otherwise.} \end{cases}$$

The constants were chosen $C_1 = 10$ and $C_2 = 5$.

## 4.2. Generation and imputation of the dropouts

The dropouts occur at the last time point (in the random variable $X_k, k = 3, 6, 12$ ) and we examine 3 cases of missingness mechanism: CRD, RD and ID (see section 1.3).

According to the definitions of the CRD, RD and ID, we delete a random value, a value in the last variable when the first variable has maximal value, and the maximal value in the last variable, respectively. Three methods of imputation based on the derived formulas (3.5), (3.9) (3.11) were used according to the correlation structure.

In both simulation studies these methods of imputation were compared with two well-known methods:

1. Imputation by the formula (3.5) vs imputation by the linear prediction, where the observation at the last time point was modelled using previous time points $X_k = \beta_0 + \beta_1 X_1 + \ldots + \beta_{k-1} X_{k-1}$.

2. Imputation by the formula (3.9) vs imputation using the *LOCF*-method (*Last Observation Carried Forward*)[2].

3. Imputation by formula (3.11) vs imputation using the *LOCF*-method.

## 4.3. Experimental design and calculations

In both simulation studies we generated $3 \times 2 \times 2 \times 3 = 36$ different data sets (for CS and AR correlation structures): $k = 3, 6, 12$ (data from 3-, 6-, 12-dimensional normal distributions), $n = 10, 20$ (small sample sizes), $\rho = 0.5,\ \rho = 0.7$, and for 3 missingness mechanisms (CRD, RD and ID). In the case of banded Toeplitz correlation structure $3 \times 2 \times 3 = 18$ different data sets were under consideration (for $\rho = 0.5$ only). For each combination formed by the above simulation factors, 1000 runs were performed.

To analyze the obtained results, the average bias was calculated as average difference between observed values and imputed values. Results were presented in units of standard deviation of given marginals.

Let $w_k$ be the observed value for the subject which drops out at the time point $k$ (i. e. $w_k = x_k$ or $w_k = z_k$ according to the simulation study), $w_{kv}$ be the corresponding imputed value using (3.5), (3.9) or (3.11) (i.e. $w_{kv} = \hat{y}_k^{CS}$, $w_{kv} = \hat{y}_k^{AR}$ or $w_{kv} = \hat{y}_k^{BT}$, respectively) and let $w_{kp}$ be the corresponding imputed value using classical well-known rules (linear prediction or *LOCF*). The standardized biases $SB_1$ (for (3.5), (3.9) or (3.11)) and $SB_2$ (for linear prediction or *LOCF* rules) are calculated as follows

$$SB_1 = \frac{w_k - w_{kv}}{S_k}, \quad SB_2 = \frac{w_k - w_{kp}}{S_k},$$

where $S_k$ is the standard deviation of observed values at last time point $k$.

Mean biases $B_1$ and $B_2$ are found by averaging values of standardized biases $SB_1$ and $SB_2$ over 1000 runs.

[2]When the main interest is the outcome at the endpoint of the study (for example in clinical trials), the *LOCF* is the most frequently used approach for dealing with missing values in continuous variables (see also pages 26–27).

The average standard deviations of biases were calculated over 1000 runs and denoted by $S_1$ and $S_2$, respectively.

## 4.4. Results

To estimate the effectiveness of new imputation rules, we compare the mean biases $B_1$ versus $B_2$, and the standard deviations $S_1$ versus $S_2$.

In the case of compound symmetry, the results show the advantage of (3.5) compared to the linear regression (see Table 1).

**Table 1:** Results of two simulation studyies in the case of compound symmetry

| I | CRD | RD | ID |
|---|---|---|---|
| $B_1$ | 0.0247 | 0.0414 | 1.5109 |
| $B_2$ | 0.0485 | 0.0961 | 1.7835 |
| $S_1$ | 0.6895 | 0.7897 | 1.0236 |
| $S_2$ | 1.0945 | 1.5627 | 2.0957 |
| II | CRD | RD | ID |
| $B_1$ | 0.0245 | 0.1173 | 1.8994 |
| $B_2$ | 0.0870 | 0.3035 | 2.0685 |
| $S_1$ | 0.6918 | 0.8216 | 1.0243 |
| $S_2$ | 1.4107 | 2.0112 | 2.0647 |

We can see that in all cases the new formula (3.5) gives better results: it has smaller bias and is more stable compared with the imputation rule based on the linear regression ($B_1 < B_2,\ S_1 < S_2$). Of course, in the case of ID both methods do not perform well; nevertheless, the new one gives smaller bias. In the case of informative dropouts, the bias is greater than in the case of random or completely random dropouts, as is usual.

In Table 2 we present the results of the simulation studies in the case of the first order autoregressive correlation structure.

**Table 2:** Results of two simulation studies in the case of autoregressive dependencies

| I | CRD | RD | ID |
|---|---|---|---|
| $B_1$ | 0.0199 | 0.0629 | 2.1261 |
| $B_2$ | 0.0213 | 0.1787 | 1.0929 |
| $S_1$ | 0.8296 | 0.8528 | 0.9599 |
| $S_2$ | 0.8776 | 0.8959 | 1.4408 |
| II | CRD | RD | ID |
| $B_1$ | 0.0426 | 0.0597 | 2.6449 |
| $B_2$ | 0.0870 | 0.3035 | 2.0685 |
| $S_1$ | 0.6918 | 0.8216 | 1.0243 |
| $S_2$ | 0.8776 | 0.8959 | 1.4408 |

Again, the new method (3.9) is more stable ($S_1 < S_2$ in all cases). In the cases CRD and RD, the new method gives smaller biases compared with the *LOCF*-method ($B_1 < B_2$ in the first two columns).
Formula (3.9) did not work well when we had informative dropouts. In this case the bias was larger compared with the *LOCF*-method, but the standard deviations were smaller ($B_1 > B_2, S_1 < S_2$ in the last column).

In Table 3 we see the results of two simulation studies in the case of the 1-banded Toeplitz correlation structure.

**Table 3:** Results of two simulation studies in the case of 1-banded Toeplitz correlation structure

| I | CRD | RD | ID |
|---|---|---|---|
| $B_1$ | 0.0045 | 0.0015 | 2.0120 |
| $B_2$ | 0.0069 | 0.2856 | 1.5027 |
| $S_1$ | 0.7879 | 0.7742 | 1.0217 |
| $S_2$ | 1.0033 | 0.9620 | 1.6687 |
| II | CRD | RD | ID |
| $B_1$ | 0.0722 | 0.0907 | 2.2320 |
| $B_2$ | 0.0143 | 0.3193 | 1.9212 |
| $S_1$ | 0.8430 | 0.8437 | 1.0915 |
| $S_2$ | 1.0212 | 1.0562 | 1.5143 |

From Table 3 it follows that in all cases imputation by formula (3.11) gives more stable solution than with LOCF-method ($S_1 < S_2$).
In the first simulation study in the cases CRD and RD the imputation formula (3.11) gives smaller bias ($B_1$) compared with bias ($B_2$) by the *LOCF*-method. In the case of ID model both imputation methods did not perform well, as expected.
The results of the second simulation study demonstrated that the formula (3.11) is sensitive to the deviations from the normal distribution; in the case of skewed distribution the estimated value was biased.

## 4.5. Analysis of dependence of experimental design

Additionally to the primary comparative analysis of mean biases the dependence of biases of the experimental design was examined. Linear regression models were fitted for biases with design parameters as independent variables.

Thus, we were interested how the changes of sample size, number of the time points or value of correlations affect the mean bias. Quite obvious result is that when the correlation coefficient increases then the mean bias decreases (negative relationship). In the case of normal distribution there were no additional dependencies, but in case of skewed distribution, the bias additionally depends on the sample size (positive relationship).

In the case of CS scheme, from the analysis of dependence of design, we got that the bias $B_1$ is smaller than $B_2$ when we had more time points (sample size $n$ is larger), and the standard deviation $S_1$ is smaller than $S_2$ when the correlation increases.

In the case of AR dependence of design analysis gave us, that if we look only at random and completely random dropouts, we can see some positive dependencies: the bias $B_1$ depends only on the missingness type, but $B_2$ (the rule *LOCF*) depends on the number of the time points as well.

In the case BT, the results of dependence analysis from the first simulation study (normal distribution) demonstrated that the bias $B_1$ decreases if the sample size $n$ increases. The bias $B_2$ did not depend on sample size $n$, either on number of measurements $k$.
Dependence analysis from the second simulation studies (skewed distribution) did not show any dependencies between results and experimental design.

## 4.6. Conclusions

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

- Bias is smaller in the case of CRD and RD missingness (smaller than 10%).
- Standard deviations are more stable.
- The formula (3.5) could be used for small data sets with several repeated measurements ($k > n$), when linear prediction does not work.
- The formulas (3.9) and (3.11) contains more information about data than the LOCF-method.
- The formula (3.11) is sensitive to the distribution, it is not good to use it for skewed marginals.

It is clear that in the case of informative dropouts we do not get good results because the dropout process is not random, and without additional information we cannot expect good results.

Thus, the new approach has essential advantages and therefore could have widely implemented in to practice.
The following advantages can be pointed out.

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

2. The simplicity of formulas (3.5) and (3.9) for calculation.

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

Certainly the Gaussian copula is not the only possibility to use in this approach. Nevertheless, since multivariate normal distribution and linear correlation coefficients form the basis for most models in data analysis, Gaussian copula is a natural starting point in this kind of research.

The copula approach is also perspective in case when we can not derive simple formulas. Copulas provide a natural approach to handle dependencies between repeated measurements. They are not difficult to apply and are reliable in many situations where the correlation structure is known.

# Bibliography


[1] Aas, K. (2005). Modelling the dependence structure of financial assets: A survey of four copulas. *Norwegian Computing Center. Applied Research and Development*, Oslo (available from www.nr.no/files/samba/bff/SAMBA2204b.pdf : May, 2006).

[2] Bouye,E., Durrleman,V., Nikeghbali, A., Riboulet, G., Roncalli, T. (2000). Copulas for Finance. A Reading Guide and Some Applications, *Working Paper*, Groupe de Recherche Operationnelle, Credit Lyonnais, Paris (available from www.gloriamundi.org/picsresources/bdnrr.pdf : Nov, 2005).

[3] Chen, J., Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16 (2), 113–131.

[4] 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.

[5] Daniels, M.J., Hogan, J.W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics*, 56, 1241–1248.

[6] Dawson, K.S., Gennings, C., Carter, W.H. (1997). Two graphical techniques useful in detecting correlation structure in repeated measures data. *The American Statistician*, 51, 3, 275–283.

[7] Demarta, S., McNeil, A.J. (2004). The $t$ Copula and Related Copulas. *International Statistical Review*, (to appear 2005) (available from www.math.ethz.ch/~mcneil/ftp/tCopula.pdf : Nov, 2005).

[8] De Michele, C., Salvadori, G. (2006). Copulas and extreme storm structure. *Geophysical Research Abstracts*, 8, 06124.

[9] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from oncomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B, 39, 1–38.

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

[11] Diggle, P.J., Liang, K.-Y., Zeger, S.L. (1994). Analysis of longitudinal data. New York: Clarendon Press, Oxford.

[12] Dupuis, D. (2006). Using copulas in hydrology: benefits, cautions, and issues. *Journal of Hydrologic Engineering*, 11, 4.

[13] Durrant, G.B. (2005). Imputation methods for handling item-nonresponse in the social sciences: a methodological review. *NCRM Working Paper Series*, 1–42
(available from www.ncrm.ac.uk/publications/documents/Methods ReviewPaperNCRM-002_000.pdf : Nov, 2005).

[14] Embrechts P., McNeil A.J. and Straumann D. (1999). Correlation and dependence in risk management: properties and pitfalls. Preprint ETH Zurich (available from www.math.ethz.ch/~embrecht : Nov, 2005).

[15] Embrechts, P., Lindskog, F., McNeil, A. (2001). Modelling dependence with copulas and applications to risk management. Preprint ETH Zurich (available from www.math.ethz.ch/finance : Nov, 2005).

[16] Fisher, N. I. (1997). Copulas. In: *Encyclopedia of Statistical Sciences*, Update Vol. 1, 159-163. New York: Wiley.

[17] Fitzmaurice, G.M. (2003). Methods for handling dropouts in longitudinal clinical trials. *Statistica Neerlandica*, 57, 1, 75–99.

[18] Frees, E.W., Valdez, E.A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2 (1), 1–25.

[19] Frees, E.W., Wang, P. (2005). Credibility using copulas. *North American Actuarial Journal*, 9 (2), 31–48.

[20] Fuller, W. A., Kim, J.K. (2005). Hot deck imputation for the response model. *Statistics Canada*, 31, 2, 139–149.

[21] Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, 74, 549–555.

[22] Genest, C., MacKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *JASA*, 40 (4), 280–283.

[23] Genest, C., Rivest, L.P. (1993). Statistical inference procedure for bivariate Archimedean copulas. *JASA*, 88 (423), 1034–1043.

[24] Gomez, E.V., Schaalje, G.B., Fellingham, G.W. (2005). Perfomance of the Kenward-Roger Method when the covariance structure is selected using AIC and BIC. *Communications in Statistics – Simulation and Computation*, 34, 377–392.

[25] Hedeker, D., Gibbons, R.D. (2006). Longitudinal data analysis. New York: Wiley. (available from www.uic.edu/classes/bstt/bstt513/ : Nov, 2005)

[26] Heitjan, D.F., Rubin, D.B. (1991). Ignorability and coarse data. *The Annals of Statistics*, 19, 4, 2244–2253.

[27] Hogan, J.W., Roy, J., Korkontzelou, C. (2004). Tutorial in biostatistics. Handling drop-out in longitudinal studies. *Statistics in Medicine*, 23, 1455–1497.

[28] Horton, N.J., Lipsitz, S.R. (2001). Multiple imputation in practice: comparison packages for regression models with missing variables. The American Statistician, 55 (3), 244–254.

[29] Joe, H. (1997). Multivariate Models and Dependence Concepts. London: Chapman and Hall.

[30] Kendall, M., Stuart, A. (1976). Design and analysis, and time-series. *The advanced theory of statistics*, **3**, 646, Moscow: Nauka (Russian).

[31] Kenward, M.G., Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13, 3, 236–247.

[32] King, G., Honaker, J., Joseph, A., Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review*, 95, 1, 49–69.

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

[34] Kruskal, W.H. (1958). Ordinal Measures of Association. *JASA*, 53, 284, 814–861.

[35] Käärik, E., Sell, A. (2004). Estimating $ED_{50}$ using the up-and-down method. In: *Proceedings in Computational Statistics. Compstat'04*, Ed. J. Antoch. Physica-Verlag, Springer, 1279–1286.

[36] Käärik, E. (2005). Handling dropouts by copulas. In: *WSEAS Transactions on Biology and Biomedicine*, Ed. N. Mastorakis. Vol **1**, 2, 93–97.

[37] Käärik, E. (2006a). Imputation algorithm using copulas. *Advances in Methodology and Statistics*, Ed. A. Ferligoj. Vol **3** (1), 109–120.

[38] Käärik, E. (2006b). Imputation by conditional distribution using Gaussian copula. In: *Proceedings in Computational Statistics. Compstat'06*, Ed. A. Rizzi and M. Vichi. Physica-Verlag, Springer, 1447–1454.

[39] Laaksonen, S. (2002). Traditional and new techniques for imputation. *Statistics in Transition*, **5**, 6, 1013–1035.

[40] Lambert, P., Vandenhende, F. (2002). 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, 3197–3217.

[41] Lindsay, J.K. (2000a). A family of models for uniform and serial dependence in repeated measurements studies. *Applied Statistics*, 49, 3, 343–357.

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

[43] Lindsey, J.K, Lindsey, P.J. (2002). Multivariate distributions with correlation matrices for nonlinear repeated measurements (available from www.luc.ac.be/˜jlindsey : Nov, 2005).

[44] Lindskog, F. (2000). Modeling dependence with copulas and applications to risk management. Master Thesis, Department of Mathematics, ETH Zürich.

[45] Lindskog, F., McNeil, A., Schmock, U. (2001). Kendall's tau for elliptical distributions.
(available from www.risklab.ch/ftp/papers/KendallsTau.pdf : May, 2006).

[46] Littell, R.C., Pendergast, J., Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. Tutorial in biostatistics. *Statistics in Medicine*, 19, 1793–1819.

[47] Little, R.J.A. (1992). Regression with missing $X$'s: a review, *JASA*, 87, 420, 1227–1237.

[48] Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. JASA, 88 (421), 125–134.

[49] Little, R.J.A. (1995). Modeling the dropout mechanism in repeated-measures studies. *JASA*, 90, 431, 1112–1121.

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

[51] Lundström, S., Särndla C.– E. (2001). Estimation in the presence of nonresponse and frame imperfections. Statistics Sweden, SCB, Örebro.

[52] Meng, X.– L. (2000). Missing data: dial M for ??? *JASA*, 95, 452, 1325–1330.

[53] Mikosch, T. (2005). Copulas: Tales and facts. *4th International Conference on Extreme Value Analysis*, Gothenburg, Sweden, 15–19 August, 2005 (available from www.math.ku.dk/˜mikosch/maphysto_extremes_2005/s.pdf : May, 2006)

[54] Molenberghs, G., Kenward, M.G., Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84 (1), 33–44.

[55] Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C. (2004). Analyzing incomplete longitudinal clinical trial data. Biostatistics, 5, 3, 445–464.

[56] Nelsen R.B. (1999). An introduction to copulas. *Lecture Notes in Statistics*, 139, New York: Springer Verlag.

[57] Nelsen R.B. (2003). Properties and applications of copulas: A brief survey, *Proceedings of the First Brazilian Conference on Statistical Modelling in Insurance and Finance* (J. Dhaene, N. Kolev, and P. Morettin, eds.), Institute of Mathematics and Statistics, University of São Paulo, 10–28.

[58] Nelsen R.B. (2005). Dependence modeling with Archimedean copulas, *Proceedings of the Second Brazilian Conference on Statistical Modelling in Insurance and Finance*, (N. Kolev and P. Morettin, editors), Institute of Mathematics and Statistics, University of São Paulo, 45–54.

[59] Rao, C. R. (1965). Linear statistical inference and its applications. New York: Wiley.

[60] Raveh, A. (1985). On the use of the inverse of the correlation matrix in multivariate data analysis. *The American Statistician*, **39**, 1, 39–42.

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

[62] Robins, J.M., Rotnitzky, A., Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. JASA, 90, 106–121.

[63] Rotnitzky, A., Robins, J.M., Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. JASA, 93 (444), 1321–1339.

[64] Rotnitzky, A., Scharfstein, D.O., Su, T.-L., Robins, J.M., (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*, 57, 103–113.

[65] Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model, *Biometrics*, 59, 829–836.

[66] Roy, J., Lin, X. (2005). Missing covariates in longitudinal data with informative dropouts: bias analysis and inference. *Biometrics*, 61, 837–846.

[67] Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, **63**, 581–592.

[68] Rubin, D.B. (1996). Multiple imputation after 18+ years. *JASA*, 91 (434), 473–489.

[69] Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapmann and Hall.

[70] Schafer, J.L., Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7, 2, 147–177.

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

[72] Zhang, L, Singh, V.P. (2006). Bivariate flood frequency analysis using the copula model. *Journal of Hydrologic Engineering*, 11, 2, 150–164.

[73] Tiit, E.-M., Käärik, E. (1996). Generation and investigation of multivariate distributions having fixed discrete marginals. *Proceedings*

*in Computational Statistics*, COMPSTAT, 12th Symposium held in Barcelona, Spain. (Ed. A. Prat). Heidelberg: Physica Varlag, 471–476.

[74] Tipa, M.A., Murphy, S.A., McLaughlin, D. (1996). Ignorable Dropout in Longitudinal Studies. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 687–692 (available from www.amstat.org/sections/srms/Proceedings/ : Nov, 2005).

[75] Troxel, A.B., Ma, G., Heitjan, D.F. (2004). An index of local sensitivity to nonignorability. *Statistica Sinica*, 14, 1221–1237.

[76] Vandenhende, F., Lambert, P. (2000). Modelling repeated ordered categorical data using copulas. *Discussion Paper* 00-25, Institut de Statistique, Université Catholique de Louvain, Belgium (available from www.stat.ucl.ac.be/ISpub/ISdp.html : Nov, 2005).

[77] 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.

[78] Vandenhende, F., Lambert, P. (2005). Local dependence estimation using semiparametric Archimedean copulas. *The Canadian Journal of Statistics*, 33, 3, 377–388.

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

[80] Wolfinger, R.D. (1996). Heterogeneous Variance: Covariance structures for repeated Measures. *Journal of Agricultural, Biological and Enviromental Statistics*, 1, 2, 205–230.

# Summary in Estonian

## Kordusmõõtmiste andmelünkade käsitlemine koopulate abil

### Kokkuvõte

Kordusmõõtmistega on tegemist juhul, kui sama objekt/subjekt on mõõdetud korduvalt ajas (ruumis). Kordusmõõtmistele on iseloomulik asjaolu, et samal objektil/subjektil teostatud mõõtmised on omavahel seotud ja seda seost ei tohi ignoreerida. Kordusmõõtmiste analüüsi kasutatakse paljudes teadusvaldkondades – biostatistikas, biomeditsiinis, sotsioloogias jm. ning enamasti on probleemiks, et ühel või teisel põhjusel pole võimalik koguda täielikke andmeid.

Puuduvate andmete ehk andmelünkade probleemiga on tegeletud juba pikka aega, süstemaatilise teooria rajajateks võib pidada D.B. Rubinit ja R.J.A. Little'i, kes esitasid ka puudumiste tüpoloogia (Rubin, 1976; Little ja Rubin, 1987). Puuduvate väärtustega andmete analüüsimiseks on välja töötatud terve rida meetodeid, kuid samas pole olemas ühtegi, mida võiks pidada universaalseks ja parimaks.

Lünkliku andmestiku käsitlemisel on põhimõtteliselt kaks ülesannet:
(a) *hindamisülesanne*, kus eesmärgiks on saada lünkliku andmestiku põhjal mudeli parameetritele hinnangud, mis on võimalikult lähedased hinnangutele, mida olnuks võimalik saada siis, kui need andmed oleksid olemas;
(b) *imputeerimisülesanne*, kus eesmärgiks on puuduva väärtuse võimalikult täpne prognoosimine.

Antud töös on vaatluse all teine ülesanne, st lünkade täitmine ehk impu-

teerimine, mis on eriti oluline praktilistes ülesannetes väikeste valimite korral. Prognoosiülesande lahendamiseks võib kasutada näiteks kas ainult vaadeldava tunnuse jaotust või kasutada tunnuse tinglikku keskväärtust, kui teiste tunnuste väärtused on teada. Viimane oleks põhimõtteliselt parim lahendus, mida aga tegelikkuses ei rakendata, sest tihti pole teada tunnuste ühisjaotust ja seega ei saa leida ka tinglikku jaotust. Võimalikuks lahenduseks sel juhul oleks leida tee ühisjaotuse lähendamiseks, seejärel leida puuduva väärtuse tinglik jaotus ja arvutada lähendatud tingliku jaotuse põhjal tinglik keskväärtus (või ka mingi muu tingliku jaotuse karakteristik). Puuduva väärtuse tingliku jaotuse kasutamise eesmärk on maksimaalselt ära kasutada kogu olemasolev informatsioon andmetes:
(1) kasutada mõõtmiste ajalugu (moodustab tingimuse);
(2) kasutada vaatlustulemuste marginaaljaotusi – tingimatuid jaotusi, mida oluliselt täpsustatakse tingimuse abil;
(3) kasutada seostestruktuuri mõõtmiste vahel.

Probleem on selles, et tuntud mitmemõõtmelised jaotused ei pruugi sobida ühisjaotuse kirjeldamiseks ja seepärast on võetud kasutusele koopulad. Uudseks aspektiks antud töös ongi tingliku jaotuse leidmine koopulate abil.

Koopula on funktsioon, mis ühendab marginaaljaotused ühisjaotuseks. Kasutades koopulat, saame eraldi hinnata marginaaljaotused ja seejärel arvestades seoste struktuuri määrata ühisjaotuse. Põhjalik teoreetiline ülevaade koopulatest on antud H. Joe ja R. B. Nelseni monograafiates (Joe, 1997; Nelsen, 1999).
Koopulad on algselt leidnud rakendust eeskätt kindlustus- ja finantsmatemaatikas, viimasel ajal ka biostatistikas (meteoroloogias), biomeditsiinis ja keskkonnastatistikas.

Kordusmõõtmiste analüüsis on koopulaid kasutanud vaid vähesed autorid. Näiteks Lindsey ja Lindsey (2002) kirjeldavad Gaussi koopulat kordusmõõtmistega andmete korral, kuid nad ei käsitle lünkadega andmestikku. Lambert ja Vandenhende (2002), Vandenhende ja Lambert (2002) on rakendanud koopulate lähenemist mudelite leidmisel korduvmõõtmistega lünklike andmestike korral, nad testisid erinevaid marginaaljaotusi (Cauchy, gamma, log-normaalne) ja kasutasid Gaussi koopulat ning Franki koopulat, kirjeldamaks seost uuritava tunnuse ja lünkade vahel.

Koopulate kasutamisel on terve rida eeliseid klassikaliste meetodite ees. Klassikalised mudelid baseeruvad mitmemõõtmelisel normaaljaotusel või mõnel teisel mitmemõõtmelisel jaotusel, mis seavad teatud nõudmised ka marginaaljaotuste kohta. Koopulamudel on paindlikum, ta lubab kombineerida erinevaid marginaaljaotusi ja rakendada nende sidumiseks erinevaid seostestruktuure. Saadud koopulamudeli sobivuse kontrollimiseks võib kasutada klassikalisi sobivuse teste ($\chi^2$, AIC, BC ja nende modifikatsioone). Võrreldes omavahel koopulaid, mis võimaldavad siduda suvalise arvu marginaaljaotusi mitmemõõtmeliseks jaotuseks, on kõige käepärasem Gaussi koopula, sest

1. Gaussi koopula puhul on avarad võimalused seoste struktuuri kirjeldamiseks: on võimalik hinnata ja arvestada kõiki $k$-mõõtmelise tunnusvektori $\frac{k(k-1)}{2}$ paarisseose kordajaid;

2. On võimalik leida seoste struktuuri kirjeldamiseks mudelid, mis sõltuvad vähesest arvust parameetritest;

3. Puuduva väärtuse tinglik jaotus on igasuguse seoste struktuuri korral leitav;

4. Lihtsate seosestruktuuride korral on võimalik tingliku keskväärtuse (st asendusväärtuse) ja standardhälbe jaoks tuletada lihtsad valemid.

Antud töös kasutataksegi Gaussi koopulat.

Töö esimeses peatükis antakse ülevaade puuduvate andmete tüpoloogiast, esitatakse lühiülevaade tuntumatest andmelünkade käsitlemismeetoditest ja formuleeritakse imputeerimisülesanne.

Teine peatükk on pühendatud koopulate teooria põhimõistetele, põhjalikumalt on käsitletud Gaussi koopulat.

Kolmandas peatükis on esitatud enamus originaaltulemusi. Kõigepealt on tutvustatud kordusmõõtmiste seostestruktuure, toodud korrelatsioonimaatriksi lahutus ja sellele vastavalt tuletatud üldine tinglikul keskväärtusel baseeruv imputeerimisvalem (3.3) (Lause 3.1) ja selle üldistus (Järeldus 3.1). Edasi on vaadeldud erijuhtudena lihtsamaid korrelatsioonistruktuure (konstantsed korrelatsioonid, esimest järku autoregressiivne sõltuvus ja

tõkestatud Toeplitzi struktuur) ja tuletatatud nende korral lihtsad imputeerimise valemid (3.5), (3.8) (3.9), (3.11) (vastavalt Lause 3.2, 3.3, 3.4 ning Järeldus 3.2–3.5). Näitena on toodud esimest järku autoregressiivse sõltuvusstruktuuriga reaalse andmestiku korral andmelünga asendamine valemi (3.9) abil. Tulemus näitab, et metoodika on igati sobiv praktiliseks kasutamiseks.

Neljandas peatükis on esitatud simulatsioonieksperimentide tulemused võrdlemaks erinevaid imputeerimismeetodeid ja selgitamaks välja koopula abil imputeerimise plusse. Simulatsioonid näitavad, et esitatud metoodika on sobiv rakendamiseks andmelünkade täitmiseks väikese valimi korral.

Enamus toodud tulemustest on avaldatud (Käärik, 2005; Käärik, 2006a; Käärik, 2006b) ja ette kantud rahvusvahelistel konverentsidel.

# Curriculum Vitae

**Name:** Ene Käärik
**Citizenship:** Estonian Republik
**Born:** September 14, 1950, Valga, Estonia
**Marital status:** married, 1 adult son
**Address:** J. Liivi 2–516, Institute of Mathematical Statistics, University of Tartu, Estonia
**Contacts:** e-mail: Ene.Kaarik@ut.ee

**Education:**
1968–1973 University of Tartu, graduated with Diploma of mathematician
1992–1994 University of Tartu, post-graduate student, MsC in Mathematics
2001-2006 University of Tartu, PhD student in mathematical statistics

**Professional employment:**
1973–1992 mathemathician at the Department of Sports Medicine UT
1993–1994 researcher at the Institute of Mathematical Statistics UT
Since 1994 lecturer at the Institute of Mathematical Statistics UT

# Curriculum Vitae

**Nimi:** Ene Käärik
**Kodakondsus:** Eesti Vabariik
**Sünniaeg ja -koht:** 14. september 1950, Valga, Eesti
**Perekonnaseis:** abielus, 1 täiskasvanud poeg
**Aadress:** J. Liivi 2–516, Tartu Ülikooli matemaatilise statistika instituut
**Kontakt:** e-mail Ene.Kaarik@ut.ee

**Hariduskäik:**
1968–1973 TRÜ Matemaatikateaduskond, matemaatiku diplom
1992–1994 TÜ matemaatikateaduskond, magistrikraad matemaatika erialal
2001-2006 TÜ matemaatika-informaatikateaduskond, doktoriõpingud matemaatilise statistika erialal

**Erialane teenistuskäik:**
1973–1992 matemaatik TRÜ Spordimeditsiini kateedris
1993–1994 teadur TÜ matemaatilise statistika instituudis
Alates 1994 lektor TÜ matemaatilise statistika instituudis