UNIVERZITA KARLOVA V PRAZE
MATEMATICKO-FYZIKÁLNÍ FAKULTA

RIGORÓZNÍ PRÁCE

Alexander Voronin

Generování náhodného výběru s předepsanými vlastnostmi s aplikacemi v bankovnictví

Katedra pravděpodobnosti a matematické statistiky

Vedoucí rigorózní práce:	RNDr. Petr Franěk, Ph.D.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie
Studijní plán:	Ekonometrie

Poděkování

Jsem velmi vděčný svému vedoucímu RNDr. Petru Fraňkovi, Ph.D. za možnost napsání této rigorózní práci v tomto fascinujícím oboru. Děkuji mu za podporu a zajimavé nápady.

Dále chci taky poděkovat Mgr. Ondřeji Vencálkovi za jeho pomoc a diskuze, a KateB. za čas, který stravila úpravováním mé angličtiny.

Rád bych ještě poděkoval mé rodině, a obzvlašť mému tatínkovi, za jejich podporu, porozumění a lásku v průběhu těchto let.

Čestné prohlášení

Prohlašuji, že jsem svou rigorózní práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.
Tato rigorózní práce je identická s diplomovou práci, kterou jsem obhájil dne 11.02.2008 na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze, až na tiskové chyby a technické připomínky recenzenta.

V Praze dne 20. srpna, 2008 Alexander Voronin

vlastnoruční podpis

Charles University in Prague
Faculty of Mathematics and Physics

Rigorous thesis

Alexander Voronin

Generating of Random Samples with Given Properties and Application to Banking

Department of Probability and Mathematical Statistics

Supervisor:	RNDr. Petr Franěk, Ph.D.
Academic program:	Mathematics
Field of study:	Probability, mathematical statistics and econometrics
Curriculum:	Econometrics

Contents

Abstrakt

Název práce: Generování náhodného výběru s předepsanými vlastnostmi s aplikacemi v bankovnictví
Autor: Alexander Voronin
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí rigorózní práce: RNDr. Petr Franěk, Ph.D., Česká Spořitelna, a.s.
e-mail vedoucího: PFranek@csas.cz

Abstrakt: Tato práce se zabývá hledáním algoritmu pro generování náhodných veličin s předem danými vlastnostmi. Jsou provedeny analýzy srovnání algoritmů a na jejích základě byl vybrán pro nás optimální algoritmus. A protože se zaměříme na generování náhodných veličin defaultů a vysvětlujících proměnných defaultů, tak je důraz při hledání algoritmu kladen na zachování závislostí těchto veličin. Dále hledáme optimální velikost generovaného vzorku při zachování vlastnosti generované veličiny. Na konci této práci aplikujeme zkoumané techniky na reálná data z banky.
Klíčová slova: copula, generování, struktura závislosti, defaulty

Abstract

Title: Generating of Random Samples with Given Properties and Application to Banking
Author: Alexander Voronin
Department: Department of Probability and Mathematical Statistics
Supervisor: RNDr. Petr Franěk, Ph.D., Česká Spořitelna, Inc.
Supervisor's e-mail address: PFranek@csas.cz

Abstract: The work concerns the searching for the algorithm for generating of the random variables with the given properties. There are made analyses of comparisons of the algorithms, and the optimal algorithm was chosen based on it. Since we focus on generating of random variables of defaults and explanatory variables of defaults, it concentrates mainly on the conservation of the dependence of these variables. Further we are looking for the optimal sample size of the generated samples under conservation of the required properties. And in the last Chapter we have applied the surveyed techniques to the real data.
Keywords: copulas, generating, dependence structure, defaults

Notation

$\mathbb{X}$	a random vector (in general, of a dimension n) modeling explanatory variable
$\overline{\mathbb{X}}$	the arithmetic mean of $\mathbb{X}$
$\mathbb{E}[\mathbb{X}]$	the mean value of random variable $\mathbb{X}$
$\mathbb{I}$	a closed interval $[0,1]$
$X \sim \mathcal{L}(\mu, \sigma^2)$	the random variable X has the distribution $\mathcal{L}(\mu, \sigma^2)$ with a mean value μ and the variance σ^2
$F^{[-1]}$	a quasi-inverse of the function F
$U(0,1)$	the standard uniform distribution
$\mathcal{E}_n(\mu, \Sigma, g)$	n-variate elliptical distribution with a mean μ, the correlation matrix Σ and the characteristic generator $g(.)$
$\equiv_d$	an equality in distribution
$(X)^t$	the transposed vector or matrix
RD	the real data
GD	the generated data
d.f.	degrees of freedom

Chapter 1

Introduction

There are many methods to describe economics. Some processes can be described exactly: time spent in a front line, chances of winning a lottery, etc. But we need to describe other processes, like granting a credit or development of prices of securities. In real world most of the processes couldn't be described exactly. Using different theories, experience and knowledge we bring in several models. Then there is a decision to make: which of the models should be used and why. For example, simulation is used to make decisions like that.

Every model has input and output data. When we fix input data - our purpose is to reach output data of a chosen model that is close to real value of the observed process. We also need to minimize the amount of input data needed to calibrate the model at the same time. For instance one of the main problems of creating reliable rating models is amount of input data. See more [7].

The final model should be very well checked and tested. First of all observed model should correspond with data from previous period, if nothing noticeable has happened that could influence a model during the observation time. If we don't have enough available data samples we cannot say anything about the model. On the contrary if real data is developing differently than the model has predicted, we have to find reasons that led to it and adjust that model according to it.

One of the mostly used means of verifications is simulation. Both input and output data can be simulated. This type of simulation is often used in different sectors.

In year 2004 a group of scientists used simulation to find out the origin of 35 meters high waves. There are many other reasons why simulation is used more and more often in all different directions of modern science.

We will estimate and improve parameters of a model using simulation. Simulation may also help us to answer question in the credit field: how much data is enough to create a reliable rating instrument?

There are different types of simulation but we are going to work with simulation of a random vector with specified properties. But we can't take simulation data from the tops of our heads. It has to correspond with reality. It means that it should have similar properties as described data in reality. As it was mentioned before we can't include all the properties, so we have to choose the most substantial ones.

In statistics every random value has following properties: specified mean value, given variance and specified correlations with other random values. A generalization of the first two values can be a marginal distribution.

Main purpose of this thesis is solving the following general problem:

generating n-variate random vectors $(X_1, \dots, X_n)$,

where $X_i \sim L_i$ has required marginal distribution function with $cor(X_i, X_j) = \sigma_{i,j}$ given correlation coefficient for $i, j \in \{1, \dots, n\}$.

Then we can add to our basic goal some other conditions, e.g. pre-described default rate dependence on different factors X_i.

After generating random vectors it may be applied to finding an optimal sample size.

There are a lot of different ways of solving this problem. We will use a tool called a *copula*. It is a statistical tool that has been popular for about 20 last years. It has very wide range of properties that allow generating n-variate vectors with a fixed correlation structure very fast and easy. If there are some specific marginal distributions, the certain kind of copulas (Archimedean copulas) may generate even more accurate vectors and sample correlation matrix.

So, we see on different algorithms using copulas, we will discuss which kinds are better or worse for our purposes and how the parameters of copulas might change the generated vectors and why. After analyzing we choose the optimal copula.

Then we estimate from the data we have from one of the Czech banks the empirical distribution functions and the dependence structure. Then we will generate vectors with given empirical distributions and dependence structure (not just correlation matrix) with the optimal copula and check whether they satisfy these conditions.

Chapter 2

Copula

In this chapter we define the copula and state its basic definitions and properties. The main book about copulas is [10]. The further information is also in [15] and [12].

In the first sections we establish a definition of the copula and show the two most important theorems. An explanation of reasons why copulas are so widely used in generating is developed in the second section. After that we get acquainted with different families of copulas and discuss advantages and disadvantages of these families. The discussion about families of copulas can be found in [11], and [4] concerns about elliptical copulas.

2.1 Definitions and Properties

We begin with the definition and some basic properties which can help us to understand the ideas of copula as a useful tool we will work with.

Theorem 2.1.1 *Let H be an univariate distribution function. Then the* quasi-inverse $H^{[-1]}$ *of H is defined as*

$$H^{[-1]}(x) = inf\{a|H(a) \geq x\}.$$

If H is continuous and strictly increasing, then $H^{[-1]} = H^{-1}$, where H^{-1} is an inverse of H.

Lemma 2.1.1 *Let H be an univariate distribution function. Then*

- *If $U \sim \mathcal{U}(0,1)$, then $H^{[-1]}(\mathcal{U}) \sim H$,*
- *If $G \sim H^{[-1]}$, then $H(G) \sim \mathcal{U}(0,1)$.*

Proof See [9].

Suppose we have n-dimensional random vector $\mathbb{X} = (X_1, \dots, X_n)$ with the joint distribution function F and with continuous marginal distributions. Denote the marginal distributions as $F_1, \dots, F_n$. If X_i has a distribution function F_i, then $F_i(X_i)$ has a standard uniform distribution (see Lemma 2.1.1). Then the joint distribution function of a random vector $\mathbb{X}$ could be written as

$$\begin{aligned} F(x_1, \dots, x_n) &= \mathbb{P}(F_1(X_1) \leq F_1(x_1), \dots, F_n(X_n) \leq F_n(x_n)) = \\ &= \mathcal{C}(F(x_1), \dots, F(x_n), \end{aligned}$$

where C is the joint distribution function with standard uniform marginals. The function $\mathcal{C}$ is called the *copula* of the random vector $\mathbb{X}$. This explains the origin of the name copula: the function $\mathcal{C}$ couples together the marginals to the joint distribution function.

After introducing the idea behind it we illustrate the formal definition of copula.

Definition 2.1.1 *An n-dimensional copula is a distribution function of random vector $\mathbb{X}_n$ with uniform marginals.*

If $n = 2$ we say $\mathcal{C}$ is a copula, and when $n > 2$ it is called an n-copula.

Another definition of a copula is:

Definition 2.1.2 *An n-dimensional copula is a function $\mathcal{C} : \mathbb{R}^n \to [0, 1]$ with following properties:*

- $\mathcal{C}(0, \ldots, 0, u_i, 0, \ldots, 0) = 0$, *for every* $u_i \in [0, 1], i \in \{1, \ldots, n\}$,
- $\mathcal{C}(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$, *for every* $u_i \in [0, 1], i \in \{1, \ldots, n\}$,
- *for every* $(u_1, \ldots, u_n), (v_1, \ldots, v_n) \in [0, 1]^n$ *such that* $u_i \leq v_i$ *for every* $i \in \{1, \ldots, n\}$

$$V_{\mathcal{C}} = \sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1+\ldots+i_n} \mathcal{C}(a_{1i_1}, \ldots, a_{ni_n}) \geq 0, \tag{2.1}$$

where $a_{j1} = u_j$ *and* $a_{j2} = b_j$ *for every* $j \in \{1, \ldots, n\}$.

The equivalent expression of the sum (2.1) is a condition $\mathbb{P}(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq F_n \leq b_n) \geq 0$.

An n-copula $\mathcal{C}$ induces a probability measure on $[0, 1]^n$ via $V_c([0, 1] \times [0, 1] \times \ldots \times [0, 1]) = C(u_1, \ldots, u_n)$. The function V_c is called the $\mathcal{C}$-volume of the rectangle $[a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$.

A function that satisfies the first property is called grounded.
The third property - the sum (2.1) - describes the analog of a non-decreasing one-dimensional function. A function with this property is thus called n-increasing. For example, the function $\max(u, v)$ is not 2-increasing function: it is enough to substitute $u_1 = u_2 = 0$ and $v_1 = v_2 = 1$ for V_c. On the other hand, the function $\min(u, v)$ is 2-increasing.

As can be easily seen, a copula in fact is a multivariate distribution function with univariate marginals restricted to the n-cube.

The following theorem, due to A. Sklar, is basic in the theory of copulas and is the foundation of the most applications to the theory of probability and statistics. Sklar's theorem shows the role copula plays in the relationship between multivariate distribution functions and their univariate marginals. It states that $\mathcal{C}$ is a distribution function and even defines its unique property.

Theorem 2.1.2 ***(Sklar's Theorem)*** *Let H is an n-dimensional distribution function with marginals $F_1, \ldots, F_n$. Then there exists an n-copula $\mathcal{C}$ such that for all $\mathbb{X} \in \overline{\mathbb{R}}^n$, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$*

$$H(x_1, \ldots, x_n) = \mathcal{C}(F_1(x_1), \ldots, F_n(x_n)). \tag{2.2}$$

If $F_1, \ldots, F_n$ are all continuous, then $\mathcal{C}$ is unique otherwise, $\mathcal{C}$ is uniquely determined on $RanF_1 \times \ldots \times RanF_n$, where $RanF_i$ is a range of the F_i. Conversely, if $\mathcal{C}$ is an n-copula and $F_1, \ldots, F_n$ are distribution functions, then the function H defined by (2.2) is a joint distribution function with marginals $F_1, \ldots, F_n$.

Proof The proof of Sklar's Theorem is quite easy, but long. For example, it can be found in [10].

If marginals are continuous, then $\mathcal{C}$ is called a unique copula, but if they are not, we say that $\mathcal{C}$ is a possible copula of distribution function F.

Now using the quasi-inverses of distribution functions we can rephrase the Sklar's Theorem as follows.

Corollary 2.1.1 *Let H be a distribution function with marginals $F_1, \ldots, F_n$ and $F_1^{[-1]}, \ldots, F_n^{[-1]}$ be quasi-inverses of $F_1, \ldots, F_n$, respectively. Then there exists an n-copula $\mathcal{C}$ such that for any $\boldsymbol{u} \in [0,1]^n$*

$$\mathcal{C}(u_1, \ldots, u_n) = H(F_1^{[-1]}(u_1), \ldots, F_n^{[-1]}(u_n)) \tag{2.3}$$

When all marginals are continuous, then above corollary gives us a method of constructing copulas from joint distribution function.

The copula from the (2.3) is called the *copula of* $F_1, \ldots, F_n$ and denoted as $\mathcal{C}$ or $\mathcal{C}_{F_1,\ldots,F_n}$ when we need to explicitly emphasize the marginals of copula.

Now we will illustrate this corollary on an example published by [5].

Example 2.1.1 ***(Gumbel's bivariate exponential distribution)***

Let H be the joint distribution function given by

$$H_\theta(x,y) = \begin{cases} 1 - \mathrm{e}^{-x} - \mathrm{e}^{-y} + \mathrm{e}^{-(x+y+\theta xy)}, & x \geq 0,\ y \geq 0\ ; \\ 0, & \textit{otherwise;} \end{cases}$$

where $\theta \in \mathbb{I}$ is a parameter. Then the marginal distribution functions are

$$F(x) = 1 - \mathrm{e}^{-x} \qquad G(y) = 1 - \mathrm{e}^{-y}$$

and the quasi-inverses functions are

$$F^{-1}(u) = -ln(1-u) \qquad G^{-1}(v) = -ln(1-v)$$

for all $(u,v) \in \mathbb{I}$. A copula according to the corollary 2.1.1 is following:

$$\begin{aligned} \mathcal{C}_\theta(u,v) &= H_\theta(F^{-1}(u), G^{-1}(v)) \\ &= u + v - 1 + (1-u)(1-v)\mathrm{e}^{-\theta ln(1-u) ln(1-v)}. \end{aligned}$$

We end this section with some examples of copulas that will be interesting for us. Some of them will be parametrical copulas. We also refer to them as families of copulas. The next chapter will provide more detailed description.

Example 2.1.2 ***(Gaussian copula)*** *The* Gaussian *or* normal *copula is expressed by*

$$\mathcal{C}_\theta^{Ga}(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi(1-\theta^2)} \exp\left\{\frac{-(u^2 - 2\theta uv + v^2)}{2(1-\theta^2)}\right\} \mathrm{d}u\,\mathrm{d}v, \tag{2.4}$$

where $\theta \in (-1,1)$ and Φ is the univariate standard normal distribution function. The Gaussian copula has a standard normal marginal distribution functions with a correlation coefficient θ. The normal copula can be constructed in any dimensions because of property of multivariate normal distribution and expression (2.4).

This section is finished with the theorem about partial derivation of copula.

Theorem 2.1.3 *Let $\mathcal{C}$ be an n-copula. Then for any $i \in \{1,\ldots,n\}$ and any $(u_1,\ldots,u_{i-1},u_{i+1},\ldots,u_n) \in \mathbb{I}$ the partial derivation $\partial\mathcal{C}/\partial u_i$ exists for almost all u_i, and for such $\boldsymbol{u}$,*

$$0 \leq \frac{\partial}{\partial u_i}\mathcal{C}(\boldsymbol{u}) \leq 1. \tag{2.5}$$

Furthermore, the functions $u_j \to \partial C(\boldsymbol{u})/\partial u_i, j \neq i$ are defined and nondecreasing almost everywhere on $\mathbb{I}$.

The word "almost" is referred to the Lebesgue measure. This result will be needed in section 3.1.

2.2 Dependence

We have mentioned basic properties of copulas. But today copulas are so popular because of their dependence properties.

Let's state lemma about Fréchet-Hoeffding bounds.

Lemma 2.2.1 *Let $\mathcal{C}$ be a copula. Then for every $\boldsymbol{u} \in [0,1]^n$,*

$$\begin{aligned} \mathcal{C}_l(\boldsymbol{u}) &= \max(u_1 + \ldots + u_n + 1 - n, 0) \leq \mathcal{C}(\boldsymbol{u}) \\ &\leq \min(u_1,\ldots,u_n) = \mathcal{C}_u(\boldsymbol{u}) \end{aligned}$$

$\mathcal{C}_l$ is called the Fréchet lower bound and $\mathcal{C}_u$ is called the Fréchet upper bound. The bounds $\mathcal{C}_l$ and $\mathcal{C}_u$ are copulas themselves for $n = 2$, whereas for $n > 2$ only $\mathcal{C}_u$ is.

The $\mathcal{C}_l$ copula is a distribution function of $(U, 1-U)^t$ and $\mathcal{C}_u$ of $(U,U)^t$, where $U \sim \mathcal{U}(0,1)$.

As it is shown on the Picture 2.1, the distribution $(U, 1-U)^t$ lies on the diagonal between points $(1,0)$ and $(0,1)$, whereas copula $\mathcal{C}_u$ has its mass on the diagonal between $(0,0)$ and $(1,1)$. Then it is called that $\mathcal{C}_l$ and $\mathcal{C}_u$ have *a perfect positive* and *a perfect negative dependence* respectively.

For $n \geq 2$ $\mathcal{C}_u$ is called Fréchet upper bound copula while for $n = 2$ $\mathcal{C}_l$ is called Fréchet lower bound copula. A third important copula is the *product copula*

$$C_\pi(u_1,\ldots,u_n) = \prod_{i=1}^{n} u_i. \tag{2.6}$$

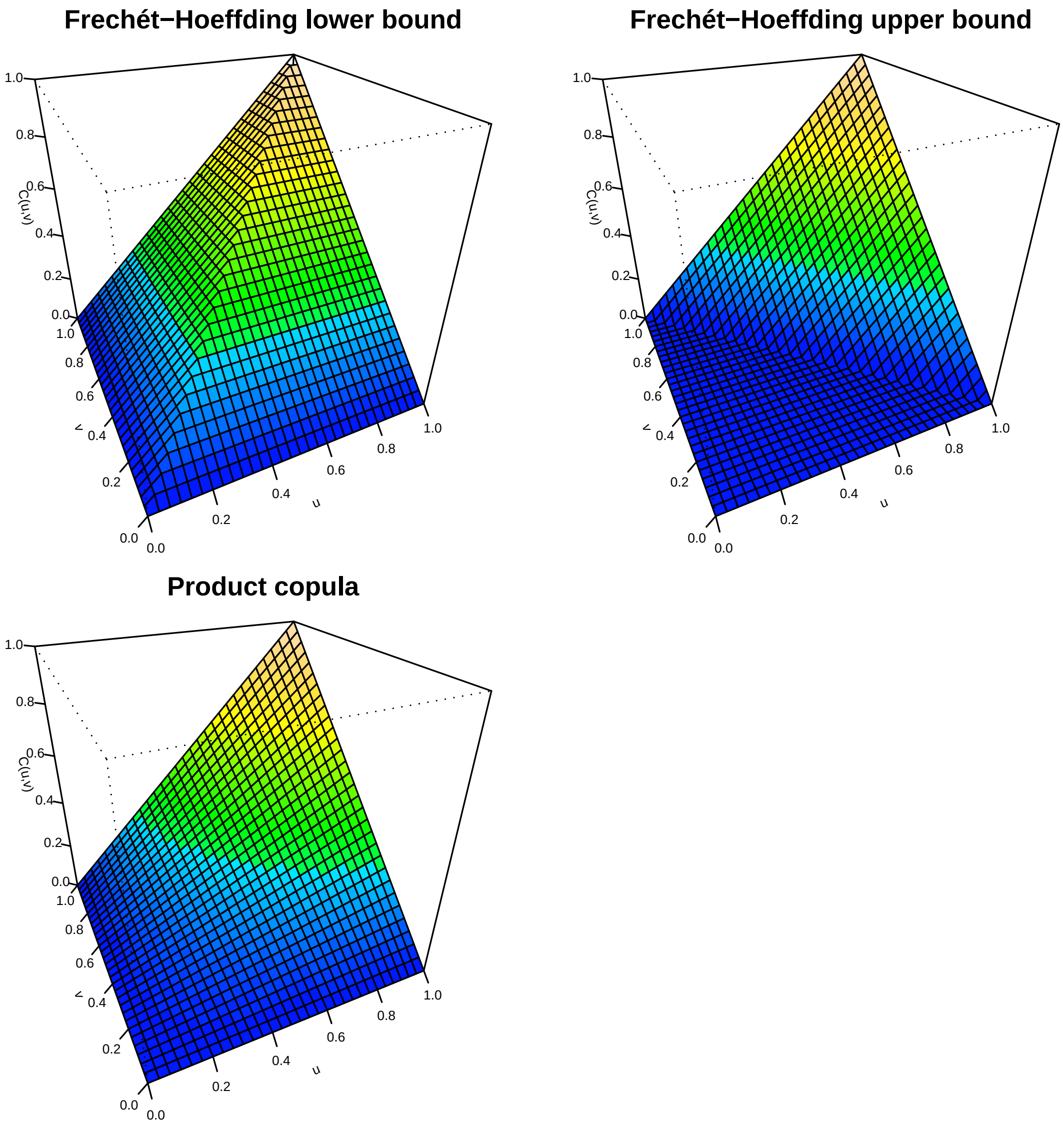

Figure 2.1: Graphs of copulas Fréchet lower $\mathcal{C}_l$ and upper $\mathcal{C}_u$ bound copulas and Product copula Π^n

Random variables generated by this copula are independent, and therefore this copula is known as *independence copula* as well and is referred to as Π^n. The product copula is the third copula on the Picture 2.1.

The following theorem is one of the most important properties of theory of copulas. The invariance property is widely used in generating, modeling and especially in study of dependence. It's probably the reason of popularity of copulas in these research areas.

Theorem 2.2.1 ***(Invariance property)*** *Let $X_1, \dots, X_n$ be continuous random variables with n-copula $\mathcal{C}_{X_1,\dots,X_n}$. Assume that $\alpha_1, \dots, \alpha_n$ are strictly increasing function on $R \to R$. Then*

$$\mathcal{C}_{X_1,\dots,X_n} = \mathcal{C}_{\alpha_1(X_1),\dots,\alpha_n(X_n)}. \tag{2.7}$$

Proof Let F_i an G_i, denote the distribution functions of X_i and $\alpha_i(X_i)$, $i \in \{1, \dots, n\}$, respectively. Since α_i are strictly increasing,

$$G_i(x) = \mathbb{P}[\alpha_i(X_i) \leq x] = \mathbb{P}[X_i \leq {\alpha_i}^{-1}(x)] = F_i({\alpha_i}^{-1}(x)),$$

for all $i \in \{1, \dots, n\}$. Therefore, for any $x_i \in \mathbb{R}^n$, $i \in \mathbb{R}^n$, $i \in \{1, \dots, n\}$, we get

$$\begin{aligned}
\mathcal{C}_{\alpha_1(X_1),\dots,\alpha_n(X_n)} &= \mathcal{C}_{\alpha_1(X_1),\dots,\alpha_n(X_n)}(G_1, \dots, G_n) = \\
&= \mathbb{P}[\alpha_1(X_1) \leq x_1, \dots, \alpha_n(X_n) \leq x_n] = \\
&= \mathbb{P}[X_1 \leq \alpha_1^{-1}(x_1), \dots, X_n \leq \alpha_n^{-1}(x_n)] = \\
&= \mathcal{C}_{X_1,\dots,X_n}(F_1({\alpha_1}^{-1}(x_1)), \dots, F_n({\alpha_n}^{-1}(x_n))) = \\
&= \mathcal{C}_{X_1,\dots,X_n}(G_1(x_1), \dots, G_n(x_n)).
\end{aligned}$$

The random variables X_i are continuous on $\mathbb{I}$, accordingly, $\mathcal{C}_{X_1,\dots,X_n} = \mathcal{C}_{\alpha_1(X_1),\dots,\alpha_n(X_n)}$ on $\mathbb{I}^n$. ■

The theorem asserts that $\mathcal{C}_{XY}$ is invariant under strictly increasing transformations of X and Y.

There is a generalized version of this property where parametrical functions are not just strictly increasing but decreasing, too. An overview of that topic could be found in [15] or [10].

Sklar's theorem and invariance property (Theorems 2.1.2 and 2.2.1) may be summarized by saying that copulas have an ability to keep their dependence structure independently on transformations of their marginals. In other words, copulas are dependent on dependence structure of its distribution only.

For example, assume 2-dimensional continuous random variable $\mathbb{X}$ with marginals F_1 and F_2. Furthermore, suppose that G_1 and G_2 are univariate distributions, and $G_1^{[-1]}$ and $G_2^{[-1]}$ are their quasi-inverses. As we know from Lemma 2.1.1, the $F_1(X_1)$ is an $(0, 1)$-uniform distributed. Then $G_1^{[-1]}(F_1(X_1))$ is a random variable with distribution G_1. Therefore $(G_1^{[-1]} \circ F_1)$ is a strictly increasing function - it's α_1 from the labels of the theorem of invariance property above, and vectors $((G_1^{[-1]} \circ F_1), (G_1^{[-1]} \circ F_1))$ and (F_1, F_2) have the same copula.

The Invariance property is a very useful property and one of the most important theoretical results in this work. It helps us in the next section which is about generating.

Example 2.2.1 ***(Frank family)*** *Further interesting example of one-parametrical copula is Frank family of copulas. It is given by*

$$\mathcal{C}_\theta(u, v) = -\frac{1}{\theta} ln\Big(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\Big), \tag{2.8}$$

where $\theta \in (-\infty, \infty)\backslash\{0\}$. This copula belongs to the Archimedean copulas, and it will be further discussed in the section 2.3.2.

2.3 Classes of copulas

Now we will discuss 2 important families of copulas: Archimedean and elliptical, their form and basic properties.

2.3.1 Elliptical copulas

In this subsection we will briefly describe elliptical distributions, and then have a closer look at elliptical copulas, their properties and methods of construction.

Definition 2.3.1 ***(Elliptical distributions)*** *The n-dimensional random vector $\mathbb{X}$ has an elliptical distribution if and only if for any $t \in \mathbb{R}^n$ the characteristic function $\varphi_{\mathbb{X}}(t) = \mathbb{E}(\exp(\mathrm{i}t'\mathbb{X}))$ has the representation*

$$t \to \phi_g(t; \mu, \Sigma, \vartheta) = \exp(\mathrm{i}t'\mu) g(t'\Sigma t; \vartheta),$$

where Σ is a symmetric positive semidefinite $n \times n$-matrix, $g(\cdot, \vartheta) : [0, \infty] \to \mathbb{R}$, $\vartheta \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$.

Here the matrix Σ denotes scales and correlations between random variables of vector $\mathbb{X}$, μ is a parameter vector and $g(\cdot, \vartheta)$ is a characteristic generator. If $\mathbb{X}$ has the elliptical distribution with these parameters, it is denoted as $\mathbb{X} \sim \mathcal{E}_n(\mu, \Sigma, g)$.

Elliptical distributions are symmetric and unimodal but are not constrained regarding kurtosis. They are called so due to contour lines having elliptical shapes. In fact, all elliptical copulas are affine extensions of normal copula, and have the elliptical contours of distributions. Differences between them are in the shapes of these elliptical contours.

The knowledge of the distribution of $\mathbb{X}$ does not completely determine the elliptical representation. It uniquely determines μ but Σ and g are only determined up to a positive constant. The matrix Σ can be chosen in such way that it is directly interpretable as the covariance matrix of $\mathbb{X}$, although this is not always so.

Let $\mathbb{X}$ have an elliptical distribution. Then $\mathbb{X} \equiv_d \mu + A\mathbb{Y}$, where $\Sigma = AA^t$ and $\mathbb{Y}$ is a random vector and characteristic generator of $\mathbb{Y}$ is g. Hence, $\mathbb{Y} \equiv_d R * \mathbf{u}$, where $\mathbf{u}$ is uniformly distributed and R is a random variable independent of $\mathbf{u}$. If $\mathbb{E}[R^2] < \infty$, then $\mathbb{E}[X] = \mu$ and $cov[\mathbb{X}] = AA^t\mathbb{E}[R^2]/n$. If we use a new characteristic generator $\hat{g}(u) := g(u/c)$ with $c = n/\mathbb{E}[R^2]$, we ensure that $cov[\mathbb{X}] = \Sigma$. More detailed proof is in [4].

Hence, in the elliptical world it is true that elliptical distributions are uniquely determined by its mean μ, covariance matrix Σ and its characteristic generator $g(u)$. Alternatively the dependence structure of the copula of a continuous elliptical distribution is fully described by the correlation matrix and its type.

For further information, see [6] or [4].

The typical representative of this family is the Gaussian copula from Example 2.1.2. It is easy to get the another expression of Gaussian n-copula with correlation matrix R from (2.4) and Theorem 2.1.2:

$$\mathcal{C}_R^{Ga}(\mathbf{u}) = \Phi_R^n(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)),$$

where Φ^n is a distribution function of n-variate standard normal distribution.

For example, from [1] we know the normal distribution is a limit case of a Student's t-distribution where degrees of freedom are going to ∞. So, that's how we are getting to the second well-known copula by creating it from t-distribution – t-copula, which takes the following form

$$\mathcal{C}^t_{\nu,R}(\mathbf{u}) = \mathcal{C}^n_{\nu,R}(t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_n)), \tag{2.9}$$

where $\mathcal{C}^n_{\nu,R}$ denotes the joint distribution function of an unbiased multivariate t-distribution with ν degrees of freedom and correlation matrix R, t_ν is the distribution function of a univariate standard t-distribution with ν degrees of freedom and t_ν^{-1} is its inverse.

As we will see later, the advantage of elliptical copulas lies in the fact that it is easy to sample from it. On the other hand, they mostly have a very difficult closed form and many parameters.

2.3.2 Archimedean copulas

In this subsection we will mention a family of copulas known as Archimedean copulas. These copulas have very wide range of applications because of their easy construction, nice properties and the great variety of families of copulas belonging to this class.

Archimedean copulas are more popular than elliptical ones, but don't have some useful properties like elliptical do.

Let's begin from the product copula. Applying log on both sides we get

$$\log \mathcal{C}_\pi(u,v) = \log(u) + \log(v).$$

If we now solve this equation for $\mathcal{C}$, we get the following copula

$$\mathcal{C}_\pi(u,v) = \varphi^{[-1]}(\varphi(u) + \varphi(v)), \tag{2.10}$$

where as φ we denoted (log) or $(-\log)$. This copula is called Archimedean copula. φ is a generator of Archimedean copula, and $\varphi^{[-1]}$ is its pseudo-inverse.

In fact, just $(-\log)$ is a generator, (log) is not and it becomes a generator by transformation of the product copula or the generator in such a way that the generator in (2.10) will be the strictly decreasing function.

A famous class of Archimedean copula is copula of the Clayton family.

$$C(u,v) = max\Big([u^{-\theta} + v^{-\theta} - 1]^{-1/\theta}, 0\Big), \tag{2.11}$$

and its generator is

$$\varphi_\theta(t) = \frac{1}{\theta}(t^{-\theta} - 1),$$

where parameter $\theta \in [-1, \infty)\backslash\{0\}$.

As we could see on 2 examples, φ is a continuous, strictly decreasing function from $\mathbb{I} \rightarrow [0,\infty]$ such that $\varphi(1) = 0$. If, further, $\varphi(0) = \infty$ the generator is called *a strict generator* and $\varphi^{[-1]} = \varphi^{-1}$.

Not every generator has as nice expression as above, therefore for solving $\varphi(\mathcal{C}(u,v)) = \varphi(u) + \varphi(v)$ we need to define the "inverse" of φ. The only problem is in point 0, so the pseudo-inverse of generator φ is the function $\varphi^{[-1]} : [0,\infty] \to \mathbb{I}$ such that

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 \leq \mathrm{t} \leq \varphi(0); \\ 0, & \varphi(0) \leq \mathrm{t} \leq \infty. \end{cases}$$

Before we establish necessary and sufficient conditions that can be used as generators of an Archimedean n-copula for $n \geq 2$, a new definition must be stated.

Definition 2.3.2 *A function $h(t)$ is* completely monotonic *on the interval J if it is continuous there and has derivatives of all orders which alternate in sign, i.e., if it satisfies*

$$(-1)^k \frac{\mathrm{d}^k}{\mathrm{d}t^k} h(t) \geq 0,$$

for all $t \in J$, and $k = 0, 1, 2, \ldots$.

If $h(t)$ is completely monotonic on $[0,\infty)$ and there exists $c > 0$ such that $h(c) = 0$, then $h(t)$ is identically 0 on $[0,\infty)$. Let φ is a generator of an Archimedean n-copula. If $\varphi^{[-1]}$ is completely monotonic, then it is positive on $[0,\infty)$, i.e. φ is a strict generator and then $\varphi^{[-1]} = \varphi^{-1}$.

Theorem 2.3.1 *Let $\varphi : [0,1] \to [0,\infty]$ be a continuous and strictly decreasing function such that $\varphi(1) = 0$ and $\varphi(0) = \infty$ and let φ^{-1} be the inverse of φ. Then the function*

$$\mathcal{C} : [0,1]^n \to [0,1]$$

given by

$$\mathcal{C}(x_1, \ldots, x_n) = \varphi^{-1}(\varphi(x_1) + \ldots + \varphi(x_n)) \tag{2.12}$$

is an n-copula if and only if φ^{-1} is completely monotonic on $[0,\infty)$.

Proof See [10].

Nelsen in [10] (page 91) notes that for $n = 2$ it is not required for a generator φ to be completely monotonic but just a convex function. As a consequence, φ doesn't need to be a strict generator, hence equality $\varphi^{[-1]} = \varphi^{-1}$ can not be assumed, and then we have to assume a pseudo-inverse of φ — $\varphi^{[-1]}$.

Example 2.3.1 *Let $\varphi(t) = -\ln t$ for $t \in \mathbb{I}$, φ satisfies the condition $\varphi(0) = +\infty$, thus it is a strict generator. Then $\varphi^{[-1]}(u) = \varphi^{-1}(u) = \exp(-u)$, and under (2.12) we get*

$$\begin{aligned} \mathcal{C}(x_1, \ldots, x_n) &= \varphi^{[-1]}(\varphi(x_1) + \ldots + \varphi(x_n)) = \\ &= \exp\Big(-[(-\ln x_1) + \ldots + (-\ln x_n)]\Big) = \prod_{i=1}^{n} x_i = \Pi(x_1, \ldots, x_n). \end{aligned}$$

Accordingly, the product copula (2.6) is a strict Archimedean copula.

Our basic purpose is to generate random vectors. There are many methods of generating with Archimedean copulas.

Here we show one with a distribution function of copula $\mathcal{C}(u,v)$. Let's begin with 2 following theorems.

Theorem 2.3.2 *Let $\mathcal{C}$ be an Archimedean copula generated by φ and let*

$$K_{\mathcal{C}}(t) = V_{\mathcal{C}}(\{(u,v) \in [0,1]^2 \mid \mathcal{C}(u,v) \leq t\}),$$

where $V_{\mathcal{C}}$ is defined in Definition 2.1.2.
Then for any $t \in [0,1]$,

$$K_{\mathcal{C}}(t) = t - \frac{\varphi(t)}{\varphi\prime(t^+)}. \tag{2.13}$$

Theorem 2.3.3 *Let φ is a generator function of the copula $\mathcal{C}$. Then the joint distribution function $H(s,t)$ of the random variables $S = \varphi(U)/[\varphi(U) + \varphi(V)]$ and $T = C(U,V)$ is given by $H(s,t) = sK_{\mathcal{C}}(t)$ for all $(s,t) \in [0,1]^2$. Hence S and T are independent, S is uniformly distributed on $[0,1]$.*

For proofs of these theorems, see [10].

The corollary of the last theorem is an instruction how to generate with Archimedean copulas, which will be studied in the next section.

But why is this family of copulas called Archimedean?

At first, let's remind an Archimedean axiom for positive real numbers:
Let $a, b \in \mathbb{R}$, then there exists $n \in \mathbb{N}$ such that $na > b$. An Archimedean copula has a very same property on the interval $[0,1]$ where it is assigned to a number $\mathcal{C}(u,v) \in [0,1]$ to the numbers $u, v \in [0,1]$.

Let $u \in [0,1]$. Then define the $\mathcal{C} -$ *powers* $u^n_{\mathcal{C}}$ as

$$\begin{aligned} u^1_{\mathcal{C}} &= u, \\ u^{n+1}_{\mathcal{C}} &= \mathcal{C}(u, u^n_{\mathcal{C}}). \end{aligned}$$

We can now state the analogue of Archimedean axiom for copulas:
Let $\mathcal{C}$ be an Archimedean copula. Then for any $u, v \in [0,1]$, there exists a positive integer n such that $u^n_{\mathcal{C}} < v$.

2.3.3 The other classes of copulas

There are other classes of copulas, like extremal, Marshall-Olkin or quasi-copulas, but they have specific conditions (for example, on marginal distributions). We need to solve a general case that's why these copulas will not be studied here. For more information, see for example [2] or [10].

2.4 Summary

In this chapter we got acquainted with copulas and discussed the two most important theorems needed to solve the problem stated in section 1 - the Sklar's Theorem 2.1.2 and Invariance property 2.2.1.

In section 2.3 were presented two basic classes of copulas, stated theirs properties and showed some typical families of each classes. Now we have the theoretical knowledge for introducing method of generation specific to different classes of copulas.

Chapter 3

Generating

In previous section we explained the theoretical background of copulas and stated their elementary properties. But as it was said at the beginning of this work, we selected copulas as a mean to generating of random samples with given properties. In this chapter we describe the effective algorithms for random variate generating with copulas. After that the numeric study is used to choose the optimal copula for generating. It is based on the main book about copulas [10] and other information sources [8] or [12].

We begin this chapter with a general algorithm of generating with all copulas. Then there are mentioned different methods of generating for different families of copulas, their advantages and disadvantages. In the further section there is a short discussion about the dependence measure - what kind is better and why. The main references about it are [12] and [3]. In the end of the chapter is made comparisons which copula is better and which distributions will we use in generating.

3.1 General algorithm

Every family of copulas has different type of algorithms depending on specific properties of the family. A general algorithm which can be used to generate all types of copulas will be presented in this section. However, this algorithm is bad in the sense of speed and efficiency in most cases.

Let $\mathcal{C}$ be an n-copula. Then denote

$$\mathcal{C}_k(u_1,\ldots,u_k) = \mathcal{C}(u_1,\ldots,u_k,1,\ldots,1),$$

for $k = 1,\ldots,n$, a k-dimensional marginals of $\mathcal{C}$, where $\mathcal{C}_1(u_1) = u_1$ and $\mathcal{C}_n(u_1,\ldots,u_n) = \mathcal{C}(u_1,\ldots,u_n)$.
Let $U_1,\ldots,U_n$ denote the uniform distributions on $[0,1]$ with joint distribution function $\mathcal{C}$. Generally, U_i can denote any distributions, but joint distribution function must be $\mathcal{C}$.

The conditional distribution of U_i given the values $U_1,\ldots,U_{i-1}$, is given by the following formula

$$\begin{aligned} \mathcal{C}_i(u_i|u_1,\ldots,u_{i-1}) &= \mathbb{P}\{U_i \leq u_i|U_1 = u_1,\ldots,U_{i-1} = u_{i-1}\} = && (3.1)\\ &= \frac{\partial^{i-1}\mathcal{C}_i(u_1,\ldots,u_i)}{\partial u_1,\ldots,\partial u_{i-1}} / \frac{\partial^{i-1}\mathcal{C}_{i-1}(u_1,\ldots,u_{i-1})}{\partial u_1,\ldots,\partial u_{i-1}}. && (3.2)\end{aligned}$$

The equality (3.1) is clear. Then by applying Sklar's theorem 2.1.2 and using the relation between the distribution function and the density, we can derive the copula:

$$f(x_1,\dots,x_n)=\frac{\partial^n(\mathcal{C}(F_1(x_1),\dots,F_n(x_n)))}{\partial F_1(x_1)\dots F_n(x_n)}\prod_{i=1}^{n}f_i(x_i), \tag{3.3}$$

where f is the density of $\mathbb{X}$ and f_i are the marginal distributions of X_i.

After the substitution (3.3) to the the equality (3.1) we get the required result (3.2).

The expression (3.2) makes sense, when numerator and denominator of it exist and the denominator is not zero.

This fact leads us to a basement of the algorithm of random variables generating called the standard construction. This algorithm generates random vector $\mathbf{u}=(u_1,\dots,u_n)$ from common copula $\mathcal{C}$.

Algorithm 1 *1. Simulate a random variate u_1 from $U(0,1)$,*

2. Simulate a random variate u_2 from $\mathcal{C}_2(.|u_1)$,

$\vdots$

i. Simulate a random variate u_i from $\mathcal{C}_i(.|u_1,\dots,u_{i-1})$,

$\vdots$

n. Simulate a random variate u_n from $\mathcal{C}_n(.|u_1,\dots,u_{n-1})$,

If we denote each random variable we get in algorithm above as $A_1,\dots,A_n$, we can make sure that the random vectors

$$\left(A_1,\mathcal{C}_2^{-1}(A_2|A_1),\dots,\mathcal{C}_n^{-1}\left(A_n|A_1,\mathcal{C}_2^{-1}(A_2|A_1),\mathcal{C}_3^{-1}(A_3|A_1,\mathcal{C}_2^{-1}(A_2|A_1)),\dots\right)\right)^t$$

have the distribution function $\mathcal{C}$.

But as it was said at the beginning of this section, this common method of generating from common copula has some disadvantages.

Let's look at the standard construction closer. In the i^{th} step (except 1^{st} one) of algorithm we know the values $u_j, j<i$, and want to find a random value u_i. After generating $q\sim U(0,1)$ we have $u_i=\mathcal{C}_i^{-1}(q|u_1,\dots,u_{i-1})$. Then we get u_i by numerical rootfinding of the equation $q=\mathcal{C}_i(u_i|u_1,\dots,u_{i-1})$.

This algorithm is useful when $\mathcal{C}_i(u_i|u_1,\dots,u_{i-1})$ has a closed form (and therefore no numerical rootfinding is required), in the other case we need to find other ways of generating. Let's show one copula with closed form.

Example 3.1.1 *We want to generate a random vector from the copula $\mathcal{C}(u,v)=(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$, for $\theta>0$, and it looks almost like the Clayton copula (2.11)), but it is not. Then using the notation of the algorithm*

$$\begin{aligned}\mathcal{C}_2(v|u) &= \frac{\partial\mathcal{C}}{\partial u}(u,v)=-\frac{1}{\theta}(u^{-\theta}+v^{-\theta}-1)^{-1/\theta-1}(-\theta u^{-\theta-1})\\ &= (u^{\theta})^{(-1-\theta)/\theta}(u^{-\theta}+v^{-\theta}-1)^{-1/\theta-1}\\ &= (1+u^{\theta}(v^{-\theta}-1))^{(-1-\theta)/\theta}.\end{aligned}$$

The solution of the equation $q = \mathcal{C}_2(v|u)$ for v is

$$\mathcal{C}_2^{-1}(q|u) = \Big((q^{\frac{-\theta}{1+\theta}} - 1)u^{-\theta} + 1\Big)^{-1/\theta},$$

and equals a variable v, i.e $\mathcal{C}_2^{-1}(q|u) = v$ too.

Hence the algorithm of generating a random vector $(u, v)^t$ from this copula $\mathcal{C}$ is

1. *Simulate 2 independent random variates $u, q \sim U(0, 1)$,*
2. *Set $v = \Big((q^{\frac{-\theta}{1+\theta}} - 1)u^{-\theta} + 1\Big)^{-1/\theta}$,*
3. *$(u, v)^t$ is required random vector.*

3.2 Using specific types of copulas for generating

There are many families of copulas having different properties that are used in different situations. Therefore there are many ways to generate a random sample with them. And because the standard construction can be a little slow, we will discuss the most useful and popular ways of generating in this section.

3.2.1 Generating with elliptical copulas

We begin with elliptical copulas. This family of copulas has multivariate elliptical distributions mentioned in the previous section.

There are many kinds of elliptical copulas but the most popular are the normal copula and the t-copula.

Normal copula

The normal copula (expression (2.1.2)) is used in cases where there is a very little information about the distributions in the problem. It may be caused by the fact that we can't find out all factors having influence on our variable. For example it's used in problems of measure of defaults, in latent variables model or estimations of Collateralized Debt Obligations (known as CDO).

The normal copula has a multivariate normal distribution. And because an elliptical distribution is uniquely determined by its mean, covariance matrix and the type of its marginals (t_ν, normal, etc.), the Gaussian copula is uniquely determined by its dependence matrix Σ and knowledge of its type (Pearson, Kendall etc.).

Suppose we have a multivariate distribution H, with marginals $F_1, \ldots, F_n$ and dependence structure defined by the covariation matrix Σ, and we want to simulate random vectors using normal copula.

First, Σ should be a strictly positive definite matrix, so there exists a matrix L, which satisfies $\Sigma = L * L^t$. If $Y_1, \ldots, Y_n \sim \mathcal{N}(0, 1)$ are independent, then

$$\mu + L\mathbb{Y} \sim \mathcal{N}_n(\mu, \Sigma).$$

One of the choices is a Cholesky decomposition of Σ. There is a lower triangular matrix L, such $LL^t = \Sigma$. The following algorithm simulates normal copula with marginals $F_1, \dots, F_n$ and covariation matrix Σ:

Algorithm 2 *1. Find Cholesky decomposition L of Σ,*

2. Simulate n independent random variable $\mathbb{Y} = (y_1, \dots, y_n) \sim \mathcal{N}(0,1)$,

3. Set $\mathbb{X} = L\mathbb{Y}$,

4. Set $\mathbb{A} = \Phi(\mathbb{X})$, where Φ is a distribution function of standard multivariate normal distribution $\mathcal{N}_n(0,1)$,

5. Set $u_i = F_i^{-1}(a_i)$, for $i \in \{1, \dots, n\}$,

Let's look at the algorithm above a little closer. The first 3 steps are making a multivariate normal distribution $\mathbb{X} \sim \mathcal{N}_n(\mu, \Sigma)$. In the 4^{th} step a sample of the normal copula is made.

There were two conditions on this algorithm: dependence structure and marginals. It is clear that condition of marginals is satisfied due to steps (4) and (5) of this algorithm.

Now we shall remind an invariance property of copulas (Theorem 2.2.1). It asserts $\mathcal{C}_{XY}$ is invariant under strictly increasing transformations of X and Y. And F_i^{-1} is strictly increasing function. And $\mathbb{A}$ is a random vector of the normal copula $\mathcal{C}_R^{Ga}$ (more (2.1.2)). Summarizing this we observe that the vector $\mathbf{u}$ in 5^{th} step of algorithm above has the same dependence structure as the normal copula, i.e. has the required dependence structure. And that is what we wanted to prove.

T-copulas

Another well-known member of elliptical distributions is the t-distribution.

We say that $\mathbb{X}$ has an n-variate t-distribution with ν degrees of freedom with mean μ, for $\nu > 1$, and with covariance matrix $\frac{\nu}{\nu-2}\Sigma$, for $\nu > 2$, if

$$\mathbb{X} \sim \mu + \frac{\sqrt{\mu}}{\sqrt{S}}\mathbb{Y}, \tag{3.4}$$

where $\mu \in \mathbb{R}^n$, $S \sim \chi^2_\nu$, $\mathbb{Y} \sim N_n(0, \Sigma)$, and S and $\mathbb{Y}$ are independent.

The t-copula $\mathcal{C}$ can be written as

$$\mathcal{C}^t_{\nu,R}(\mathbf{u}) = t^n_{\nu,R}(t_n^{-1}(u_1), \dots, t_n^{-1}(u_n)), \tag{3.5}$$

where R is the correlation matrix of Σ, i.e. $R_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$ for all $i, j \in \{1, \dots, n\}$, and $t^n_{\nu,R}$ denotes the distribution function of t-distribution (3.4). And t_n^{-1} is a margin of $t^n_{\nu,R}$, and all these marginals are equal.

If we want to generate a random vector with marginals $F_1, \dots, F_n$ and the correlation matrix R (or the covariation matrix Σ), then the expression (3.4) gives us an easy idea of algorithm of random generating with t-copulas

Algorithm 3 *1. Find Cholesky decomposition L of Σ,*

2. Simulate n independent random variables $\mathbb{Y} = (y_1, \dots, y_n) \sim \mathcal{N}(0,1)$,

3. *Set* $\mathbb{X} = L\mathbb{Y}$,

4. *Simulate random variables* $\mathbb{Z} \sim \chi^2_\nu$ *independent of* $\mathbb{X}$,

5. *Set* $\mathbb{A} = \frac{\mathbb{X}}{\sqrt{\frac{\mathbb{Z}}{\nu}}}$,

6. *Set* $\mathbb{B} = t_\nu(\mathbb{A})$,

7. *Set* $u_i = F_i^{-1}(b_i)$, *for* $i \in \{1, \dots, n\}$.

The vector in the 6^{th} step of the algorithm above is a random vector of t-copula $C^t_{\nu,R}$ with ν degrees of freedom and the covariation matrix Σ. The 7^{th} step uses the invariance property 2.2.1 to make a required marginal distribution.

The reason, why this algorithm generates vectors with given correlation structure, is the invariance property 2.2.1. It was explained more in the previous subsection.

It is also very important to realize that the Gaussian and t-copulas are copulas of elliptical distributions, *but* they are *not* elliptical distributions themselves. Therefore we can use the t-copula with 1 and 2 degrees of freedom for generating, though the t-distribution with the same degrees of freedom has no variance.

Before some basic graphs and differences between different kinds of copula will be shown, we conclude a theoretical part of the algorithms of elliptical copulas described in this subsection by saying that generating with elliptical n-copulas is fast and easy to implement and generally doesn't require any special conditions on correlation matrix or on the shape of marginal distributions.

3.2.2 The other types of elliptical copulas

There are other types of elliptical copulas, but they are mostly created by transformations of the two basic types of copulas (i.e. normal or t-copula). We don't need them in our work, the further information about nonstandard types of elliptical copulas is in [4].

Examples

We conclude this section by showing graphics of a normal copula (2.1.2), t-copulas with 3 and 10 degrees of freedom.

All copulas are standard copulas (2-dimensional copulas) having the normal marginal distribution $N(1, 3)$ and the exponential marginal distribution with parameter 1. The correlation coefficient we want generate with is $0, 7$.

On the first picture 3.1 is the normal copula. The sample correlation coefficient is $0, 69$.

On the second picture 3.2 is a t-copula with 3 degrees of freedom. It looks almost like the previous one. However, this one has more extreme values. That is the big advantage (or sometimes disadvantage) of t-copula comparing to the normal copula. The sample correlation coefficient is $0, 64$.

On account of the algorithms in sections 3 and 2, the generated copula does not depend on marginal distributions.

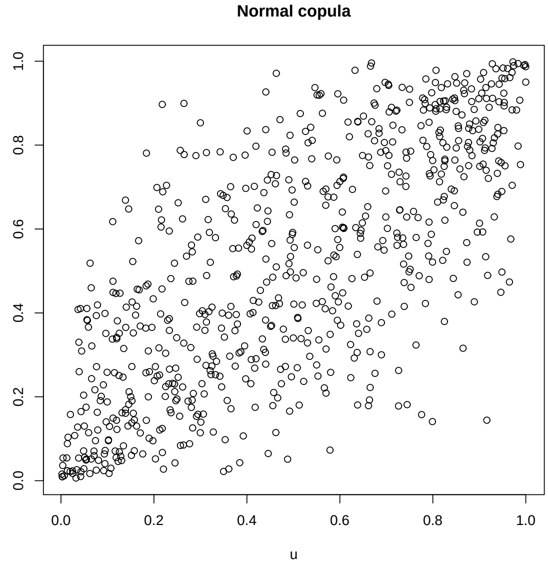

Figure 3.1: The bivariate normal copula with correlation $0,7$.

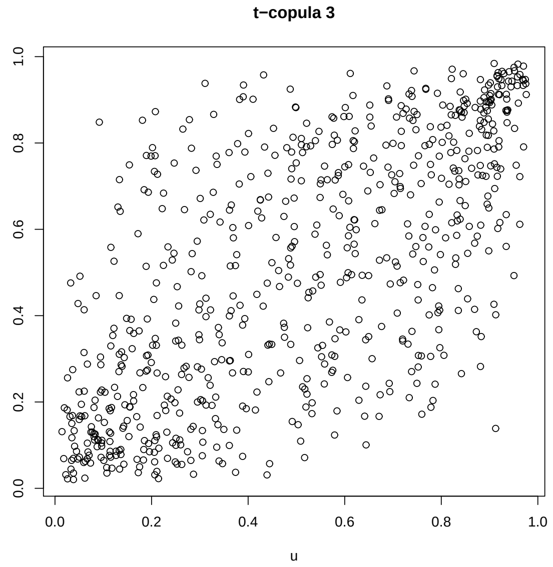

Figure 3.2: The bivariate t-copula copula with 3 degrees of freedom with correlation $0,7$.

3.2.3 Generating with Archimedean copulas

There are many different algorithms of generating with Archimedean copulas. They have one common property - they are easy to generate. On the other side, there are different algorithms of generating for different families of Archimedean copulas. Of course there are some common methods for the families with the same specific character.

Here we will show one way of using the application of the theorem 2.3.3. We want to generate $(u, v)^t$ with joint distribution function $\mathcal{C}$, which is an Archimedean copula with a generator φ.

Algorithm 4 *1. simulate $s, q \sim \mathcal{U}(0,1)$, s and q are independent,*

2. set $t = K_{\mathcal{C}}^{-1}(q)$, where $K_{\mathcal{C}}$ is a distribution function of $\mathcal{C}(u,v)$ defined in theorem, 2.3.2

3. $u = \varphi^{[-1]}(s\varphi(t))$, $v = \varphi^{[-1]}((1-s)\varphi(t))$.

The proof that this algorithm yields the desired result can be found in [12]. This algorithm is simple, but generating the general n-variate Archimedean copula with it is very complicated because of the explicit expression of the distribution function of copula.

Generating with Archimedean copulas requires at first knowledge of a generator function, and then a knowledge of certain marginal distributions. But we want find an algorithm at first and then use it for generating different distributions. That is the reason, why the Archimedean copulas can't be used for our purpose.

More information about this or other kinds of generating will appear in [8] or [12].

3.3 Measure of dependence's problem

The main purpose of this work is solving the following problem:
generating n-variate random vectors $(X_1, \ldots, X_n)$ with required marginal distribution and given correlation coefficients $cor(X_i, X_j)$ for $i, j \in \{1, \ldots, n\}$.

Until now we have studied methods how to generate vectors with fixed marginal distribution and correlation coefficients of some kind. And there are many different kinds of correlation, the most popular are Pearson's, Kendall's or Spearman's.

The theory about copulas, for example the invariance property (2.2.1) states that generated vectors have the required correlation. But in practise it doesn't work so well.

In statistics the linear correlation (Pearson's coefficient) is frequently used as a measure of dependence. It is explained in [12] that

> () ... most random variables are not jointly elliptically distributed and using linear correlation as a measure of dependence in such situation might prove very misleading..

In general we can say that since copulas use non-linear dependence structure, using the linear correlation coefficient can lead to the wrong results. Furthermore, linear correlation is not defined if variance of X and Y is infinite.

Due to the quote and notes above, we need to work with more robust correlation coefficient. That can be, for example, Kendall's τ or Spearman's ρ. There are many books and articles devoted to this problem - refer for example to [3].

A big advantage of rank correlations matrices is an invariance under strictly increasing transformations of the vector components (not just strictly increasing *linear* transformations as by Pearson's). But for using these more modern correlation matrix, we need to be able to compute Kendall's or Spearman's coefficients or at least relationships between them and Pearson's correlation. In sense of copula, Kendall's τ and Spearman's ρ can be expressed only in terms of $\mathcal{C}$ of (X, Y) like ([10])

$$\begin{aligned}
\tau_K(X,Y) = \tau_K(\mathcal{C}) &= 4\iint_{[0,1]^2} \mathcal{C}(u,v)\mathrm{d}\mathcal{C}(u,v) - 1, \\
\rho_S(X,Y) = \rho_S(\mathcal{C}) &= 12\iint_{[0,1]^2} uv\mathrm{d}\mathcal{C}(u,v) - 3 \\
&= 12\iint_{[0,1]^2} \mathcal{C}(u,v)\mathrm{d}u\mathrm{d}v - 3.
\end{aligned}$$

Let X and Y be random variables and F and G be strictly increasing functions. From Theorem 2.2.1 follows that $\mathcal{C}(X,Y) = \mathcal{C}(F(X), G(Y))$. It is equivalent to $(F(X), G(Y)) \sim \mathcal{C}$, and then we get

$$\begin{aligned}
\tau_K(X,Y) &= \tau_K(F(X), G(Y)), \\
\rho_S(X,Y) &= \rho_S(F(X), G(Y)).
\end{aligned}$$

The following theorem asserts a relation between Kendall's correlation matrix and linear correlation matrix R for nondegenerate elliptical distributions, so Kendall's can be estimated from R.

Theorem 3.3.1 *Let $\mathbb{X} \sim \mathcal{E}_n(\mu, R, \phi)$ is a elliptical multivariate random variable with $\mathbb{P}(X_i = \mu_i) < 1$ and $\mathbb{P}(X_j = \mu_j) < 1$. Then*

$$\tau_K(X_i, X_j) = (1 - \sum_{x \in R} (\mathbb{P}(X_i = x))^2)\frac{2}{\pi}\arcsin(R_{ij}), \tag{3.6}$$

where the sum extends over all atoms of the distribution of X_i. If rank$(R) \geq 2$, then the expression (3.6) simplifies to

$$\tau_K(X_i, X_j) = (1 - (\mathbb{P}(X_i = x))^2)\frac{2}{\pi}\arcsin(R_{ij}).$$

Proof See [12].

For us it is an important corollary saying that if $\mathbb{P}(X_i = x) = 0$ for all i, what is true for e.g. multivariate t and normal distributions with strictly positive correlation matrix Σ, then for all i, j

$$\tau(X_i, X_j) = \frac{2}{\pi}\arcsin(R_{ij}). \tag{3.7}$$

For elliptical copulas the Kendall's τ is an efficient estimator of covariance.

All theoretical considerations above may be summarized by saying that Kendall's τ is better for our purpose.

3.4 Numeric study

Now we know the different algorithms for generating random samples with given marginal distributions and given correlation matrix. But for solving our basic problem we need to know the only algorithm that is the best or optimal in some way. In this section we will make comparisons that will help us to choose the best algorithm.

The statistical software R version 2.6.0 (more in [13] or [14]) is used for all the numeric studies in this work.

3.4.1 The choice of a training set

We have discussed algorithms with two groups of copulas - Archimedean and elliptical copulas. We explained that both of these families have their advantages and disadvantages. And our goal is to decide which one has more advantages, and which one we will choose to solve the basic problem.

So let's remind some basic facts about these two families.

Elliptical copulas don't require very much information about the situations they are used in. In general, they are used in situations when we don't know anything else. That is one of the reasons why they have recently been so popular in generating of courses and defaults of funds or other commercial papers recently. In addition to that their simplicity and analytical manageability are other important reasons of its popularity.

On the other side, generating with Archimedean copulas means knowing further details of the model. Thus, we may use a better model for generating which can include more properties closer to the reality. And that is the main advantage of this family of copula. It may be a disadvantage too, because we must exactly know the initial conditions and in terms of it choose a particular kind of Archimedean copula. It means that everything depends on choosing a generator.

There is one more special property of Archimedean copula - for generating we use an inverse of the distribution function of copula $\mathcal{C}$, and for general n-multivariate copula it could have very non-trivial inverse.

On account of the remark above, we can state that Archimedean copulas don't come in useful for generating a random sample with required dependence structure and marginal distributions. It means we should concentrate on elliptical copula.

We know that there are two famous groups of elliptical copulas - normal and t-copula. Normal copula $\mathcal{C}_R^{Ga}$ has multivariate normal distribution with mean equal to 0 and covariation matrix Σ. There are no other options. In contrast to Gaussian copula, the t-copula $\mathcal{C}_{\nu,R}^t$ can have many representations because it depends on its degrees of freedom. We know that normal and t-copulas have a relation between them - $\mathcal{C}_{\nu,R}^t$ converges to $\mathcal{C}_R^{Ga}$ if $\nu \to \infty$. In practice it means that if a number of degrees of freedom is higher than 10, the t-copula behaves as a Gaussian one.

Now our purpose is to decide which elliptical copula is the best or optimal one. Before we begin with comparisons, we need to make clear which meaning of a copula is better than the other one? What is the criterion for choosing? To explain this, let's go back to our basic goal. We want to generate a random sample with given marginals and correlation structure. Hence, these two conditions could be our criteria that assess the effectiveness of all algorithms.

Algorithms of generating with elliptical copulas were made in a way that guarantees required marginals for sufficient number of repetitions (the theoretical proof follows from the last two steps of the algorithms in section 3.2.1). The statistical verification of this claiming is in the further sections.

Another possibility is to choose by a correlation match, i.e. the correlation matrix of samples generated by the copula is the closest to the correlation matrix we generate from.

The required marginals are guaranteed, therefore a correlation matrix is our criterion.

For our purpose, we will generate the first data set (all marginals and parameters are fictitious): a 7 dimensional vector $\mathbb{X}$ with marginals Multinomial distribution with values $(0, 5, 10, 15)$ and probabilities $(0.22, 0.6, 0.08, 0.01)$, Multinomial distribution with values $(-5, 2, 8, 14, 20, 26, 32, 38, 46, 52)$ and probabilities $(10, 10, 20, 5, 5, 5, 30, 5, 5, 5)$, *Gamma*$(1)$, *Beta*$(1, 0.5)$, *Exp*$(10)$, *N*$(0, 5)$, *Logistic*$(1, 5)$.

The table 3.1 shows the Pearson correlation matrix used for generating.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1.00	−0.50	−0.41	−0.19	0.42	−0.06	−0.40
X_2	−0.50	1.00	0.79	0.75	−0.15	0.68	0.77
X_3	−0.41	0.79	1.00	0.88	−0.44	0.86	0.97
X_4	−0.19	0.75	0.88	1.00	−0.19	0.92	0.88
X_5	0.42	−0.15	−0.44	−0.19	1.00	−0.32	−0.45
X_6	−0.06	0.68	0.86	0.92	−0.32	1.00	0.86
X_7	−0.40	0.77	0.97	0.88	−0.45	0.86	1.00

Table 3.1: The Pearson correlation matrix of data

Each result in this chapter is made for 5 different correlation matrices and for 5 different sets of marginals. We are going to talk just about the results of input data (the random vector and correlation matrices) described above. The results of the other input data are mentioned only if they are quite different.

Before we go further, let R denote the given Pearson correlation matrix and ρ_{ij} is its correlation coefficient at the place i, j. Let $cor(\mathbb{X})$ denote the sample Pearson correlation matrix and $cor(\mathbb{X})_{ij}$ is its correlation coefficient at the place i, j. The lower index G, resp. t, on $\mathbb{X}$ means the vectors generated with Gaussian, resp. with t-copulas, sometimes if it is necessary, numbers - degrees of freedom - will be added to the index.

3.4.2 Comparison criteria

As we have just said we need to use some theoretical instruments to compare the accuracy of the sample correlation matrix and decide which copula is better. So, we use the 3 following comparison criteria:

1. the 1-norm distance defined by

$$D_1(\mathbb{X}) = \sum_{i,j} |cor(\mathbb{X})_{ij} - \rho_{ij}|$$

2. the Euclidean metric defined by

$$D_2(\mathbb{X}) = \sqrt{\sum_{i,j} (cor(\mathbb{X})_{ij} - \rho_{ij})^2}$$

3. infinity-norm distance defined by

$$D_3(\mathbb{X}) = \max_{i,j} |cor(\mathbb{X})_{ij} - \rho_{ij}|$$

All these criteria are metrics in $\mathbb{R}^n$.

Let A and B be some different copulas, then let $\mathbb{D}_i^{A,n}, i \in \{1, 2, 3\}$ denotes the vector of n different values of $D_i(\mathbb{X})$, where $\mathbb{X}$ is generated with copula A. We want to find out, whether the vector $\mathbb{D}_i^{A,n}$ is same as a vector $\mathbb{D}_i^{B,n}$ or not for $i \in \{1, 2, 3\}$.

It is used the Student's t-test (with equal sample sizes). Statistically speaking, we test a $H_0 : \mathbb{E}[\mathbb{D}_i^A] = \mathbb{E}[\mathbb{D}_i^B]$ against $H_1 : \mathbb{E}[\mathbb{D}_i^A] > \mathbb{E}[\mathbb{D}_i^B]$, where $\mathbb{D}_i^A$ is a random variable having the distribution function of $\mathbb{D}_i^{A,n}$. It means, we think the correlation matrices generated with A copulas are closer in sense of metric to R than correlation matrices generated with B copulas. We reject the null hypothesis on the significance level $\alpha = 0.05$, if

$$\frac{\overline{\mathbb{D}_i^{A,n}} - \overline{\mathbb{D}_i^{B,n}}}{\sqrt{S_{A_n}^2 + S_{B_n}^2}} \sqrt{n} \geq t_{2(n-1)}(1 - \alpha), \tag{3.8}$$

where n is a sample size, $t_n(\alpha)$ is a quantile of t-distribution with n d.f. and $S_{A_n}^2$ is a sample variance of $\mathbb{D}_i^{A,n}$.

The conditions of the t-test must be satisfied: $\mathbb{D}_i^{A,n}$ and $\mathbb{D}_i^{B,n}$ are independent, $\mathbb{D}_i^{A,n} \sim \mathcal{L}(\mu_1, \sigma^2)$, $\mathbb{D}_i^{B,n} \sim \mathcal{L}(\mu_2, \sigma^2)$, $\sigma > 0$. Further $\sigma^2 < \infty$, since $D_i(\mathbb{X})$ is the metric.

Note that every test in this work is made on the significance level 0.05, if it is not stated otherwise.

3.4.3 Normal or t-copula

In this subsection we present advantages and disadvantages of sample correlation matrices with normal and t-copulas. That's why we begin our study with comparisons of sample correlation matrices of vectors generated with normal and t-copulas. The Pearson correlation coefficient is considered as a correlation coefficient here.

The 1-norm distance

We begin our comparison with the 1-norm distance. The histograms of this metrics for different families of copula are in the picture 3.3. There are no big visual differences between them.

Let's study $H_0 : \mathbb{E}[\mathbb{D}_1^{\mathbb{Y}^G}] = \mathbb{E}[\mathbb{D}_1^{\mathbb{X}_{t_1}}]$ against $H_1 : \mathbb{E}[\mathbb{D}_1^{\mathbb{X}_{t_1}}] < \mathbb{E}[\mathbb{D}_1^{\mathbb{Y}^G}]$, we get $20.2 \geq 1.64$, hence we reject the H_0. Therefore the correlation structure of $\mathbb{Y}_G$ is further (in sense of the first metric) from the R than $\mathbb{X}_{t_1}$ from the R.

Testing the correlation structure of $\mathbb{Y}_G$ with the correlation structures of $\mathbb{X}_{t_k}$, $k \in \{1, 2, 3, 5, 7, 9, 11\}$ gives the same result. So, generating with t-copula is more accurate than with normal copula.

If we test which t-copula is better, we get that there is no significant difference between generating with t-copula with 1 and 2 d.f. But if we test the hypothesis $\mathbb{E}(\mathbb{D}_1^{\mathbb{Y}_{t_9},n} - \mathbb{D}_1^{\mathbb{X}_{t_1},n}) = 0$, we reject H_0. The results of other tests look very similar - it means that correlation matrices of vectors generated with t-copula with lower degrees of freedom are closer to R than the other ones.

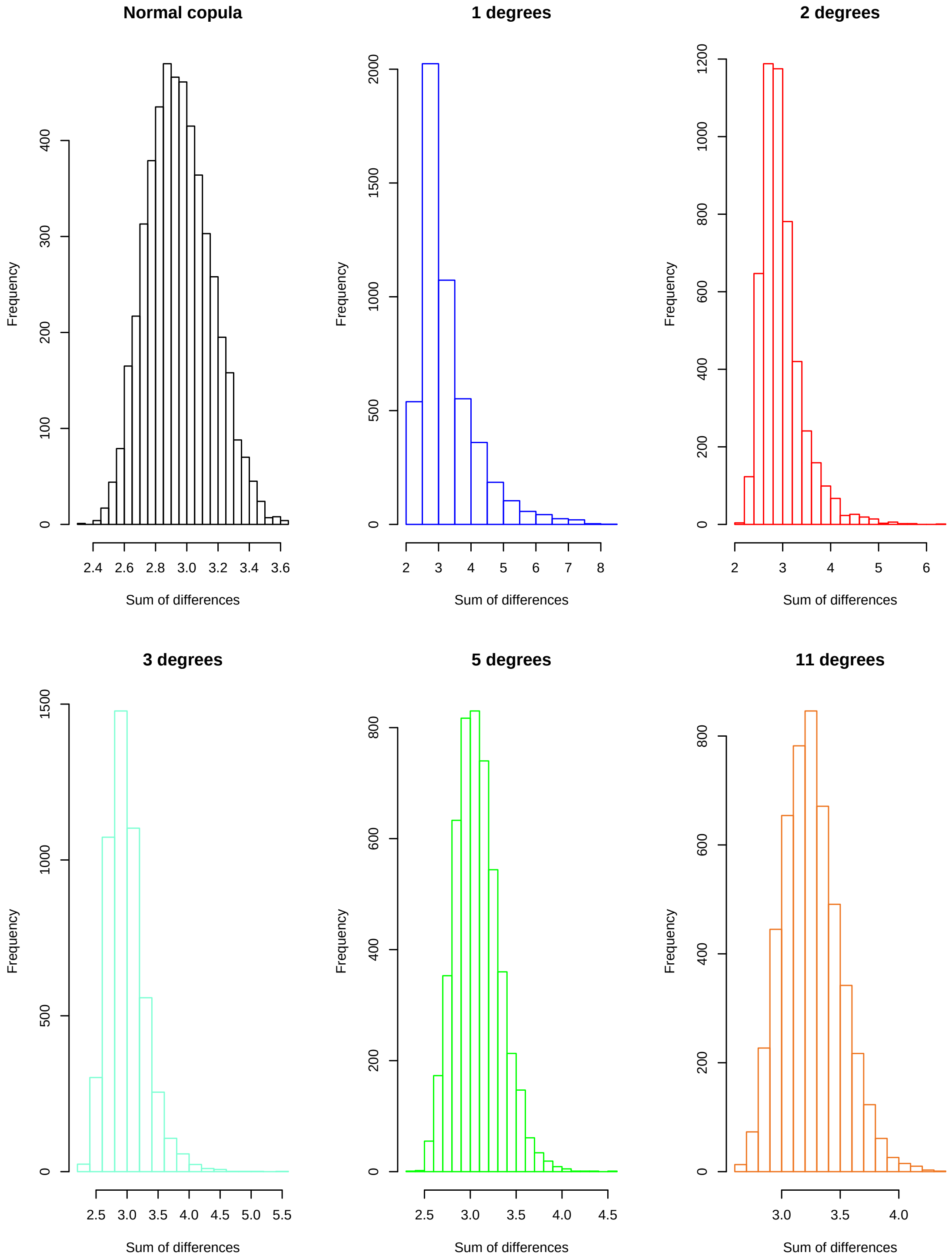

Figure 3.3: The histograms of $\sum_{i,j} |cor(\mathbb{X})_{i,j} - \rho_{ij}|$ for different copulas

The Euclidean metric

The second metric is the Euclidean metric. At first, we will show the histograms of this metric for different families of copula - the picture 1 in the appendix. We may have doubts again that Euclidean metric gives the same results for different copulas. Therefore we make the t-test.

At first, we test the hypothesis $H_0 : \mathbb{E}[\mathbb{D}_2^{\mathbb{Y}^G,n}] = \mathbb{E}[\mathbb{D}_2^{\mathbb{X}_{t_1},n}]$ against $H_1 : \mathbb{E}[\mathbb{D}_2^{\mathbb{X}_{t_1},n}] < \mathbb{E}[\mathbb{D}_2^{\mathbb{Y}^G,n}]$. We get $23.6 > t_{10000}(0.95) = 1.64$, where $t_n(0.95)$ is the quantile function of the t-distribution. So, we reject the H_0 and confirm the claim that the Euclidean metric of $\mathbb{Y}^G$ has bigger mean value than of $\mathbb{Y}^{t_1}$.

The other test may show that the vectors generated with t-copula with one and two degrees of freedom have the significantly smaller mean values. And there is no significant difference between means of metric of correlation matrices t-copula with one and two d.f.

The infinity-norm distance

The third comparison criterion is the maximal absolute value of a difference between R and correlation matrices generated with different families of copulas. The histograms of these values of normal and t-copulas with $1, 2, 3, 5, 11$ degrees of freedom are in the appendix in the picture 2.

There are some interesting results: we see that $cor(\mathbb{X})_G$ has big mean value. And the best results of the criterion gives us the correlation matrix generated with t-copula with 1 degrees of freedom. As the degrees of freedom of t-copulas grow, the results decrease.

For confirming this claim we use the t-test. Let's test $H_0 : \mathbb{E}[\mathbb{D}_3^{\mathbb{Y}^G,n}] = \mathbb{E}[\mathbb{D}_3^{\mathbb{X}_{t_1},n}]$ against $H_1 : \mathbb{E}[\mathbb{D}_3^{\mathbb{X}_{t_1},n}] < \mathbb{E}[\mathbb{D}_3^{\mathbb{Y}^G,n}]$. On the significant level 0.05 we reject the null hypothesis against alternative hypothesis that the mean value of $\mathbb{D}_3^{\mathbb{X}_{t_1},n}$ is smaller than the mean value of $\mathbb{D}_3^{\mathbb{Y}^G,n}$. In the same way, we can statistically show that every $cor(\mathbb{X}^{t_k}), k \in \{1, 2, 3, 5\}$ has smaller mean value than $cor(\mathbb{Y}^G)$. But there is no significant difference between $cor(\mathbb{Y}^G)$ and $cor(\mathbb{X}^{t_k})$ for $k > 6$.

Naive metric

The last "metric" we use is rather naive than mathematical. The results are in the table 3.2. The correlation matrices of the vectors generated with the normal and t-copulas with different degrees of freedom are compared using different count of repetitions.

For example, the red value 29% in this table means that 2000 pairs of vectors $\mathbb{X}_{\mathbb{G}}$ with normal and vectors $\mathbb{X}_t$ with t-copula with 3 degrees of freedom were generated. Then we compare the matrices $\mid cor(\mathbb{X}_G) - R \mid$ and $\mid cor(\mathbb{X}_t) - R \mid$ (where R is positive definite matrix we generate from), and compute how many elements of the first matrix are smaller than the relevant elements of the second matrix. And the red value 29% is an average value.

In other words, 29% says that in mean there are just 29% percents of all elements of correlation matrix of $\mathbb{X}_G$ which are closer to R than elements of $\mathbb{X}_t$ to R.

Moreover, the table 3.2 shows that the higher the sample size is - the more accurate the vectors generated with t-copulas are. The next observation may be the fact that $cor(\mathbb{X}_t)$ generated with t-copulas with higher degrees of freedom are more accurate than the one with lower d.f.

		Sample size						
		100	500	2000	5000	10000	20000	50000
Degrees of freedom of t-copula	1	37	35	35	34	34	31	28
	2	36	32	30	28	26	26	23
	3	36	32	29	25	25	22	21
	5	38	32	29	26	24	23	20
	7	39	34	30	27	26	23	21
	8	37	34	30	27	24	23	21
	11	39	36	32	30	27	26	22

Table 3.2: All numbers inside of a table are percentages that denote how often are the sample correlation matrices generated with normal copula closer to R than correlation matrices generated with t-copula (with a view to degrees of freedom and sample sizes)

Conclusion

Summarizing this, we can say that the random vectors generated with t-copula with lower degrees of freedom (like 1 or 2) have the correlation matrices which are the closest (in meaning of described metrics) to the given correlation matrix. But this result is not new. We have talked about that in the subsection 3.2.1. A more complete background and results may be obtained in [12].

Author has found no hypothesis that tests whether a correlation matrix is equal to a given matrix under general different distributions. We use the comparisons above for deciding the optimal copula for generating.

Since the other data sets gave the very same results, we decide for *the t-copula with* 1 *degrees of freedom.*

3.4.4 The optimal sample size

The optimal sample size for generating vectors with t-copula with 1 d.f. is left to find . Our purpose is to generate it in a such way that the sample correlation matrix will be close to the given R and increasing of sample size doesn't improve their distance. There were generated vectors with 7 different sample sizes that were measured by 3 metrics: in picture 3 is 1-norm metric, in picture 4 is the Euclidean metric and in the picture 5 is the infinity norm distance. All pictures are in the appendix. All histograms show that the minimum for generating is 5000.

I remind that there are two discrete marginal random variables in our training set. And the minimum 5000 samples were established in case of discrete or more exactly, multinomial distributions. And why are we talking about this? Because the multinomial distribution makes trouble in sample correlation matrix - however, the binomial distribution behaves the same way as other distributions (it is also approximately normal for large n and p not too close to 1 or 0).

If there are generated vectors without multinomial marginal distributions, then we get the following histograms of metrics: picture 7 for the Euclidean metric, picture 8 for the infinity norm and picture 6 for the 1-norm distance. All pictures are in the appendix.

Now can the means of metrics of correlation matrices generated with t-copula with 1 d.f.

be compared to the given correlation matrix R. This comparison has been made in the table 3.3 depending on the sample sizes and on whether one of the marginal distributions is the multinomial one.

Summarizing this, we can recommend that if $\mathbb{X}$ does not have any multinomial marginal distributions, the sample size can be from 1000 to 1500.

Maximal distance

		Sample sizes						
		100	500	2000	5000	10000	20000	50000
mean value of $\mathbb{X}$	with multinomial	0.20	0.15	0.13	0.13	0.13	0.13	0.13
	without multinomial	0.17	0.10	0.09	0.09	0.09	0.09	0.09

Euclidean distance

		Sample sizes						
		100	500	2000	5000	10000	20000	50000
mean value of $\mathbb{X}$	with multinomial	0.62	0.47	0.44	0.43	0.43	0.43	0.43
	without multinomial	0.51	0.32	0.27	0.25	0.25	0.25	0.24

1-norm distance

		Sample sizes						
		100	500	2000	5000	10000	20000	50000
mean value of $\mathbb{X}$	with multinomial	2.85	2.18	2.00	1.97	1.94	1.94	1.92
	without multinomial	2.26	1.48	1.25	1.19	1.15	1.14	1.12

Table 3.3: The comparisons of the mean values of the metrics $cor(\mathbb{X}) - R$, where the vectors $\mathbb{X}$ are generated with t-copula with 1 d.f. with different sample sizes, with multinomial marginal distributions and without them.

3.5 The choice of random distributions

There are many random distributions that are used in different situations, for different purposes and functions. For solving the problem, described in the first chapter, we should decide which one will we use.

At first, we can begin with distributions that are good, necessary or useful.

For solving our problem, we will surely need the normal distribution. It belongs to the most common distributions that are used everywhere. It is reasonable to suppose that the normal distribution will play a very big role in generating of "random people".

Of course, we should add to our "portfolio of distributions we use" the continuous uniform distribution. It is a basic distribution.

The third distribution will be the gamma distribution. It may be used for expressing wages, heights of book-debt of population or many other things. This kind of distribution is useful for another interesting property: applying it, we may easily get a few related distributions, for example, the most important for us are the exponential and χ^2_ν.

The next distribution that is needed, is the binomial one (or in general, multinomial). It is very important distribution, since it allows us to work with discrete random variables. Some variables, like sex, a number of children or city aren't continuous, hence it is useful to include it to the distributions we can use.

These were necessary distributions, we must be able to work with. But our goal is to generate random samples as well as we can. Therefore we choose a few more.

The beta and the logistic distributions are the two remaining distributions that we will use for generating.

¿From further results of our numeric study follows that the other distributions, like lognormal or Cauchy, give bad results - a big dispersion of sample correlation coefficients or, moreover, unbiased, i.e. the mean values of 10000 sample correlations of lognormal and the other random variables are about $\sim \mid 0.4 \mid$ higher than a specified correlation.

Note that the multinomial distribution gives the biggest dispersion from all distributions. So, the less multinomial distributions are generated, better accuracy we have.

Summarizing this, we can use the following distributions:

- uniform distribution,
- normal distribution,
- exponential distribution,
- beta distribution,
- gamma distribution,
- logistic distribution,
- discrete binomial (multinomial) distribution.

These distributions are the only ones we will use for our further work.

3.6 Summary

Now we summarize all information of our study.

The best for our purpose is the generating with t-copula with 1 degrees of freedom. The allowed marginal distributions are uniform, normal, exponential, beta, gamma, logistic and discrete multinomial distributions.

It is enough to generate 1500 vectors to stabilize the sample correlation matrix if there is no marginal multinomial distribution. Otherwise, it is recommended to generate at least about 5000 vectors.

Chapter 4

Application

In this section we will study the real data we have. At first we will discuss them and make an analysis of an expected potential of every variable. On the basis of these results we choose about 6 the most potential variables. Then we will use some transformation and helpful functions to find marginal distributions. At the end the dependence structure between them will be calculated.

4.1 Data

The data we have are related to applicants for a credit product. It came from one of the Czech banks.

We have 1225 samples with 15 parameters. The parameters are the age, number of children, number of other dependents, an existence of a home phone, applicant's income, spouse's income, applicant's employment status (i.e. private or public sector, or military or student, etc.), residential status (owner, tenant furnished etc.), value of home, mortgage balance outstanding, outgoings on mortgage or rent, outgoings on loans, outgoings on hire purchase, outgoings on credit cards, and the last parameter is good/bad indicator that denotes the defaults.

Now we can start studying the data and try to find any outlying points or find out normality or non-normality.

At first if we realize what kind of data we have, we can't expect normality of our data and normality of some joint distributions neither. But we don't know anything about existing outlying points. In the 1^{st} chapter we have determined our basic goal which is generating with the given correlation matrix and marginal distributions. But outlying points could make our decision about sampling marginal distribution difficult.

In the next subsection we will study every marginal distribution and theirs outlying points separately.

For completeness the basic descriptive statistics of all variables are shown in the appendix (in the table 1).

4.1.1 The choice of generated variables

The data have 15 parameters, including the successful credit repayment. But not every parameter gives us new information about the possibility to repay. For every parameter will be made analysis about its potential. This analysis can appear either from the data or we can decide basing on an intuition or a guess. If the parameter is chosen, then the deeper analysis of it will be made in the next subsection. If it is not, then the reasons will be stated.

The age of the applicant

The first parameter is The age of the applicant. It is widely known that it is one of the most important variables determining repayment ability. So, it is our first variable which will be generated.

The number of children

The second surveyed parameter is The number of children. That is one of five discrete parameters in the data.

The basic summary can be seen in the contingency table 2 in the appendix that shows how much cases of different number of children we have in our data, and how it depends on the result of credit application.

Without any loss of generality we may combine the applicants with 5 and 4 children.

We are now able to test a hypothesis, whether Good/bad indicator and Number of Children are independent. After using the χ^2-test of independence we get that a critical value is $\chi^2_4(0.05) \doteq 9.49$ and it is bigger than χ^2-statistic $\doteq 4.18$. Therefore, we don't reject the null hypothesis that there is no relationship between repayment and the number of children.

The visual data validation is clearly seen from the table 3, where expected values are calculated on the basis of marginal distributions of surveyed two variables (i.e. under the null hypothesis) and the table 2, which are in the appendix.

It means that there is no reason to generate this parameter because it gives no information explaining the defaults. The probability of a success is 0.74 no matter what the number of children is.

The number of other dependents

The next parameter is The number of other dependents.

We claim again that this variable bring us no new information: the correlation between Number of other dep. and Number of children is > 0.98. Hence, a generating of this variable is useless.

Phone owner

The third discrete parameter is Phone owner.

We make a contingency table again, when a dependence between Phone owner and the result is - the table 4 in the appendix.

The χ^2-test testing, whether Phone owner is independent on Good/bad indicator, does not reject the null hypothesis on the significance level 0.05. So, that parameters is useless too.

Spouse's income and Applicant's income

Now we have two continuous parameters Spouse's income and Applicant's income. The income is surely one of the most important (according to the risk managers of the banks it is the most important) factor in the repayment of a credit. Therefore both parameters should be generated.

Applicant's employment status

One of two last discrete parameter is Applicant's employment status. It can be one of the following statuses: government, housewife, military, private sector, public sector, retired, self employed, student, unemployed, others and no response.

In the table 4.1 is shown the comparison between the Employment status and Good/bad indicator. After the adding the status Others to Private and statuses No response and Unemployed to Retired (statuses with the similar rate (Good indicator \ bad indicator)), our table satisfies the condition $n_{i.}n_{.j}/n > 5$, where $n_{i.}$ denotes the column marginal totals and $n_{.j}$ denotes the row marginal totals, $n = 1225$ is the grand total. Now the condition about the limit distribution is satisfied, and it can be made the χ^2-test. The statistic of it is $45 > 15 = \chi^2_8(0.95)$, so the table shows contingency between the two variables. It means, this variable is useful for us and should be generated.

Residential status

The last discrete parameter is Residential status. There are 5 statuses in our data: owner, tenant, furnished, tenant, unfurnished, with parents.

The dependence Good/bad indicator on Residential status is on the following contingency table 5 in the appendix.

The p-value of χ^2-test is 0.32, and we don't reject the independence of the variables on significance level 0.05. So, we may claim that generating of this variable gives us nothing new.

Value of home

The ninth parameter is Value of home. The p-value of the Kolmogorov-Smirnov test about, whether a distribution of the parameter is the uniform distribution, is 0.134. Based on the discussion with risk management in a bank, we found that the value of home has an influence on the capital liability, but not on the repayment ability. Hence, this variable is for our purpose useless.

Mortgage balance outstanding

How much debt do people have? The answer on this question gives us this parameter.

If we see in the appendix in the table 6, and make the χ^2-test, we get the p-value = 0.26. Hence, we don't reject the hypothesis about the independence of Mortgage balance outstanding and the Good/bad indicator.

And moreover, at the first glance it can be strange, but it has almost the same distribution like the previous parameter, and the correlation coefficient between them is 0.64. It could be caused by the dependence of a house value on a mortgage. In advanced countries the most people buy houses on a mortgage. So, the generating of this variable gives no information explaining the defaults.

Outgoings

The further parameters are Outgoings on mortgage or rent, Outgoings on Loans, Outgoings on Hire Purchase and Outgoings on credit cards. Of course they use for a calculation of applicant's creditworthiness as his potential costs. On the other hand, they belong to the less important factors. The most values of these outgoings are equal zero in our data, and distributions of outgoings also look very similar

That were the reasons, why only 1 parameter of all outgoings will be generated. It is Outgoings on mortgage or rent. At first, it has the smallest number of zero values of all of outgoings. Secondly, it is very often the biggest cost of the household budget and thirdly it says much about a solvency of applicant as a regular cost.

Good/bad indicator

The last parameter is an indicator saying whether the credit were payed. All parameters before now were explanatory variables, and this indicator is a response variable. Good means the paid-up credit and Bad is a default.

4.1.2 The fitting of marginal distributions

In the last subsection we dealt with each parameter separately. We said that the following 6 variable will be generated: Good/bad indicator, Age, Applicant's income, Spouse's income, Applicant's employment status and Outgoings on mortgage and rent. We will study these variables now.

We will try to find theoretical distribution functions being up to our samples, make tests that will help us with verifying it, and may be make small discussions about variables.

For determination of the parameters of cumulative distribution functions, like a rate in the exponential distribution, or a shape and a scale in the gamma distribution, the MLE-fitting (Maximum-Likelihood Fitting of Univariate Distributions) is used, more information can be found in [1].

Unfortunately, most of surveyed marginal distributions have the shape of some "standard" distribution (like normal, beta etc.), but its domain is bigger than a domain of a "standard" distribution.

For example: the variable $X \sim B(1,2)$ has Beta distribution, so $X \in [0,1]$. The surveyed variable Y has the same shape of density like X, but $Y \in [0,c]$. So $Y = c * X$, where $c > 0$ is a constant. The distribution of the Y is $F_Y(y) = F_X(y/c)$. The question is how to find the constant c?

There are two methods:
if X has the Beta distribution (like in our example), then we denote $c = max_i(X_i)$, where X_i are the samples. But if X has other distribution, like gamma or exponential, then we can not use this method.

Hence, we make a little trick - we use a helpful function: it finds such c that the expression $D_n = \sup_x |F_{Xn}(x/c) - F_Y(x)|$ takes the minimum for all parameters of Y, where $F_Y(x)$ is a distribution function of Y, and $F_{Xn}(x)$ denotes a sample distribution function of X. The D_n

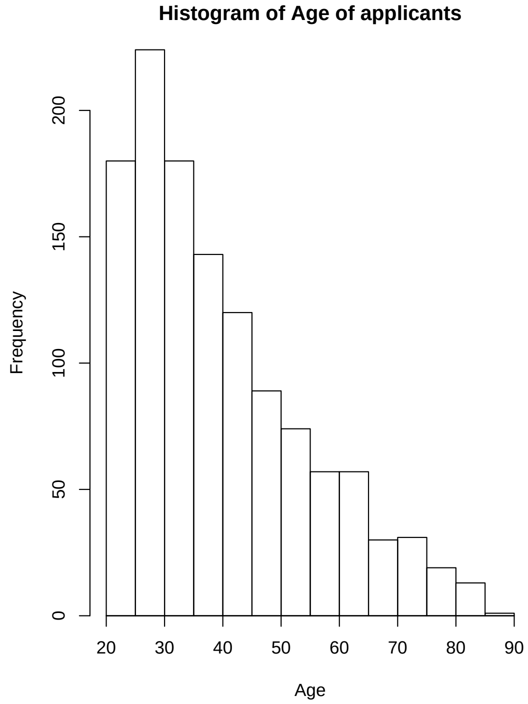

Figure 4.1: The histogram of Age of applicants

is the Kolmogorov-Smirnov statistic. By the Glivenko-Cantelli theorem, if D_n converges to 0 almost surely, then the sample $F_{Xn}(x/c)$ comes from the distribution $F_Y(x)$.

After the c is known, we use the MLE-fitting for the variable $F_{Xn}(x/c)$.

Before we go any further, let $X_i, i \in 1, \ldots, 6$ denote the random variable with the empirical distribution function $F_n(X_i)$, let $Y_i, i \in 1, \ldots, 6$ denote the random variable with the distribution function $F(Y_i)$, which is the estimated $F_n(X_i)$. Then let the estimate that is got as a result of our function, will be called as the KSFE (Kolmogorov-Smirnov Function Estimate) in this work. We have 4 numbers as a result of this function: 1^{st} and 2^{nd} mean the chosen parameters of the distribution function, 3^{rd} is the Kolmogorov-Smirnov distance for the chosen parameters and 4^{th} number is c.

Age

The first variable is Age of applicant.

At first, the histogram is on the picture 4.1.

Let's try to find the CDF. The histogram predicts possible beta distribution with bigger first parameter α and smaller second β. We need to find the constant c described above. With a view to a discussion above, let $c = max_i(X_i) = 87$

The parameters of $X_1/87$ estimated by the maximum likelihood function is $(2.3, 0.9)$, or the equivalent statistical expression:

$$X_1 = 20 + 87 * Z, \tag{4.1}$$

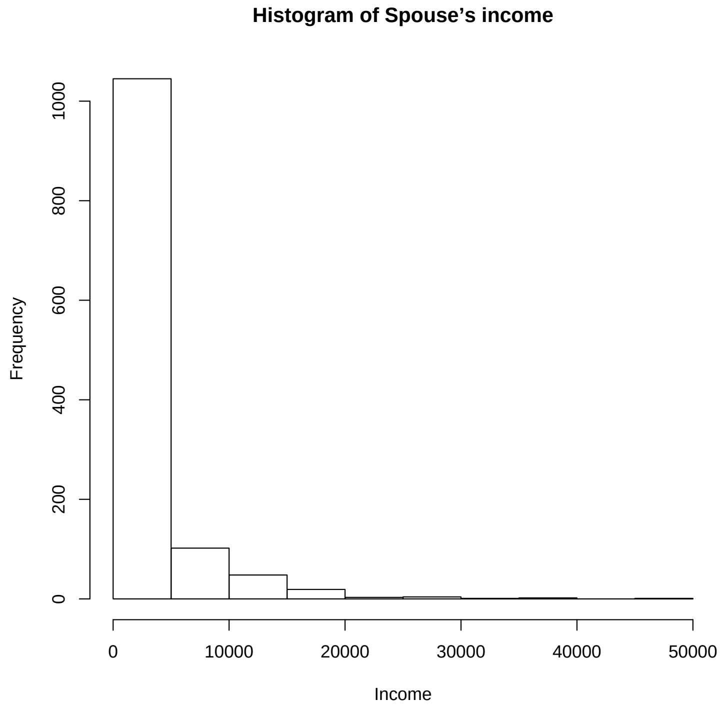

Figure 4.2: The histogram of Spouse's income

where $Z \sim Beta(0.9, 2.3)$.

What is left is to show statistically that X_1 and Y_1 don't differ. We know that 0.034 is the greatest discrepancy between the observed X_1 and expected cumulative distributions Y_1, and the approximative critical value of Kolmogorov-Smirnov test on significance level 0.05 is $D^*_{1225}(0.05) \doteq 0.038$ and it is greater than 0.034. Hence, Kolmogorov-Smirnov test doesn't reject the null hypothesis that the variable X_1 and the expected variable Y_1 have the same distribution function.

Spouse's income

The second variable is Spouse's income.

The histogram of the income is on the picture 4.2. The most values are less than 5000, and we have few values bigger than 20000.

There are 908 zero-values of Spouse's income, and 6 values bigger than 20000, so if we make another histogram of random variable X_2, but with constraints $0 < X_2 \leq 20000$, we get another much nicer histogram 4.3. The shape of 4.3 reminds us one of famous distribution - the gamma distribution.

Using the KSFE we get 4 following numbers:

```
[1] 1.400000e+00 7.000000e-01 4.832866e-02 3.750000e+03
```

c is equal 3750, and MLE gives us the following result:

$$X_2 = 3750 * Z_{21} * Z_{22}, \tag{4.2}$$

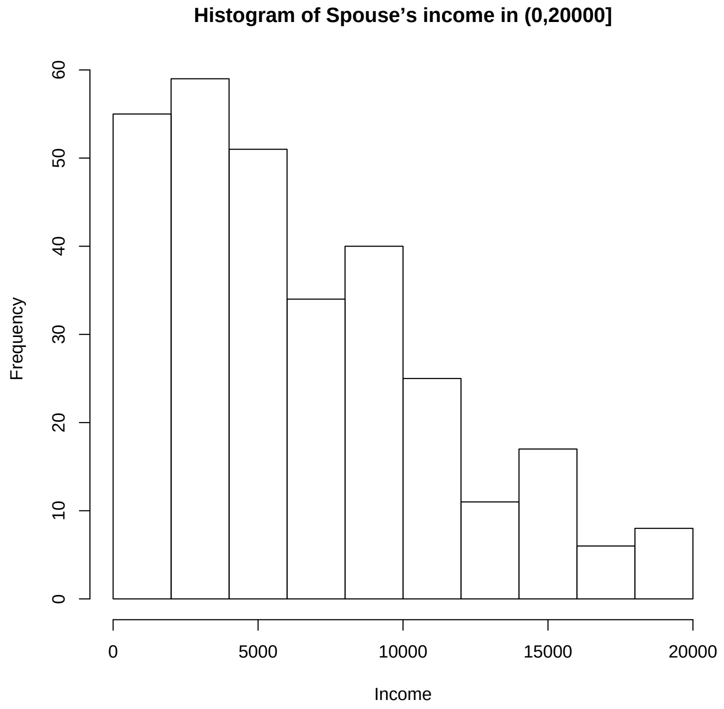

Figure 4.3: The histogram of Spouse's income in interval $(0, 20000]$

where $Z_{21} \sim Bi(1, 0.74)$, $Z_{22} \sim Gamma(1.6, 0.7)$ and Z_{21}, Z_{22} are independent. Also, X_2 could we analyze as a multiple of two variables and constant c: a binomial distribution Z_1 with values $(0, 1)$ and $\mathbb{P}(Z_{21} = 0) = \mathbb{P}(Z_{21} = 0) = 0.74 \doteq 908/1225$, multiplied by $c = 3750$ and by gamma distribution Z_{22} with parameters $k = 1.6$ and $\theta = 0.7$.

The Kolmogorov-Smirnov statistic of ML estimate is 0.101 and the approximative critical value of Kolmogorov-Smirnov test on significance level 0.05 is $D^*_{317}(0.05) \doteq 0.078$. Hence, Kolmogorov-Smirnov test reject the null hypothesis that a given data set could have been drawn from a given distribution. The Kolmogorov-Smirnov statistic for KSFE is 0.048 that is not enough to reject the null hypothesis. But the MLE is more important for us.

In the picture 4.4 we can see the established and empirical distribution functions.

Note that an explicit form of distribution function of X_2 is very complicated, hence there is used a little trick in generating. A vector $\mathbb{Z} \sim 3750 * Gamma(1.6, 0.7)$ is generated and then the each element of $\mathbb{Z}$ is multiplied by $Z_2 \sim Bi(1, 0.74)$. The resultant variable has a required distribution.

Applicant's income

The next variable is Applicant's income. It is the most important data in application of credit.

We can assert that studying of this variable will be the very similar as studying of Spouse's income, so we may try to make the same steps.

The histogram 4.5 looks like the histogram of Spouse's income 4.2, but doesn't have so much zero income - just 206.

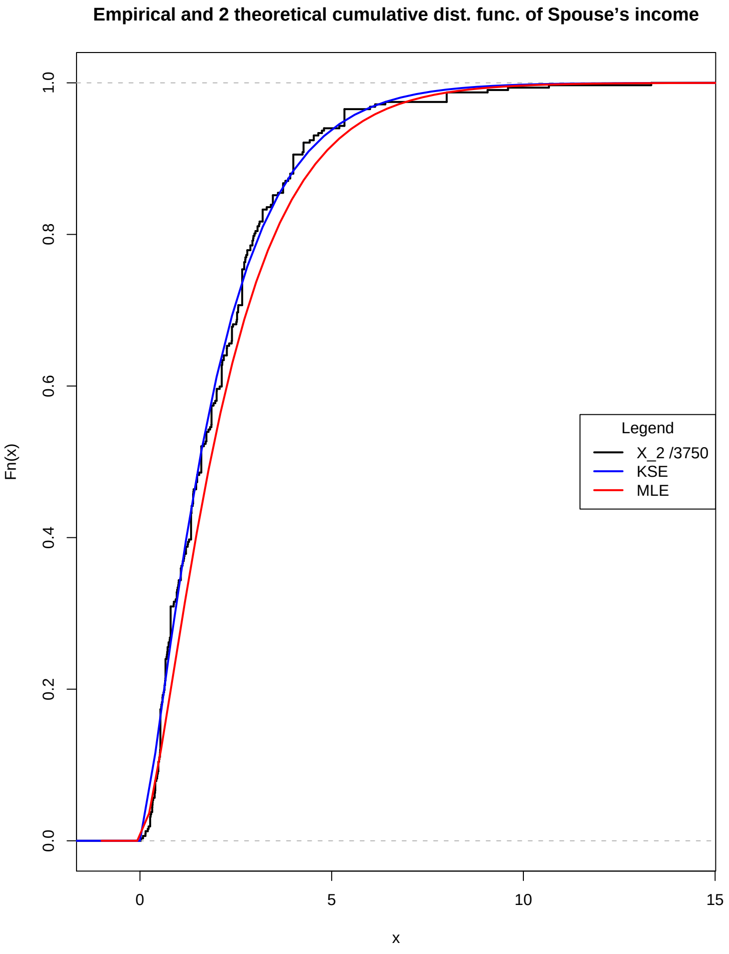

Figure 4.4: The EDF and established CDF of Spouse's income

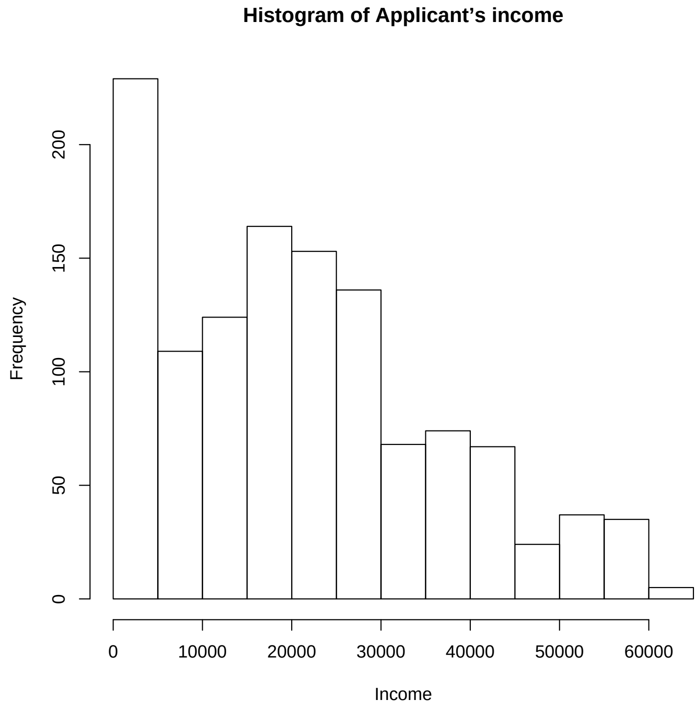

Figure 4.5: The histogram of Applicant's income

Applying KSFE on the Spouse's income data without zero points, we get:

```
[1] 3.000000e+00 9.000000e-01 2.741837e-02 7.800000e+03
```

Then the maximal likelihood estimate with $c = 7800$ is $X_3 = 7800 * Y_3$, where $Y_3 \sim Gamma(2.9, 1)$. The MLE and KSFE are in the picture 4.6.

The Kolmogorov-Smirnov statistic of MLE is 0.11, and a critical value on the significance level 0.05 is 0.043. So, in this case the test again rejects the null hypothesis. If it were tested the KSFE, it would have not been rejected on the significance level 0.05.

Applicant's employment status

One of two lase discrete variable is Applicant's employment status. It can be one of the following statuses: government, housewife, military, private sector, public sector, retired, self employed, student, unemployed, others and no response.

In the table 4.1 we can see that we have too little values of N (=No response), U (=Unemployed) and O (=Others) and some others. It will be good to reduce a number of statuses.

After combining statuses with similar probability of default we get 3 groups:

1. Government and Others,
2. Public sector, Private sector, Self employed, Military and Student,
3. Housewife, Unemployed, No Response and Retired.

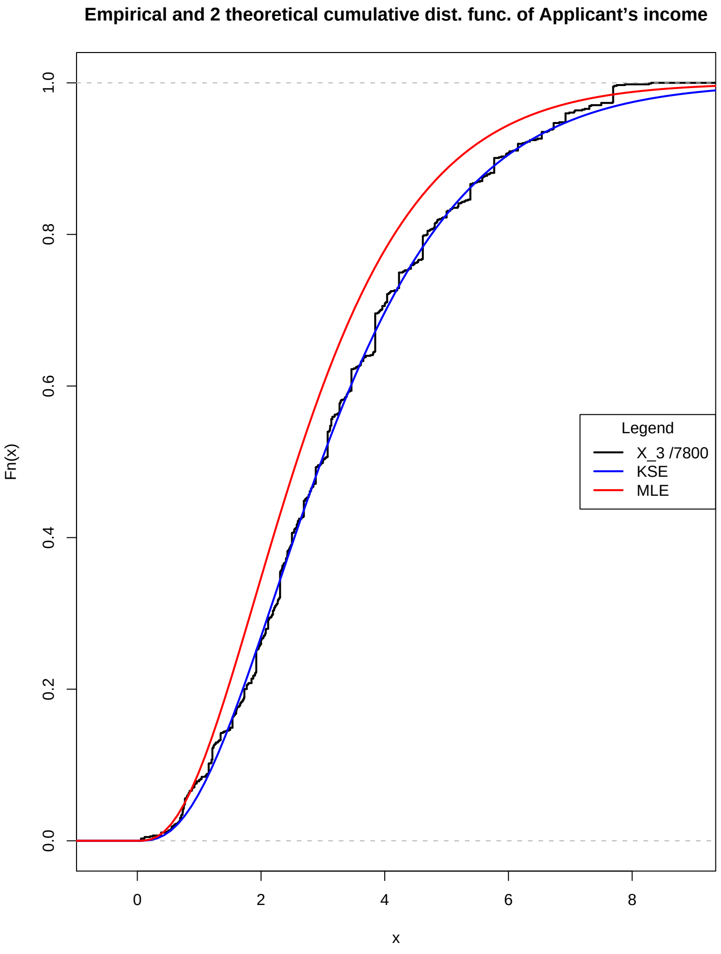

Figure 4.6: The EDF and established CDF of Applicant's income

		Employment status											Total
		P	Self	M	O	Priv	Retired	St	Unempl	Gov	HW	N	
Good/bad indicator	Good	22	88	17	5	413	55	86	4	187	21	4	902
	Bad	8	36	6	1	118	49	37	4	44	16	4	323
Total		30	124	23	6	531	104	123	8	231	37	8	1225

Table 4.1: The Employment status vs. Good/bad indicator

The result of configuration is interesting, but not surprising.

In first group are 831 members with mean probability of a successful repayment 0.75, in second - 157 with 0.53, and in third - 237 with 0.82.

Summarizing, the variable X_4 has the multinomial distribution with probabilities $(0.68, 0.12, 0.2)$.

Outgoings on mortgage or rent

The further variable is Outgoings on mortgage or rent.

As it is usual in outgoings, there are almost half zero-points - 526. A lot of people (or even most of them) have no outgoings on mortgage or any other loans because they don't want to be in debt.

After eliminating zero values, we use our programme to find a distribution that is close to empirical one. KSFE gives us the following result:

```
[1] 2.600000e+00 1.140000e+01   0.04738909 3.150000e+03
```

This time we will approximate with beta distribution. $c = 3150$, and then MLE is $X_5 = 3150 * Z_{51} * Z_{52}$, where $Z_{51} \sim Bi(1, 0.57)$, $Z_{52} \sim Beta(1.7, 8)$ and Z_{51}, Z_{52} are independent.

The graphical illustration is on the picture 4.7.

The approximation with Y_5 distribution has the Kolmogorov-Smirnov statistic 0.079, it is bigger than the critical value 0.051, hence the Kolmogorov-Smirnov tests rejects that this two variables have the same distribution.

Good/bad indicator

There is one variable left. This is the binomial distribution with the probability of success $(\mathbb{P}(X_7 = 0))$ is 0.74. The Good/bad indicator could be expressed as $X_6 \sim B(1, 0.26)$.

4.1.3 Estimate of the correlation matrix R

Now we know, which variables X_i will be generated and also which marginal distributions have X_i. Therefore a sample correlation matrix of X_i is left to find.

In the previous chapter it was noticed that we will generate with the Pearson correlation matrix - it is shown on the table 4.2.

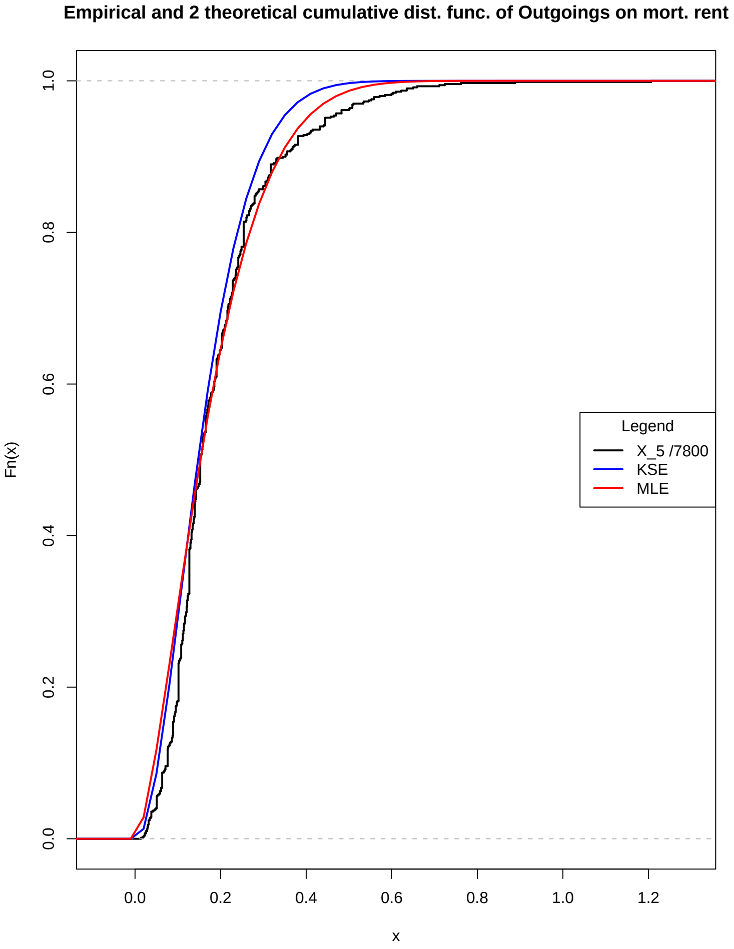

Figure 4.7: The empirical and two Beta distributions

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.00	−0.11	−0.18	−0.06	0.05	−0.11
X_2	−0.11	1.00	0.12	0.01	0.13	−0.06
X_3	−0.18	0.12	1.00	0.06	−0.03	−0.01
X_4	−0.06	0.01	0.06	1.00	0.39	−0.19
X_5	0.05	0.13	−0.03	0.39	1.00	−0.07
X_6	−0.11	−0.06	−0.01	−0.19	−0.07	1.00

Table 4.2: The estimated Pearson correlation matrix of chosen variables

4.2 Summary

We have found 6 marginal distributions and the correlation matrix. Now we know what kind of marginal distributions will be generated, what dependence structure is between them, and in the previous chapter we have chosen the *t*-copula with 1 d.f. for generating. Moreover, because here we have two discrete variables about 5000 sample sizes are recommended.

Chapter 5

Generation and verification

In the previous chapters we concerned the methods of generating with copulas, then we found the optimal copula in the sense of metrics and the sample size. Further we analyzed the real data. And in this chapter will we combine our present results: at first, samples with the marginal distributions and correlation matrix from the last chapter are generated. Then become verifications, whether these samples have the required marginal distributions, whether they also have the required correlation matrix and whether the generated defaults have the same distribution as the training set.

The most results in this chapter are based on the methods or technics from the previous chapters.

5.1 Generating samples

Before we begin, let $Y_i, i \in \{1, \ldots, 6\}$ be random variables with required marginal distribution, $X_i, i \in \{1, \ldots, 6\}$ be random variables the sample marginal distributions created by simulation. Further, let R be the required correlation matrix.

We want to generate the sample using the t-copula with one degree of freedom. The marginal distributions of $\mathbb{Y}$ are more described in the section 4.1.2. The correlation matrix R is in the table 4.2.

The sample size of each sample is 5000. There are also made 10000 samples.

We repeat that the significance level is 0.05, if it is not stated otherwise.

5.2 Verification

There are three conditions of samples that should be satisfied:

1. distribution of X_i is equal to Y_i, for all i
2. the sample correlation matrix $cor(\mathbb{X})$ is "close" enough to R
3. the distribution of defaults (variable Good/bad indicator) is for every X_i the same as for every Y_i

The first condition is absolutely clear. The second one says that some metric of $(cor(\mathbb{X}) - R)$ should be close to zero. We use the same metrics as before: the 1-norm distance, the Euclidean metric and the infinity-norm distance.

The third condition requires the same amount of defaults for Y_i and X_i on the same segment of a domain of i^{th} distribution function.

5.2.1 Equality of marginal distributions

The first condition that generated samples have to be satisfied is the equality of marginal distributions. The Kolmogorov-Smirnov test is used to verify it. And because we made 10000 samples, there is showed the means and medians of p-values of tests for every variable on the table 5.1.

	Variable					
	X_1	X_2	X_3	X_4	X_5	X_6
Mean	0.51	0.52	0.50	0.89	0.50	0.94
Median	0.51	0.52	0.51	0.99	0.50	0.99

Table 5.1: The mean value and median of p-values of Kolmogorov-Smirnov test for every variable

The most of p-values of Kolmogorov-Smirnov test for X_4, resp. X_6, were 0.999. These ones are the multinomial variables with three, resp. two, possible outcomes and for a sample size 5000 is the Kolmogorov-Smirnov statistic incredible small.

The results of this table confirm our claim that the generated samples have the required marginal distributions.

5.2.2 Correlation matrix

The second condition we have is an accuracy of sample correlation matrix.

We have generated 10000 samples, hence there are 10000 sample correlation matrices, and therefore we may make the confidence intervals of each correlation coefficient.

In the table 5.2 the confidence intervals and the mean of the sample correlation coefficients are compared with the given correlation coefficients. We see that the most given correlations are in the confidence interval, and just 4 are out. And just 2 of them are far enough from it - it is the correlation coefficients the $cor(X_4, X_5)$ and $cor(X_4, X_6)$. I remind that X_4 is the multinomial distribution, and we have said that multinomial distributions make bad results in correlations.

And since the distribution X_6 is the multinomial too, the results of the table 5.2 are better, than we could expected.

We can also make the histograms of metrics of differences of correlation matrices we use before and compare it with the results of the metrics of differences correlation matrices generated in the section 3.4.

The histogram of the 1-norm metric is in the picture 5.1. The mean value is 1.65. When we made research about which copula is better, we made the histograms of 1-norm metric for t-copula with 1 d.f. and the sample size 5000 (picture 6 in the appendix). But if these histograms are compared, the first one has smaller value. Moreover, if we tested the hypothesis whether mean values of both distributions are the same or the first one is smaller, then on the significant level 0.05 the t-test would rejected the null hypothesis.

	X_2	X_3	X_4	X_5	X_6
X_1	$(-0.12, -0.5)$ $-0.08; -0.11$	$(-0.18, 0.09)$ $-0.13; -0.18$	$(-0.13, -0.06)$ $-0.9; -0.06$	$(-0.03, 0.05)$ $0.01; 0.05$	$(-0.15, -0.09)$ $-0.12; -0.11$
X_2		$(0, 0.08)$ $0.05; 0.12$	$(0, 0.6)$ $0.02; 0.01$	$(0.2, 0.9)$ $0.05; 0.13$	$(-0.06, 0.01)$ $-0.01; -0.06$
X_3			$(0.04, 0.1)$ $0.07; 0.06$	$(-0.03, 0.04)$ $0.01; -0.03$	$(-0.01, 0.6)$ $0.03; -0.01$
X_4				$(0.18, 0.24)$ $0.21; 0.39$	$(-0.07, -0.12)$ $-0.1; -0.19$
X_5					$(-0.07, 0)$ $-0.03; -0.07$

Table 5.2: The sample confidence intervals (upper line), the mean of the sample correlation coefficients (first number in the lower line) and the given correlation coefficients (the second number in the lower line). The confidence interval is red marked, if the required correlation coefficient is out the confidence interval

It means, in the sense of this metric are the correlation matrices of samples closer to R than the samples we have generated in the section 3.4.

The next metric is the Euclidean metric. The histogram of it is on the picture 5.2. The mean value and a median are equal 0.22. This histogram could be compared with one of the histogram on the picture 7 in the appendix, where is written Sample size 5000. This histogram has the mean value 0.24 and a median 0.22. Because $1.55 \leq 1.64 = t_n^{-1}(0.95)$, the t-test does not reject that mean values of two metrics are the same.

The last metric is the infinity-norm metric. The histogram is on the picture 5.3. The mean value is 0.19. Again we can compare it with the respective variable on the picture 8 in the appendix, the mean value of it is 0.10. It is clear that the t-test rejects the hypothesis about the equality of both variables.

Summarizing this, the given correlation coefficients are at most in the confidence intervals. The multinomial marginal distribution X_4 has two of four correlation coefficients that are out of the confidence intervals.

We have also three metrics and three different results: one significantly proved that correlation matrices of vectors generated in this chapter are better than correlation matrices generated in the section 3.4. The second metric shows no difference between them, and the third metric gives the opposite result than the first metric gives. On the other hand,

5.2.3 The frequencies of defaults

At last, there is required that the relationship between explanatory variables and default indicator is kept in the generated samples.

For a verification of it the contingency table and then the χ^2-test are used.

Age

Let's see on the first variable The age. In the table 5.3 is the dependence of defaults in the real and generated data.

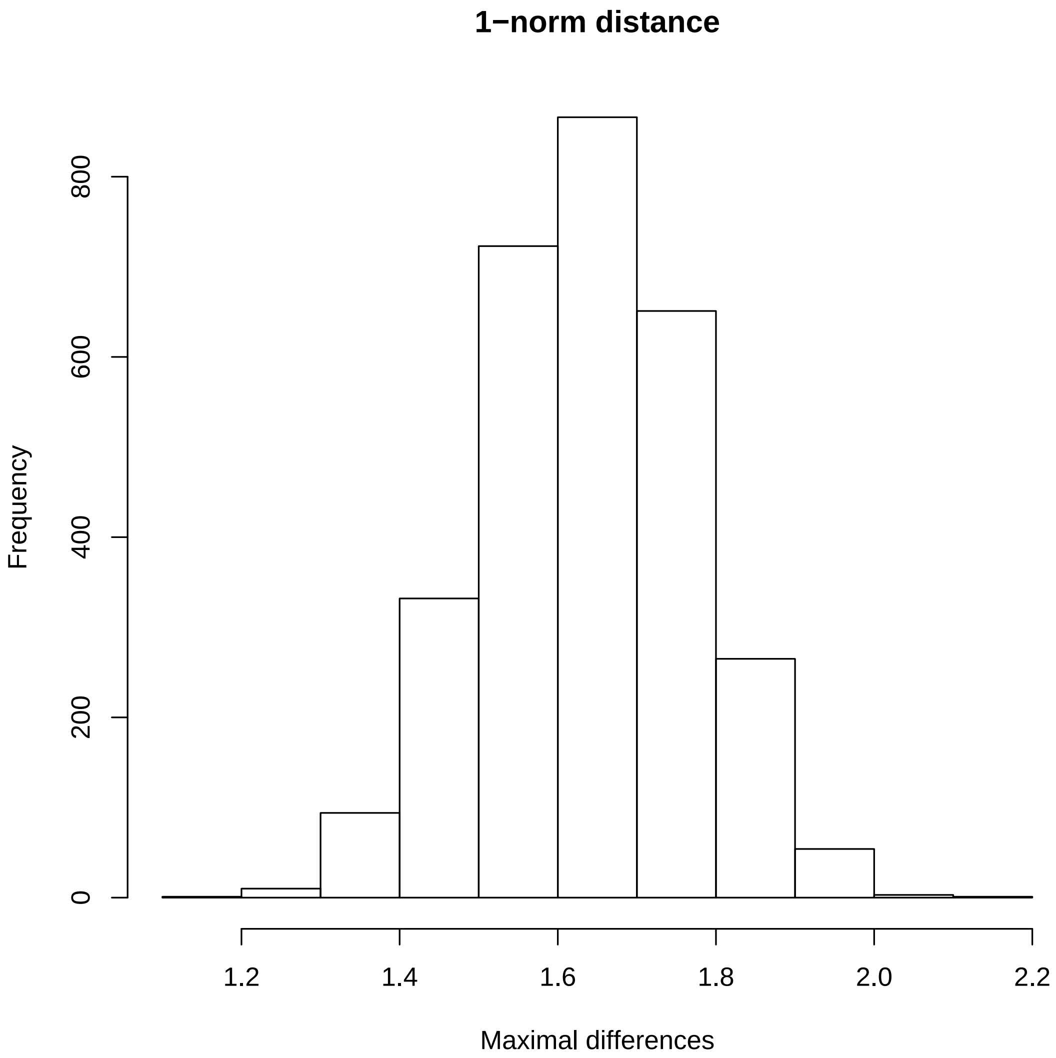

Figure 5.1: The histograms of $\sum_{i,j} |cor(\mathbb{X})_{i,j} - \rho_{ij}|$ for samples generated with t-copula with 1 d.f.

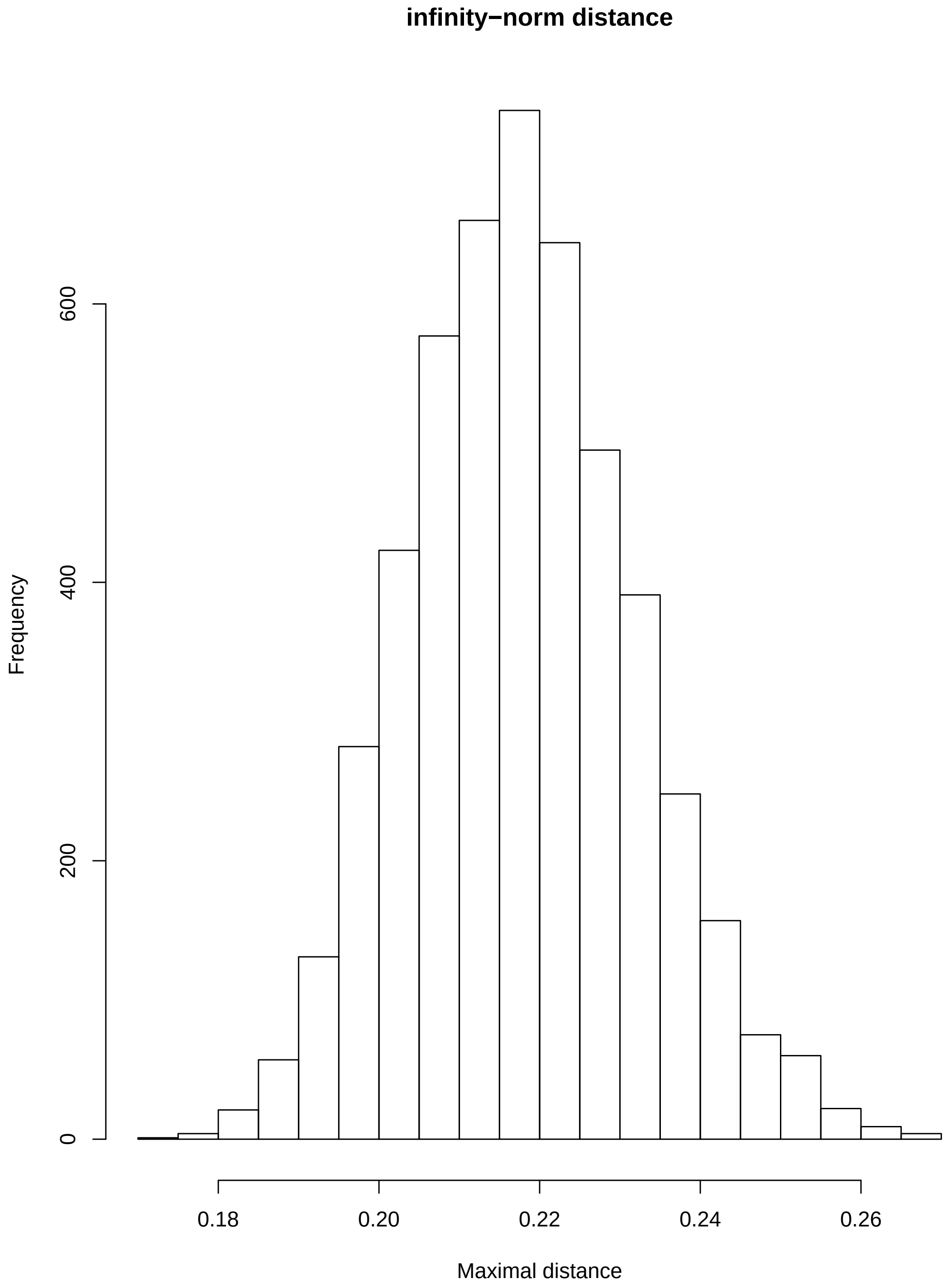

Figure 5.2: The histograms of $\sqrt{\left(\sum_{i,j}\left(cor(\mathbb{X})_{i,j}-\rho_{ij}\right)^2\right)}$ for samples generated with t-copula with 1 d.f.

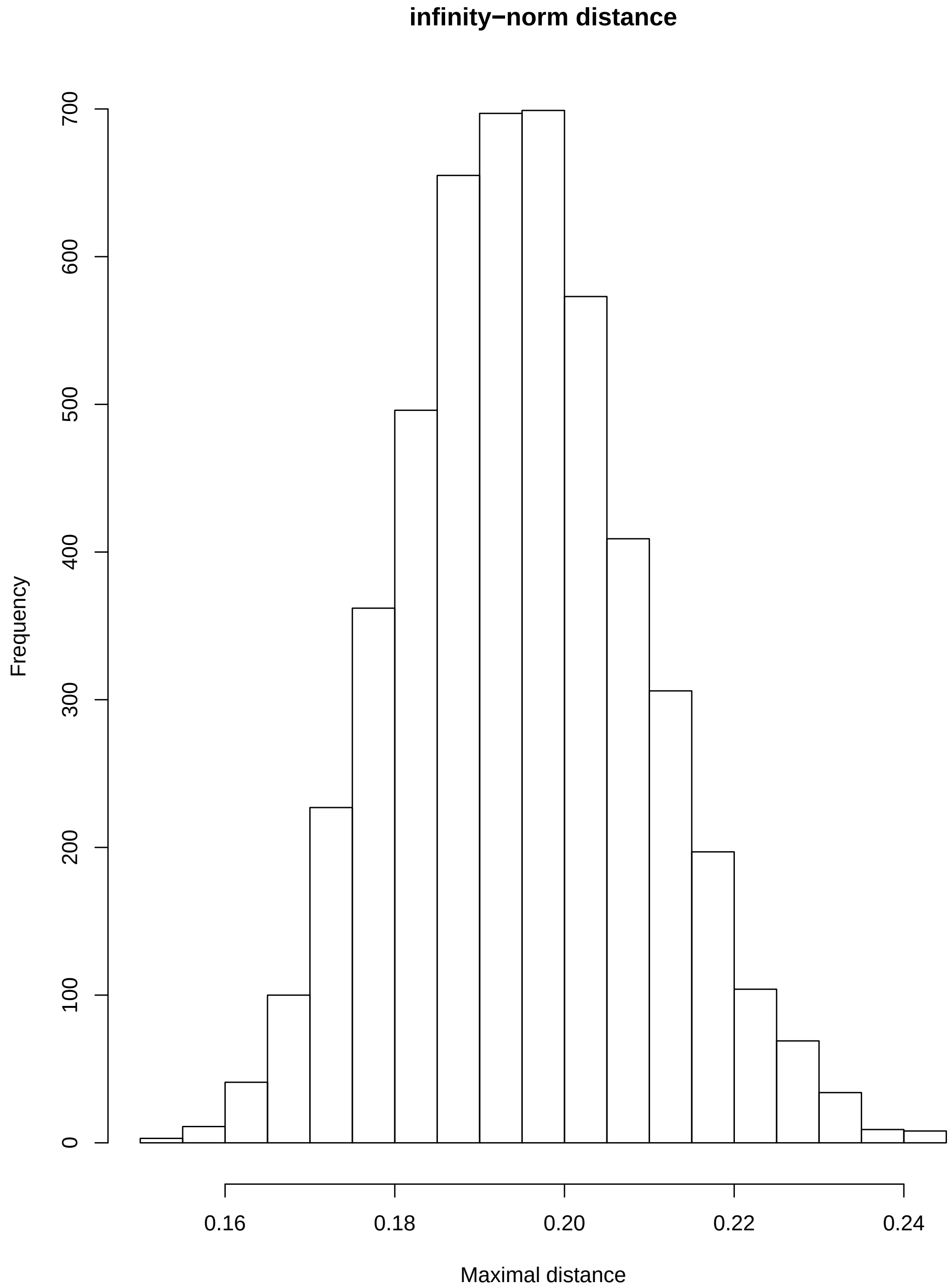

Figure 5.3: The histograms of $\max_{i,j} |(cor(\mathbb{X})_{i,j} - \rho_{ij}|$ for samples generated with t-copula with 1 d.f.

	The age					
	≤ 30	$30-40$	$40-50$	$50-65$	$65-75$	$75 \leq$
% of defaults in real data	21	10	16	35	17	1
mean % of defaults in generated data	31	15	14	25	12	3

Table 5.3: The percentages of defaults in the real and generated data depending on the age of an applicant

The p-value of the χ^2-test is 0.25, it means we don't reject the null hypothesis that the the percentages of defaults are the same in the real and generated data.

Spouse's income

The second variable is Spouse's income. If we see whether the percentages of defaults in the real and generated data are dependent, we get that the χ^2-test's p-value is 0.86. Therefore, the null hypothesis about independence of defaults can not be rejected. The contingency table is the table 5.4.

	Spouse's income					
	$=0$	$1-4000$	$4001-8000$	$8001-14000$	$14001-20000$	$20000 \leq$
% of defaults in real data	78	8	7	4	2	1
mean % of defaults in generated data	74	10	5	5	3	3

Table 5.4: The percentages of defaults in the real and generated data depending on the spouse's income

Applicant's income

The next variable we study is the Applicant's income. The table of defaults is in the appendix, table 7. And the p-value of the χ^2-test is 0.14, it means we don't reject the null hypothesis.

Applicant's employment status

The only discrete variable is the Applicant's employment status. The p-value of the χ^2-test is 0.61. So the null hypothesis hasn't been rejected in this case either. The table of defaults 8 is in the appendix again.

Outgoings on mortgage or rent

The last variable is the Outgoings on mortgage or rent. The χ^2-test doesn't reject the null hypothesis about the independence of the defaults in real and in generated data. The p-value is 0.6. The table of defaults 9 is in the appendix.

	1^{st} gr.	2^{nd} gr.	3^{rd} gr.	$1^{st}+2^{nd}$ gr.	$1^{st}+3^{rd}$ gr.	$2^{nd}+3^{rd}$ gr.	all gr.
RD	25 %	47%	18%	28%	24%	29%	26%
GD	27 %	30%	26%	26%	27%	25%	26%

Table 5.5: The measure of defaults for different groups in RD and GD. An abbreviation "gr" in the table means group

5.2.4 The distribution of defaults

The last comparison the generated samples with the given data consists in the comparison the relation the defaults to the total number of credits for each variable for the given and the generated data. Mathematically speaking, we study the values

$$\frac{\#(X_6|X_i < n)}{\#(X_i < n)},$$

where $n \in \{\text{domain of } X_i\}$, and $\#(Y > 0)$ denotes the amount of values Y, which are bigger than zero.

If the these values for given data and the generated one will be close to each other for all n, then the generated data reflect the default's dependence too. If these values for one data set will generate the line, then this data set gives no new information about the defaults. One more remark, all histograms in this subsection don't have zero points - just for a graphical lucidity.

The first graph 5.4 shows that the most risk group of applicants is between $20 - 25$ years old. The applicants, which are older than 30, have the same risk rate. The generated data describe the defaults very exactly.

In the second graph 5.5 are the histograms of spouse's income for RD and GD and the measure of defaults to the total number of credits for RD and GD. We see that the measures for RD and GD are almost the lines, which are very similar. That is very surprising result, because it means that the spouse's income does *not* depend on the repayment.

The next graph 5.6 describes the measures for the applicant's income. It is seen that here the measure of defaults depends on the level of the income. On the other side, the measure of defaults in RD is not so similar with the measure of defaults in GD as in the previous variables. But it is still close enough.

The further variable is the employment status. It is a discrete variable, so here we show the table 5.5, which describes us, what happen with the defaults for different groups of statuses (more about the groups of statuses is in the section 4.1.2). There are some bigger differences in some groups. That can be caused by that that employment status is a discrete variable, and as we said before, discrete variables may cause problems.

The last variable is the outgoings on mortgage/rent. The measures of defaults to the total number of credits is on the picture 5.7.

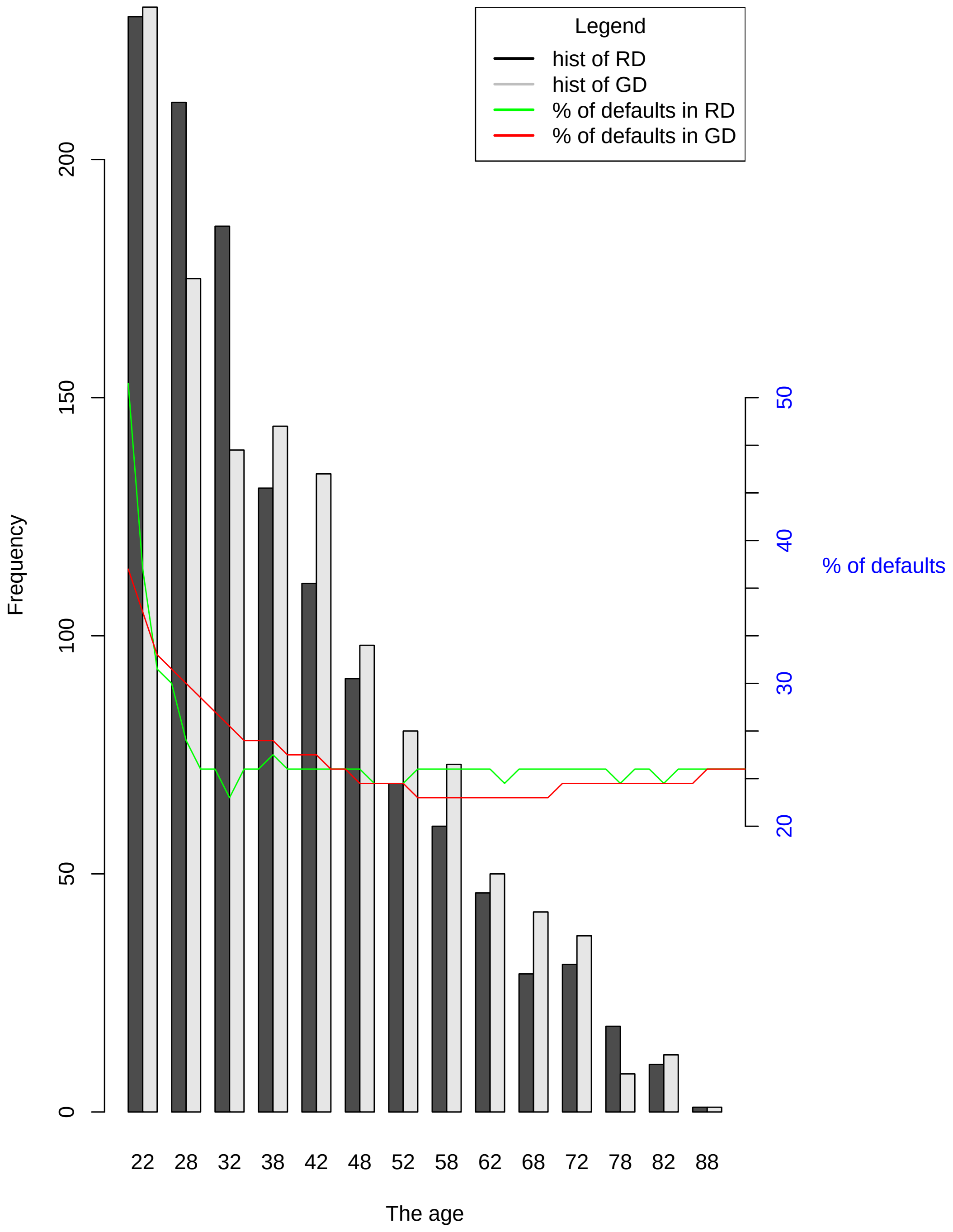

Figure 5.4: There are histograms of the age of the real data (RD) and the generated data (GD) and the curves, meaning the relation the defaults to the the total number of credits, for RD and GD

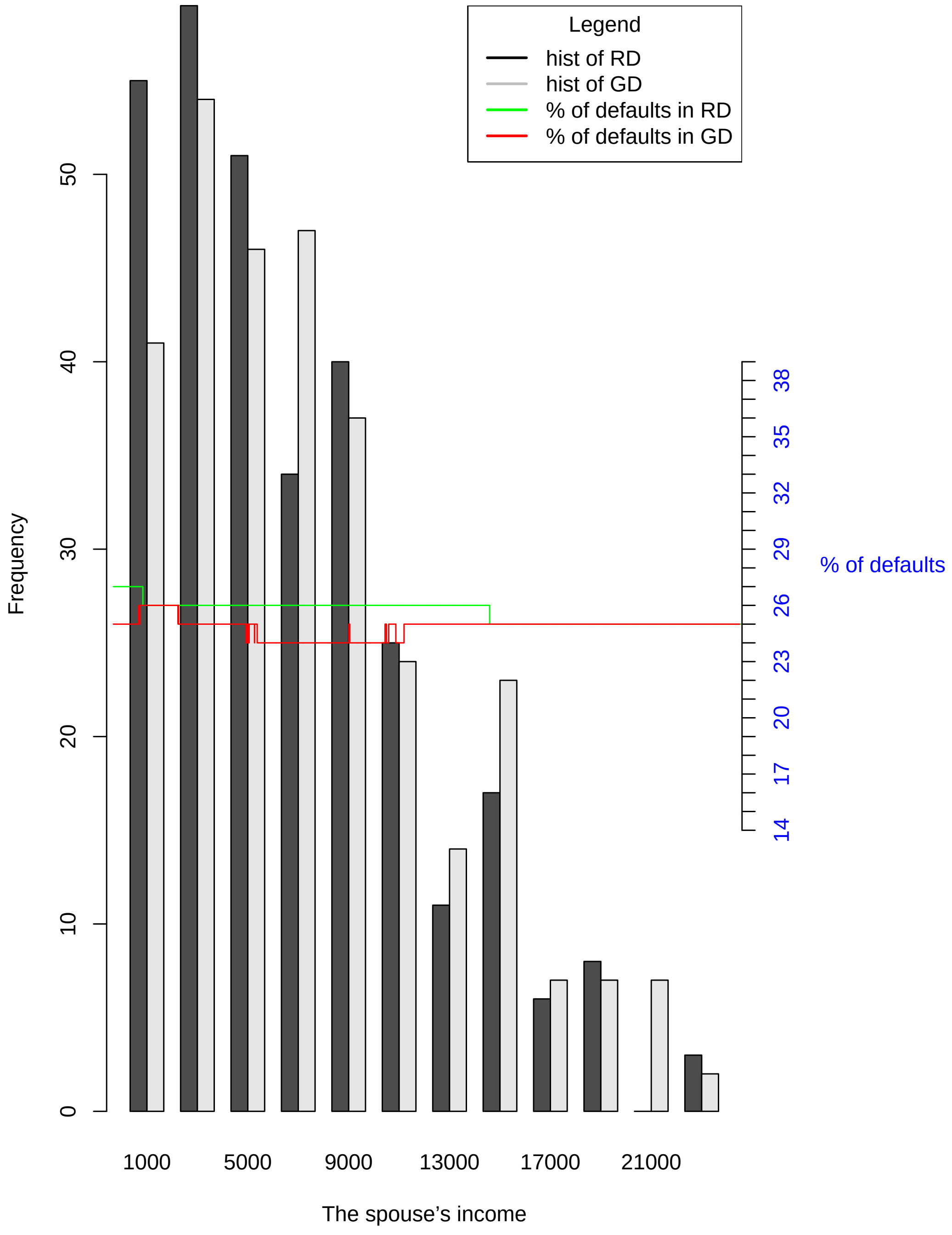

Figure 5.5: There are histograms of the spouse's income of the real data (RD) and the generated data (GD) and the curves, meaning the relation the defaults to the the total number of credits, for RD and GD

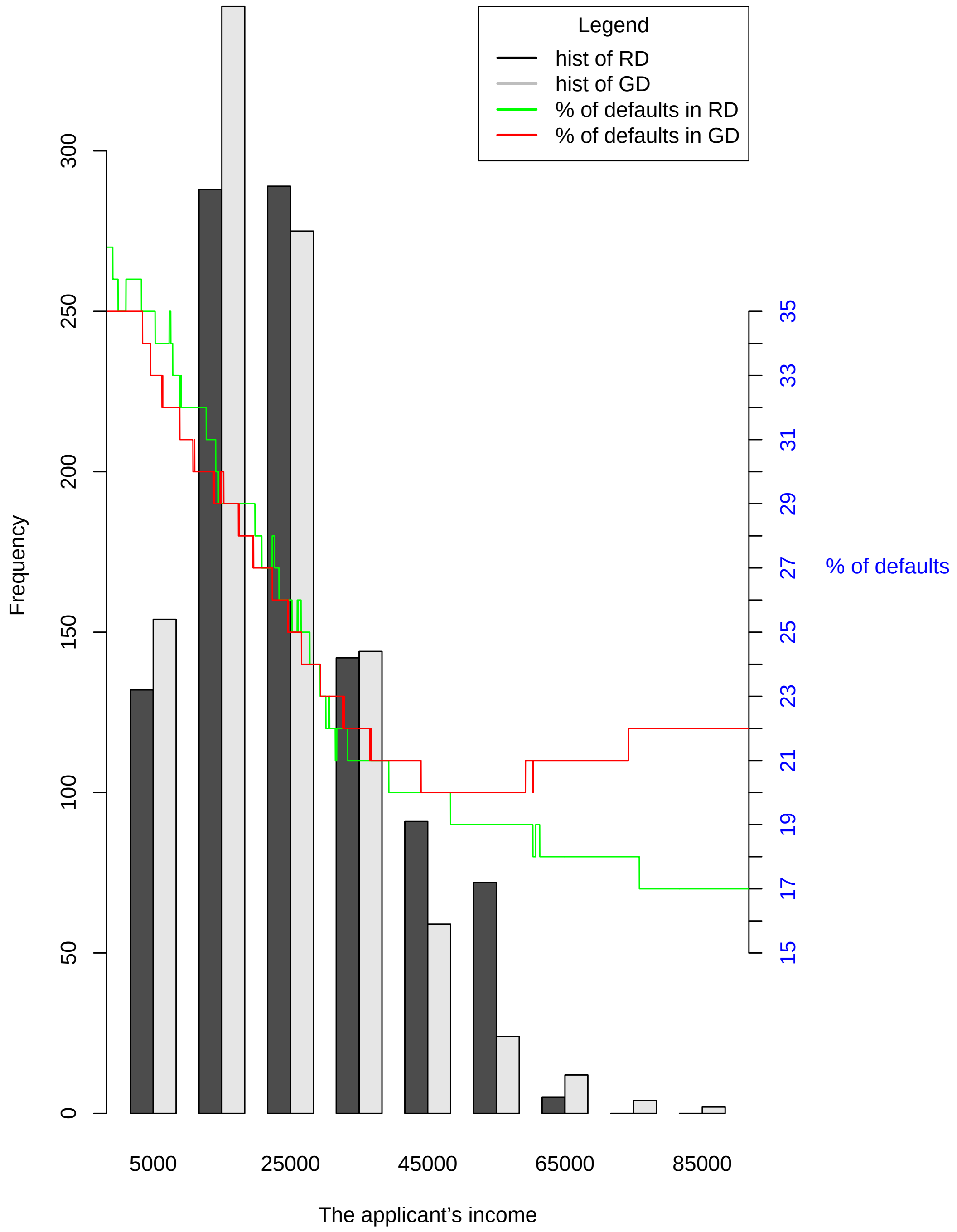

Figure 5.6: There are histograms of the applicant’s income of the real data (RD) and the generated data (GD) and the curves, meaning the relation the defaults to the the total number of credits, for RD and GD

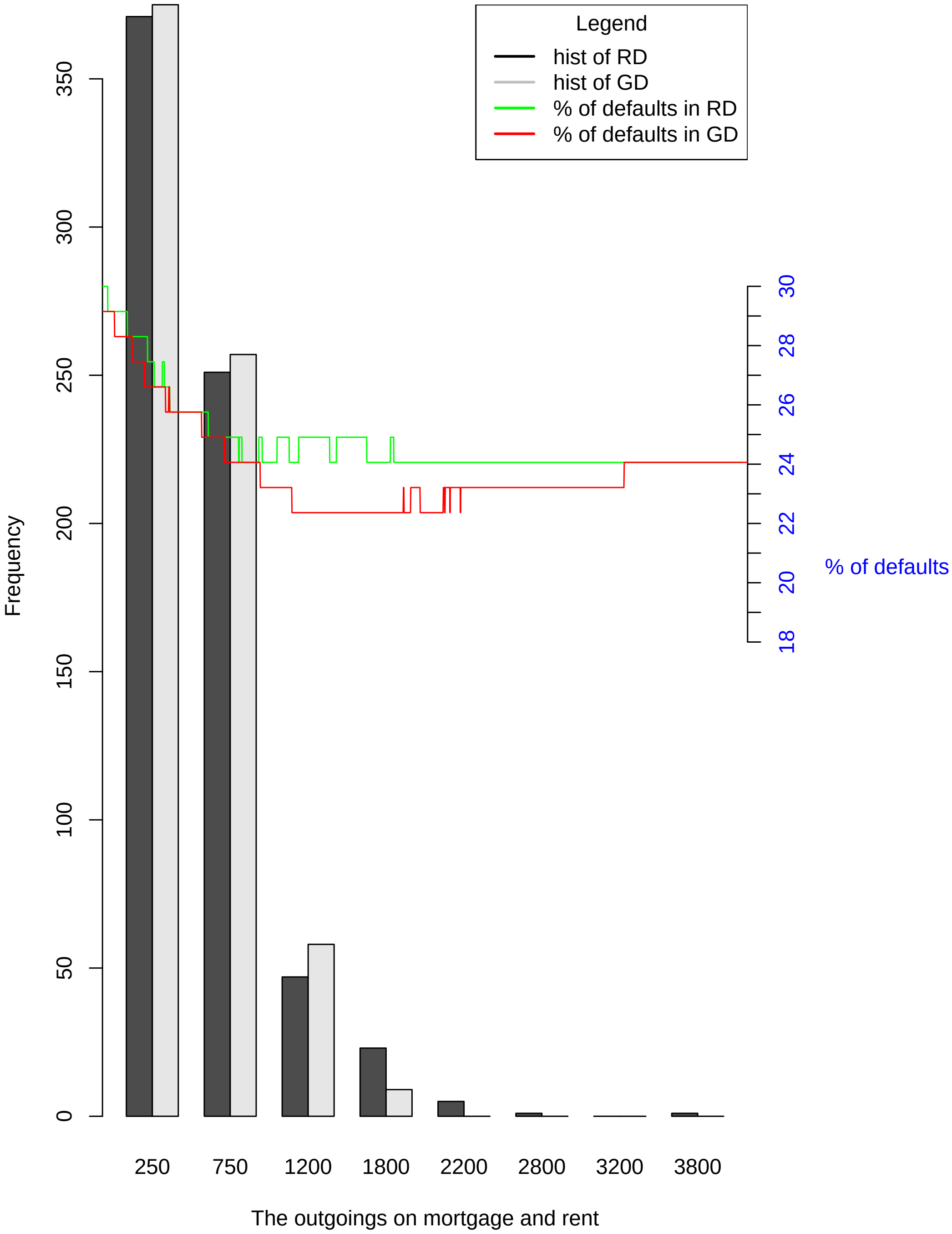

Figure 5.7: There are histograms of the outgoings on mortgage or rent of the real data (RD) and the generated data (GD) and the curves, meaning the relation the defaults to the the total number of credits, for RD and GD

The results of this section shows that the generated vectors have the very similar default's structure as the real data.

5.3 Summary

At the beginning of this work we said that we want to generate the vectors that will have the required marginal distributions and the given correlation matrix. In this section we have done it and verified that these two conditions are satisfied.

Moreover, the generated vectors satisfy two more conditions: both of them concern to the required defaults.

Chapter 6

Conclusion

Through our work we have shown many different methods of generating with copulas. In all the mentioned approaches we have concentrated on the study of different kinds of dependence structures - the correlation dependence between the variables and the relationship between explanatory variables and default.

In Chapter 2 we explained what are copulas and stated some properties and different families of copulas. The two important theoretical results (the Sklar's theorem 2.1.2 and invariance property 2.2.1) which were referred in this Chapter, allowed us the formulation of the two central results of this work: the algorithm of generating vectors with normal copula and with t-copulas with ν degrees of freedom, which have the required marginal distributions and the given correlation coefficients.

Then in Section 3.4 we made comparisons of results of these algorithms to find out the only one algorithm.

Further, Chapter 4 was concerned by the analysis of the real data of defaults and the variables, which influence the defaults.

Finally, using the copula we have generated these variables and checked if they have the same dependence structures as the real data.

The results of this work may be applied in the many fields, especially in credit field to generate vectors of defaults or non-defaults. On the other hand, we should be careful on the minimal sample size and on the marginal distributions - as we have seen the dependence structure can be biased due to multinomial marginals.

Acknowledgments

I am very grateful to RNDr. Petr Franěk, PhD. for giving me the great opportunity of writing this thesis in this fascinating area. I would also like to express my sincere appreciation to him for his helpful suggestions and support.

Further, I wish to thank Mgr. Ondřej Vencálek for his help and interesting discussions, and KateB. for her time devoted to editing my English.

I want to thank my family, especially to my father, for their support, understanding and love during these years.

Appendix

Histograms to the 3^{rd} chapter

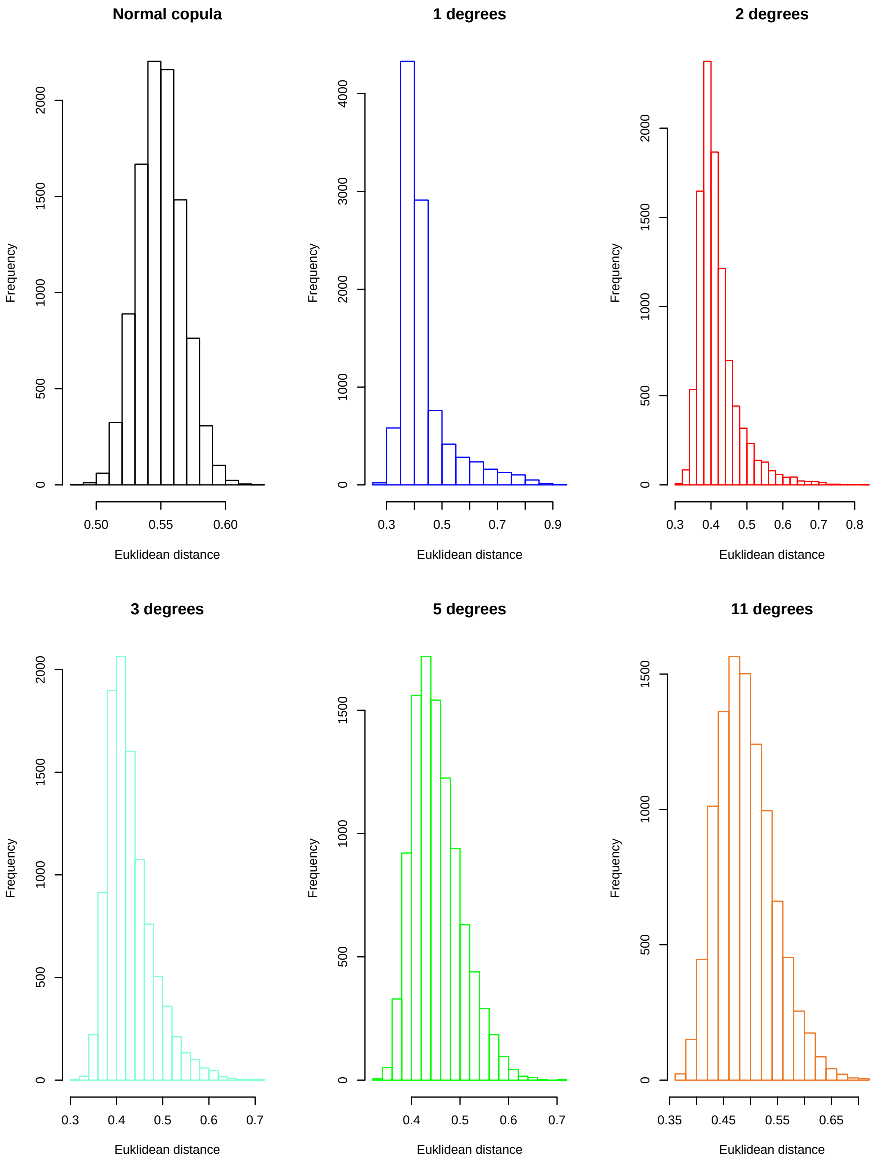

Figure 1: The histograms of $\sqrt{\left(\sum_{i,j}(cor(\mathbb{X})_{i,j}-\rho_{ij})^2\right)}$ for different copulas

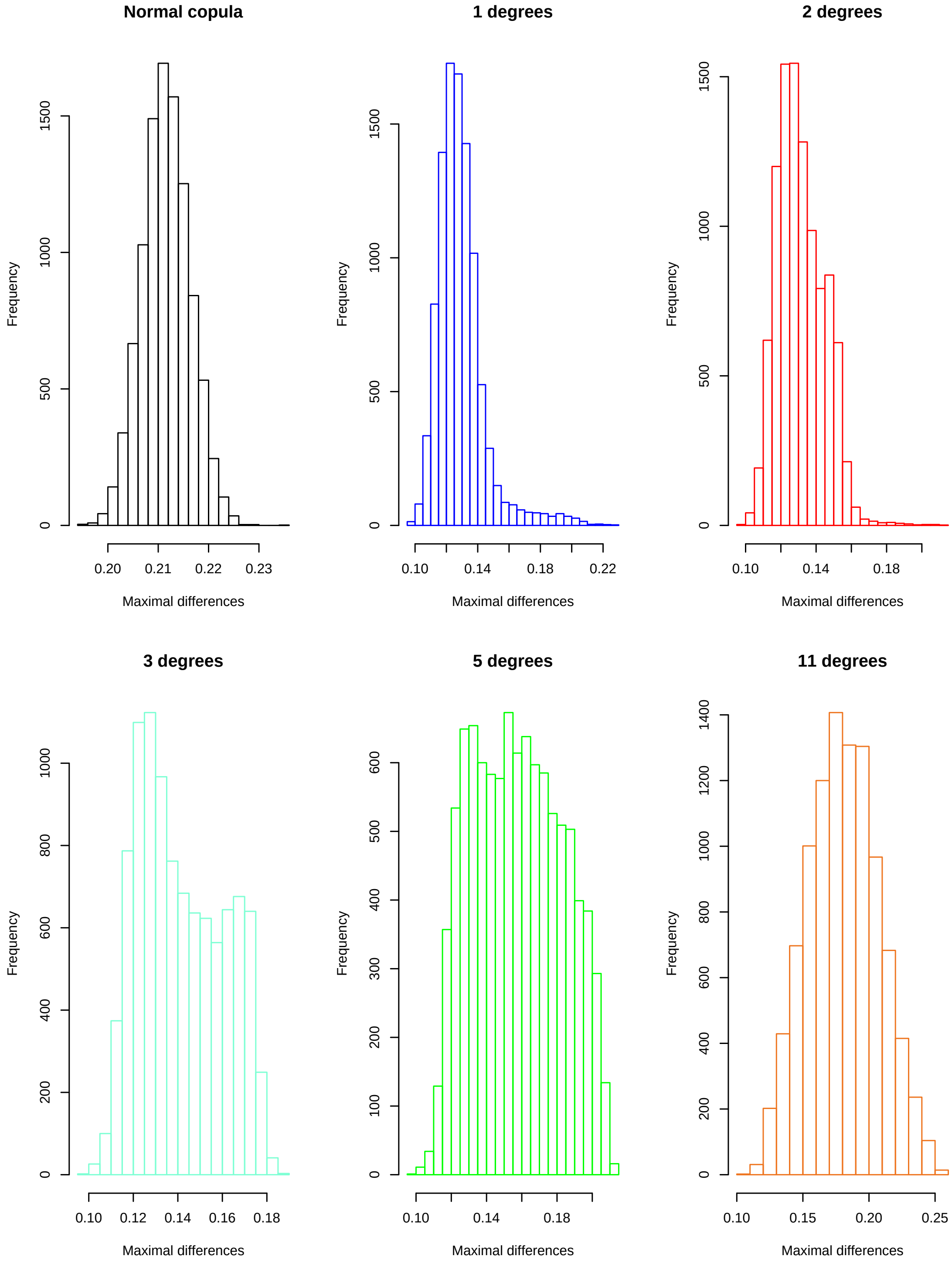

Figure 2: The histograms of $\max_{i,j} |cor(\mathbb{X})_{i,j} - \rho_{ij}|$ for different families of copulas

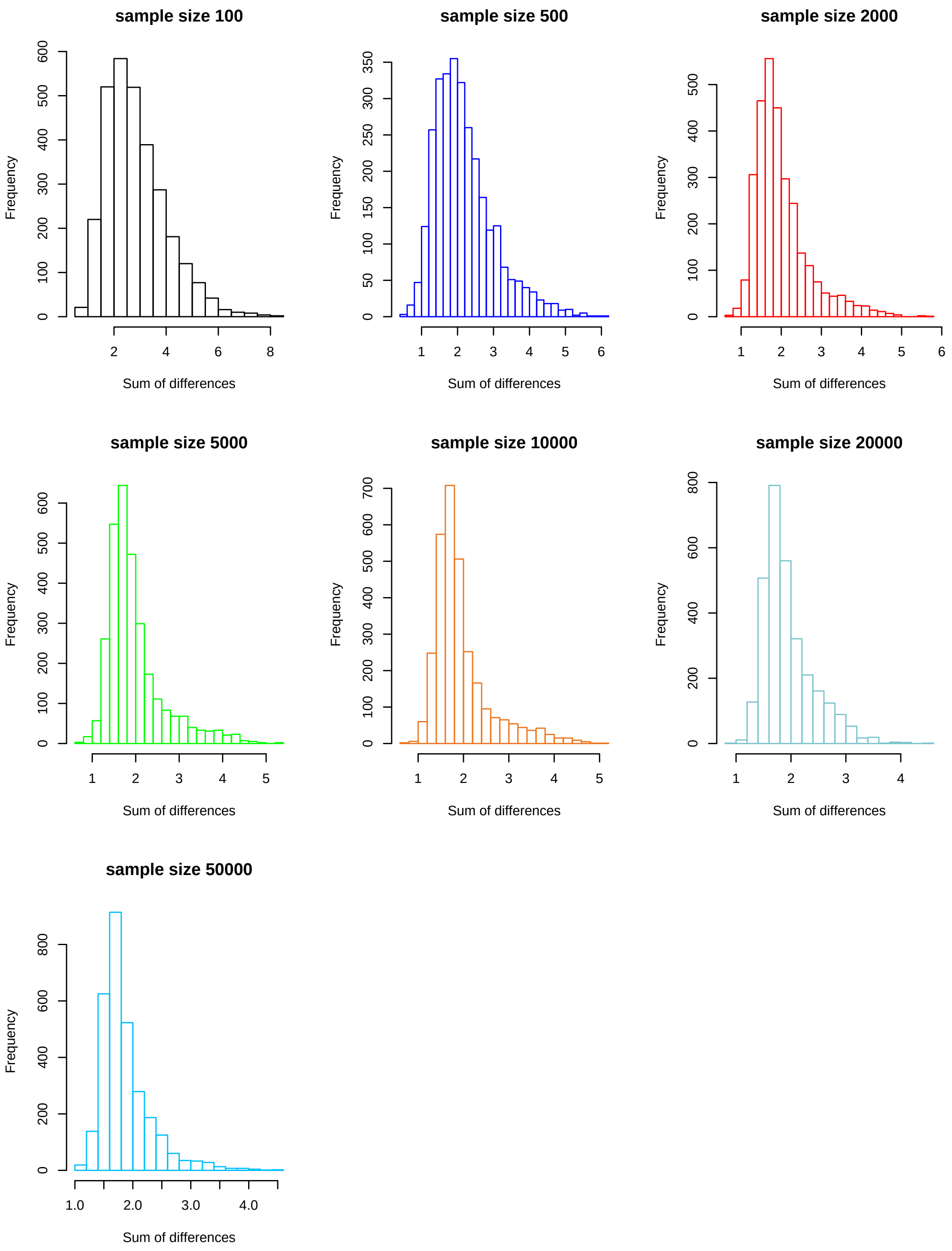

Figure 3: The histograms of $\sum_{i,j} |cor(\mathbb{X})_{i,j} - \rho_{ij}|$ for t-copula with 1 d.f. and different sample sizes

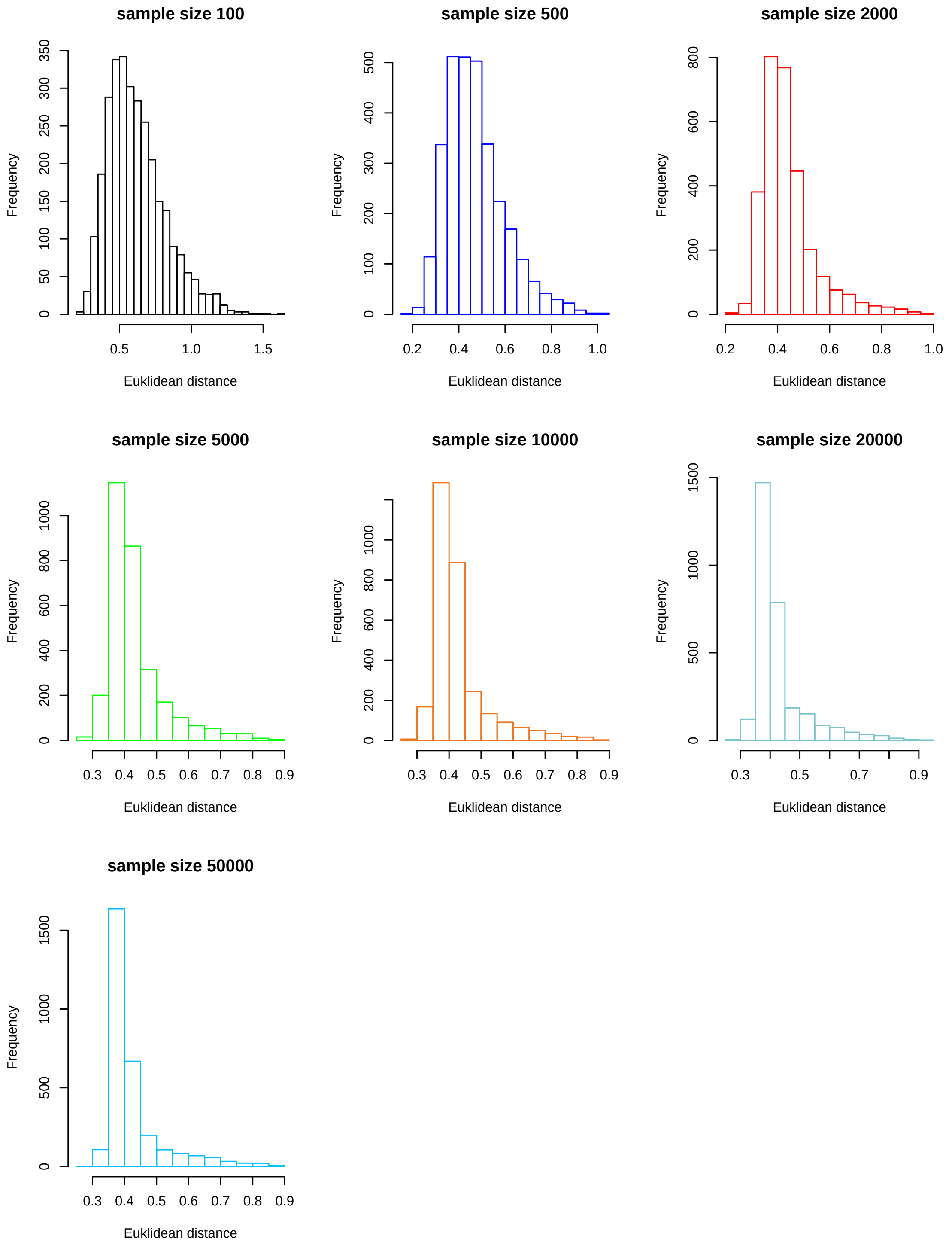

Figure 4: The histograms of $\sqrt{\left(\sum_{i,j}(cor(\mathbb{X})_{i,j}-\rho_{ij})^2\right)}$ for t-copula with 1 d.f. and different sample sizes

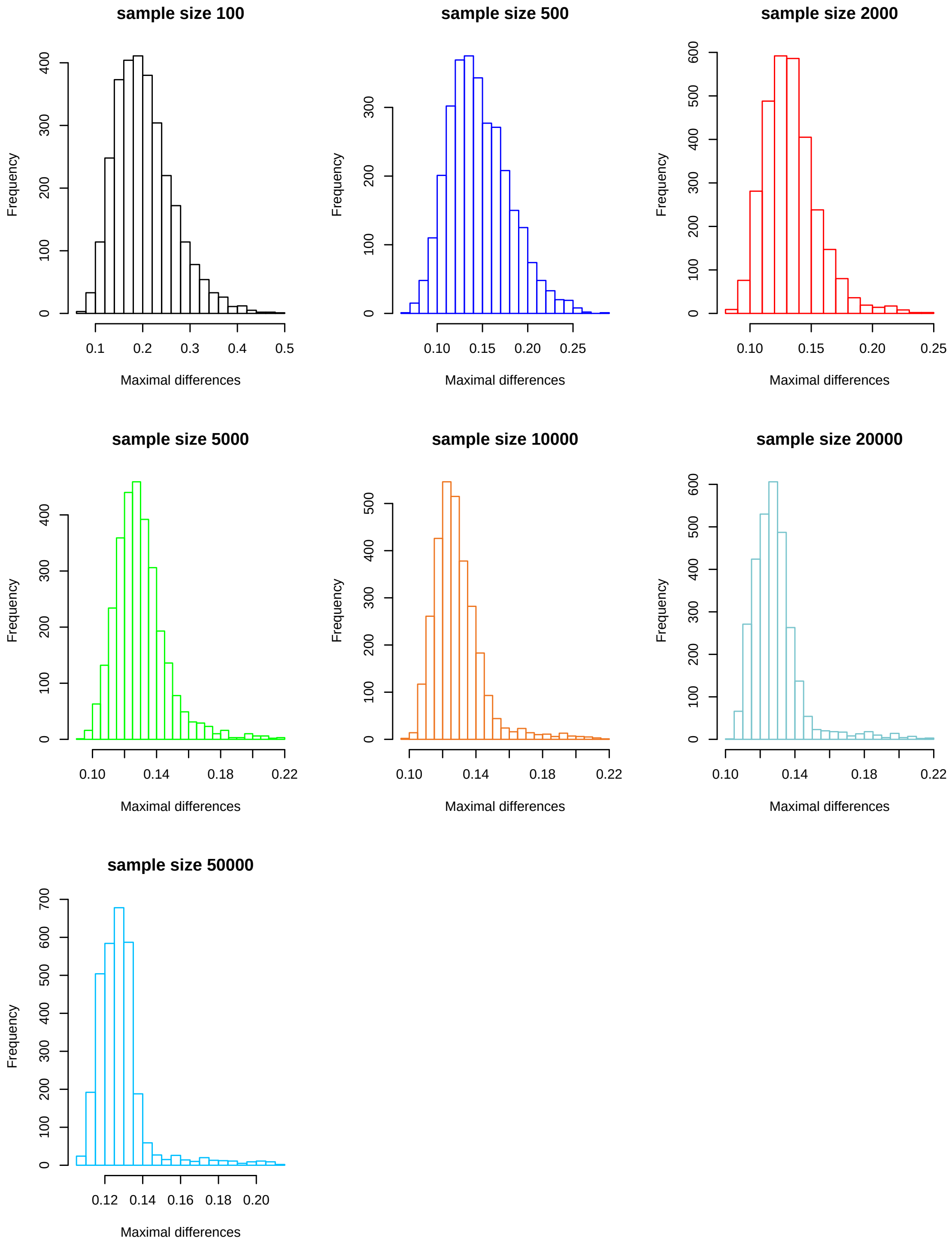

Figure 5: The histograms of $\max_{i,j} |cor(\mathbb{X})_{i,j} - \rho_{ij}|$ for t-copula with 1 d.f. and different sample sizes

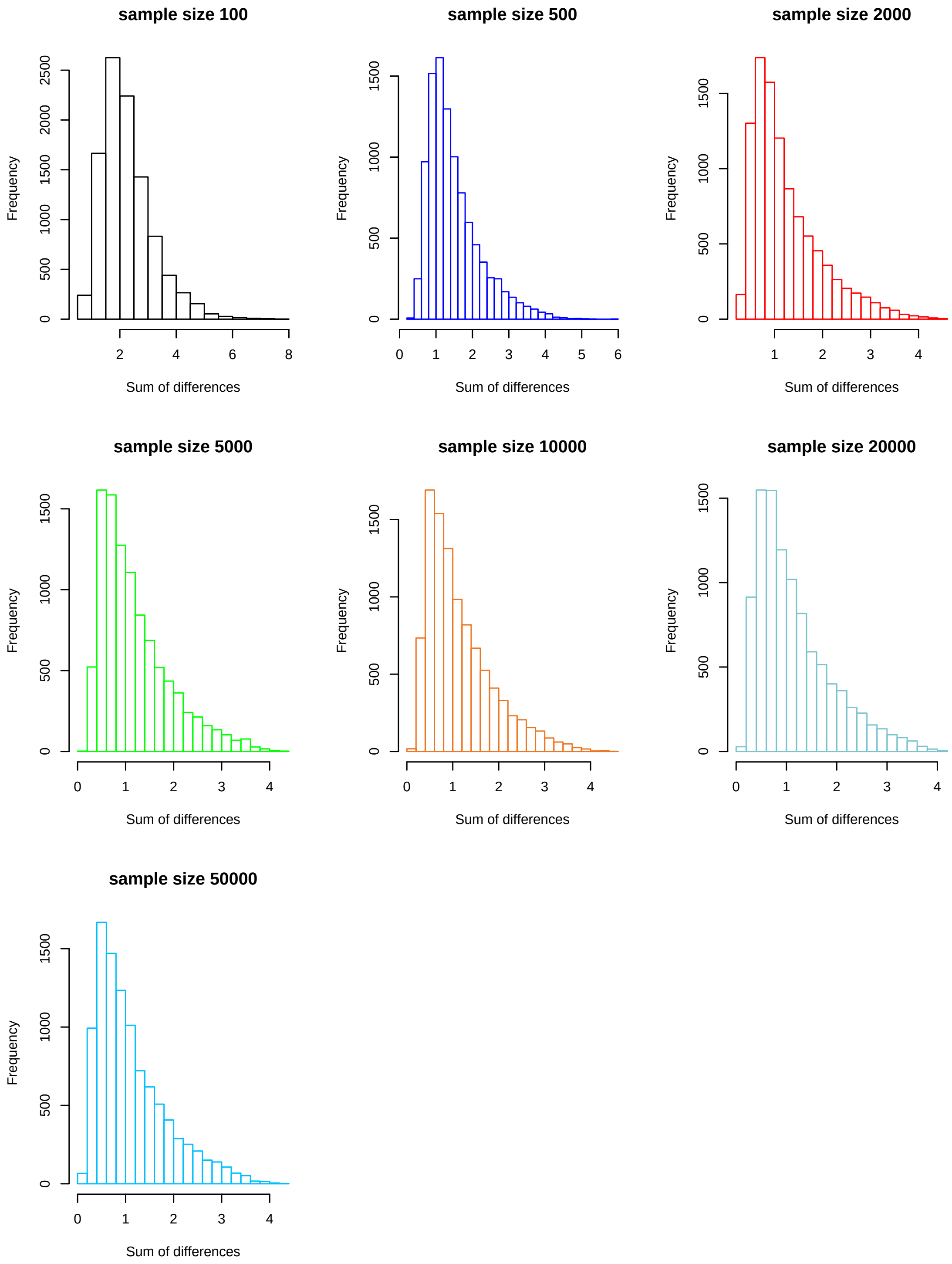

Figure 6: The histograms of $\sum_{i,j} |cor(\mathbb{X})_{i,j} - \rho_{ij}|$ for t-copula with 1 d.f. and different sample sizes (without multinomial marginal distributions)

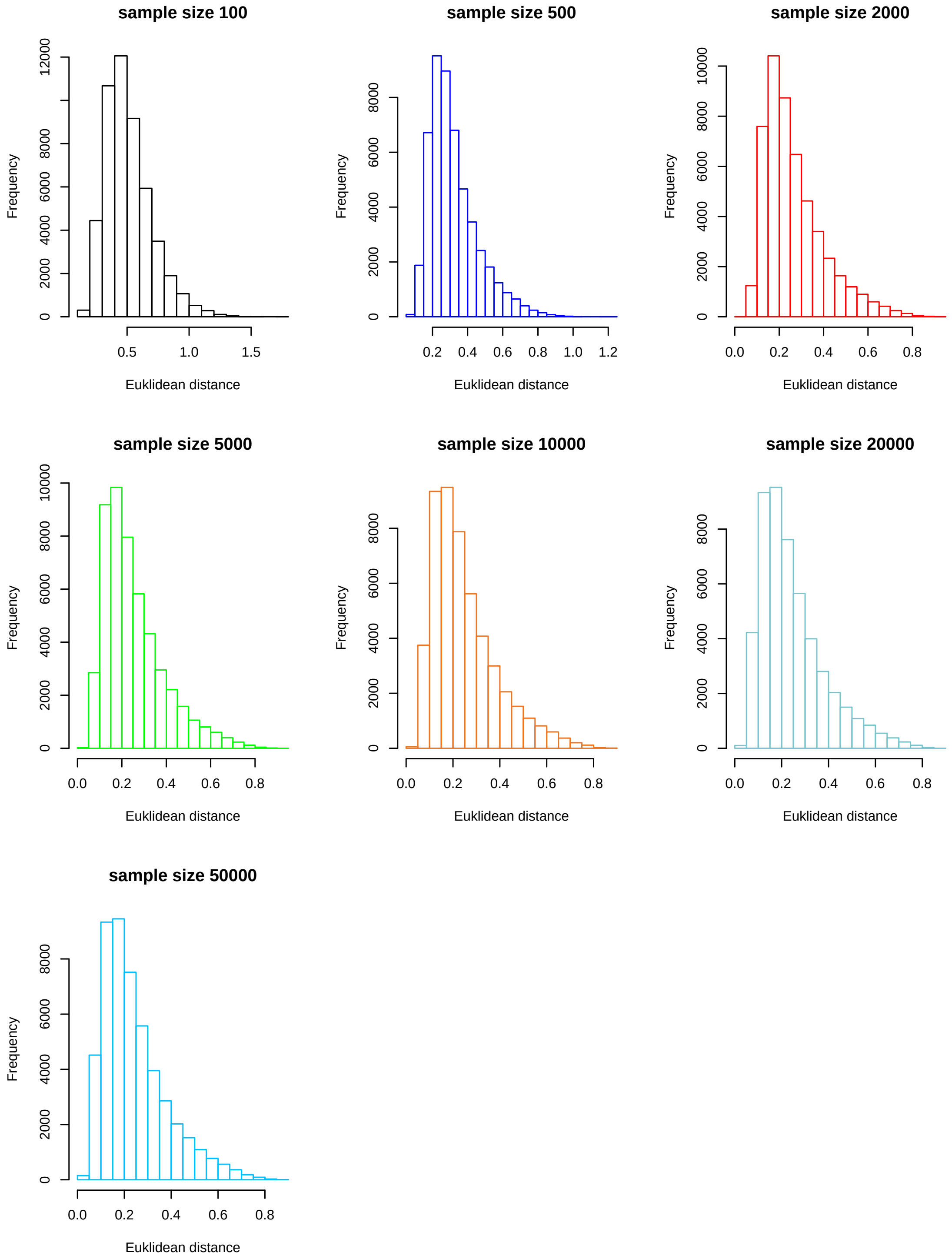

Figure 7: The histograms of $\sqrt{\left(\sum_{i,j}\left(cor(\mathbb{X})_{i,j}-\rho_{ij}\right)^2\right)}$ for t-copula with 1 d.f. and different sample sizes (without multinomial marginal distributions)

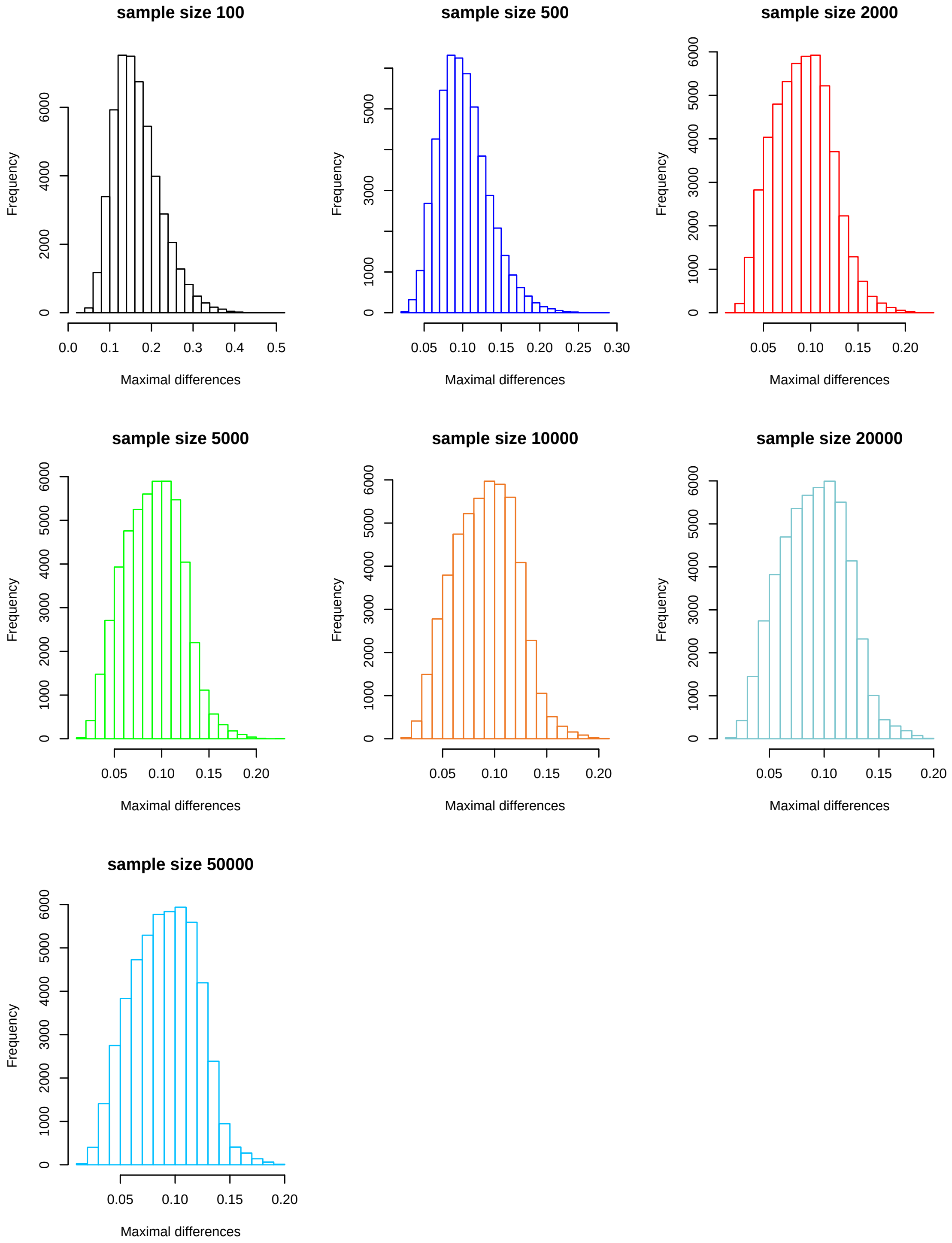

Figure 8: The histograms of $\max_{i,j} |cor(\mathbb{X})_{i,j} - \rho_{ij}|$ for t-copula with 1 d.f. and different sample sizes (without multinomial marginal distributions)

Tables for the 4^{th} chapter

	The age	Number of children	Number of other dependents	Spouse's income	Applicant's income	Value of home
Min	20	0	0	0	0	0
1st Qu.	27	0	0	0	9000	0
Median	35	0	0	0	19500	0
Mean	38.96	0.62	0.04	1990	21240	15690
3rd Qu.	48	1	0	1040	30600	28930
Max	87	5	2	50000	64800	64930

	Mortgage balance outstanding	Outgoings on mortgage or rent	Outgoings on loans	Outgoings on hire purchase	Outgoings on credit cards
Min	0	0	0	0	0
1st Qu.	0	0	0	0	0
Median	0	256	0	0	0
Mean	11230	342	121.90	28.72	39.60
3rd Qu.	20000	528	0	0	0
Max	64000	3800	28000	1600	2800

Phone owner

Yes	No
1107	118

Good/bad indicator

Yes	No
902	323

Applicant's employment status

	Amount
Government	231
Housewife	37
Military	23
Private sector	531
Public sector	30
Retired	104
Self employed	124
Student	123
Unemployed	8
Others	23
No response	8

Residential status

	Amount
Owner	624
Tenant furnished	129
Tenant unfurnished	154
With parents	252
Other	66

Table 1: The descriptive statistics of every parameter

		Number of children						Total
		0	1	2	3	4	5	
Good/bad indicator	Good	596	110	137	39	17	3	902
	Bad	224	37	45	15	2	0	323
Total		820	147	182	54	19	3	1225

Table 2: Contingency table of Number of Children vs. Good/bad indicator and theirs marginal distributions

		Number of children				
		0	1	2	3	more than 4
Good/bad indicator	Good	604	108	134	39.8	16.2
	Bad	216	38.8	48	14.2	5.8

Table 3: The *expected* values under the independence of The number of children and Good/bad indicator

		Phone owner		Total
		Yes	No	
Good/bad indicator	Good	822	80	902
	Bad	285	38	323
Total		1107	118	1225

Table 4: The contingency table of Phone owner and Good/bad indicator and its marginal distribution

		Residential status				Total
		Tenant furnished	Owner	With parents	Tenant furnished	
Good/bad indicator	Good	134	460	192	116	902
	Bad	61	164	60	38	323
Total		195	624	252	154	1225

Table 5: The Residential status vs. Good/bad indicator

		Mortgage balance outstanding					
		$=0$	$1-15000$	$15001-30000$	$30001-40000$	$40001-50000$	$50000 \leq$
Good/bad indicator	Good	510	144	78	50	38	82
	Bad	199	49	20	13	19	23

Table 6: The Mortgage balance outstanding vs. Good/bad indicator

Tables for the 5^{th} chapter

	Applicant's income					
	$= 0$	$1 - 4000$	$4001 - 8000$	$8001 - 14000$	$14001 - 20000$	$20000 \leq$
% of defaults in real data	30	22	21	15	8	4
mean % of defaults in generated data	17	33	17	14	10	9

Table 7: The percentages of defaults in the real and generated data depending on the applicant's income

	applicant's employment status		
	$1^{st}group$	$2^{nd}group$	$3^{rd}group$
% of defaults in real data	68	13	19
mean % of defaults in generated data	69	9	22

Table 8: The percentages of defaults in the real and generated data depending on the applicant's employment status

	Outgoings on mortgage or rent					
	$= 0$	$1 - 500$	$501 - 1000$	$1001 - 1500$	$1501 - 2000$	$2000 \leq$
% of defaults in real data	53	24	17	4	1	1
mean % of defaults in generated data	43	30	16	9	1	1

Table 9: The percentages of defaults in the real and generated data depending on the outgoings on mortgage or rent

Bibliography

[1] Anděl, Jíří . *Základy matematické statistiky.* Charles University in Prague, Faculty of Mathematics and Physics, 2002. Preprint.

[2] Beatriz Vaz de Melo Mendes and Marco Aurélio Sanfins. The limiting copula of the two largest order statistics of independent and identically distributed samples. *Brazilian Journal of Probability and Statistics*, 21:85–101, 2007.

[3] Conover, W.J. *Practical Non-Parametric Statistics.* John Wiley and Sons, second edition, 1980.

[4] G. Frahm, M. Junker and A. Szimayer. Ellipcital copulas: applicability and limitations. *Statistics & Probability Letters*, 63:275–286, 2003.

[5] Gumbel, E.J. Bivariate exponential distributions. *J. Amer. Statist. Assoc.*, 55:698–707, 1960.

[6] Kai-Tai Fang, Samuel Kotz and Kai-Wang Ng. *Symmetric Multivariate and Related Distributions.* Chapman and Hall Ltd, 1990.

[7] Marhoun, P. Skóringové a klasifikacní metody v bankovnictví. Master's thesis, Department of Probability and Mathematical statistics, Faculty of Mathematics and Physics, Charles University in Prague, 2005.

[8] Mario R. Melchiori CPA. Tools for sampling multivariate archimedean copulas. Technical report, Universidad Nacional del Litoralch, Santa Fe, Argentina, March 2006.

[9] McNeil, Alexander J. Multivariate Models: Theory. 2000.

[10] Nelsen, Roger B. *An introduction to Copulas.* Springer Science+Business Media, 1999.

[11] P. Embrechts, A. McNeil and D. Straumann. Correlation and dependence in risk management: properties and pitfalls. July 1999.

[12] P. Embrechts, F. Lindskog and A. McNeil. Modelling Dependence with Copulas and Applications to Risk Management. Technical report, Department of Mathematics, ETHZ, Zurich, Switzerland, September 10, 2001, and edition in 2003.

[13] R Foundation. *The R Project for Statistical Computing.* http://www.r-project.org/.

[14] School of Mathematical and Physical Sciences, University of Newcastle. *Forum of R.* http://tolstoy.newcastle.edu.au/R/.

[15] Verschuere, B. On Copulas and their Application to CDO Pricing. working paper, jan 2006.

[16] Zvára, K. R & Regrese. 2007.

Index