Geir Drage Berentsen

Thesis for the degree of Master of Science
Financial and Insurance Mathematics
University of Bergen, Norway
2nd June 2009

Analysis of left truncated data with an application to insurance data

This thesis is written in LaTeX 2ε with the 'uib-mi-master' document class, developed by Karl Ove Hufthammer. It was compiled using pdfTeX-1.40.4 on 2nd June 2009. The body text is 11 point URW Palladio with small caps. The maths font is URW Palladio and Pazo Math, the heading font is HV Math, and the computer code font is Bera Mono.

Acknowledgements

First of all, I would like to thank my supervisor Jostein Paulsen for providing me with an interesting topic and for giving me many valuable comments and suggestions. I am also greatly indebted to Karl Ove Hufthammer and Arne Johannes Holmin who helped me solve many of the technical problems encountered in this thesis.

Key words: Random truncation, quasi independence, Pearson's correlation coefficient, Kendall's tau, U-statistics, Product-limit estimator, Copulas.

Contents

Notation

$TN_2(\mu_x, \mu_t, \sigma_x^2, \sigma_t^2, \rho)$	the truncated bivariate normal distribution
$X_{(1)}, \ldots, X_{(n)}$	the ordered variables satisfying $X_{(1)} \leq \cdots \leq X_{(n)}$
$\mathbb{R}$	the set of real numbers
$\overline{\mathbb{R}}$	the extended set of real numbers $\mathbb{R} \cup \{-\infty, \infty\}$
$\mathbb{R}^p$	the p-dimensional space $\underbrace{\mathbb{R} \times \mathbb{R} \cdots \times \mathbb{R}}_{p}$
.	unspecified set
$F(z-)$	$\lim_{x \uparrow z} F(x)$, limit of $F(x)$, letting x increase towards z
$F(z+)$	$\lim_{x \downarrow z} F(x)$, limit of $F(x)$, letting x decrease towards z
card	cardinality
$\lvert x \rvert$	the absolute value of x
x^T	x transposed
$\text{sgn}(x)$	the sign of x, i.e. 1 if $x > 0$, 0 if $x = 0$, -1 if $x < 0$
sup	supremum, the least upper bound
inf	infemum, the greatest lower bound
max	maximum
min	minimum
i.i.d.	independent and identically distributed
$:=$	defined as
$\xrightarrow{P}$	convergence in probability
$\xrightarrow{a.s.}$	convergence almost surely
$\xrightarrow{d}$	convergence in distribution
χ_v^2	the Chi-square distribution with v degrees of freedom
ϕ	the standard normal cumulative distribution function
$AVE(\boldsymbol{x})$	the sample average of $\boldsymbol{x} = (x_1, \ldots, x_n)$
$VAR(\boldsymbol{x})$	the sample variance of $\boldsymbol{x} = (x_1, \ldots, x_n)$
$\forall$	for all
$\nabla f(\boldsymbol{a})$	gradient vector whose components are the partial derivatives of f at $\boldsymbol{a}$, i.e. $\nabla f(\boldsymbol{a}) = (\frac{\partial f}{\partial a_1}, \ldots, \frac{\partial f}{\partial a_n})$

1 Introduction

1.1 Topics covered in the thesis

This thesis discusses different ways of analysing left truncated data when the lower bound itself is a stochastic variable. We will consider the possible dependence between the variable of interest and the truncating variable, and how the dependency structure between these variables influence estimation of the underlying distribution.

1.1.1 Random left truncation

In a sample subject to left truncation by some lower bound, all the values below this bound are entirely omitted. Opposed to the concept of *left censoring*[1], we have no record of how many observations are omitted, nor what the lower bound may be (unless this is prior knowledge). In random left truncation

[1] In the left censoring case we are given an observation or a note that the observation is below the bound. In addition, we know the value of this bound.

the lower bound is a random variable. We call this variable ”the truncating variable”, while we call the variable subject to left truncation ”the variable of interest”.

1.1.2 Dependent truncation

When the variable of interest is larger than the truncating variable we assume both variables are observed. In some cases there is a relation between these two variables, and we say that the data are subject to a dependent truncation. The assumption of independence between these variables can in general not be tested with a truncated dataset. The reason is that we do not know anything about the behaviour of the unobserved data.

In chapter 2 we will approach this problem by introducing a weaker assumption called quasi independence, which can be interpreted as independence between the variables we do observe. This assumption can be tested with a truncated dataset. For this purpose we will consider two different measures of dependence for truncated data. The asymptotic properties of the sample version of these measures will be studied and used to approximate the distribution of finite sample test-statistics.

1.1.3 Reconstruction of the unconditional distribution

In chapter 3 we will consider a nonparametric maximum likelihood estimator called the Product-limit estimator. This estimator aims to reconstruct the unconditional distribution of the variable of interest using truncated data. This method depends heavily on the assumption of quasi independence and is therefore not suitable for data subject to a dependent truncation. Analysis of such data will be the primary subject in the rest of the thesis.

The problem of reconstructing the joint distribution between the variable of interest and the truncating variable have been considered by very few authors, and only just recently. A warm up to this subject is given in chapter 4, where we assume that these variables follow a bivariate normal distribution. The observed data will then follow the so-called truncated bivariate normal distribution. Under this assumption, estimates of the unknown parameters

can be obtained by maximum likelihood estimation. In chapter 5 we will consider the more general parametrisation done with copulas. For this model we will consider maximum likelihood based estimation and a semi-parametric approach proposed in recent literature.

1.2 Examples

Example 1.2.1: Retirement House: Klein and Moeschberger (2003)

In a retirement centre subjects are observed only if they live long enough to enter the retirement house. The lifetime X is then left truncated by the retirement house entry age, T. There is reason to believe that these variables are dependent. People who enter the retirement house earlier may get better medical attention and therefore live longer. On the other hand, people with poor health and shorter expected lifetime may retire earlier.

Example 1.2.2: AIDS study: Kalbfleisch and Lawless (1989)

Let Y be the infection time where 1 represents January 1978 and let T be the incubation time in months for people who were infected by contaminated blood transfusions and developed AIDS by 1 July 1986. Since the total study period is 102 months only individuals with $T + Y < 102$ were included in the sample. Then, letting $X = 102 - Y$ yields the model described: (X, T) is observed only if $T < X$. Kalbfleisch and Lawless (1989), amongst others, analyse these data based on the assumption that X and T are independent. Later, Tsai (1990) pointed out that this assumption fails to hold.

1.3 Applications to insurance

In casualty insurance, claims are only observed if they are larger than the corresponding deductible. In many cases, the insurance companies assign individual deductibles for each object. Though it may seem strange to think of deductibles as random variables (since we more or less control these values),

such a consideration could provide useful information. If there is a significant association between the claims and deductibles it can be reasonable to use deductibles as an additional covariate when estimating claims. This consideration can also be used to estimate the number of unreported claims. For this purpose the joint distribution of claims and deductibles must be estimated. In chapter 6 we will apply some of the methods considered in this thesis on insurance data from ships.

All numerical procedures and graphical displays in this thesis are carried out using the statistical program R.

2

Measures of dependence for truncated data

Many methods concerning truncated data depend on the assumption of independence between the variable of interest and the truncating variable. Therefore, to use these methods one would have to investigate the dependence between these variables. In this chapter we will consider two different quantities designed to measure the dependence in truncated data. The first quantity is a generalisation of the Pearson product-moment correlation coefficient proposed by Chen *et al.* (1996). The second is a generalisation of Kendall's Tau proposed by Tsai (1990).

2.0.1 The general case

Let X^* be the variable of interest subject to left truncation by the truncating variable T^*. That is, the sampling mechanism is such that (X^*, T^*) is included in the sample if and only if $X^* > T^*$ (See figure 2.1 on page 8).

We assume that there are n such pairs amongst the original sample of unknown size N. When (X^*, T^*) is included in the sample we denote it (X, T), i.e. $(X, T) = (X^*, T^* | X^* > T^*)$.

Let $H(x,t)$ be the joint distribution of (X^*, T^*) with marginals $F(x) = H(x, \infty)$ and $G(t) = H(\infty, t)$. Let $H^c(x,t)$ denote the conditional cumulative distribution of (X^*, T^*), given that $X^* > T^*$. Thus

$$\begin{aligned} H^c(x,t) &= P(X^* \leq x, T^* \leq t \mid X^* > T^*) \\ &= \frac{P(X^* \leq x, T^* \leq t, X^* > T^*)}{P(X^* > T^*)} = \int\int_{\Delta(x,t)} \mathrm{d}H(u,v)/\alpha, \end{aligned} \tag{2.1}$$

where

$$\begin{aligned} &\alpha = P(X^* > T^*) = \int\int_{u \geq v} \mathrm{d}H(u,v) \quad \text{and} \\ &\Delta(x,t) = \{(u,v); v < u \leq x, v \leq t\}. \end{aligned}$$

The conditional cumulative distribution of X and T are given by $F^c(x) = H^c(x, \infty)$ and $G^c(t) = H^c(\infty, t)$, respectively. Given the density $h(x,t)$ of (X^*, T^*) the conditional density is

$$h^c(x,t) = \begin{cases} h(x,t)/\alpha, & x > t, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

In the continuous case, given the density h we have that

$$\begin{aligned} H^c(x,t) &= \int\int_{\Delta(x,t)} h(u,v)\,\mathrm{d}u\,\mathrm{d}v/\alpha, \\ \alpha &= \int\int_{u>v} h(u,v)\,\mathrm{d}u\,\mathrm{d}v. \end{aligned}$$

Below is a graphical depiction of the sampling mechanism.

$$\underbrace{(X_1^*, T_1^*), \ldots, (X_N^*, T_N^*)}_{i.i.d. H(x,t)} \xrightarrow{\textit{Truncation}} \underbrace{(X_1, T_1), \ldots, (X_n, T_n)}_{i.i.d. H^c(x,t)}, \quad n \leq N.$$

$$\underbrace{X_1^*, \ldots, X_N^*}_{i.i.d. F(x)} \xrightarrow{Truncation} \underbrace{X_1, \ldots, X_n}_{i.i.d. F^c(x)}, \qquad \underbrace{T_1^*, \ldots, T_N^*}_{i.i.d. G(t)} \xrightarrow{Truncation} \underbrace{T_1, \ldots, T_n}_{i.i.d. G^c(t)}$$

This is the general setup in the left truncation case and the notations will be kept throughout the thesis. Later we will consider the estimation of the distribution function of X^* using the so called *Product-limit estimator*. However, the consistency of this estimator depends heavily on the assumption of *quasi independence*, which we will consider in the following section.

2.1 Quasi independence

Since we are unable to observe data in the region $X^* \leq T^*$, and thus do not know anything about the dependence in that region, we can't decide whether or not X^* and T^* are independent. However, there is a weaker definition of independence called quasi independence.

Definition 2.1.1: Quasi independence

Let the marginal distributions of X^* and T^* be $F(x) = H(x, \infty)$ and $G(t) = H(\infty, t)$ respectively. The variables X and T in the observable vector (X, T) are said to be quasi independent if the corresponding distribution $H^c(x, t)$ has the following property:

$$\begin{aligned} H_0: \quad & H^c(x,t) = \int\int_{\Delta(x,t)} \mathrm{d}F(u)\,\mathrm{d}G(v)/\alpha_0, \\ \text{where} \quad & \alpha_0 = \int\int_{u>v} \mathrm{d}F(u)\,\mathrm{d}G(v). \end{aligned} \tag{2.3}$$

Given the densities g and f corresponding to G and F, this assumption is equivalent to

$$H_0': \quad h^c(x,t) = \begin{cases} f(x)g(t)/\alpha_0, & x > t, \\ 0, & \text{otherwise.} \end{cases}$$

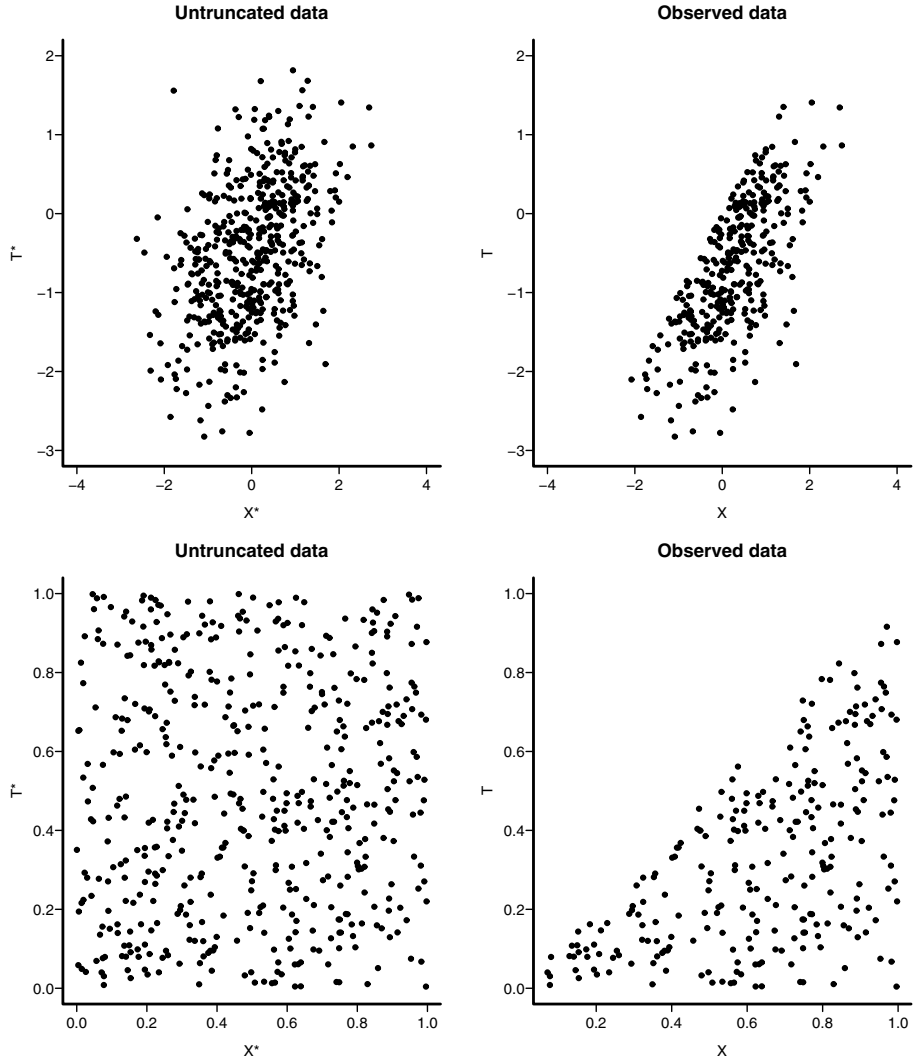

Figure 2.1: The sampling mechanism: The first two plots illustrates the sampling mechanism when (X^*, T^*) follows a bivariate normal distribution with $\rho = 0.4$. The last two plots illustrates the sampling mechanism when X^* and T^* are independent uniformly distributed on $[0, 1]$.

We will denote the assumption of quasi independence H_0. The term quasi independence was first used in the contingency table literature to describe variables which behaved as independent variables in certain subsets of the table. In our setting the interpretation is quite similar. The assumption of quasi independence imply that (X^*, T^*) behaves as independent variables in the region $\{(X^*, T^*)|X^* > T^*\}$. It is easily seen that independence between the two variables imply quasi independence. The converse statement is not true, as illustrated by the following example.

Example 2.1.2: Quasi independent variables which are not independent

Consider the contingency table 2.1. In this case both X^* and T^* are discrete and uniform on $\{0, 1, 2, 3\}$. We see that the requirements for quasi independence holds since

$$\alpha = \sum_{i>j}\sum h(i,j) = \alpha_0 = \sum_{i>j}\sum f(i)g(j) = \frac{6}{16},$$

$$h^c(i,j) = \begin{cases} f(i)g(j)/\alpha_0, & i > j, \\ 0, & \text{otherwise.} \end{cases}$$

However, the variables are not independent for $X^* \leq T^*$.

Table 2.1

		T^*				
	$h(x,t)$	0	1	2	3	$f(x)$
X^*	0	1/16	1/16	0	2/16	1/4
	1	1/16	1/16	2/16	0	1/4
	2	1/16	1/16	1/16	1/16	1/4
	3	1/16	1/16	1/16	1/16	1/4
	$g(t)$	1/4	1/4	1/4	1/4	1

Unfortunately, this means that even though we can establish that the observations in a truncated dataset are quasi independent, the variables may still be dependent. However, Tsai (1990) pointed out that many methods which originally are stated to work only under the assumption of independence between X^* and T^* will also work under the assumption of quasi indepen-

dence. Motivated by this, we will consider measures which can help us decide whether or not the data depart from the hypothesis of quasi independence. We will first consider a generalisation of the well-known Pearson product-moment correlation proposed by Chen *et al.* (1996), which we will denote by ρ_c.

2.2 The generalised product-moment correlation coefficient

Definition 2.2.1

Given the random samples (X_1, T_1) and (X_2, T_2) from the truncated vector (X, T), we define the conditional product-moment correlation by

$$\rho_c = \frac{E[(X_1 - X_2)(T_1 - T_2) \mid A]}{\{E[(X_1 - X_2)^2 \mid A]E[(T_1 - T_2)^2 \mid A]\}^{1/2}},$$

where $A = \{\max(T_1, T_2) < \min(X_1, X_2)\}$. Alternatively we can write:

$$\rho_c = \frac{E[(X_1 - X_2)(T_1 - T_2)I_A]}{\{E[(X_1 - X_2)^2 I_A]E[(T_1 - T_2)^2 I_A]\}^{1/2}} = \frac{\sigma_{XT}}{\{\sigma_{XX}\sigma_{TT}\}^{1/2}}, \tag{2.4}$$

were I_A is the indicator function of the set A.

The last representation of ρ_c is valid since

$$\rho_c = \frac{E[(X_1 - X_2)(T_1 - T_2)I_A]/P(A)}{\{(E[(X_1 - X_2)^2 I_A]/P(A))\,(E[(T_1 - T_2)^2 I_A]/P(A))\}^{1/2}}.$$

And we see that $P(A)$ in the numerator cancel the $P(A)$s in the denominator. Obviously, ρ_c is only defined when $P(A) \neq 0$. Note that by conditioning on the event A, the two points (X_1, T_1) and (X_2, T_2) become "comparable" under a truncation since the point $(\min(X_1, X_2), \max(T_1, T_2))$ given A always is located in the observable region.

The natural thing to do next is to investigate the relation between ρ_c and H_0. The following theorem holds for every distribution of (X, T).

Theorem 2.2.2

Given quasi independence between X and T, it follows that $\rho_c = 0$. That is:

$$H_0 \quad \Rightarrow \quad \rho_c = 0. \tag{2.5}$$

Proof: A proof is given in section A.1 on page 89

So at least in some sense $\rho_c = 0$ indicate no relation between X and T. However, one can't conclude quasi independence between them except in one special case. We know from classical statistics that independence is equivalent to zero correlation in the multivariate normal case. The next result shows the corresponding relations between quasi independence and $\rho_c = 0$ in the *truncated bivariate normal*[1] case:

Theorem 2.2.3

If (X, T) follows a truncated bivariate normal distribution, then

$$H_0 \quad \Leftrightarrow \quad \rho_c = 0 \tag{2.6}$$

Proof: A proof is given in (Chen *et al.*, 1996).

The theorem tells us that if we are able to establish that our data follows a truncated bivariate normal distribution, a good estimate of ρ_c could help us decide whether or not our data are quasi independent.

We continue with a computational procedure for ρ_c.

2.2.1 Monte Carlo approximation of ρ_c

From equation 2.1 on page 6 we know that given the unconditional distribution $H(x, t)$ we also know the conditional distribution $H^c(x, t)$. Hence, in such a situation, we can compute the exact value of ρ_c. The computation of ρ_c can rarely be done analytically, so we need a numerical method to do this. As an alternative to numerical integration, we can use Monte Carlo Integration. This procedure and many other problems in this thesis requires simulations from the conditional distribution H^c. If we know how to simulate from the unconditional distribution H, then a simulation procedure to obtain n i.i.d. variables distributed according to H^c is as follows:

[1] Whenever (X^*, T^*) is bivariate normal distributed, i.e. $(X^*, T^*) \sim N_2(\mu_x, \mu_t, \sigma_x^2, \sigma_t^2, \rho)$, we say that (X, T) follows a *truncated bivariate normal distribution*. The shorthand notation will be $(X, T) \sim TN_2(\mu_x, \mu_t, \sigma_x^2, \sigma_t^2, \rho)$

1. Simulate $\tilde{N} = 1\,000\,000$ i.i.d. pairs (X_i^*, T_i^*) from H and let

$$\alpha_{MC} = \text{card}\{i | X_i^* > T_i^*\}/\tilde{N}.$$

2. Put $N = n/\alpha_{MC}$ and repeat the simulation of N i.i.d. pairs (X_i^*, T_i^*) until a sample with $card\{i|X_i > T_i\} = n$ is obtained.
3. Let $(X_1, T_1), \ldots, (X_n, T_n)$ be the n pairs in the sample obtained by 2. where $X_i^* > T_i^*$.

Then $(X_1, T_1), \ldots, (X_n, T_n)$ will be i.i.d. according to H^c.

Procedure 1. is an easy way of estimating α by MC integration if we know the unconditional distribution H. Also note that $N = n/\alpha_{MC}$ is the optimal initial value of N if we want to form a subset of $(X_1^*, T_1^*), \ldots, (X_N^*, T_N^*)$ according to 3. with size n (see section 6.6 on page 83).

Using the above procedure we can simulate two large samples A and B independently from H^c:

$$A = \{(X_1^A, T_1^A), (X_2^A, T_2^A), \ldots, (X_n^A, T_n^A)\},$$
$$B = \{(X_1^B, T_1^B), (X_2^B, T_2^B), \ldots, (X_n^B, T_n^B)\},$$

and let:

$$\hat{\sigma}_{XT} = \frac{1}{n}\sum_{i=1}^{n}(X_i^A - X_i^B)(T_i^A - T_i^B)I_{A_i},$$
$$\text{where} \quad A_i = \{\max(T_i^A, T_i^B) < \min(X_i^A, X_i^B)\}.$$

Note that the elements in the above sum are independent. Therefore, by the *Strong Law of Large Numbers*[2], we know that $\hat{\sigma}_{XT}$ converges *almost surely*[3] to $E[(X_1 - X_2)(T_1 - T_2)I_A] = \sigma_{XT}$. Hence, for a sufficiently large n, this is a good approximation of σ_{XT}. Using the same sample, similar approximations can be done for σ_{XX} and σ_{TT} giving an approximation of $\rho_c = \sigma_{XT}/\{\sigma_{XX}\sigma_{TT}\}^{1/2}$. Note

[2]SLLN: Let $\bar{X}_n$ be the average of the first n of a sequence of independent, identically distributed random variables $X_1, X_2, \ldots$. If $E|X_1| < \infty$ then $\bar{X}_n \xrightarrow{a.s} EX_1$ by the strong law of large numbers.

[3]a.s: The sequence X_n is said to converge almost surely to X if $d(X_n, X) \to 0$ with probability one for a proper norm d. This is denoted $X_n \xrightarrow{a.s.} X$.

that this can be a time consuming process if the truncated proportion is large. We will now use this method to make a visual inspection of the behaviour of ρ_c.

Consider the case where (X, T) follows a truncated bivariate normal distribution, that is $(X, T) \sim TN_2(\mu_x, \mu_t, \sigma_x^2, \sigma_t^2, \rho)$. In this example we keep the parameters $(\mu_x, \mu_t, \sigma_x^2, \sigma_t^2)$ fixed, while varying ρ. We then calculate ρ_c using the method described on the previous page. In each case, the number of simulations were $n = 200000$.

Figure 2.2 illustrates the relations between ρ and ρ_c for three different truncated bivariate normal distributions. The same plot is given in (Chen *et al.*, 1996) for the same distributions, where the calculation of ρ_c was done by numerical integration. The result is the same, so we trust the accuracy of our Monte Carlo approximation. To compare the difference, $|\rho_c - \rho|$, a straight line was included in the plot. Notice that the difference $|\rho_c - \rho|$ is small in the $TN_2(0, -1, 1, 1/16, \rho)$ case, while $|\rho_c - \rho|$ is rather large in the $TN_2(0, 0, 1, 1, \rho)$. In the first case the truncated proportion is small, so one would not expect ρ_c to deviate much from ρ. However, in the second case the truncated proportion is relatively high, making $|\rho_c - \rho|$ larger.

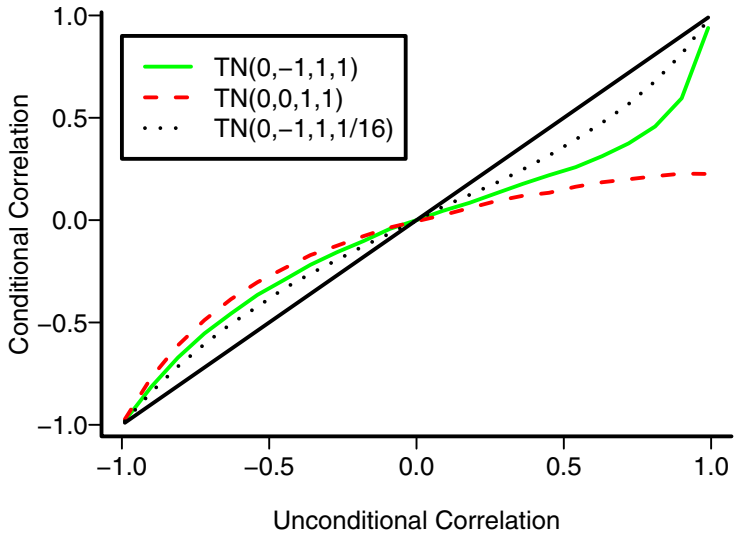

Figure 2.2: Unconditional correlation versus Conditional correlation in the truncated bivariate normal case where ρ varies from -1 to 1.

2.2.2 Sample conditional product-moment correlation coefficient

To utilise theorem 2.2.2 on page 10 and theorem 2.2.3 on page 11 we need a good estimate of ρ_c. A consistent estimate is as follows:

Definition 2.2.4

Let $(X_1, T_1), \ldots, (X_n, T_n)$ be i.i.d random vectors following the same distribution as (X, T). A pair (X_i, T_i) and (X_j, T_j) is called comparable if $\max(T_i, T_j) < \min(X_i, X_j)$. Using these pairs, the sample association between X and T in the observable region can be measured by

$$r_c = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)(T_i - T_j)I_{ij}}{\{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2 I_{ij}\}^{1/2}\{\sum_{i=1}^{n}\sum_{j=1}^{n}(T_i - T_j)^2 I_{ij}\}^{1/2}}, \quad \text{where} \quad I_{ij} = I\{\max(T_i, T_j) < \min(X_i, X_j)\}. \tag{2.7}$$

For simplicity, we will denote this in the following two ways:

$$r_c = \frac{\sum_{i=1}^{n} S_{XT_i}}{\{\sum_{i=1}^{n} S_{XX_i}\}^{\frac{1}{2}}\{\sum_{i=1}^{n} S_{TT_i}\}^{\frac{1}{2}}} = \frac{S_{XT}}{\{S_{XX}S_{TT}\}^{\frac{1}{2}}}.$$

Note that this is not the same estimate as the Monte Carlo approximation of ρ_c done in section 2.2.1 on page 11. In practice we do not enjoy the luxury of two independent datasets from the same distribution. And as we will see in section A.2 on page 91, it is harder to derive the asymptotic properties of r_c. Still, if we want to know which values of r_c which imply significant departure from H_0, these properties must be investigated. Three desirable properties[4] of r_c are given in the following theorem:

[4] A sequence of random variables X_n is said to converge in probability to X if for all $\epsilon > 0$ $P(d(X_n, X) > \epsilon) \to 0$ for a proper norm d. This is denoted $X_n \xrightarrow{P} X$. The sequence X_n is said to converge in distribution to X if $P(X_n \leq x) \to P(X \leq x)$ for every x which the limit distribution function $P(X \leq x)$ is continuous. This is denoted $X_n \xrightarrow{d} X$.

Theorem 2.2.5

Let

$$\widehat{\text{var}}(r_c) = r_c^2 \sum_{i=1}^{n} \left(\frac{S_{XX_i}}{S_{XX}} + \frac{S_{TT_i}}{S_{TT}} - 2\frac{S_{XT_i}}{S_{XT}} \right)^2.$$

Then:

$$\begin{aligned} & r_c \xrightarrow{P} \rho_c, \\ & n\widehat{\text{var}}(r_c) \xrightarrow{P} \sigma^2, \\ & \sqrt{n}\,(r_c - \rho_c) \xrightarrow{d} N(0, \sigma^2). \end{aligned} \tag{2.8}$$

Proof: A proof is given in section A.2 on page 91.

We do not give an explicit expression for the asymptotic variance σ^2 because it depends on the distribution of the data, and because it is difficult to derive. For practical purposes we only need to know how to estimate σ^2 consistently so that we can form a statistic capable of determining significant departure from H_0. Such a statistic is given in the following lemma.

Lemma 2.2.6

$$\frac{r_c - \rho_c}{\sqrt{\widehat{\text{var}}(r_c)}} \xrightarrow{d} N(0,1) \tag{2.9}$$

Proof: The proof is straightforward:

$$\frac{r_c - \rho_c}{\sqrt{\widehat{\text{var}}(r_c)}} = \left\{ \frac{\sqrt{n}\,(r_c - \rho_c)}{\sigma} \right\} \left\{ \frac{\sigma}{\{n\widehat{\text{var}}(r_c)\}^{1/2}} \right\} = a_n b_n$$

By theorem 2.2.5

$$\begin{aligned} a_n &\xrightarrow{d} N(0,1) \\ b_n &\xrightarrow{P} 1 \end{aligned}$$

Hence by *Slutsky's Theorem*[5]

$$a_n b_n \xrightarrow{d} N(0,1) \quad \text{and the proof is complete.}$$

2.2.3 Testing the assumption of quasi independence with r_c

Lemma 2.2.6 on the preceding page provides the means for testing the hypothesis $H_R : \rho_c = 0$ versus $H_R^c : \rho_c \neq 0$. For sufficiently large n, reject H_R whenever

$$|R| = \left| \frac{r_c}{\sqrt{\widehat{\text{var}}(r_c)}} \right| > Z_{\epsilon/2}, \tag{2.10}$$

where ϵ denotes the significance level of the test and $Z_{\epsilon/2}$ the corresponding normal critical value. In general, when H_R is rejected, we can only conclude that there is no linear relationship between the variables in the observable area. However, assume further investigation implies that the data follows a truncated bivariate normal distribution. Then rejecting H_R is, according to theorem 2.2.3 on page 11, equivalent to rejecting the hypothesis of quasi independence H_0. In section 6.5 on page 79 we consider a goodness of fit test which can be used to test whether or not the data follows a truncated bivariate normal distribution. This test and ρ_c are together useful tools when we wish to test the hypothesis of quasi independence.

2.2.4 Simulation result

To support Theorem 2.2.5 on the previous page a simulation was carried out in R. The following routine was repeated 400 times for every fixed combination of $n = 30, 80, 150$ and $\rho = 0, 0.3, 0.7$:

- n pairs were drawn from the truncated bivariate normal distribution $TN_2(0, -1, 1, 1/4, \rho)$.
- From these n pairs, r_c and $\widehat{\text{var}}(r_c)$ were computed.

[5]Slutsky: Let X_n, X and Y_n be random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} a$ for a constant a, then
(i) $X_n + Y_n \xrightarrow{d} X + a$
(ii) $X_n Y_n \xrightarrow{d} Xa$
(iii) $X_n / Y_n \xrightarrow{d} X/a$, if $a \neq 0$.

For each fixed combination of ρ and n the samples $r_c^1, \ldots, r_c^{400}$ and $\hat{\text{var}}(r_c^1), \ldots, \widehat{\text{var}}(r_c^{400})$ were obtained. We then calculated:

$$AVE(r_c) = \frac{1}{400}\sum_{i=1}^{400} r_c^i$$

$$AVE(\hat{V}) = \frac{1}{400}\sum_{i=1}^{400} \widehat{\text{var}}(r_c^i)$$

$$VAR(r_c) = \frac{1}{399}\sum_{i=1}^{400} (r_c^i - AVE(r_c))^2,$$

The results can be seen in table 2.2 on the following page. The motivation of the simulation was to support the following:

- *Consistency of r_c*
  For every ρ, the theoretical value of ρ_c is computed using the Monte Carlo approximation described earlier. We then compare $AVE(r_c)$ against ρ_c as n increases. This procedure will detect bias.

- *Consistency of $n\widehat{\text{var}}(r_c)$*
  This evaluation is more complex since we do not know the real value of σ^2. However, by repeating the routine 400 times we can compute the empirical variance $VAR(r_c)$ which should be an accurate approximation of $\text{var}(r_c)$. We can then compare $AVE(\hat{V})$ against $VAR(r_c)$. Note that both quantities should decrease when n increase.

- *Normality of r_c*
  To investigate the normality of r_c the p-value of the Shapiro-Wilks statistic was computed in each case. For small p-values this test rejects the hypothesis that the 400 computed values of r_c follows a normal distribution.

Results We get that $AVE(r_c)$ is close to ρ_c in all cases, and the estimate improves as n increases. For $n = 80$ and $n = 150$ the mean of the estimated asymptotic variances $AVE(\hat{V})$ is close to the empirical variance $VAR(r_c)$, and they both approach zero as n increase. A closer investigation of the $\rho = 0.7$ case showed a few outliers of r_c. As can be seen in table 2.2 on the next page, these outliers greatly affected the Shapiro-Wilks test of normality. Removing

the few outliers improved the p-values significantly, though such a procedure is considered to be one of the "deadly sins" amongst statisticians. In the cases when $\rho = 0$ and $\rho = 0.3$, the assumption of normality is not rejected. A similar simulation was carried out by Chen *et al.* (1996) with similar results.

Table 2.2: Simulation results of r_c from truncated samples sized n of a bivariate normal distribution with $\mu_x = 0$, $\mu_t = -1$, $\sigma_x^2 = 1$ and $\sigma_t^2 = 1/4$.

ρ	ρ_c		n=30	n=80	n = 150	Truncated proportion
0	0	$AVE(r_c)$	0.0050	0.0024	-0.0023	0.1858
		$VAR(r_c)$	0.0323	0.0112	0.0054	
		$AVE(\hat{V})$	0.0248	0.0101	0.0054	
		Normal p	0.3802	0.7028	0.4453	
0.3	0.1772	$AVE(r_c)$	0.1761	0.1768	0.1787	0.1524
		$VAR(r_c)$	0.0265	0.0082	0.0055	
		$AVE(\hat{V})$	0.0222	0.0091	0.0048	
		Normal p	0.8097	0.8206	0.6519	
0.7	0.4633	$AVE(r_c)$	0.4882	0.4648	0.4646	0.0882
		$VAR(r_c)$	0.0139	0.0054	0.0036	
		$AVE(\hat{V})$	0.0134	0.0054	0.0032	
		Normal p	0.0081	0.0247	0.0531	

Notice how $AVE(\hat{V})$ and $VAR(r_c)$ in table 2.2 on the facing page both decrease when ρ increases. We know from classical statistics that the sampling variance of the sample correlation is approximately

$$\frac{(1-\rho^2)^2}{n}.$$

Thus the sample correlation becomes more accurate as $|\rho| \to 1$. As seen in figure 2.3, $|\rho_c - r_c|$ is smaller and vary less when $|\rho_c| \to 1$, so there seem to be a similar relation between r_c and ρ_c. When the original data comes from the bivariate normal distribution, the value of ρ influence the truncated proportion $(N-n)/N$. As seen in table 2.2 on the facing page, increasing ρ decreases the truncated proportion. This will also affect r_c. We conclude that the dependency structure of the observed data influence the accuracy of r_c.

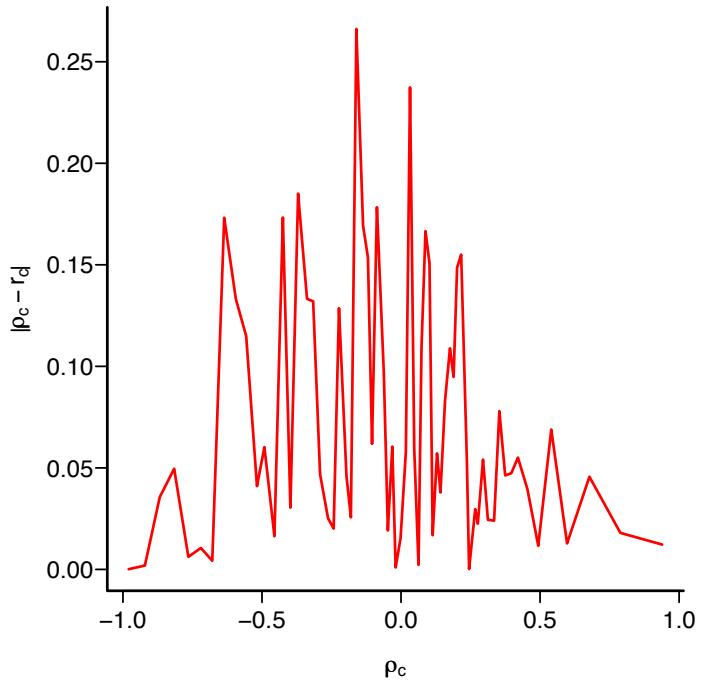

Figure 2.3: Plot of 67 computations of $|\rho_c - r_c|$. Every r_c was computed from the truncated bivariate normal distribution $TN_2(0, -1, 1, 1, \rho)$ where $n = 100$ and ρ varied from -1 to 1 (causing ρ_c to vary from -1 to 1).

2.3 Conditional Kendall's Tau

Similar to the standard Pearson correlation, ρ_c measures the linear relationship between the variables in the observable region. In addition, it also depends on the marginal distribution of X and T, e.g. ρ_c is only defined when $E[(X_1 - X_2)^2 I_A]$ and $E[(T_1 - T_2)^2 I_A]$ are finite. This can pose as a problem if we are dealing with infinite-variance distributions. In these cases the conditional Kendall's tau is a more suitable measure of dependence.

The standard Kendall's tau is a measure of concordance for bivariate random vectors. Consider two points in $\mathbb{R}^2$, denoted (x_1, t_1) and (x_2, t_2). We say the points are concordant if $(x_1 - x_2)(t_1 - t_2) > 0$ and discordant if $(x_1 - x_2)(t_1 - t_2) < 0$. Let (X_1, T_1) and (X_2, T_2) be independent random vectors from the same distribution. If T tends to increase with X we expect the probability of concordance to be high relative to the probability of discordance. We expect the opposite if T tends to decrease with increasing X. Motivated by this, Kendall's tau is just the probability of concordance minus the probability of discordance for these pairs. The conditional version is defined in the same way for a truncated vector (X, T), only conditioned on the event A, that the two pairs are comparable. Applications and a generalised Kendall's tau statistic are discussed in (Tsai, 1990). Let us begin with the definition.

Definition 2.3.1

Given the random samples (X_1, T_1) and (X_2, T_2) from the truncated vector (X, T), we define the conditional Kendall's tau:

$$\tau_c = 2P\{(X_1 - X_2)(T_1 - T_2) > 0 | A\} - 1,$$

where as before, $A = \{\max(T_1, T_2) < \min(X_1, X_2)\}$.

In the unconditional case, when X and T are independent, we have that $P\{(X_1 - X_2)(T_1 - T_2) > 0) = 1/2$ and $\tau = 0$. Similarly, we have the following relation between τ_c and the assumption of quasi independence H_0.

Theorem 2.3.2

Given quasi independence between X and T, it follows that $\tau_c = 0$. That is:

$$H_0 \quad \Rightarrow \quad \tau_c = 0$$

Proof: Rewrite τ_c in the following way

$$\begin{aligned}\tau_c &= 2P\{(X_1 - X_2)(T_1 - T_2) > 0|A\} - 1 \\ &= P\{(X_1 - X_2)(T_1 - T_2) > 0|A\} + P\{(X_1 - X_2)(T_1 - T_2) > 0|A\} - 1 \\ &= P\{(X_1 - X_2)(T_1 - T_2) > 0|A\} + 1 - P\{(X_1 - X_2)(T_1 - T_2) < 0|A\} - 1 \\ &= E[\text{sgn}(X_1 - X_2)(T_1 - T_2)|A] = E[\text{sgn}(X_1 - X_2)(T_1 - T_2)I_A]/P(A),\end{aligned}$$

and consider the last expectation. Under the assumption $P(A) \neq 0$, the proof is completely analogous to that of theorem 2.2.2 on page 10 given in section A.1 on page 89, so the details are omitted.

2.3.1 Sample conditional Kendall's Tau and asymptotic properties of the corresponding test-statistic T

Definition 2.3.3

Let $(X_1, T_1), \ldots, (X_n, T_n)$ be i.i.d. random vectors following the same distribution as (X, T). Then the sample conditional Kendall's Tau is given by

$$t_c = \frac{1}{k} \sum\sum_{i<j} \text{sgn}\left((X_i - X_j)(T_i - T_j)\right) I_{ij}, \tag{2.11}$$

where $I_{ij} = I\{\max(T_i, T_j) < \min(X_i, X_j)\}$ and $k = \sum\sum_{i<j} I_{ij}$.

To test the assumption $\tau_c = 0$ we must consider the properties of a simplified version of t_c. Let K be the number of concordant comparable pairs minus the number of discordantly comparable pairs. Thus

$$K = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{sgn}\left((X_i - X_j)(T_i - T_j)\right) I_{ij}.$$

To establish the asymptotic properties of K we need to rewrite it. We define the set $\mathscr{R}_i$ and the cardinality of $\mathscr{R}_i$ by

$$\mathscr{R}_i = \{j|T_j \leq X_i \leq X_j\}, \quad R_i = \sum_{j=1}^{n} I(T_j \leq X_i \leq X_j) = \text{card}(\mathscr{R}_i).$$

In the expression of K all elements appear twice since

$$\operatorname{sgn}\left((X_i - X_j)(T_i - T_j)\right) I_{ij} = \operatorname{sgn}\left((X_j - X_i)(T_j - T_i)\right) I_{ji}.$$

By summing over $j \in \mathscr{R}_i$ we avoid this and we do not need to divide by one half. In addition, when $j \in \mathscr{R}_i$, the indicator function I_{ij} will be 1, so we can omit the indicator function as well. If we assume that the distribution of (X, T) is continuous we can ignore the probability of ties[6]. For every X_j for which $j \in \mathscr{R}_i$ we then have that $\operatorname{sgn}(X_j - X_i) = 1$. This leads to the following representation of K

$$K = \sum_{i=1}^{n} \sum_{j \in \mathscr{R}_i} \operatorname{sgn}(T_j - T_i) = \sum_{i=1}^{n} S_i.$$

The rewriting of K is motivated by the following nice result about the random variables S_i.

Theorem 2.3.4

Assume that the distribution of (X, T) is continuous so that the probability of ties can be ignored. Under H_0 the conditional distribution of S_i given the set $\mathscr{R}_i$ is uniform. The probability mass function is given by

$$\begin{aligned} f_i(j) &= P(S_i = j \mid R_i = r_i) \\ &= \begin{cases} \dfrac{1}{r_i} & j = r_i - 1, r_i - 3, \ldots, -r_i + 3, -r_i + 1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

$$\text{Hence} \quad E(S_i | R_i = r_i) = 0, \quad \operatorname{Var}(S_i | R_i = r_i) = \frac{1}{3}(r_i^2 - 1).$$

Proof: A proof in the case $R_i = 3$ is given in section A.3 on page 96.

A visual inspection of this theorem can be seen in figure 2.4 and figure 2.5 on page 24. Note that $\rho = 0$ imply quasi independence in the truncated bivariate normal case. Figure 2.4 was generated by drawing a sample from $TN_2(0, -1, 3, 1, \rho = 0)$ of size $n = 50$. If the sample contained a set $\mathscr{R}_i$ with cardinality $R_i = 10$, the corresponding value of S_i were calculated. This routine

[6] A tie is when the realisation of two variables are equal.

was repeated until 12 000 values of $S_i|R_i = 10$ were attained.

At first eyesight this result does not seem to help us decide significant departure from H_0. In practice we only have one data set, so we can't e.g. make a QQ-plot to check if $S_i|R_i = r_i$ actually is uniform. However, the application of this result becomes clear in the next theorem.

Theorem 2.3.5

Assume that the distribution of (X, T) is continuous and that the assumption H_0 holds, then

$$T = \frac{K}{\{\frac{1}{3}\sum_{i=1}^{n}(r_i^2 - 1)\}^{\frac{1}{2}}} \xrightarrow{d} N(0,1).$$

Sketched proof:
It can be shown, see (Tsai, 1990, page 173), that conditioned on $(R_1 = r_1, \ldots, R_n = r_n)$, $S_1, \ldots, S_n$ are mutually independent. Hence K is the sum of conditionally independent variables S_i. By theorem 2.3.1 on the preceding page it then follows that

$$E(K|R_1 = r_1, \ldots, R_n = r_n) = \sum_{i=1}^{n} E(S_i|R_i = r_i) = 0,$$
$$\mathrm{Var}(K|R_1 = r_1, \ldots, R_n = r_n) = \sum_{i=1}^{n} \mathrm{Var}(S_i|R_i = r_i) = \frac{1}{3}\sum_{i=1}^{n}(r_i^2 - 1).$$

And since K is a sum of independent variables it is possible to use the *central limit theorem*[7] on

$$T = \frac{\sum_{i=1}^{n}(S_i - E(S_i|R_i = r_i))}{\sum_{i=1}^{n}\mathrm{Var}(S_i|R_i = r_i)} = \frac{K}{\{\frac{1}{3}\sum_{i=1}^{n}(r_i^2 - 1)\}^{\frac{1}{2}}}.$$

We can't apply the classical central limit theorem since the variances $\mathrm{Var}(S_i|R_i = r_i)$ are not equal. However, the result follows from Lindebergs central limit theorem if the Lindeberg condition holds. In Tsai (1990) this is verified by evaluating the stronger Lyapunov condition.

[7] Let $\bar{X}_n$ be the average of the first n variables of a sequence of independent, identically distributed random variables $X_1, X_2, \ldots$. If $E|X_1|^2 < \infty$ the central limit theorem asserts that $\sqrt{n}(\bar{X}_n - EX_1) \xrightarrow{d} N(0, varX_1)$

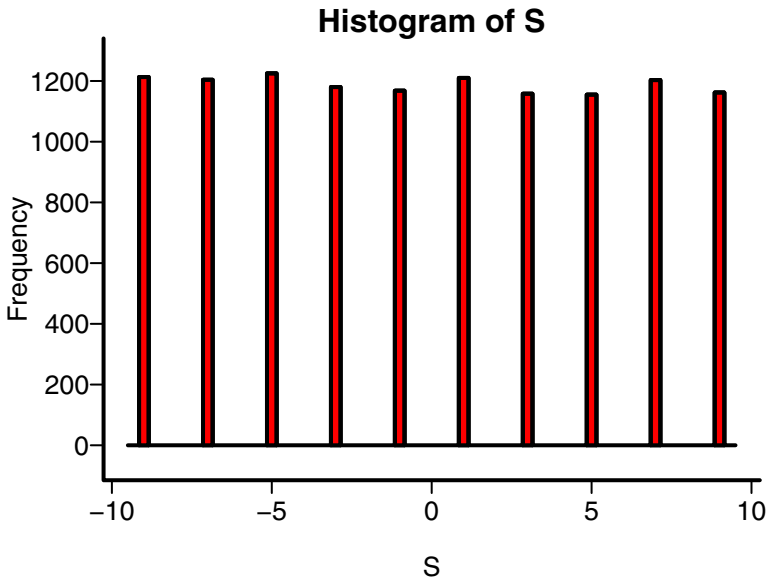

Figure 2.4: 12000 i.i.d. $S_i|R_i = 10$ drawn from 12000 samples from the truncated bivariate normal distribution $TN_2(0, -1, 3, 1, \rho = 0)$, each of size $n = 50$.

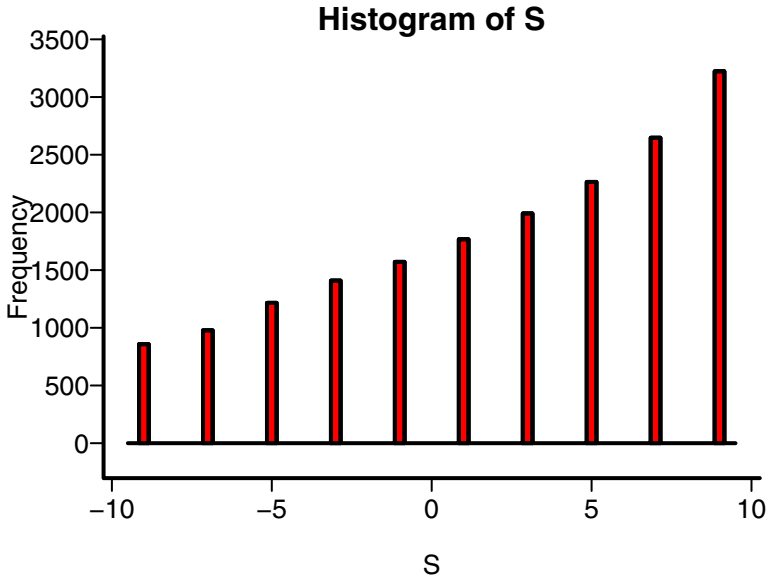

Figure 2.5: The effect when the variables are not quasi independent: The same procedure as above only with $TN_2(0, -1, 3, 1, \rho = 0.7)$.

2.3.2 Testing the assumption of quasi independence with T

Using theorem 2.3.5 on page 23 we can test the hypothesis $H_T : \tau_c = 0$ versus $H_T^c : \tau_c \neq 0$. For sufficiently large n, reject H_T whenever

$$|T| = \left| \frac{K}{\{\frac{1}{3}\sum_{i=1}^{n}(r_i^2 - 1)\}^{\frac{1}{2}}} \right| > Z_{\epsilon/2}\ , \tag{2.12}$$

where ϵ denotes the significance level of the test and $Z_{\epsilon/2}$ the corresponding normal critical value. Analogous to accepting the hypothesis $\rho_c = 0$, accepting the hypothesis $\tau_c = 0$ do not imply quasi independence between the variables in general. And opposed to ρ_c, there is no direct link between τ_c and H_0 when the data follows a truncated bivariate normal distribution. However, this can be a more suitable test when data do not follow a truncated bivariate normal distribution. In chapter 5 we will see how the conditional Kendall's tau can be used to estimate copula parameters.

The following example illustrates that T is invariant to strictly increasing transformations of the data.

Example 2.3.6

Table 2.3 was made by first calculating the statistics R and T using a sample from $TN_2(0, -1, 2, 2, \rho = 0)$ of size $n = 100$. Afterwards, the same statistics were calculated from the exponentially transformed data. We see that the R statistic is not invariant for such a transformation of the data, while T is.

Table 2.3: Invariance of T

	data	transformed data
R	-0.087	-0.437
P-value	0.465	0.331
T	-0.22	-0.22
P-value	0.41	0.41

2.3.3 Monte Carlo approximation of τ_c

The calculation of τ_c can be done similar to the Monte Carlo approximation of ρ_c. With the same notations as in section 2.2.1 on page 11 let

$$\hat{\tau}_c = \frac{1}{n_A} \sum_{i=1}^{n} \operatorname{sgn}\left((X_i^A - X_i^B)(T_i^A - T_i^B)\right) I_{A_i},$$

$$\text{where} \quad A_i = \{\max(T_i^A, T_i^B) < \min(X_i^A, X_i^B)\}, \quad n_A = \sum_{i=1}^{n} I_{A_i}\,.$$

This is the average of the function $\operatorname{sgn}((X_1 - X_2)(T_1 - T_2))$ amongst the comparable pairs. By the strong law of large numbers we know that $\hat{\tau}_c$ converges almost surely to $E\left[\operatorname{sgn}((X_1 - X_2)(T_1 - T_2))\,|A\right] = \tau_c$. Hence, for a sufficiently large n, this is a good approximation of τ_c. Note that $\hat{\tau}_c$ is not the same as the sample conditional Kendall's tau t_c. In figure 2.6 this method is used to make a visual inspection of the relations between the unconditional correlation and τ_c when (X, T) follows a truncated bivariate normal distribution. For comparison, we include ρ_c in the plot.

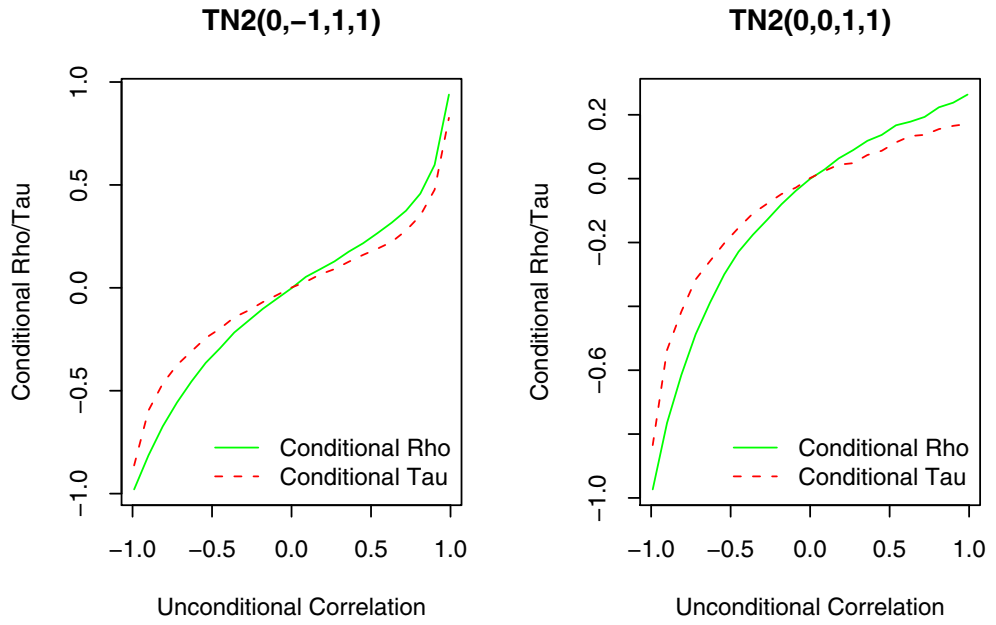

Figure 2.6: Unconditional correlation versus ρ_c and τ_c in two different truncated bivariate normal cases, where ρ varies from -1 to 1.

3

The Product-limit estimator

In this and the two following chapters we will consider the problem of reconstructing the distribution of (X^*, T^*) using data from the observed vector $(X, T) = (X^*, T^* | X^* > T^*)$. The focus of this chapter will be on the reconstruction of the marginal distributions F and G under the assumption that X and T are quasi independent. For this purpose we will consider a well established estimator known as the Product-limit estimator.

3.1 Definition and conditions

The sampling mechanism in the left truncated case will cause the distribution of X to be centred to the right of the distribution of X^*. This is natural since the sampling mechanism removes points in the lower domain of F. The Product-limit estimator (PLE) was derived by Lynden-Bell (1971) and is a nonparametric MLE of F. When data are given, the PLE attempts to correct for this bias by assigning higher weights to the smaller values of $(X_1, \ldots, X_n)$. A detailed discussion of the PLE weights can be found in (Stute and Wang,

2008). In this chapter we will give a short description of this estimator based on the work of Woodroofe (1985). However, we will define the PLE using a quantity called the number at risk, $R(\cdot)$.

Definition 3.1.1: Number at risk

Given the observations $(X_1, T_1), \ldots, (X_n, T_n)$ the number at risk at (x, t) is given by

$$R(x,t) = \sum_{i=1}^{n} I(X_i \geq x, T_i \leq t), \quad \tilde{R}(z) = R(z,z).$$

In a survival analysis setting, where X is truncated lifetime, $\tilde{R}(z)$ is the individuals at risk at time z ($T \leq z$) that have not yet died ($X \geq z$).

Definition 3.1.2: The Product-limit estimator

Let F and G be the distributions of X^* and T^* respectively. For the i.i.d. truncated observations $(X_1, T_1), \ldots, (X_n, T_n)$ let

$$r(z) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = z), \quad s(z) = \frac{1}{n}\sum_{i=1}^{n} I(T_i = z).$$

Then the Product-limit estimator of F is given by

$$\hat{F}_{PL}(z) = 1 \quad - \prod_{\text{distinct } X_i \leq z} \left(\frac{\tilde{R}(X_i) - r(X_i)}{\tilde{R}(X_i)} \right), \tag{3.1}$$

and the Product-limit estimate of G is

$$\hat{G}_{PL}(z) = \prod_{\text{distinct } T_i > z} \left(\frac{\tilde{R}(T_i) - s(T_i)}{\tilde{R}(T_i)} \right). \tag{3.2}$$

In both cases, an empty product is to be interpreted as one.

Consider the conditional likelihood of the observed data points $x_1, \ldots, x_n$ and $t_1, \ldots, t_n$, written as a function of F and G:

$$L_n = \prod_{i=1}^{n} \left(\mathrm{d}F(x_i)\, \mathrm{d}G(t_i) / \int\int_{u \geq v} \mathrm{d}F(u)\, \mathrm{d}G(v) \right).$$

Maximising this likelihood with respect to F and G results in the Product-limit estimator of F and G. Such a procedure is described in Wang *et al.* (1986) in the right truncation case. However, we will not study these details. Instead we will try to derive the PLE of F under some assumptions about its properties.

Assume that there are no ties in our observations. Hence $X_{(1)} < X_{(2)}, \ldots, < X_{(n)}$ and $r(X_i) = 1$, $1 \leq i \leq n$. We will assume that our estimator $\hat{F}_{PL}$ is a right continuous function with jumps at $X_1, \ldots, X_n$ and with support on $[X_{(1)}, X_{(n)}]$ so that $\hat{F}_{PL}(X_{(1)}-) = 0$. Similarly we assume that $\hat{G}_{PL}$ is a right continuous function with jumps at $T_1, \ldots, T_n$ and with support on $[T_{(1)}, T_{(n)}]$.

When X^* and T^* are independent we have that

$$P(T \leq z \leq X) = P(T^* \leq z \leq X^* | X^* > T^*) = \frac{1}{\alpha_0} G(z)(1 - F(z-)),$$

where $\alpha_0 = P(X^* > T^*)$. An estimation equation for F and G is therefore given by

$$\frac{\alpha_0}{n} \tilde{R}(z) = \hat{G}_{PL}(z)(1 - \hat{F}_{PL}(z-)), \tag{3.3}$$

which, when written in logarithm form is

$$\log\left(\frac{\alpha_0}{n}\tilde{R}(z)\right) = \log\left(\hat{G}_{PL}(z)\right) + \log\left(1 - \hat{F}_{PL}(z-)\right). \tag{3.4}$$

Next, let $X \in (X_1, \ldots, X_n)$ and remember that $X > T$ for the corresponding $T \in (T_1, \ldots, T_n)$. This means that jumps for the functions $\hat{G}_{PL}$ and $\hat{F}_{PL}$ will not occur at the same points. Consequently, we have that $\tilde{R}(X+) = \tilde{R}(X) - 1$ and $\hat{G}_{PL}(X+) = \hat{G}_{PL}(X)$. Hence, by subtracting 3.4 at $X+$ from 3.4 at X we get

$$\begin{aligned} \log\left(\frac{\alpha_0}{n}\tilde{R}(X)\right) - \log\left(\frac{\alpha_0}{n}(\tilde{R}(X) - 1)\right) \\ = \log(1 - \hat{F}_{PL}(X-)) - \log(1 - \hat{F}_{PL}(X)). \end{aligned} \tag{3.5}$$

Since $\hat{F}_{PL}$ is a right continuous step function with jumps at $X_1, \ldots, X_n$, we have that

$$\log\left(1 - \hat{F}_{PL}(X_{(i)})\right) = \log\left(1 - \hat{F}_{PL}(X_{(i+1)}-)\right), \quad 1 \leq i \leq n-1.$$

As we have assumed $\hat{F}_{PL}(X_{(1)}-) = 0$ we get that

$$\log\left(1 - \hat{F}_{PL}(X_{(1)}-)\right) = \log(1) = 0.$$

Therefore, if we let $X_{(a)} = \max(X_1 \ldots, X_n | X_i \leq z)$ and sum equation 3.5 over all X_i where $X_i \leq z$, we get

$$\begin{aligned}
&\sum_{X_i \leq z} \left[\log\left(\alpha_0 \frac{\tilde{R}(X_i)}{n}\right) - \log\left(\alpha_0 \frac{\tilde{R}(X_i) - 1}{n}\right)\right] = \underbrace{\log\left(1 - \hat{F}_{PL}(X_{(1)}-)\right)}_{0} \\
&+ \underbrace{\left[-\log\left(1 - \hat{F}_{PL}(X_{(1)})\right) + \log\left(1 - \hat{F}_{PL}(X_{(2)}-)\right)\right]}_{0} \\
&+ \underbrace{\left[-\log\left(1 - \hat{F}_{PL}(X_{(2)})\right) + \log\left(1 - \hat{F}_{PL}(X_{(3)}-)\right)\right]}_{0} + \ldots \\
&+ \underbrace{\left[-\log\left(1 - \hat{F}_{PL}(X_{(a-1)})\right) + \log\left(1 - \hat{F}_{PL}(X_{(a)}-)\right)\right]}_{0} - \log\left(1 - \hat{F}_{PL}(X_{(a)})\right) \\
&= -\log\left(1 - \hat{F}_{PL}(X_{(a)})\right).
\end{aligned}$$

It follows from the definition of $X_{(a)}$ that $\hat{F}_{PL}(X_{(a)}) = \hat{F}_{PL}(z)$, so the above equation becomes

$$\begin{aligned}
\log(1 - \hat{F}_{PL}(z)) &= -\sum_{X_i \leq z} \left[\log\left(\alpha_0 \frac{\tilde{R}(X_i)}{n}\right) - \log\left(\alpha_0 \frac{\tilde{R}(X_i) - 1}{n}\right)\right] \\
&= \sum_{X_i \leq z} \log\left(\frac{\tilde{R}(X_i) - 1}{\tilde{R}(X_i)}\right) \\
&= \log\left(\prod_{X_i \leq z} \left(\frac{\tilde{R}(X_i) - 1}{\tilde{R}(X_i)}\right)\right).
\end{aligned}$$

Thus, an estimator of F is given by

$$\hat{F}_{PL}(z) = 1 - \prod_{X_i \leq z} \left(\frac{\tilde{R}(X_i) - 1}{\tilde{R}(X_i)}\right),$$

which equals the definition given by equation 3.1 on page 28 when there are no ties amongst $(X_1, \ldots, X_n)$. Analogous procedures can be done for the

truncating variables $(T_1, \ldots, T_n)$ to get

$$\hat{G}_{PL}(z) = \prod_{T_i > z} \left(\frac{\tilde{R}(T_i) - 1}{\tilde{R}(T_i)} \right).$$

Note that the definition of $\hat{F}_{PL}$ given in equation 3.1 on page 28 also implies that $\hat{F}_{PL}$ is supported on $[X_{(1)}, X_{(n)}]$. This may seem strange since the support of F most likely extends below $X_{(1)}$ and above $X_{(n)}$. However, values within the interval $[X_{(1)}, X_{(n)}]$ are assigned approximately the correct mass (See figure 3.1 on page 36). Under the conditions given in the following theorem (following Woodroofe (1985)), the interval $[X_{(1)}, X_{(n)}]$ will extend to the support of F as $n \to \infty$.

Theorem 3.1.3: Conditions for consistent reconstructions of *F* and *G*

Put

$$a_F = \inf\{x : F(x) > 0\}, \quad b_F = \sup\{x : F(x) < 1\},$$
$$a_G = \inf\{t : G(t) > 0\}, \quad b_G = \sup\{t : G(t) < 1\}.$$

If X^* and T^* are independent, then $\hat{F}_{PL}$ is a consistent reconstruction of F only if $a_G \leq a_F$. Similar, $\hat{G}_{PL}$ is a consistent reconstruction of G only if $b_G \leq b_F$.

Proof: A proof can be found in (Woodroofe, 1985).

These conditions are reasonable: If $a_G > a_F$, we will never get information about the region at which $X^* < a_G$ because of the sampling mechanism. Consequently, F can't be fully recovered. Similar, if $b_G > b_F$, we will never get information about the region at which $T^* > b_F$. Note that if F and G are continuous, the conditions are $G^{-1}(0) \leq F^{-1}(0)$ and $G^{-1}(1) \leq F^{-1}(1)$.

Originally, the Product-limit estimator is said to be applicable when X^* and T^* are independent. Tsai (1990) pointed out that the asymptotic properties would remain the same if we assume quasi independence, though he did not give any explicit arguments of why this is so. We leave this verification open, but point out that the estimation equation 3.3 on page 29 is valid also under the assumption of quasi independence.

3.2 A problematic property of the PLE

Assume that there are no ties among $X_1, \ldots, X_n$ so that $r(X_{(i)}) = 1$ for $1 \leq i \leq n$. If

$$\tilde{R}(X_{(i)}) = 1 \quad \text{for some } i, \quad 1 \leq i < n, \tag{3.6}$$

then

$$\frac{\tilde{R}(X_{(i)}) - r(X_{(i)})}{\tilde{R}(X_{(i)})} = 0, \quad \text{hence} \quad F(x_{(i)}) = 1.$$

The following example illustrates how this may lead to unreasonable results

Example 3.2.1

Consider the pair $(x_{(1)}, t)$ in a sample of n pairs (x_i, t_i). Assume all the other truncating variables t_i in the sample are larger than $x_{(1)}$. Then $\tilde{R}(x_{(1)}) = \sum_{i=1}^{n} I(t_i \leq x_{(1)}) = I(t \leq x_{(1)}) = 1$, Thus

$$\hat{F}_{PL}(x_{(1)}) = 1.$$

We see how important independence is since such samples are most likely obtained when there is a positive association between X^* and T^*. However, 3.6 may occur in the independent case as well. Woodroofe (1985) showed that the probability of 3.6 occurring is asymptotically negligible in the independent case. As a precautionary measure, one should always check the values of $\tilde{R}(x_{(i)})$ for $1 \leq i \leq n$. Hopefully, only $\tilde{R}(x_{(n)})$ should be equal to 1. If not, Woodroofe (1985) suggested replacing $\tilde{R}$ by

$$\tilde{R}^*(z) = \max\{\tilde{R}(z), nk_n(z)\}, \quad 0 \leq z \leq x_{(n)},$$

where k_n is a non-increasing function for which $k_n(z) > k_n(x_{(n)}) = \frac{1}{n}$ for all $z < x_{(n)}$. In this way $\tilde{R}^*(x_i) > 1$ for $1 \leq i < n$, and $\tilde{R}^*(x_{(n)}) = 1$.

3.3 Applications of the PLE

Given the Product-limit estimators of F and G we have the following nonparametric maximum likelihood estimates[1]:

$$\hat{\mu}_x = \int_{-\infty}^{\infty} u \, \mathrm{d}\hat{F}_{PL}(u) = \sum_{i=1}^{n} x_{(i)} \hat{F}_{PL}\{x_{(i)}\}, \tag{3.7}$$

$$\hat{\mu}_t = \int_{-\infty}^{\infty} v \, \mathrm{d}\hat{G}_{PL}(v) = \sum_{i=1}^{n} t_{(i)} \hat{G}_{PL}\{t_{(i)}\},$$

$$\hat{\sigma}_x^2 = \int_{-\infty}^{\infty} (u - \hat{\mu}_x)^2 \, \mathrm{d}\hat{F}_{PL}(u) = \sum_{i=1}^{n} (x_{(i)} - \hat{\mu}_x)^2 \hat{F}_{PL}\{x_{(i)}\},$$

$$\hat{\sigma}_t^2 = \int_{-\infty}^{\infty} (v - \hat{\mu}_t)^2 \, \mathrm{d}\hat{G}_{PL}(v) = \sum_{i=1}^{n} (t_{(i)} - \hat{\mu}_t)^2 \hat{G}_{PL}\{t_{(i)}\},$$

where $\hat{F}_{PL}\{x_{(i)}\}$ and $\hat{G}_{PL}\{t_{(i)}\}$ are the Product-limit density functions given by:

$$\begin{aligned}
&\hat{F}_{PL}\{x_{(i)}\} = \hat{F}_{PL}(x_{(i)}) - \hat{F}_{PL}(x_{(i-1)}), \quad 2 \leq i \leq n, \\
&\hat{F}_{PL}\{x_{(1)}\} = \hat{F}_{PL}(x_{(1)}), \\
&\hat{G}_{PL}\{t_{(i)}\} = \hat{G}_{PL}(t_{(i)}) - \hat{G}_{PL}(t_{(i-1)}), \quad 2 \leq i \leq n, \\
&\hat{G}_{PL}\{t_{(1)}\} = \hat{G}_{PL}(t_{(1)}).
\end{aligned}$$

By using Product-limit estimates of F and G it is also possible to estimate $\alpha_0 = P(X^* > T^*)$. First, note that

$$\alpha_0 = \int \int_{u>v} \mathrm{d}F(u) \, \mathrm{d}G(v) = \int_{-\infty}^{\infty} \Big[\int_{-\infty}^{u} \mathrm{d}G(v) \Big] \, \mathrm{d}F(u) = \int_{-\infty}^{\infty} G(u) \, \mathrm{d}F(u).$$

Hence the nonparametric maximum likelihood estimator of α_0 is

$$\hat{\alpha}_0 = \int_{-\infty}^{\infty} \hat{G}_{PL}(u) \, \mathrm{d}\hat{F}_{PL}(u) = \sum_{i=1}^{n} \hat{G}_{PL}(x_{(i)}) \hat{F}_{PL}\{x_{(i)}\}. \tag{3.8}$$

[1] The parameters μ_x, μ_t, σ_x^2 and σ_t^2 denotes $E(X^*)$, $E(T^*)$, $\mathrm{Var}(X^*)$ and $\mathrm{Var}(T^*)$ respectively.

3.4 Simulation result

To evaluate the performance of the PLE, two separate simulations were carried out in R. The first simulation is simply a re-do of the example given in (Woodroofe, 1985). In the second simulation the PLE is tested in the independent truncated bivariate normal case.

3.4.1 Simulation 1: The independent uniform case

Assume X^* and T^* are independent and uniformly distributed on the unit interval. Obviously, $\alpha_0 = 1/2$. The joint distribution of the truncated variables X and T is then given by equation 2.3 on page 7:

$$\begin{aligned} F^c(x,t) &= \int\int_{\Delta(x,t)} \mathrm{d}F(u)\,\mathrm{d}G(v)/\alpha_0 = 2\int_0^x \int_0^{\min(t,u)} \mathrm{d}v\,\mathrm{d}u \\ &= 2\int_0^x \min(t,u)\,\mathrm{d}u = 2(xt - \frac{1}{2}t^2), \quad 0 \le x \le 1, \quad 0 \le t < 1, \quad x > t. \end{aligned} \tag{3.9}$$

Note that the last equality is only valid when $x > t$. So marginally, since $F^c(x) = H^c(x,1)$, we have to put $t = 1$ in the last integral. We then get

$$\begin{aligned} &F^c(x) = 2\int_0^x u\,\mathrm{d}u = x^2, \quad 0 \le x \le 1, \\ &\text{hence} \quad E(X) = \int_0^1 P(X > u)\,\mathrm{d}u = \int_0^1 (1-u^2)\,\mathrm{d}u = \frac{2}{3}\,. \end{aligned}$$

In the simulation, $n = 10$ pairs of (x,t) were drawn from the distribution given in 3.9. It turned out in this case that $\tilde{R}(x_{(i)}) > 1$ for $1 \le i < n$, so we did not have to use $\tilde{R}^*$ in the calculation of the PLE. The result can be viewed in table 3.1. Here $\hat{F}_{PL}(x_{(i)})$ should be compared with $x_{(i)}$ since $F(x) = x$, $0 \le x \le 1$. The sample average was $\bar{x} = 0.6437$, which is close to $E(X) = 2/3$. The MLE of the mean given by equation 3.7 on the previous page was $\hat{\mu}_x = 0.5603$, which is quite close to $E(X^*) = 1/2$. The MLE of α_0 given by equation 3.8 on the preceding page was $\hat{\alpha}_0 = 0.5925$, also somewhat close to $\alpha_0 = 1/2$. It would be optimistic to hope for a better result when the calculation was done using only ten data points. In fact, repetition of the simulation revealed a rather erratic behaviour of the PLE and the resulting estimates of α_0 and μ_x. By increasing n

the estimates became more accurate, supporting the consistency of the PLE.

i	$x_{(i)}$	t	$\tilde{R}(x_{(i)})$	$\hat{F}_{PL}(x_{(i)})$	$\hat{F}_{PL}\{x_{(i)}\}$
1	0.2575	0.2363	5	0.2000	0.2000
2	0.4087	0.1695	6	0.3333	0.1333
3	0.4357	0.3765	5	0.4666	0.1333
4	0.6438	0.0420	6	0.5555	0.0888
5	0.6658	0.6285	5	0.6444	0.0888
6	0.6724	0.5971	4	0.7333	0.0888
7	0.7225	0.3251	3	0.8222	0.0888
8	0.8203	0.1389	3	0.8814	0.0592
9	0.8970	0.7317	2	0.9407	0.0592
10	0.9129	0.0816	1	1.0000	0.0592

Table 3.1: Calculations of the PLE in the independent uniform case. The estimated MLE of the mean and the sample average are $\hat{\mu}_x = 0.5603$ and $\bar{x} = 0.6437$.

3.4.2 Simulation 2: The independent normal case

In this simulation, $n = 100$ pairs of (X, T) where drawn from a $TN_2(0, -1, 2, 2, 0)$ distribution. Also in this simulation, $\tilde{R}(x_{(i)}) > 1$ for $1 \leq i < n$, so we did not need to use $\tilde{R}^*$ in the calculation of the PLE. A visual representation of the estimated PLE is displayed in figure 3.1. The real distribution of X^* is included in the figure. In the interval $[x_{(1)}, x_{(n)}]$ the PLE is quite close to the real distribution, but the estimation error is a little larger in the lower end of this interval. This is the penalty we receive when we assign larger weights to the smaller values of $(x_1, \ldots, x_n)$.

By the definition of a truncated bivariate normal distribution, $X^* \sim N(0, 2)$. The MLE of the mean is $\hat{\mu}_x = 0.0534$, which is quite close to $EX^* = 0$. This should be compared to the sample average $\bar{x} = 0.5372$. The MLE of the variance is 1.7168, also quite close to $\text{Var}\, X^* = 2$. By Monte Carlo approximation, the true value of α_0 is 0.6922. The MLE of α_0 given by equation 3.8 on page 33 is 0.6763.

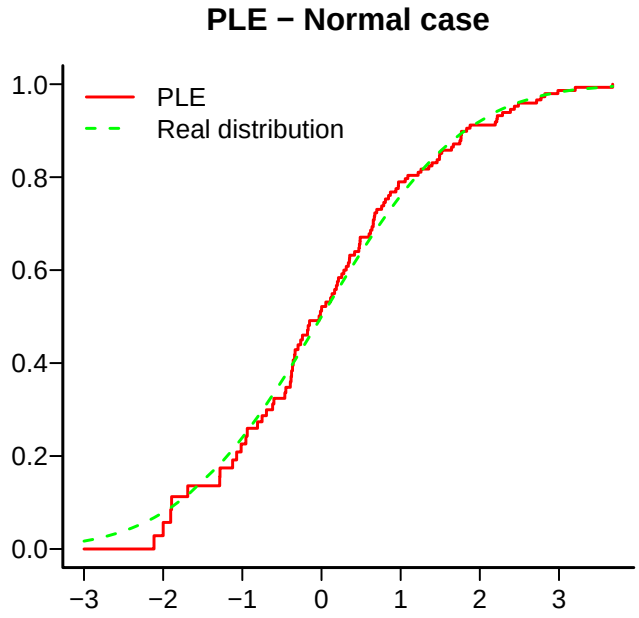

Figure 3.1: Calculation of the PLE of F in the independent, truncated bivariate normal case $TN_2(0, -1, 2, 2, 0)$, $n = 100$. Note that $\hat{F}_{PL}(z) = 0$ for $z < x_{(1)} = -2.11$.

3.5 The generalised inverse of the PLE

Assume that we wish to simulate from the PLE, by generating a random variable X^* with the distribution $\hat{F}_{PL}$. The standard procedure is then to let $X^* = \hat{F}_{PL}^{-1}(U)$ where $U \sim U[0,1]$ and $\hat{F}_{PL}^{-1}(u) = \inf\{x : \hat{F}_{PL}(x) \geq u\}$. In the continuous case this is valid since

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

Simulating from the generalised inverse of the PLE can be done using the following algorithm.

- Let $\{x_{(i)}\}$, $1 \leq i \leq n$ be the ordered values of the sample,
- Apply $\hat{F}_{PL}$ on the sequence $\{x_{(i)}\}$,
- Draw $U \sim U[0,1]$ and let $X^* = \min\{x_{(i)} : \hat{F}_{PL}(x_{(i)}) \geq U\}$.

Then X^* is distributed according to $\hat{F}_{PL}$. Note that we may only draw values equal to values in the observed data. If we want to simulate from the conditional empirical distribution $F_n^c(z) = 1/n \sum_{i=1}^{n} I(X_i \leq z)$ we may simply draw uniformly from the observed data.

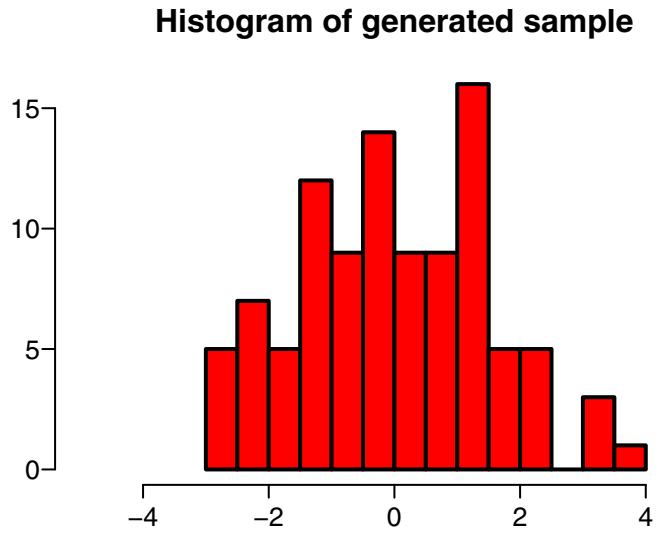

Figure 3.2: Histogram of 100 random variables drawn from $\hat{F}_{PL}$ using the algorithm described above. $\hat{F}_{PL}$ was made using $n = 100$ i.i.d. $TN_2(0, -1, 2, 2, 0)$.

4

Maximum likelihood estimation in the truncated bivariate normal case

For data subject to a dependent truncation the PLE is not applicable. For such data we must model the dependence between X^* and T^* in some way. For example, we can assume that (X^*, T^*) follows some joint distribution function. We can then derive the conditional distribution of (X, T) and choose parameters that fits the observed data. This can be done by using the maximum likelihood method.

4.1 Basic properties and definitions

Let $\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n$ be i.i.d. $f_X(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{X}_i \in \mathbb{R}^p$. Then the likelihood function is defined by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(\boldsymbol{x}_i|\boldsymbol{\theta}).$$

The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}$ which maximises the likelihood function, hence the name. For such estimates we have the following result

Theorem 4.1.1

Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$. Then under suitable regularity conditions on $f_X(\boldsymbol{x}|\boldsymbol{\theta})$

$$\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}.$$

That is, $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$.

The "suitable regularity conditions" can be found in (Casella and Berger, 1990, page 516) for a scalar parameter. For a vector parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ more assumptions are required, but they are usually satisfied in reasonable problems.

Definition 4.1.2

The Fisher-information is defined by

$$I(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log(L(\boldsymbol{\theta})).$$

The Fisher-information is a $d \times d$ matrix. When evaluated at $\hat{\boldsymbol{\theta}}$ we call it the observed Fisher-information.

Theorem 4.1.3

Assume $\hat{\boldsymbol{\theta}}$ is a consistent MLE of $\boldsymbol{\theta}$. Then

$$I^{1/2}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(0, \boldsymbol{I}_d),$$

where $\boldsymbol{I}_d$ is the identity matrix of size d.

As a result we have that for a finite sample estimate $\hat{\boldsymbol{\theta}}$, $sd(\theta_i) \approx \sqrt{I^{ii}}$ where I^{ii} is the i'th element in the diagonal of $I^{-1}(\hat{\boldsymbol{\theta}})$. In addition, for a sufficiently large n, $\hat{\theta}_i$ is approximately normally distributed with expectation θ_i and variance I^{ii}. The following theorem provides the Fisher information of a continuous transformation of the MLE .

Theorem 4.1.4

For the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ let $\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta}))$ be such that the $p \times p$ matrix J with elements $J_{ij} = \partial g_j / \partial \theta_i$ is invertible. Then the Fisher information of $\mathbf{g}(\boldsymbol{\theta})$ is

$$I\left(\mathbf{g}(\boldsymbol{\theta})\right) = J^{-1} I(\boldsymbol{\theta})(J^T)^{-1}, \quad \text{hence} \quad I^{-1}\left(g(\boldsymbol{\theta})\right) = J^T I^{-1}(\boldsymbol{\theta}) J.$$

The following theorem shows the invariance property of MLE.

Theorem 4.1.5

Assume $\hat{\boldsymbol{\theta}}$ is a consistent MLE of $\boldsymbol{\theta}$. Then for a continuous function $g : \mathbb{R}^p \to \mathbb{R}^p$ we have that

$$I^{1/2}\left(\mathbf{g}(\hat{\boldsymbol{\theta}})\right)\left(\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta})\right) \xrightarrow{d} N_d(0, \boldsymbol{I}_d).$$

4.2 Estimation with a truncated dataset

Consider the model where $(X_1, T_1), \ldots, (X_n, T_n)$ are i.i.d. $TN_2(\mu_x, \mu_t, \sigma_x^2, \sigma_t^2, \rho)$. The density of (X^*, T^*), which we want to reconstruct, is then given by

$$h(x,t) = \frac{1}{2\pi\sigma_x\sigma_t\sqrt{1-\rho^2}} \times \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{t-\mu_t}{\sigma_t}\right) + \left(\frac{t-\mu_t}{\sigma_t}\right)^2\right]\right\},$$

and the density of (X, T) is given by equation 2.2 on page 6. Let $\boldsymbol{\theta} = (\mu_x, \mu_t, \sigma_x^2, \sigma_t^2, \rho)$ and $(\mathbf{x}, \mathbf{t}) = ((x_1, t_1), \ldots, (x_n, t_n))$. The probability of the observed data $(\mathbf{x}, \mathbf{t})$ is then the likelihood function

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} h^c(x_i, t_i|\boldsymbol{\theta}) = \frac{1}{\alpha(\boldsymbol{\theta})^n} \prod_{i=1}^{n} h(x_i, t_i|\boldsymbol{\theta}),$$

where $\alpha(\boldsymbol{\theta}) = P(X^* > T^*|\boldsymbol{\theta})$. Maximising $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is equivalent to maximising the log likelihood function:

$$\begin{aligned}\log(L(\boldsymbol{\theta})) &= -n\log\alpha(\boldsymbol{\theta}) + \sum_{i=1}^{n}\log h(x_i, t_i|\boldsymbol{\theta}) \\ &= C - n\log\alpha(\boldsymbol{\theta}) - n\log(\sigma_x\sigma_t) - \frac{n}{2}\log(1-\rho^2) \\ &\quad - \frac{1}{2(1-\rho^2)}\sum_{i=1}^{n}\left[\left(\frac{x_i-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{t_i-\mu_t}{\sigma_t}\right) + \left(\frac{t_i-\mu_t}{\sigma_t}\right)^2\right]\end{aligned}$$

In this particular case $\alpha(\boldsymbol{\theta})$ can be calculated quite easy:

$$\alpha = P(X^* > T^*) = P(T^* - X^* < 0) = P(U < 0).$$

And since (X^*, T^*) is bivariate normal distributed it follows that

$$U \sim N(\mu_t - \mu_x, \sigma_x^2 + \sigma_t^2 - 2\rho\sigma_x\sigma_t).$$

Hence

$$\alpha(\boldsymbol{\theta}) = \phi\left(\frac{\mu_t - \mu_x}{\sqrt{\sigma_x^2 + \sigma_t^2 - 2\rho\sigma_x\sigma_t}}\right),$$

where ϕ denotes the standard normal cumulative distribution function.

The function $\log(L(\boldsymbol{\theta}))$ can be maximised numerically. In this optimisation it is important to let ρ be constrained to the open interval $(-1, 1)$. To evaluate the method, four samples of different size were drawn from $TN_2(0, -1, 1, 1, 0.3)$ with sample size $n = 50, 100, 200$ and 500. The optimisation was done using a Quasi-Newton method in R. The result can be viewed in table 4.1 on the following page. We see that the estimates improve as n increase. Different samples gave of course different answers, but all were close to the real parameters for large n. We will continue the evaluation of this method in section 4.3 on page 43.

Table 4.1: MLE of the parameters in a truncated bivariate normal distribution where the real parameters are $\mu_x = 0$, $\mu_t = -1$, $\sigma_x^2 = 1$, $\sigma_t^2 = 1$ and $\rho = 0.3$.

	n=50	n=100	n=200	n=500	True value
$\hat{\mu}_x$	-0.2529	-0.0729	0.0286	-0.0128	0
$\hat{\mu}_t$	-1.0130	-1.0744	-0.9276	-0.9669	-1
$\hat{\sigma}_x^2$	1.0733	0.9505	0.9699	1.0050	1
$\hat{\sigma}_t^2$	0.7925	0.9911	0.9848	1.0572	1
$\hat{\rho}$	0.1202	0.2583	0.1741	0.2732	0.3

4.2.1 Normal linear model

A natural extension of the i.i.d. $TN_2(\mu_x, \mu_t, \sigma_x^2, \sigma_t^2, \rho)$ model, that allows some modelling of the relationship with covariates, is to drop the identical requirement from the i.i.d. This can be done by modelling the means μ_{xi} and μ_{ti} as functions of the covariates. The simplest structure of such a function is the linear model:

$$\mu_{xi} = \alpha_x + y_i^T \beta_x \,, \quad \mu_{ti} = \alpha_t + y_i^T \beta_t \,,$$

where y_i is a vector of covariates. If we assume that $(X_1, T_1), \ldots, (X_n, T_n)$ are independent $TN_2(\mu_{xi}, \mu_{ti}, \sigma_x^2, \sigma_t^2, \rho)$ the log likelihood function becomes

$$\begin{aligned}
\log(L(\boldsymbol{\theta})) &= -\sum_{i=1}^{n} \log \alpha_i(\boldsymbol{\theta}) + \sum_{i=1}^{n} \log h(x_i, t_i|\boldsymbol{\theta}) \\
&= C - \sum_{i=1}^{n} \log \alpha_i(\boldsymbol{\theta}) - n \log(\sigma_x \sigma_t) - \frac{n}{2} \log(1-\rho^2) - \\
&\quad \frac{1}{2(1-\rho^2)} \sum_{i=1}^{n} \left[\left(\frac{x_i - \mu_{xi}}{\sigma_x} \right)^2 - 2\rho \left(\frac{x_i - \mu_{xi}}{\sigma_x} \right) \left(\frac{t_i - \mu_{ti}}{\sigma_t} \right) + \left(\frac{t_i - \mu_{ti}}{\sigma_t} \right)^2 \right],
\end{aligned}$$

where

$$\alpha_i(\boldsymbol{\theta}) = \phi\left(\frac{\mu_{ti} - \mu_{xi}}{\sqrt{\sigma_x^2 + \sigma_t^2 - 2\rho\sigma_x\sigma_t}} \right).$$

This model is used in section 6.7 on page 86 to investigate if the dependence between X and T can be explained by a common covariate.

4.3 Testing dependence using the MLE of ρ

It is especially interesting that theorem 4.1.3 on page 39 provides a method for testing $\rho = 0$ versus $\rho \neq 0$. In the optimisation, we only need to include the "hessian=TRUE" command in R to produce the so called observed Fisher information matrix $I(\hat{\boldsymbol{\theta}})$. In our case, we have that $\text{sd}(\hat{\rho}) \approx \sqrt{I^{55}}$. We should therefore reject the hypothesis $\rho = 0$ whenever

$$|Z^*| = \left| \frac{\hat{\rho}}{\sqrt{I^{55}}} \right| > |Z_{\epsilon/2}|.$$

Moreover, when (X^*, T^*) follows a bivariate normal distribution then rejecting $\rho = 0$ is equivalent to rejecting the hypothesis of independence. Unfortunately, there is no way of testing the assumption that (X^*, T^*) follows a bivariate normal distribution with a truncated dataset. We will address this problem in section 6.5 on page 79.

To investigate the above approximation a simulation was carried out in R. The following routine was repeated 400 times for every fixed combination of $n = 50, 100, 200$ and $\rho = -0.5, 0, 0.5$:

- n pairs were drawn from the truncated bivariate normal distribution $TN_2(0, -1, 1, 1, \rho)$,
- Using these n pairs, the MLEs of ρ and I^{55} were computed.

For each fixed combination of ρ and n we then obtain the samples $\hat{\rho}_1, \ldots, \hat{\rho}_{400}$ and $I_1^{55}, \ldots, I_{400}^{55}$. We then calculate:

$$AVE(\hat{\rho}) = \frac{1}{400} \sum_{i=1}^{400} \hat{\rho}_i,$$
$$AVE(I^{55}) = \frac{1}{400} \sum_{i=1}^{400} I_i^{55},$$
$$VAR(\hat{\rho}) = \frac{1}{399} \sum_{i=1}^{400} (\hat{\rho}_i - AVE(\hat{\rho}))^2,$$

Result

The result can be seen in table 4.2 on the next page. In all cases, $AVE(\hat{\rho})$

is close to ρ. We see that $AVE(I^{55})$ is quite close to $VAR(\hat{\rho})$ for $n = 100$ and $n = 200$, indicating that the approximation $\text{sd}(\hat{\rho}) \approx \sqrt{I^{55}}$ is not so bad. Both quantities decreases with increasing n, hence the accuracy of the estimates is increasing in n. The Shapiro Wilks statistic for testing normality in the sample of 400 ρ's asserts that the assumption of normality hold when $n = 100$ and $n = 200$, but not when $n = 50$. However, the computed p-values are all low, and when we tested the normality of $z_i = (\hat{\rho} - \rho)/\sqrt{I^{55}}$ we got a rejection in all cases.

Table 4.2: Simulation results of $\hat{\rho}$ from truncated samples sized n of a bivariate normal distribution with $\mu_x = 0$, $\mu_t = -1$, $\sigma_x^2 = 1$, and $\sigma_t^2 = 1$.

ρ		n=50	n=100	n = 200	Truncated proportion
-0.5	$AVE(\hat{\rho})$	-0.4886	-0.4983	-0.4996	0.2818
	$VAR(\hat{\rho})$	0.0323	0.0141	0.0054	
	$AVE(I^{55})$	0.0295	0.0151	0.0054	
	Normal p	0.0113	0.0822	0.1031	
0	$AVE(\hat{\rho})$	0.0051	0.0046	0.0020	0.2393
	$VAR(\hat{\rho})$	0.0577	0.0224	0.0114	
	$AVE(I^{55})$	0.0476	0.0244	0.0123	
	Normal p	0.0421	0.0666	0.1340	
0.5	$AVE(\hat{\rho})$	0.4972	0.4995	0.4986	0.1592
	$VAR(\hat{\rho})$	0.0236	0.0120	0.0053	
	$AVE(I^{55})$	0.0251	0.0118	0.0058	
	Normal p	0.0237	0.0712	0.1022	

The main problem in this case is that ρ is restricted to the interval $(-1, 1)$. So if e.g. we have that $\rho = 0.7$ the normal approximation will be very poor because of the negative skewness (see figure 4.1 on the following page). Therefore, consider the so-called Fisher's z transform

$$g(\hat{\rho}) = \frac{1}{2} \log \frac{1+\hat{\rho}}{1-\hat{\rho}} \, .$$

This transformation was originally proposed for the standard sample correlation. Note that the domain of this function is $(-\infty, \infty)$, so this transformation will spread out the shorter tail of $\hat{\rho}$. By the invariance properties of MLE, $g(\hat{\rho})$ is the MLE of $g(\rho)$ so this function will also converge towards a normal distribution. The idea is that it will do so faster than $\hat{\rho}$. It follows from theorem 4.1.4 and theorem 4.1.5 on page 40 (or the Delta method) that

$$\frac{g(\hat{\rho}) - g(\rho)}{\sqrt{I^{55}}|g^{'}(\hat{\rho})|} \xrightarrow{d} N(0,1), \quad \text{where} \quad g^{'}(\hat{\rho}) = \frac{1}{(1+\hat{\rho})(1-\hat{\rho})}.$$

We can utilise this to test $H_Z : \rho = 0$ versus $\rho \neq 0$. Reject H_Z whenever

$$|Z| = |\frac{g(\hat{\rho})}{\sqrt{I^{55}}|g^{'}(\hat{\rho})|}| > Z_{\epsilon/2} \, , \tag{4.1}$$

where ϵ denotes the significance level of the test and $Z_{\epsilon/2}$ the corresponding normal critical value.

To see if Fisher's z transform improves the normal approximation, we computed 400 of the following quantities

$$A = \frac{\hat{\rho} - \rho}{\sqrt{I^{55}}} \, , \qquad B = \frac{g(\hat{\rho}) - g(\rho)}{\sqrt{I^{55}}|g^{'}(\hat{\rho})|} \, ,$$

based on 400 independent samples of size $n = 100$ of the truncated bivariate normal distribution with correlation 0.7. A histogram of the 400 MLEs of $\hat{\rho}$ and the Fisher's z transform of these values can be seen in figure 4.1 on the following page. The Shapiro Wilks test produced a p-value equal to $4.764e^{-06}$ for the normality of the 400 computed values of A, while the p-value for the normality of the corresponding values of B was 0.3757. The statistic Z is therefore preferable compared to Z^*.

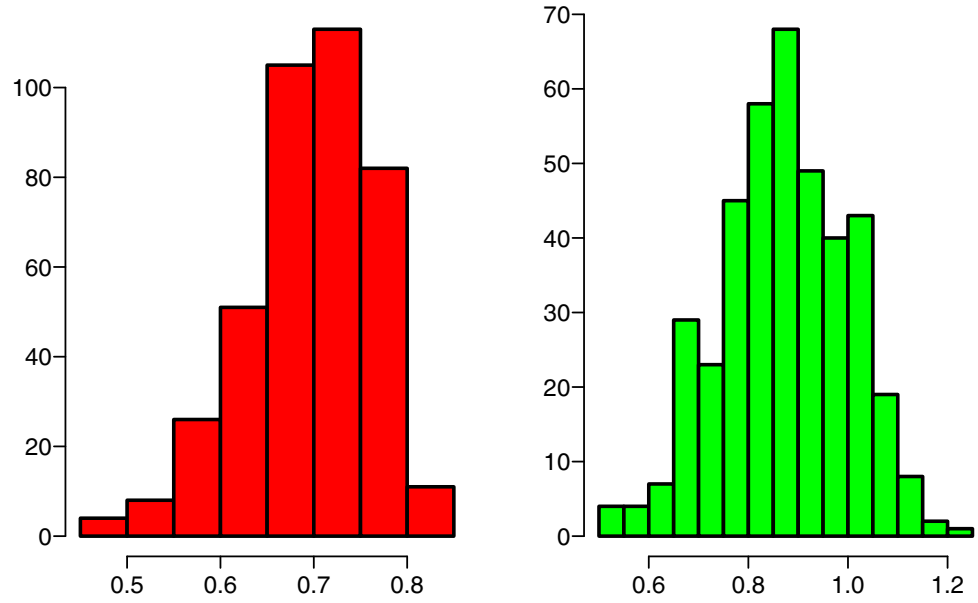

Figure 4.1: To the left we have a histogram of 400 MLE of ρ based on 400 independent samples of size $n = 100$ of the truncated bivariate normal distribution with correlation 0.7, and to the right a histogram of the Fisher's z transformed of these values.

5

Copula models for dependently truncated data

In recent years there has been a growing interest in copulas and their applications in statistics. Copulas provide a method of describing the relationship between a multivariate distribution function and its margins. The idea is to form joint distributions by *coupling* together marginal distributions using dependent uniform distributions.

We will start by giving the formal definition and some basic results about copulas. In the rest of the chapter we will consider how to model the dependency structure in truncated data using the concept of copulas.

5.1 Basic properties and results

We will follow the notation suggested by McNeil, Frey and Embrechts in (McNeil *et al.*, 2005) and define a copula function C in the following way:

Definition 5.1.1: Copula

Let C: $[0,1]^d \to [0,1]$, and assume the following three properties hold

1. $C(u_1, \ldots, u_d)$ is increasing in each component u_i.
2. $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i \quad \forall\, i \in \{1, \ldots, d\}$, $u_i \in [0,1]$.
3. For all $(a_1, \ldots, a_d)$, $(b_1, \ldots, b_d) \in [0,1]^d$ with $a_i \leq b_i$ we have

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1,\ldots,i_d} C(u_{1i_1}, \ldots, u_{di_d}) \geq 0,$$

where $u_{j1} = a_j$ and $u_{j2} = b_j \; \forall\, j \in \{1, \ldots, d\}$.

Then C is a copula.

The first property is required for any distribution function. The second property gives uniform marginals which is the main idea of copulas. The last property is the so called rectangle inequality which ensures non-negative values of $P(a_1 \leq U_1 \leq b_1, \ldots, a_d \leq U_d \leq b_d)$ when $(U_1, \ldots, U_d)^T$ is distributed according to C.

The following theorem states that all multivariate density functions contain copulas.

Theorem 5.1.2: Sklar's theorem

Let F be a joint distribution function with margins $F_1, \ldots, F_d$. Then there exists a copula C: $[0,1]^d \to [0,1]$ such that, for all $x_1, \ldots, x_d$ in $\bar{\mathbb{R}} = [-\infty, \infty]$,

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{5.1}$$

If the margins are continuous, then C is unique; otherwise C is uniquely determined on $\operatorname{Ran} F_1 \times \operatorname{Ran} F_2 \times \cdots \times \operatorname{Ran} F_d$, where $\operatorname{Ran} F_i = F_i(\bar{\mathbb{R}})$ denotes the range of F_i. Conversely, if C is a copula and $F_1, \ldots, F_d$ are univariate distribution functions, then the function F defined in 5.1 is a joint distribution function with margins $F_1, \ldots, F_d$.

Proof: A proof can be found in (Nelsen, 1999, page 18).

The converse statement of Sklar's theorem tells us that we can construct multivariate distributions using a copula and arbitrary marginal distributions. Such a distribution is often called *meta distribution*.

Before we continue with the definition of Archimedean copulas, we need to define the pseudo-inverse.

Definition 5.1.3: Pseudo-inverse

Let $\phi : [0,1] \to [0,\infty]$ be a continuous, strictly decreasing function such that $\phi(1) = 0$. The *pseudo-inverse* of ϕ is the function $\phi^{[-1]}$ with $\operatorname{Dom}\phi^{[-1]} = [0,\infty]$ and $\operatorname{Ran}\phi^{[-1]} = [0,1]$ given by

$$\phi^{[-1]}(t) = \begin{cases} \phi^{-1}(t), & 0 \leq t \leq \phi(0), \\ 0, & \phi(0) \leq t \leq \infty. \end{cases} \tag{5.2}$$

The following theorem provides a method for constructing 2-dimensional copulas:

Theorem 5.1.4: Archimedian Copulas

Let $\phi : [0,1] \to [0,\infty]$ be a continuous, strictly decreasing function such that $\phi(1) = 0$ and let $\phi^{[-1]}$ be the pseudo-inverse of ϕ defined by equation 5.2. Let the function $C : [0,1]^2 \to [0,1]$ be given by

$$C(u_1, u_2) = \phi^{[-1]}(\phi(u_1) + \phi(u_2)). \tag{5.3}$$

Then the function C is a copula if and only if ϕ is convex.

Proof: A proof can be found in (Nelsen, 1999, page 91).

The function ϕ is often called the generator of the copula and a copula constructed in this way is called an Archimedean copula. Table 5.1 on the following page list the three generators which we will use in chapter 6. For variables following a meta distribution constructed with an Archimedean copula there is a special relation between Kendall's tau and the generator function. This relation is formulated in the following theorem:

Theorem 5.1.5

Let X and T be random variables following a meta distribution with an Archimedean copula C generated by ϕ. Then Kendall's tau for X and T is given by:

$$\tau = 1 + 4 \int_0^1 \frac{\phi(t)}{\phi'(t)} \, \mathrm{d}t. \tag{5.4}$$

Proof: A proof can be found in (Nelsen, 1999, page 130).

Copula	Generator $\phi(t)$	τ	$\theta \in$
Gumbel	$(-\ln t)^\theta$	$1 - 1/\theta$	$[1, \infty)$
Clayton	$\frac{1}{\theta}(t^{-\theta} - 1)$	$\theta/(\theta + 2)$	$[-1, \infty) \setminus \{0\}$
Frank	$-\ln\left(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right)$	$1 - 4\theta^{-1}(1 - D_1(\theta))$	$(-\infty, \infty) \setminus \{0\}$

Table 5.1: Table summarising the generators, the relation between Kendall's tau and the copula parameter, and permissible parameter values for the Archimedean copulas Gumbel, Clayton and Frank. $D_1(\theta)$ is the Debye function $D_1(\theta) = \theta^{-1} \int_0^\theta t/(\exp(t) - 1) \, \mathrm{d}t$.

5.1.1 Simulating from meta distributions

In R, the package "copula" provides a series of tools when dealing with copulas. In particular, it features simulations from the most common copulas such as the Gaussian and Archimedean copulas. It is also possible to simulate from meta distributions, but only for a few kinds of marginal distributions. Assume we want to simulate from the meta distribution

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)),$$

and we know how to simulate from C. We can then use the following procedure:

1. Simulate $\boldsymbol{U} = (U_1, \ldots, U_d)$ from C.
2. Let $X_i = F_i^{-1}(U_i), \quad 1 \leq i \leq d$.

Then $\boldsymbol{X} = (X_1, \ldots, X_d)$ is distributed according to the meta distribution:

$$\begin{aligned} P(X_1 \leq x_1, \ldots, X_1 \leq x_d) &= P(F_1^{-1}(U_1) \leq x_1, \ldots, F_d^{-1}(U_d) \leq x_d) \\ &= P(U_1 \leq F_1(x_1), \ldots, U_d \leq F_d(x_d)) \\ &= C(F_1(x_1), \ldots, F_d(x_d)) \end{aligned}$$

5.2 Maximum likelihood based estimation

Let us assume the distribution of the unconditional vector (X^*, T^*) is given by the meta distribution

$$H(x,t) := C\left(F(x), G(t)\right), \tag{5.5}$$

where C is an arbitrary copula, and F and G are arbitrary distribution functions. The corresponding density function is then given by

$$h(x,t) = \frac{\partial}{\partial x \partial t} H(x,t) = c\left(F(x), G(t)\right) f(x) g(t),$$

where c denotes the copula density function, and f and g are the density functions of F and G, respectively. According to equation 2.2 on page 6 the conditional density function of $(X,T) = (X^*, T^* | X^* > T^*)$ is

$$h^c(x,t) = \begin{cases} c(F(x), G(t)) f(x) g(t) / \alpha, & x > t, \\ 0, & \text{otherwise}, \end{cases} \tag{5.6}$$

where

$$\alpha = P(X^* > T^*) = \int \int_{u>v} c(F(u), G(v)) f(u) g(v) \, \mathrm{d}u \, \mathrm{d}v.$$

Assume that $(X_1, T_1) = (x_1, t_1), \ldots, (X_n, T_n) = (x_n, t_n)$ are i.i.d. according to equation 5.6. For simplicity let $\boldsymbol{\theta}$ denote the unknown parameters specifying the copula and the marginals and let $(\mathbf{x}, \mathbf{t}) = ((x_1, t_1), \ldots, (x_n, t_n))$. We then

obtain the following log likelihood function:

$$\begin{aligned}\log(L(\boldsymbol{\theta})) &= -n\log\alpha(\boldsymbol{\theta}) + \sum_{i=1}^{n}\log h(x_i, t_i|\boldsymbol{\theta}) \\ &= -n\log\alpha(\boldsymbol{\theta}) + \sum_{i=1}^{n}\log c(F(x_i|\boldsymbol{\theta}), G(t_i|\boldsymbol{\theta})|\boldsymbol{\theta}) \\ &\quad + \sum_{i=1}^{n}\log f(x_i|\boldsymbol{\theta}) + \sum_{i=1}^{n}\log g(t_i|\boldsymbol{\theta}).\end{aligned}$$

In this setting, the traditional separation of first fitting the marginals and then the copula is not possible. This is because $\alpha(\boldsymbol{\theta})$ depends on both the copula parameter and the parameters of the marginals. This means we have to optimise the log likelihood directly. The difficult part of optimising this likelihood is the computation of $\alpha(\boldsymbol{\theta})$. The available package on multidimensional integration in R did not work properly for varying parameters. The other option was Monte Carlo integration:

1. Simulate $(X_1^*, T_1^*), \ldots, (X_k^*, T_k^*)$ according to 5.5 for each set of parameters $\boldsymbol{\theta}$.
2. Let

$$\alpha(\boldsymbol{\theta}) = \frac{1}{k}\text{card}\{i|X_i^* > T_i^*\}.$$

   In the optimisation we let $k = 1\ 000\ 000$.

However, including a simulation in a function subject to optimisation will in general be problematic. Such a function is not deterministic and the maximum of the function will vary for every simulation, though not necessarily by much. There are two solutions to this problem, one good and one bad.

A bad solution is to lower the convergence tolerance so that the iterations in the optimisation stop when the function is "close" to its maximum.

A good solution is to fix the seed which R uses to generate random numbers. The simulation is then deterministic for varying parameters causing the function to be deterministic. In addition, to make the optimisation more efficient, one should choose reasonable initial values. When we have normal marginals,

natural initial values for μ_x and μ_t are the empirical means $\bar{x}$ and $\bar{t}$, while the empirical variance s_x^2 and s_t^2 are natural initial values for σ_x^2 and σ_t^2. A simple, but somewhat ad hoc way of choosing initial values for θ, is to choose θ so that the right side of equation 5.4 on page 50 equals the sample Kendall's tau. This is a bad estimate of θ, but will work fine as an initial value.

5.2.1 Performance of the optimisation

Due to the Monte Carlo integration, the optimisation became rather time consuming. We also had to increase the default convergence tolerance and increase the total number of iterations allowed, to make the optimisation work.

Based on 100 independent truncated samples of size $n = 100$ from the Clayton copula with parameter θ and with normal marginals we computed 100 estimates of the parameters $(\mu_x, \mu_t, \sigma_x^2, \sigma_t^2, \theta)$. The true parameter values were set to $(0, -1, 1, 1, 3)$. We then computed the sample average and the sample variance of these estimates denoted $AVE(\cdot)$ and $VAR(\cdot)$, respectively. This routine was repeated, but with the exponential marginals $F(x) = 1 - e^{-\alpha x}$ and $G(t) = 1 - e^{-\beta t}$. In this case the true parameter values of (α, β, θ) were set to $(0.5, 1, 2)$.

The result can be seen in table 5.2. In all cases, $AVE(\cdot)$ is close to the real parameter value suggesting that the optimisation works. Compared to the marginal parameters, the variance amongst the estimates of θ is larger.

Table 5.2: Simulation result.

		μ_x	μ_t	σ_x^2	σ_t^2	θ
Normal	$AVE(\cdot)$	-0.0055	-1.0023	0.9742	0.9787	3.0062
marginals	$VAR(\cdot)$	0.0106	0.0123	0.0153	0.0208	0.3202
		α	β	θ		
exponential	$AVE(\cdot)$	0.5074	1.0281	2.0911		
marginals	$VAR(\cdot)$	0.0029	0.0168	0.1981		

5.3 Estimation based on the conditional Kendall's tau and the copula-graphic estimator

The goal of the following analysis is to estimate the copula parameter in an Archimedean family and the marginal distributions, without making any parametric assumptions about the marginal distributions. Such an estimation procedure was proposed by Lakhal Chaieb *et al.* (2006) and is an application of the conditional Kendall's tau and the copula-graphic estimators.

5.3.1 Model and assumptions

As for the MLE estimation we will assume that (X^*, T^*) is distributed according to a meta distribution. However, for this method it is convenient to write the joint distribution of (X^*, T^*) as

$$P(X^* > x, T^* \leq t) = C^*(S(x), G(t)). \tag{5.7}$$

Where $F(x) = 1 - S(x)$ and $G(t)$ are the marginal distributions of X^* and T^*, respectively, and C^* is a copula function which we will call the semisurvival copula. Care should be taken not to confuse C^* with the copula C that we try to estimate in section 5.2 on page 51. If we assume that the relations given by 5.5 and 5.7 both hold we have

$$\begin{aligned} C^*(S(x), G(t)) &= P(X^* > x, T^* \leq t) \\ &= G(t) - H(x, t) \\ &= G(t) - C(F(x), G(t)) \\ &= G(t) - C(1 - S(x), G(t)) \ , \end{aligned}$$

hence the relation between C and C^* is given by $C^*(u, v) = v - C(1 - u, v)$.

When the joint distribution of (X^*, T^*) is given by 5.7, the conditional distribution of $(X, T) = (X^*, T^* | X^* > T^*)$ can be written as

$$\pi(x, t) = P(X > x, T \leq t) = C^*(S(x), G(t))/\alpha \ , \qquad x > t, \tag{5.8}$$

where

$$\alpha = P(X^* > T^*) = \int \int_{u>v} c^*(S(u), G(v)) f(u) g(v) \, \mathrm{d}u \, \mathrm{d}v,$$

and c^* is the copula density corresponding to C^*. Note that 5.8 is valid since $x > t$ implies that

$$P(X^* > x, T^* \leq t, X^* > T^*) = P(X^* > x, T^* \leq t) = C^*(S_x(x), G(t)).$$

For this estimation procedure we will only consider semisurvival copulas which are members of the Archimedean family. Hence

$$\pi(x,t) = \phi^{-1}[\phi\{S(x)\} + \phi\{G(t)\}]/\alpha, \qquad x > t, \tag{5.9}$$

where ϕ is a non-increasing convex function defined on $[0,1]$, with $\phi(1) = 0$. We will only consider generator functions with one parameter denoted by θ.

It is assumed that for some $t_0 > x_0$[1]

$$G(x_0) > 0, \quad S(x_0) = 1, \quad G(t_0) = 1, \quad \text{and} \quad S(t_0) > 0.$$

In (Lakhal Chaieb *et al.*, 2006), these conditions are used to derive the asymptotic theory of the estimation procedure we will describe in the following sections. However, we will not study these details.

5.3.2 The copula-graphic estimator

The motivation for using the semisurvival copula is that for points (x,t) satisfying $x > t$ we have that $\{(X^*, T^*) | X^* > x, T^* \leq t\} \subset (X, T)$. So when we observe i.i.d pairs $(X_1, T_1), \ldots, (X_n, T_n)$ from the observable region (X, T) we have the following empirical estimate of $\pi(x,t)$:

$$\hat{\pi}(x,t) = \frac{R(x,t)}{n}, \tag{5.10}$$

[1] Lakhal Chaieb *et al.* (2006) write their article in a survival analysis setting and thus only consider positive defined variables. However, we see no reason why this estimation procedure can't be applied to variables which can take on negative values as well.

where $R(\cdot)$ is the "number at risk" function given by definition 3.1.1 on page 28:

$$R(x,t) = \sum_{i=1}^{n} I(X_i \geq x, T_i \leq t).$$

Lakhal Chaieb *et al.* (2006) utilise this estimate of $\pi(x,t)$ to derive semi-parametric estimates of S and G. These estimators are known as the copula-graphic estimators. We will start by giving the definition.

Definition 5.3.1: The copula-graphic estimator

Assume that $(X_1, T_1), \ldots, (X_n, T_n)$ are i.i.d. according to 5.9 and let $\tilde{R}(z) = R(z,z)$. Given the copula parameter θ and $\alpha = P(X^* > T^*)$ we define the copula-graphic estimators of S and G as

$$\hat{S}_{CG}(z) = \phi^{-1}\left\{ \sum_{X_i \leq z} \left[\phi\Big(\alpha \frac{\tilde{R}(X_i) - 1}{n}\Big) - \phi\Big(\alpha \frac{\tilde{R}(X_i)}{n}\Big) \right] \right\}, \tag{5.11}$$

$$\hat{G}_{CG}(z) = \phi^{-1}\left\{ \sum_{T_i > z} \left[\phi\Big(\alpha \frac{\tilde{R}(T_i) - 1}{n}\Big) - \phi\Big(\alpha \frac{\tilde{R}(T_i)}{n}\Big) \right] \right\}. \tag{5.12}$$

In both cases, an empty sum is to be interpreted as 0.

For the independent copula the generator function is $\phi(t) = -\log(t)$, hence $\phi^{-1}(t) = e^{-t}$. We therefore get

$$\begin{aligned}
\hat{F}_{CG}(z) &= 1 - \phi^{-1}\left\{ \sum_{X_i \leq z} \left[\phi\left(\alpha \frac{\tilde{R}(X_i) - 1}{n}\right) - \phi\left(\alpha \frac{\tilde{R}(X_i)}{n}\right) \right] \right\} \\
&= 1 - \exp\left\{ \sum_{X_i \leq z} \left[\log\left(\alpha \frac{\tilde{R}(X_i) - 1}{n}\right) - \log\left(\alpha \frac{\tilde{R}(X_i)}{n}\right) \right] \right\} \\
&= 1 - \exp\left\{ \sum_{X_i \leq z} \left[\log\left(\frac{\tilde{R}(X_i) - 1}{\tilde{R}(X_i)}\right) \right] \right\} \\
&= 1 - \exp\left\{ \log\left(\prod_{X_i \leq z} \left(\frac{\tilde{R}(X_i) - 1}{\tilde{R}(X_i)}\right) \right) \right\} \\
&= 1 - \prod_{X_i \leq z} \left(\frac{\tilde{R}(X_i) - 1}{\tilde{R}(X_i)}\right).
\end{aligned}$$

This shows that the copula-graphic estimator of F reduces to the Product-limit estimator in the independent case. The derivation of the copula-graphic estimator is also very similar to the derivation of the Product-limit estimator which was done in section 3.1 on page 27.

Let us for the time being assume that the copula parameter θ is known and that we observe i.i.d. pairs $(X_1, T_1), \ldots, (X_n, T_n)$ from the observable region (X, T). If we let $x = t = z$, and put the empirical estimate of $\pi(x, t)$ given by equation 5.10 on page 55 equal to the model given in 5.9, we get the following estimating equation for G and S

$$\alpha \frac{\tilde{R}(z)}{n} = \phi^{-1}[\phi\left(\hat{S}_{CG}(z-)\right) + \phi\left(\hat{G}(z)\right)]. \tag{5.13}$$

Similar to the derivation of the PLE, we postulate some properties about $\hat{S}_{CG}$ and $\hat{G}_{CG}$. Assume that $\hat{S}_{CG}$ is a decreasing right-continuous function with jumps at $X_1, \ldots, X_n$ and that $\hat{G}_{CG}$ is an increasing right-continuous function with jumps at $T_1, \ldots, T_n$. In addition, assume that $\hat{S}_{CG}$ is supported on $[X_{(1)}, X_{(n)}]$ so that $\hat{S}_{CG}(X_{(1)}-) = 1$. Applying ϕ on both sides of 5.13 yields

$$\phi\left(\alpha \frac{\tilde{R}(z)}{n}\right) = \phi\left(\hat{S}_{CG}(z-)\right) + \phi\left(\hat{G}(z)\right). \tag{5.14}$$

Next, let $X \in (X_1, \ldots, X_n)$ and remember that $X > T$ for the corresponding $T \in (T_1, \ldots, T_n)$. This means that jumps for the functions $\hat{G}_{CG}$ and $\hat{S}_{CG}$ will not occur at the same points. Consequently, we have that $\tilde{R}(X+) = \tilde{R}(X) - 1$ and $\hat{G}_{PL}(X+) = \hat{G}_{PL}(X)$. By subtracting 5.14 at $X+$ from 5.14 at X we get

$$\phi\left(\hat{S}_{CG}(X-)\right) - \phi\left(\hat{S}_{CG}(X)\right) = \phi\left(\alpha \frac{\tilde{R}(X)}{n}\right) - \phi\left(\alpha \frac{\tilde{R}(X) - 1}{n}\right). \tag{5.15}$$

Since $\hat{S}_{CG}$ is a right continuous step function with jumps at $X_1, \ldots, X_n$, we have that

$$\phi\left(\hat{S}_{CG}(X_{(i)})\right) = \phi\left(\hat{S}_{CG}(X_{(i+1)}-)\right), \quad 1 \leq i \leq n-1,$$

where as before, $X_{(1)}, \ldots, X_{(n)}$ is the ordered values of $X_1, \ldots, X_n$. As we have assumed $\hat{S}_{CG}(X_{(1)}-) = 1$, we get that,

$$\phi\left(\hat{S}_{CG}(X_{(1)}-)\right) = \phi(1) = 0.$$

Therefore, if we let $X_{(a)} = \max(X_1 \ldots, X_n | X_i \leq z)$ and sum equation 5.15 over all X_i where $X_i \leq z$, we get

$$\begin{aligned}
&\sum_{X_i \leq z}\left[\phi\left(\alpha\frac{\tilde{R}(X_i)}{n}\right) - \phi\left(\alpha\frac{\tilde{R}(X_i)-1}{n}\right)\right] = \underbrace{\phi\left(\hat{S}_{CG}(X_{(1)}-)\right)}_{0} \\
&+ \underbrace{\left[-\phi\left(\hat{S}_{CG}(X_{(1)})\right) + \phi\left(\hat{S}_{CG}(X_{(2)}-)\right)\right]}_{0} \\
&+ \underbrace{\left[-\phi\left(\hat{S}_{CG}(X_{(2)})\right) + \phi\left(\hat{S}_{CG}(X_{(3)}-)\right)\right]}_{0} + \ldots \\
&+ \underbrace{\left[-\phi\left(\hat{S}_{CG}(X_{(a-1)})\right) + \phi\left(\hat{S}_{CG}(X_{(a)}-)\right)\right]}_{0} - \phi\left(\hat{S}_{CG}(X_{(a)})\right) = -\phi\left(\hat{S}_{CG}(X_{(a)})\right).
\end{aligned}$$

It follows from the definition of $X_{(a)}$ that $\hat{S}_{CG}(X_{(a)}) = \hat{S}_{CG}(z)$, so the above equation becomes

$$\begin{aligned}
\phi\left(\hat{S}_{CG}(z)\right) &= -\sum_{X_i \leq z}\left[\phi\left(\alpha\frac{\tilde{R}(X_i)}{n}\right) - \phi\left(\alpha\frac{\tilde{R}(X_i)-1}{n}\right)\right] \\
&= \sum_{X_i \leq z}\left[\phi\left(\alpha\frac{\tilde{R}(X_i)-1}{n}\right) - \phi\left(\alpha\frac{\tilde{R}(X_i)}{n}\right)\right].
\end{aligned} \tag{5.16}$$

Analogous procedures can be done for the truncating variables $T_1, \ldots, T_n$ to get

$$\phi\left(\hat{G}(z)\right) = \sum_{T_i > z}\left[\phi\left(\alpha\frac{\tilde{R}(T_i)-1}{n}\right) - \phi\left(\alpha\frac{\tilde{R}(T_i)}{n}\right)\right]. \tag{5.17}$$

Applying ϕ^{-1} on 5.16 and 5.17 leads to the definition of the copula-graphic estimators of S and G.

The copula-graphic estimators of S and G also provides an estimation equation

for α. If we plug 5.16 and 5.17 into 5.14 we get the following equation

$$\sum_{X_i<z}\left[\phi\left(\alpha\frac{\tilde{R}(X_i)}{n}\right)-\phi\left(\alpha\frac{\tilde{R}(X_i)-1}{n}\right)\right]$$
$$+\sum_{T_i>z}\left[\phi\left(\alpha\frac{\tilde{R}(T_i)}{n}\right)-\phi\left(\alpha\frac{\tilde{R}(T_i)-1}{n}\right)\right]+\phi\left(\alpha\frac{\tilde{R}(z)}{n}\right)=0.$$

In particular, we can choose $z = T_{(n)}$ so the equation simplifies to

$$H_1(\alpha,\theta)=\sum_{X_i<T_{(n)}}\left[\phi\left(\alpha\frac{\tilde{R}(X_i)}{n}\right)-\phi\left(\alpha\frac{\tilde{R}(X_i)-1}{n}\right)\right]+\phi\left(\alpha\frac{\tilde{R}(T_{(n)})}{n}\right)=0.$$

In practice we do not know the copula parameter so we need a second estimating equation to estimate α and θ . To obtain a second equation we will revisit the conditional Kendall's tau and consider its relation to a generalised cross ratio function.

5.3.3 The cross-ratio function and its relation to conditional Kendall's tau

Oakes (1989) describes a measure of local dependence called the cross-ratio function, which can be used in this setting. A local dependence at (x,t) between X and T is then defined, for $x > t$, as

$$\psi^*(x,t)=\frac{\pi(x,t)D_1D_2\pi(x,t)}{D_1\pi(x,t)D_2\pi(x,t)},$$

where the operators D_1 and D_2 are given by $D_1=-\dfrac{\partial}{\partial x}$ and $D_2=\dfrac{\partial}{\partial t}$.

In section A.4 on page 97 we show that the cross-ratio can be rewritten as

$$\psi^*(x,t)=\frac{P\{(X_1-X_2)(T_1-T_2)<0|\tilde{X}_{1,2}=x,\tilde{T}_{1,2}=t\}}{P\{(X_1-X_2)(T_1-T_2)>0|\tilde{X}_{1,2}=x,\tilde{T}_{1,2}=t\}}, \tag{5.18}$$

where (X_1,T_1) and (X_2,T_2) are independently distributed as 5.8 and

$$\tilde{X}_{1,2}=\min(X_1,X_2),\qquad \tilde{T}_{1,2}=\max(T_1,T_2).$$

From 5.18 we see that ψ^* is the ratio between the probability of discordance and the probability of concordance, given that $\tilde{X}_{1,2} = x$ and $\tilde{T}_{1,2} = t$. Hence $\psi^*(x,t) < 1$ indicates a positive association at (x,t) while $\psi^*(x,t) > 1$ indicates a negative association at (x,t). Note that $\psi(x,t)$ is only defined when $x > t$, since $\pi(x,t)$ is only defined when $x > t$. Consequently, the pairs (X_1, T_1) and (X_2, T_2) are comparable when we condition on $\tilde{X}_{1,2} = x$ and $\tilde{T}_{1,2} = t$. If we assume that X and T are quasi independent, then $\psi^*(x,t) = 1$.

The following theorem states that if $\pi(x,t)$ is Archimedean, then ψ^* can be expressed in terms of the generator function ϕ.

Theorem 5.3.2

Suppose that $\pi(x,t)$ is Archimedean so that 5.9 holds. Then $\psi^*(x,t)$ depends on x and t only through $v = \alpha\pi(x,t)$ with

$$\psi^*(x,t) = \psi(v) = \frac{-v\phi''(v)}{\phi'(v)}.$$

Proof: A proof analogous to this case can be found in (Oakes, 1989, page 488).

For the independent copula the generator function is $\phi(t) = -\log(t)$. So when $\pi(x,t) = S_x(x)G(t)$, theorem 5.3.2 states that $\psi(v) = 1$.

Theorem 5.3.3

Suppose that $\pi(x,t)$ is Archimedean so that 5.9 holds. Then we have the following relation between the cross-ratio function and conditional Kendall's tau:

$$\tau_c = E\left[\frac{1 - \psi\{\alpha\pi(\tilde{X}_{1,2}, \tilde{T}_{1,2})\}}{1 + \psi\{\alpha\pi(\tilde{X}_{1,2}, \tilde{T}_{1,2})\}}\middle| A\right],$$

where $A = \{\tilde{T}_{1,2} < \tilde{X}_{1,2}\}$.

Proof: For simplicity let

$$a = P\{(X_1 - X_2)(T_1 - T_2) > 0 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t\},$$
$$b = P\{(X_1 - X_2)(T_1 - T_2) < 0 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t\}.$$

Assume that X and T are continuous variables. We may then ignore the possibility of ties so that $a + b = 1$. Then 5.18 and theorem 5.3.2 on the preceding page implies that

$$\frac{1-\psi\{\alpha\pi(x,t)\}}{1+\psi\{\alpha\pi(x,t)\}} = \frac{a-b}{a+b} = E[\text{sgn}(X_1 - X_2)(T_1 - T_2)|\tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t].$$

Again, note that we only consider points (x,t) where $x > t$ so $\tilde{X}_{1,2} = x$ and $\tilde{T}_{1,2} = t$ imply that $A = \{\tilde{T}_{1,2} < \tilde{X}_{1,2}\}$. Consequently, we have that

$$E[\text{sgn}(X_1 - X_2)(T_1 - T_2)|A, \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t] = \frac{1-\psi\{\alpha\pi(x,t)\}}{1+\psi\{\alpha\pi(x,t)\}}\,.$$

In the proof of theorem 2.3.2 on page 20 we showed that τ_c could be written as $\tau_c = E[\text{sgn}(X_1 - X_2)(T_1 - T_2)|A]$. Using this definition and conditioning on $\tilde{X}_{1,2}$ and $\tilde{T}_{1,2}$ we get

$$\begin{aligned}
\tau_c &= E[\text{sgn}(X_1 - X_2)(T_1 - T_2)|A] \\
&= E\left[E[\text{sgn}(X_1 - X_2)(T_1 - T_2)|A, \tilde{X}_{1,2}, \tilde{T}_{1,2}]\middle|A\right] \\
&= E\left[\frac{1-\psi\{\alpha\pi(\tilde{X}_{1,2}, \tilde{T}_{1,2})\}}{1+\psi\{\alpha\pi(\tilde{X}_{1,2}, \tilde{T}_{1,2})\}}\middle|A\right].
\end{aligned}$$

5.3.4 Estimating the copula parameter using the conditional Kendall's tau

Let $(X_1, T_1), \ldots, (X_n, T_n)$ be i.i.d. random vectors following 5.8. In view of theorem 5.3.3 on the preceding page and equation 5.10 on page 55 an estimate of τ_c is given by

$$\hat{\tau}_c = \frac{1}{k}\sum\sum_{i<j} \frac{1-\psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}}{1+\psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}} I_{ij}, \tag{5.19}$$

where $\tilde{X}_{i,j} = \min(X_i, X_j)$, $\tilde{T}_{i,j} = \max(T_i, T_j)$, $I_{ij} = I\{\tilde{T}_{i,j} > \tilde{X}_{i,j}\}$ and $k = \sum\sum_{i<j} I_{ij}$. The second estimating equation is then obtained by setting $\hat{\tau}_c$ equal to the sample conditional Kendall's tau, t_c, given by equation 2.11 on page 21:

$$\frac{1}{k}\sum\sum_{i<j} \frac{1-\psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}}{1+\psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}} I_{ij} = \frac{1}{k}\sum\sum_{i<j} \text{sgn}(X_i - X_j)(T_i - T_j) I_{ij}\,.$$

This equation is equivalent to

$$
\begin{aligned}
0 &= \sum_{i<j}\sum \left[\operatorname{sgn}(X_i - X_j)(T_i - T_j) - \frac{1 - \psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}}{1 + \psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}} \right] I_{ij} \\
&= \sum_{i<j}\sum \left[I\{(X_i - X_j)(T_i - T_j) > 0\} - \frac{1}{1 + \psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}} \right] I_{ij} \\
&\quad + \sum_{i<j}\sum \left[\frac{\psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}}{1 + \psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}} - I\{(X_i - X_j)(T_i - T_j) < 0\} \right] I_{ij}.
\end{aligned}
$$

When $\hat{\tau}_c \neq 0$, the solution $(\hat{\alpha}, \hat{\theta})$ must be such that the last two sums both equal 0. We can therefore choose either one of these sums as a second estimating equation. We will follow Lakhal Chaieb *et al.* (2006) and use

$$
\begin{aligned}
&H_2(\alpha, \theta) \\
&= \frac{1}{n^2} \sum_{i<j}\sum \left[I\{(X_i - X_j)(T_i - T_j) > 0\} - \frac{1}{1 + \psi\{\alpha R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}} \right] I_{ij} = 0 \,,
\end{aligned}
$$

as the second estimating equation. The estimates $\hat{\alpha}$ and $\hat{\theta}$ are therefore obtained by solving

$$
H(\alpha, \theta) = \begin{pmatrix} H_1(\alpha, \theta) \\ H_2(\alpha, \theta) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.
$$

5.3.5 Estimating procedure for the Frank and Clayton Copulas

Before we can proceed with an evaluation of the estimation procedure discussed in the previous sections we need to know how to simulate from a meta distribution on the form:

$$P(X > x, T \leq t) = C^*(S(x), G(t)) = C^*(1 - F(x), G(t)).$$

If we know how to simulate from C^* we can do the following procedure:

1. Simulate (U, V) from C^*.
2. Let $X = F^{-1}(1 - U), \quad T = G^{-1}(V)$.

Then (X, T) will be distributed according to $C^*(S(x), G(t))$:

$$\begin{aligned} P(X > x, T \leq t) &= P(F^{-1}(1 - U) > x, G^{-1}(V) \leq t) \\ &= P(1 - U > F(x), V \leq G(t)) \\ &= P(U < 1 - F(x), V \leq G(t)) \\ &= C^*(1 - F(x), G(t)). \end{aligned}$$

Let us considered the case where C^* is the Frank's copula, but with a small modification of the generator ϕ:

$$\phi(t) = \log \frac{(1 - e^{\theta})}{(1 - e^{\theta t})}, \quad \phi'(t) = \frac{\theta e^{\theta t}}{1 - e^{\theta t}}, \quad \phi''(t) = \frac{\theta^2 e^{\theta t}}{(1 - e^{\theta t})^2}. \tag{5.20}$$

The only change from the standard generator function is $-\theta \to \theta$. For this parametrisation we have that positive values of θ corresponds to a positive dependence between X and T. According to theorem 5.3.2 on page 60 the cross-ratio function can be written as

$$\psi(v) = \frac{-v\phi''(v)}{\phi'(v)} = \frac{\theta v}{e^{\theta v} - 1}.$$

For this cross-ratio function the estimation equation $H_2(\alpha, \theta)$ becomes

$$H_2(\alpha,\theta) = \frac{1}{n^2}\sum_{i<j}\sum\Bigg[I\{(X_i - X_j)(T_i - T_j) > 0\} \\ - \frac{\exp\{\alpha\theta R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\}}{\alpha\theta R(\tilde{X}_{i,j}, \tilde{T}_{i,j}) + \exp\{\alpha\theta R(\tilde{X}_{i,j}, \tilde{T}_{i,j})/n\} - 1}\Bigg] I_{ij} = 0.$$

Note that $H_2(\alpha, \theta)$ only depends on α and θ through $\gamma = \alpha\theta$. An estimation procedure can then be given by

1. Solve $H_2(\alpha, \theta) = H_2(\gamma) = 0$ to obtain $\hat{\gamma}$.
2. Solve $H_1(\alpha, \hat{\gamma}/\alpha) = 0$ to obtain $\hat{\alpha}$.
3. Let $\hat{\theta} = \hat{\gamma}/\hat{\alpha}$.

In figure 5.1 on the next page the copula-graphic estimator is plotted against the true marginal distribution. The estimation is based on a sample of size $n = 100$, simulated from the frank copula with different types of marginals. For comparison we included the PLE.

We see that the copula-graphic estimator performes significantly better as an estimator of the marginal distribution than the PLE. Notice how the PLE underestimates when $\theta = -5$ and overestimates when $\theta = 5$. For the chosen parametrisation of the Frank copula these values correspond to negative and positive dependence, respectively.

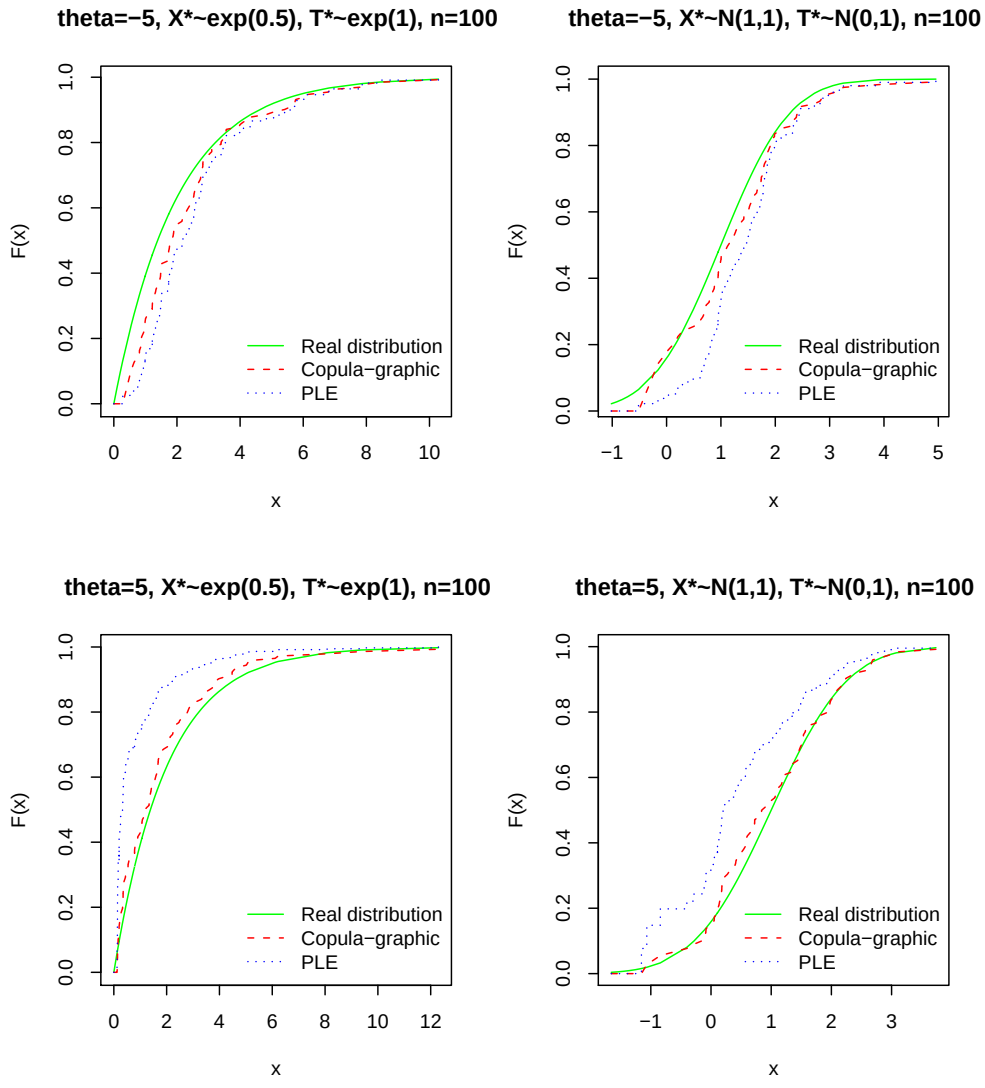

Figure 5.1: Copula-graphic estimator of $F(x)$ based on a sample of size $n = 100$, simulated from the Frank copula with parameter $\theta = 5, -5$ and with varying marginals.

For the Clayton copula the generator function is given by

$$\phi(t) = \frac{1}{\theta}(t^{-\theta} - 1), \quad \phi'(t) = -t^{-(\theta+1)}, \quad \phi''(t) = (\theta+1)t^{-(\theta+2)}.$$

The cross-ratio function is therefore given by

$$\psi(v) = \frac{-v\phi''(v)}{\phi'(v)} = \theta + 1.$$

For this cross-ratio function the estimation equation $H_2(\alpha, \theta)$ becomes

$$H_2(\alpha, \theta) = \frac{1}{n^2} \sum_{i<j} \sum \left[I\{(X_i - X_j)(T_i - T_j) > 0\} - \frac{1}{\theta + 2} \right] I_{ij} = 0.$$

Here $H_2(\alpha, \theta)$ is completely independent of α so solving $H_2(\alpha, \theta) = H_2(\theta) = 0$ gives us $\hat{\theta}$. Then $\hat{\alpha}$ can be obtained by solving $H_1(\alpha, \hat{\theta}) = 0$. In figure 5.2 the copula-graphic estimator is plotted against the true marginal distribution. The estimation is based on a sample of size $n = 100$, simulated from the Frank copula with normal and exponential marginals.

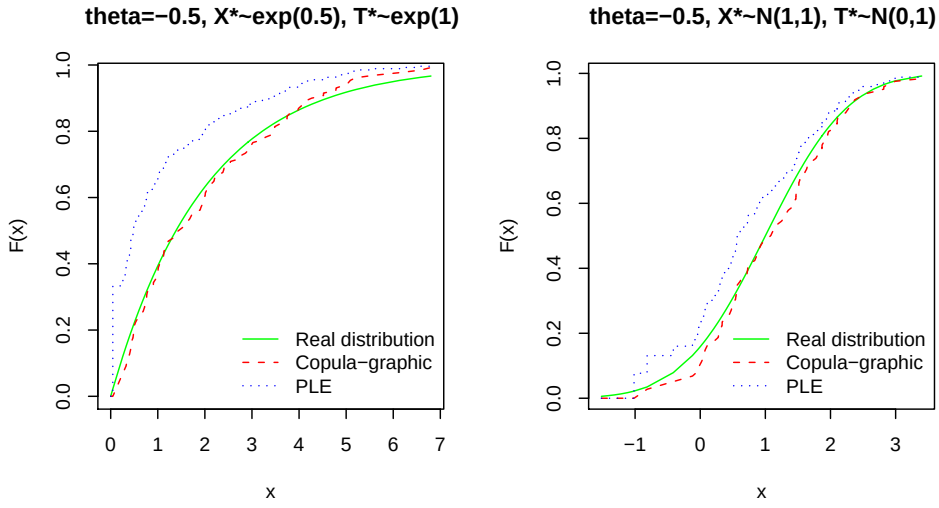

Figure 5.2: Copula-graphic estimator of $F(x)$ based on a sample of size $n = 100$, simulated from the Clayton copula with parameter $\theta = -0.5$ and with varying marginals.

5.3.6 Simulation

To evaluate the precision of this estimation procedure a simulation was carried out in R. The following routine was repeated 100 times for every fixed combination of $n = 100, 300$ and $\theta = 2, 5$:

- n truncated pairs were drawn from $\pi(x,t) = C^*(S(x), G(t))/\alpha$ with C^* as the Frank's copula with parameter θ and with marginals $S(x) = e^{-\frac{1}{2}x}$ and $G(t) = 1 - e^{-t}$.

- From these n pairs, $\hat{\theta}$ and $\hat{\alpha}$ were computed using the described method. These estimates where then used in the copula-graphic estimator of $\hat{S}_{CG}$ to compute $\hat{S}_{CG}(q_1)$, $\hat{S}_{CG}(q_2)$, $\hat{S}_{CG}(q_3)$ and $\hat{S}_{CG}(q_4)$. Where q_1, q_2, q_3 and q_4 are quantiles in the $\exp(1/2)$ distribution corresponding to $S(q_1) = 0.2$, $S(q_2) = 0.4$, $S(q_3) = 0.6$ and $S(q_4) = 0.8$

Let λ represent the parameters $\theta, \alpha, S(q_1), S(q_2), S(q_3)$ and $S(q_4)$, and let $\hat{\lambda}_i$ represent estimate number i. We then calculated:

$$AVE(\lambda) = \frac{1}{100}\sum_{i=1}^{100}\hat{\lambda}_i, \qquad VAR^*(\lambda) = \frac{1}{100}\sum_{i=1}^{100}(\hat{\lambda}_i - \lambda)^2.$$

Note that $VAR^*(\cdot)$ is the estimation error and not the sample variance. By Monte Carlo integration, $\theta = 2$ and 5 corresponds to $\alpha = 0.7052$ and 0.7725. In all cases, $AVE(\lambda)$ is close to the true value suggesting that the estimation procedure works. In most cases the estimators improved when n was increased from 100 to 300. When we increased θ, which is equivalent to increasing the dependence between X and T, there was a reduction in the estimation error. Remember that this also happened in our simulation of the sample conditional product-moment correlation coefficient r_c. This change increases the value of α and thus reduces the truncated propotion. Consequently, the n we observe represents a larger proportion of the original N, which will result in better estimates. Also notice that $VAR^*((S(q_i))$ increase when i increases, thus the copula-graphic estimator and the PLE share the property of larger estimation error in the lower domain of F. A similar simulation was carried out by Lakhal Chaieb *et al.* (2006) with similar results.

Table 5.3: Simulation results.

θ, α		n=100	n=300
2, 0.7052	$AVE(\hat{\alpha})$	0.6893	0.6920
	$AVE(\hat{\theta})$	1.8816	1.9422
	$AVE(\hat{S}(q_1))$	0.1927	0.1956
	$AVE(\hat{S}(q_2))$	0.3912	0.3934
	$AVE(\hat{S}(q_3))$	0.5901	0.5890
	$AVE(\hat{S}(q_4))$	0.7873	0.7893
	$VAR^*(\hat{\alpha})$	0.0170	0.0053
	$VAR^*(\hat{\theta})$	1.3406	0.4320
	$VAR^*(\hat{S}(q_1))$	0.0027	0.0007
	$VAR^*(\hat{S}(q_2))$	0.0065	0.0021
	$VAR^*(\hat{S}(q_3))$	0.0105	0.0034
	$VAR^*(\hat{S}(q_4))$	0.0109	0.0037
5, 0.7725	$AVE(\hat{\alpha})$	0.7563	0.7629
	$AVE(\hat{\theta})$	4.9732	4.9670
	$AVE(\hat{S}(q_1))$	0.1958	0.1971
	$AVE(\hat{S}(q_2))$	0.3936	0.3950
	$AVE(\hat{S}(q_3))$	0.5910	0.5954
	$AVE(\hat{S}(q_4))$	0.7997	0.7975
	$VAR^*(\hat{\alpha})$	0.0079	0.0025
	$VAR^*(\hat{\theta})$	1.5025	0.4591
	$VAR^*(\hat{S}(q_1))$	0.0023	0.0005
	$VAR^*(\hat{S}(q_2))$	0.0037	0.0012
	$VAR^*(\hat{S}(q_3))$	0.0030	0.0013
	$VAR^*(\hat{S}(q_4))$	0.0033	0.0014

6

Analysing the dependence between deductibles and claim sizes in shipping data

We will now consider twodimensional data where the variable of interest is the claim size and the truncating variable is the deductible of insured ships. These data are subject to the truncating sampling mechanism since we do not observe claims smaller than the corresponding deductible.

6.1 About the data

The ships are divided into the four categories: Cargo-, Bulk-, Container- and Tankships. All the ships in our analysis have reported a claim larger than the corresponding deductible. Even though many of the ships are given the same deductible, we will treat this quantity as a random variable.

It is in the insurance companies interest to estimate claims based on different covariates like age, engine type and total sum insured. The total sum insured plays an important role in our analysis of the data. We will denote this quantity by z. Within the four categories there is a wide range in the total sum insured. Therefore, to obtain an i.i.d. model, the claim size and deductible are divided by total sum insured. In this way we can compare ships of different value. Because of a rather large difference between the standardised deductibles and claim sizes we will consider the log transformed data.

To formalise, let $(X_1', T_1'), \ldots, (X_n', T_n')$ denote the original claim sizes and deductibles and let $z_1, \ldots, z_n$ denote the corresponding total sum insured. Then the variables subject to our analysis are given by

$$X_i = \log \frac{X_i'}{z_i} \quad \text{and} \quad T_i = \log \frac{T_i'}{z_i}, \quad 1 \leq i \leq n. \tag{6.1}$$

Note that $X_i' > T_i'$ imply $X_i > T_i$, so the sampling mechanism is the same for these variables. A summary of these quantities for the dataset Cargo can be seen in table 6.1. Notice how the transformation given above reduces the differences between the size of X and T.

Table 6.1: Summary of (X', T'), (X, T) and z for the dataset Cargo.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
X'	89000	174177	248420	448071	477760	8871006
T'	37450	75000	100000	134267	150000	500000
z	2450000	6300000	10880000	13790000	18000000	51400000
X	-5.7194	-4.1696	-3.6626	-3.5607	-2.9552	-0.8853
T	-6.7806	-5.0752	-4.7444	-4.5633	-4.0279	-2.4850

6.2 Testing the assumption of quasi independence

In view of equation 6.1 on the facing page, we have reason to believe that X is positively related to T since both variables depend on the value of total sum insured. Since we suspect a positive dependency structure we will consider the following assumptions:

1. $H_R : \rho_c = 0$ against $\rho_c > 0$.
2. $H_T : \tau_c = 0$ against $\tau_c > 0$.
3. $H_Z : \rho = 0$ against $\rho > 0$.

We put the significance level at 5%.

The two first assumptions can be properly tested by means of the statistics R and T given by equation 2.10 on page 16 and equation 2.12 on page 25, respectively. In section 6.5 on page 79 we will see that the truncated bivariate normal distribution fits the data relatively well, which is an indication that (X^*, T^*) may be bivariate normally distributed. This means that the third assumption may be tested by the statistic Z given by equation 4.1 on page 45.

For each dataset we computed the test statistics R, T and Z and the corresponding p-values. Note that the 5% critical value for these test statistics is 1.6448. The result can be seen in table 6.2. For the datasets Cargo, Bulk and Container, all three assumptions are rejected. For the dataset Tank we reject H_T but not H_R and H_Z. Since (X, T) may follow the truncated bivariate normal distribution, a rejection of H_R is a strong indication that the data are not quasi independent. We conclude that X and T are most likely not quasi independent in the datasets Cargo, Bulk and Container, while the assumption of quasi independence is questionable in the dataset Tank.

Table 6.2: Test results.

	R	T	Z	p-value(H_R)	p-value(H_T)	p-value(H_Z)
Cargo	2.1254	3.3078	4.1906	0.0167	0.0005	1.39e-05
Bulk	3.3138	3.4707	3.5733	0.0004	0.0003	0.0002
Container	7.4520	6.9828	10.6086	4e-14	1 e-12	0
Tank	1.2134	2.7565	1.5209	0.1124	0.0029	0.0641

6.3 Reconstruction of the joint distribution

We will now try to reconstruct the unconditional joint distribution of the claim size and the deductible. For this purpose we will consider the following models:

1. SE - Sample Estimates
   For comparison we include the sample mean and variance of $x_1, \ldots, x_n$ and $t_1, \ldots, t_n$, and the sample correlation between them. Because of the sampling mechanism we expect $\bar{x}$ to overestimate μ_x and $\bar{t}$ to underestimate μ_t. Hopefully, the other methods considered will adjust their means compared to these estimates.

2. PLE - The Product-limit estimator
   We know that X and T are probably not quasi independent in the datasets Cargo, Bulk and Container, so the PLE is not recommendable for these datasets. Still, we include the estimated parameters for comparison. The assumption of quasi independence may hold in the dataset Tank, so the PLE may be applicable for this dataset.

3. TBN - MLE assuming the data are Truncated Bivariate Normal distributed
   The first pure parametric candidate. A scatterplot from a truncated bivariate normal distribution should resemble an ellipse with the part $X < T$ "cut off". Figure B.1 on page 100 do in fact exhibit this characteristics of the truncated bivariate distribution, so this is a promising parametrisation.

4. Copula - MLE assuming the data follows a truncated meta distribution
   With this model we can try to fit different types of dependency structures. In section 6.4 on page 75 we will argue that normal marginals is a good choice for these datasets. The considered models can be seen in table 6.3 on the facing page.

5. CGE - Copula graphic estimator
   Using the procedure described in section 5.3 on page 54 we can estimate the parameters α and θ. This is done under the assumption that C^* is the modified Frank copula given by equation 5.20 on page 63 .

Table 6.3

	Copula	Marginal F	Marginal G
Copula1	Clayton	Normal	Normal
Copula2	Frank	Normal	Normal
Copula3	Gumbel	Normal	Normal

For the parametric distributions Copula1, Copula2 and Copula3 we also computed Monte Carlo estimates of α. The estimated parameters for the different models can be found in table 6.4 on the next page.

6.3.1 Results

As we expected, compared to the SE all the other methods gives a smaller estimate of μ_x and a larger estimate of μ_t. The TBN method gives a smaller estimate of ρ than the SE. For the datasets Cargo, Bulk and Container the PLE gives quite different estimates compared to the other methods, while for the dataset Tank the PLE estimates are somewhat closer to the others. This is reasonable since the assumption of quasi independence may hold for this dataset. This is also reflected when we try to fit a meta distribution to this dataset using the Gumbel copula: The optimisation tends towards the "illegal" copula parameter $\theta = 1$, so no parameter estimates are reported for this model. When $\theta \to 1$ in Gumbel copula we get the independence copula.

Because of the results in section 6.2 on page 71 we do not recommend the PLE for these datasets. To proceed with the CGE model, we would have to know if the selected Archimedean copula fits the data. This problem is addressed by Beaudoin and Lakhal-Chaieb (2008), but the topic is not covered in in this thesis. We will therefore proceed with an evaluation of the fit of the pure parametric models TBN, Copula1, Copula2 and Copula3. In the following section we will evaluate the marginal fit of these models. In section 6.5 on page 79 we will evaluate the joint fit of these models.

Table 6.4: Estimated parameters

Model	μ_x	μ_t	σ_x^2	σ_t^2	ρ	θ	α
			Cargo				
SE	-3.5607	-4.5629	0.90106	0.8303	0.7295	NA	NA
PLE	-4.5372	-2.7780	1.0286	0.2345	NA	NA	0.0721
TBN	-3.8224	-4.3698	1.0455	0.9066	0.5291	NA	0.7154
Copula1	-3.6199	-4.3766	0.9037	1.0809	NA	1.4131	0.8007
Copula2	-3.6888	-4.4706	1.0382	0.9698	NA	5.7277	0.8438
Copula3	-3.9991	-4.3842	1.2163	0.8306	NA	1.4207	0.6547
CGE	NA	NA	NA	NA	NA	3.8780	0.6828
			Bulk				
SE	-3.6604	-4.8383	0.7958	0.4250	0.5506	NA	NA
PLE	-4.1311	-3.9287	0.9632	0.1866	NA	NA	0.4089
TBN	-3.9570	-4.7729	1.0776	0.4365	0.3722	NA	0.7922
Copula1	-3.8245	-4.7693	0.9192	0.5024	NA	0.8351	0.8628
Copula2	-3.9337	-4.7887	1.0413	0.4519	NA	2.7953	0.8066
Copula3	-4.1711	-4.7016	1.3279	0.4421	NA	1.1094	0.6710
CGE	NA	NA	NA	NA	NA	2.3918	0.8444
			Container				
SE	-3.8240	-4.9982	0.9824	0.6924	0.7760	NA	NA
PLE	-4.3958	-3.4928	0.8860	0.4816	NA	NA	0.1836
TBN	-3.9073	-4.9855	1.0618	0.6907	0.7280	NA	0.9354
Copula1	-3.8476	-4.8800	1.0664	0.8174	NA	1.7851	0.9147
Copula2	-3.8888	-4.9459	1.0723	0.7169	NA	6.2805	0.9250
Copula3	-3.9848	-4.9901	1.2083	0.7233	NA	1.8561	0.8872
CGE	NA	NA	NA	NA	NA	5.1417	0.8853
			Tank				
SE	-3.9651	-5.1648	0.7709	0.3084	0.3968	NA	NA
PLE	-4.2455	-4.5011	0.7369	0.2427	NA	NA	0.5698
TBN	-4.4778	-5.0628	1.2814	0.3276	0.1401	NA	0.6882
Copula1	-4.2983	-5.0841	1.0753	0.3587	NA	0.3051	0.7690
Copula2	-4.3040	-5.0810	1.1149	0.3294	NA	1.1247	0.7641
Copula3	NA	NA	NA	NA	NA	NA	NA
CGE	NA	NA	NA	NA	NA	1.4002	0.8564

6.4 Monte Carlo estimated QQ-plots for truncated data

For a random variable X with distribution function F we define the q-quantile of F as

$$\pi_q(F) = \inf\{x : F(x) \geq q\}.$$

Assume we observe i.i.d. $X_1, \ldots, X_n$ from an unknown distribution and we want to check if the data follows a specific distribution F. Let $\hat{F}_n$ denote the empirical distribution function, i.e. $\hat{F}_n(x) = 1/n \sum_{i=1}^{n} I(X_i \leq x)$. Then a visual inspection can be done with the QQ-plot given by

$$(\pi_q(F), \hat{\pi}_q(\hat{F}_n)),$$

where a straight line is an indication that F fits the data. Since $\hat{\pi}_q(\hat{F}_n) = X_{(min\{i:q\leq \frac{i}{n}\})}$ the points of the plot are given by

$$(F^{-1}(i/n), X_{(i)}), \quad 1 \leq i \leq n.$$

In our case, we want to check if $X_1, \ldots, X_n$ follow $F^c(x) = H^c(x, \infty)$ for some estimated H^c. In section 3.4 on page 34 we managed to find a closed form of F^c when X^* and T^* were uniformly distributed on the unit interval. However, in most cases F^c and G^c are given by integrals which must be solved numerically. Simulating from F^c on the other hand is easy (See section 2.2.1 on page 11). We therefore propose a QQ-plot using the empirical distribution of $X_1, \ldots, X_n$ and $\tilde{X}_1, \ldots, \tilde{X}_{\tilde{n}}$, where $\tilde{X}_i$ is simulated from F^c and $\tilde{n}$ is large. Let $\tilde{F}_n^c$ denote the empirical distribution of $\tilde{X}_1, \ldots, \tilde{X}_{\tilde{n}}$. The resulting QQ-plot is then given by

$$(\tilde{F}_n^{c\,-1}(i/n), X_{(i)}) = (\tilde{X}_{(\min\{j:j\geq i\frac{\tilde{n}}{n}\})}, X_{(i)}).$$

Similarly, if we want to check if $T_1, \ldots, T_n$ follow $G^c(t) = H^c(\infty, t)$ we can consider the QQ-plot

$$(\tilde{G}_n^{c\,-1}(i/n), T_{(i)}) = (\tilde{T}_{(\min\{j:j\geq i\frac{\tilde{n}}{n}\})}, T_{(i)}).$$

For a sufficiently large $\tilde{n}$ these plots should be approximately the same as the one we would get with F^c and G^c. A straight line in this QQ-plot indicates that F^c and G^c fits the data. And if F^c and G^c fits the data it is reasonable to believe

F and G represents the unobserved data as well.

As can be seen figure 6.1 on the facing page and figure 6.2 on page 78, the conditional normal distribution fits the data quite well. These plots were made using the method described on the previous page. The variables $\tilde{X}_1, \ldots, \tilde{X}_{\tilde{n}}$ where simulated from the estimated truncated bivariate normal distribution. This model asserts that X^* and T^* both follow a normal distribution. Notice that several of the ships in the dataset Container have the same total sum insured and deductible. For this dataset there is little hope of finding any parametric distribution that will fit. We conclude that the normal distribution is a good choice of marginal distributions for X^* and T^*.

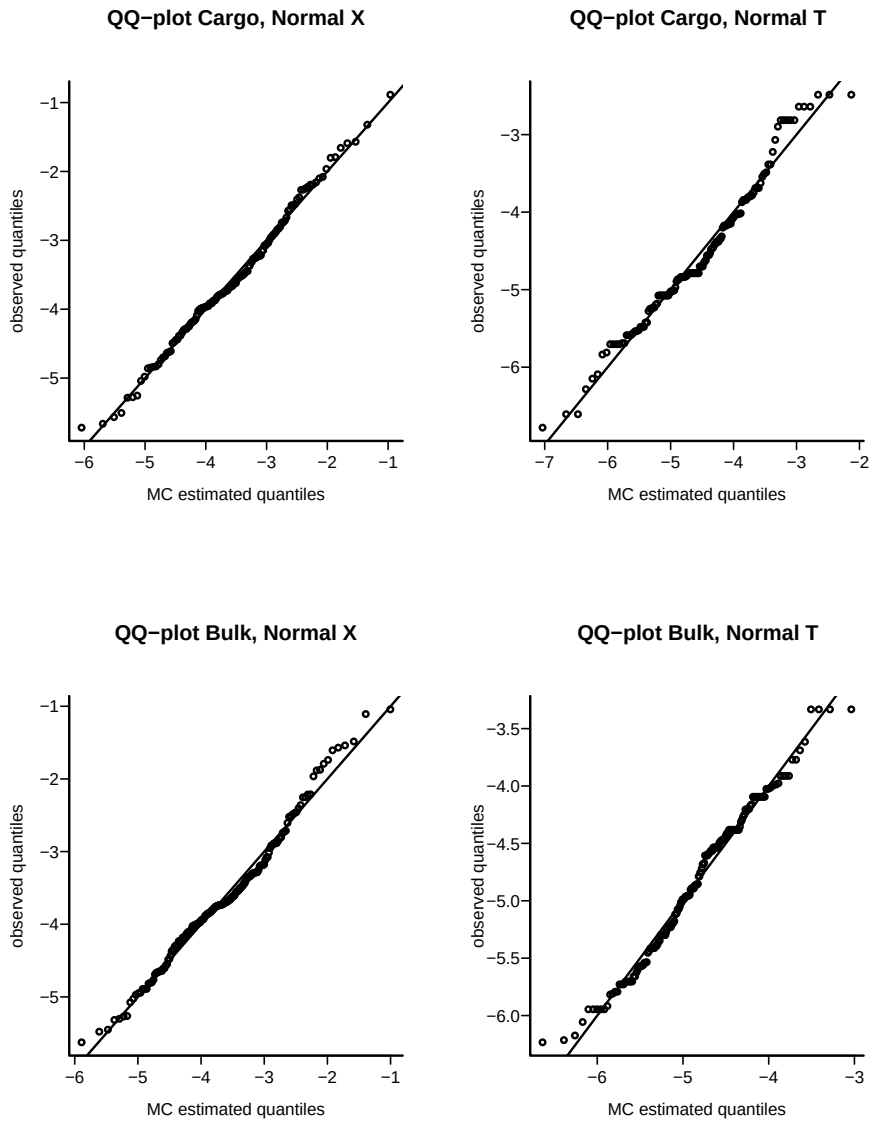

Figure 6.1: QQ-plot for the observed data versus the conditional normal distribution for the datasets Cargo and Bulk. The conditional normal quantiles was MC estimated from 300 000 pairs (X, T) drawn from the estimated TBN.

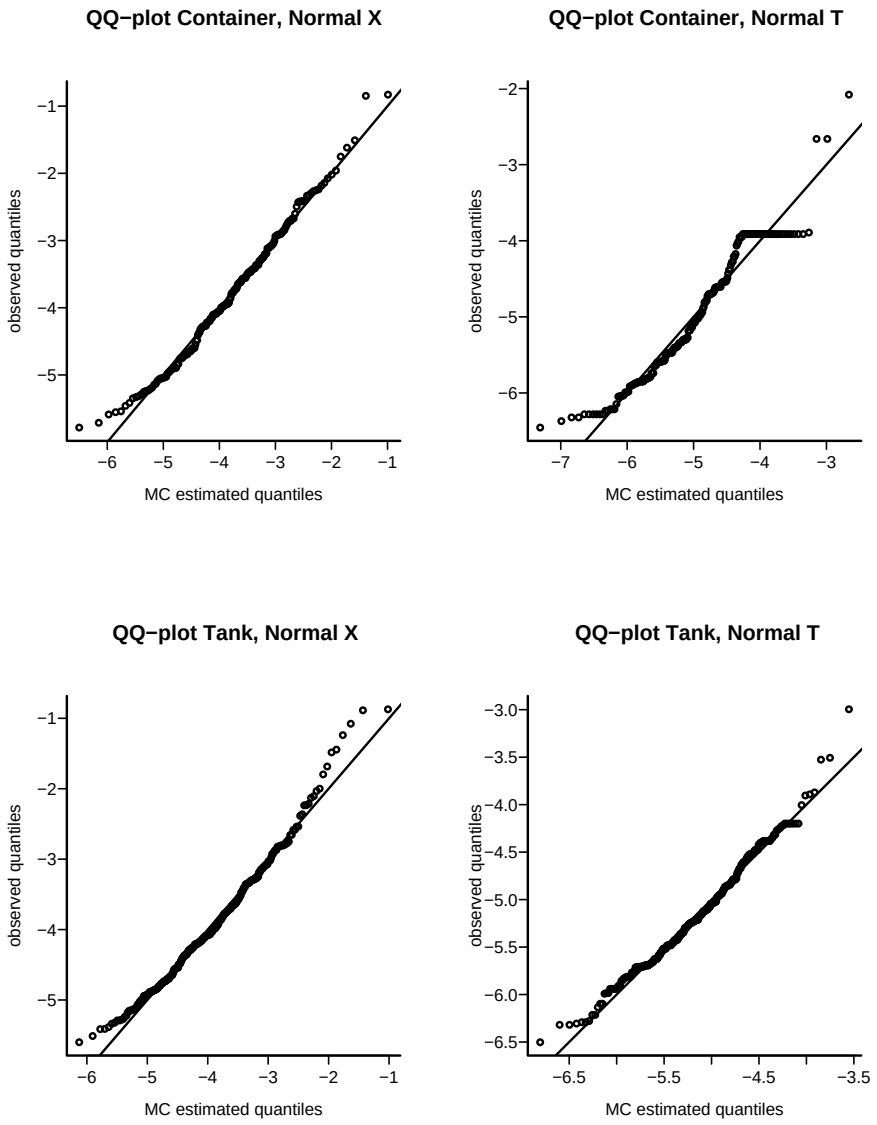

Figure 6.2: QQ-plot for the observed data versus the conditional normal distribution for the datasets Container and Tank. The conditional normal quantiles was MC estimated from 300 000 pairs (X, T) drawn from the estimated TBN. Notice that several of ships in the dataset Container is given the same deductible.

6.5 Goodness of fit test

Since we aim to reconstruct the distribution of (X^*, T^*), it would be nice to test

$$H(x,t) = \tilde{H}(x,t) \quad \forall\ (x,t),$$

where $\tilde{H}$ is one of the suggested parametric distributions. But such a procedure is not possible since the data we observe belong to the truncated vector $(X,T) = (X^*, T^* | X^* > T^*)$. However, if H^c fits the observed data it is reasonable to believe H represent the unobserved data as well. So the best we can do is to test

$$H_0 : H^c(x,t) = \tilde{H}^c(x,t) \quad \forall\ (x,t), \tag{6.2}$$

where $H^c(x,t)$ is given by equation 2.1 on page 6. A test for H_0 was suggested by Jostein Paulsen during his lectures in Risk Management.

First we simulate a substantial number of data from the estimated distribution $\tilde{H}^c$. We then divide $\mathbb{R}^2$ into k disjoint rectangles $I_1, \ldots, I_k$. The idea is that if the fit is good the relative number of data in each of the rectangles should be approximately the same for the observed data as for the simulated data. To be more specific, let $(X_1, T_1), \ldots, (X_n, T_n)$ be i.i.d. with distribution $H^c(x,t)$ and let $(\tilde{X}_1, \tilde{T}_1), \ldots, (\tilde{X}_{\tilde{n}}, \tilde{T}_{\tilde{n}})$ be i.i.d. with distribution $\tilde{H}^c(x,t)$. Let

$$\begin{aligned} n_j &= \text{card}\{i | (X_i, T_i) \in I_j\}, \quad j = 1, \ldots, k\,, \\ \tilde{n}_j &= \text{card}\{i | (\tilde{X}_i, \tilde{T}_i) \in I_j\}, \quad j = 1, \ldots, k\,. \end{aligned}$$

Let $\boldsymbol{p} = (p_1, \ldots, p_k)$, where $p_i = P((X,T) \in I_i)$. We can then consider $(n_1, \ldots, n_k)'$ to be multinomially distributed with n trials and k classes having probabilities $\boldsymbol{p}$. The Pearson statistic for testing the null hypothesis $H_0 : \boldsymbol{p} = \boldsymbol{a}$ is given by

$$C_n(\boldsymbol{a}) = \sum_{j=1}^{k} \frac{(n_j - n a_j)^2}{n a_j}.$$

It can be shown, see (van der Vaart, 1998), that under H_0 the sequence $C_n(\boldsymbol{a})$ converges to the χ^2_{k-1}-distribution as $n \to \infty$. An extension of this test is to

replace $\boldsymbol{a}$ by an estimate of $\boldsymbol{p}$, where the estimator $\hat{\boldsymbol{p}}$ is constructed so that it is a good estimator if the null hypothesis is true. Under the originally null hypothesis 6.2 the SLLN assures us that

$$\hat{p}_j = \frac{\tilde{n}_j}{\tilde{n}} \xrightarrow{a.s} p_j \quad \text{as} \quad \tilde{n} \to \infty.$$

Hence, this estimator is a good estimator if 6.2 is true. Deviation from 6.2 should then be reflected by the corresponding test statistic

$$C_{n,\tilde{n}}(\hat{\boldsymbol{p}}) = \sum_{j=1}^{k} \frac{(n_j - n\hat{p}_j)^2}{n\hat{p}_j}.$$

However, this procedure results in a reduction in the number of degrees of freedom. It can be shown, see e.g. (van der Vaart, 1998), that under H_0

$$C_{n,\tilde{n}}(\hat{\boldsymbol{p}}) \xrightarrow{d} \chi^2_{k-1-r} \quad \text{as} \quad n, \tilde{n} \to \infty \quad \text{at a rate so that} \quad \frac{n}{\tilde{n}} \to 0,$$

where r is the number of parameters estimated in $\tilde{H}^c(x,t)$. In the truncated bivariate case we need to estimate $(\mu_x, \mu_t, \sigma_x^2, \sigma_t^2, \rho)$, so $r = 5$.

The program we made simulates $\tilde{n} = 1\ 000\ 000$ pairs (X_i, T_i) from $\tilde{H}^c(x,t)$. For $C_{n,\tilde{n}}(\hat{\boldsymbol{p}})$ to be approximately chi-squared distributed it is advised, as a rule of thumbs, to choose the rectangles so that

$$\min\{n_1, \ldots, n_k, \tilde{n}_1, \ldots, \tilde{n}_k\} \geq 10.$$

To obtain such rectangles the program splits both datasets into a 9×9 grid, and then concatenates neighbouring cells with too few observation, so that each concatenated cell has at least 10 observations. The problem of too few observations in cells is mainly a problem in the original data set. The concatenation process is displayed in figure 6.3 on the next page.

6.5.1 Results

With significance level 0.01 we accept that three of the datasets may be truncated bivariate normally distributed. In appendix B we see that the scatterplot drawn from this distribution resembles the originally dataset better than the

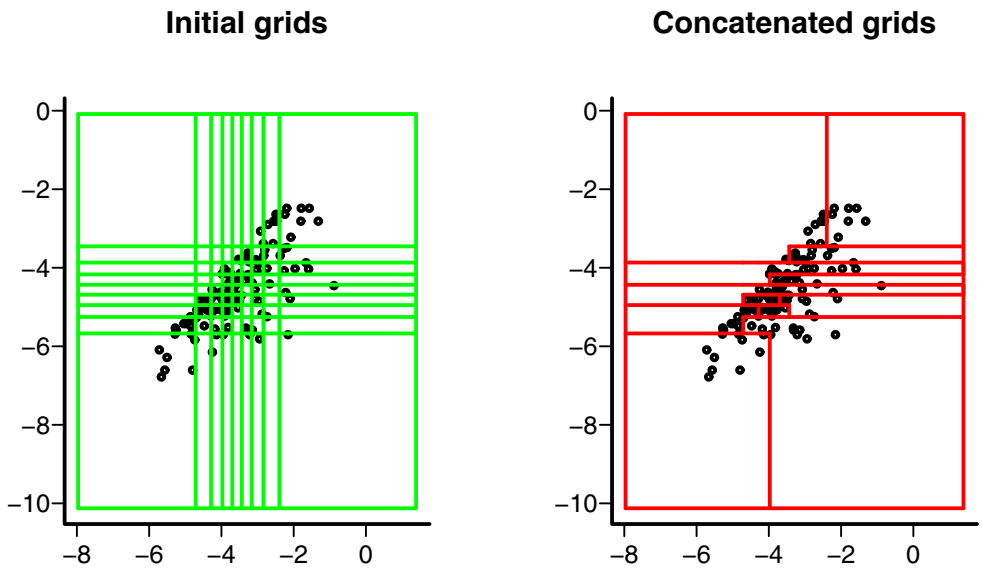

Figure 6.3: The process of concatenating grids when testing the fit of copula3 to the data set Cargo.

others. For the dataset Container there is little hope of finding a good fit no matter how many distributions we try. This is because marginally, no parametric estimate of G^c will be able to fit the data. None of the copula models fitted the data particularly well so we conclude that the overall best parametrisation (amongst the considered) is done with a truncated bivariate normal distribution.

One should note that this is a very powerful test, so if the data deviate just a little from $\tilde{H}^c(x,t)$ we get a rejection of H_0. Therefore, even though the p-value for the truncated bivariate normal distribution is small, we choose this model in our further investigation of the data.

Table 6.5: Table summarising the performance of different methods. k denotes the number of cells left after the concatenation.

Method	Dataset	p-value	1% critical value	$C_{n,\tilde{n}}$	k
TBN	Cargo	0.0100	15.08	15.07	11
	Bulk	0.0107	21.66	21.46	15
	Container	0.0002	20.09	29.72	14
	Tank	0.0607	34.80	28.08	24
Copula1	Cargo	0.0296	15.08	12.40	11
	Bulk	0.0051	21.66	23.52	15
	Container	3.75e-07	23.20	49.21	16
	Tank	1.83e-07	33.40	64.58	23
Copula2	Cargo	7.52e-05	15.08	26.38	11
	Bulk	5.76e-05	23.20	36.97	16
	Container	0.0004	21.66	30.20	15
	Tank	0.0002	34.80	46.56	24
Copula3	Cargo	0.0035	16.81	19.37	12
	Bulk	0.0096	21.66	21.77	15
	Container	0.0001	23.20	34.33	16
	Tank	NA	NA	NA	NA

6.6 Some applications of the estimated distribution

Let us assume that our datasets follow the estimated truncated bivariate normal distribution. From section 5.2 on page 51 we have that an estimate of $\alpha = P(X^* > T^*)$ is given by

$$\alpha(\hat{\boldsymbol{\theta}}) = \phi\left(\frac{\hat{\mu}_t - \hat{\mu}_x}{\sqrt{\hat{\sigma}_x^2 + \hat{\sigma}_t^2 - 2\hat{\rho}\hat{\sigma}_x\hat{\sigma}_t}}\right).$$

Having estimated α, we can estimate the population size N of the original sample. If we consider $(X_1^*, T_1^*), \ldots, (X_N^*, T_N^*)$ as independent trials where the event $X_i^* > T_i^*$ is a success, then $n \sim \text{Binomial}(N, \alpha)$ for all $N \geq 1$. We then have

$$\text{by SLLN} \quad \frac{n}{N} \xrightarrow{a.s} \alpha\,, \quad \text{hence} \quad \hat{N} = \frac{n}{\hat{\alpha}}\,.$$

Note that this is an approximation of an approximation, so the estimate is not necessarily accurate. An estimate of the number of unreported claims is then given by $\hat{n}_{ur} = \hat{N} - n$. The estimates for our datasets can be seen in table 6.6 on the next page.

When (X^*, T^*) is bivariate normally distributed it is straightforward to verify that the conditional distribution of X^* given $T^* = t$ is

$$N\big(\mu_x + \rho\frac{\sigma_x}{\sigma_t}(t - \mu_t), \sigma_x^2(1 - \rho^2)\big).$$

Let $f_{X|T}(x)$ be the density function of $X^*|T^* = t$ and consider the function

$$e^*(t) = E[\exp(X^*)|T = t] = \int_{-\infty}^{\infty} e^x f_{X|T}(x)\,\mathrm{d}x. \tag{6.3}$$

In view of equation 6.1 on page 70, it is natural to consider the function

$$e(u) = E[\exp(X^*)|\exp(T) = u] = e^*(\log u),$$

so that we can observe the dependency structure between X'/z and T'/z. In practice the deductible is known, so for a given deductible t for a ship with corresponding total sum insured z the expected claim size is $ze(t/z)$.

In figure 6.4 on the next page we plotted $e(u)$ using the estimated parameters from our datasets. For the dataset Container we expect a claim equal to the total sum insured when the deductible is set to 40% of total sum insured. Remember that the TBN model did not fit this dataset particularly well, so this result is most likely misleading. For the dataset Tank, there is little change in the expected claim size when we vary the deductible.

Table 6.6: Estimates of the population size of the original sample and the number of unreported claims.

	n	$\hat{\alpha}$	$\hat{N}$	$\hat{n}_{ur}$
Cargo	140	0.7154	196	56
Bulk	176	0.7922	222	46
Container	191	0.9354	204	13
Tank	299	0.6882	434	135

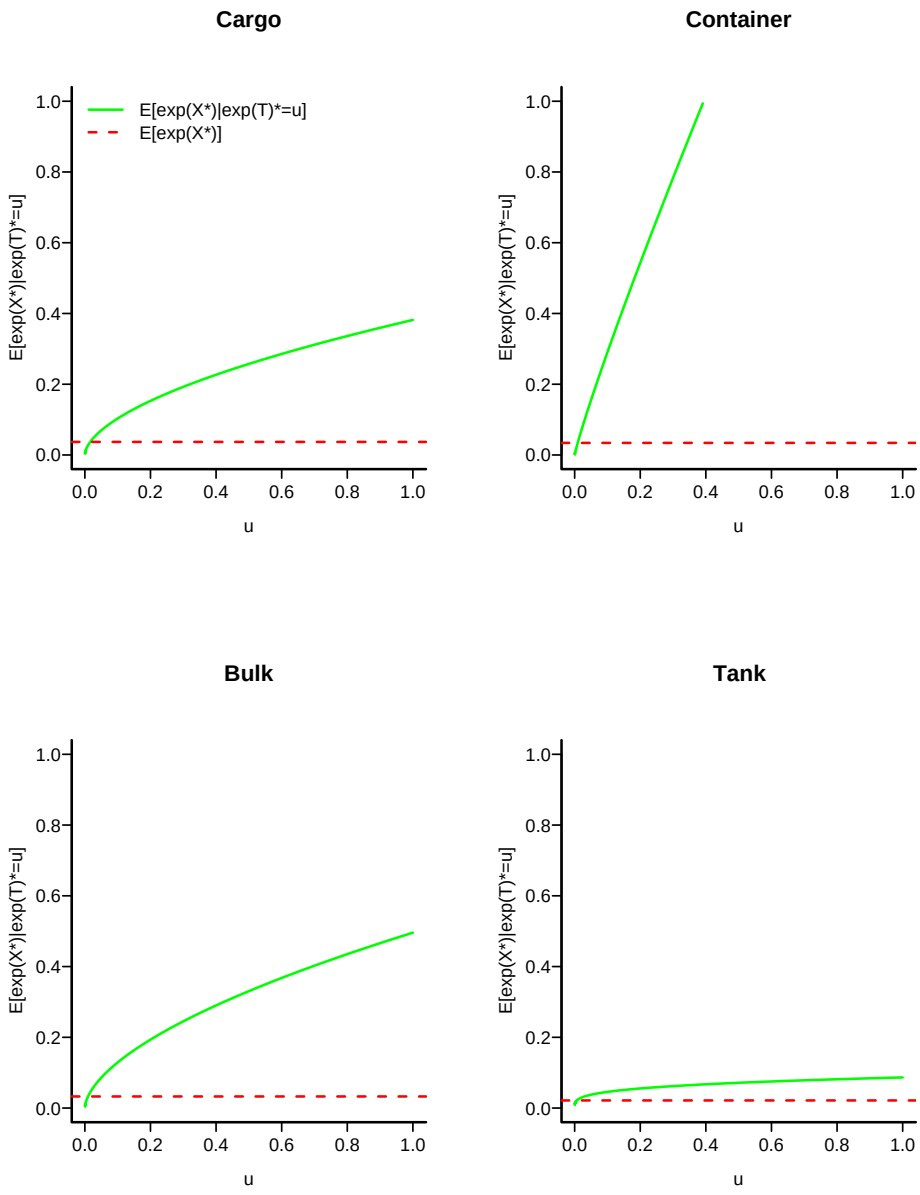

Figure 6.4: The relation between the standardised claim sizes and deductible under the estimated TBN model.

6.7 Further investigation of the dependence

As mentioned we believe some of the association between X and T can be explained by the standardisation done using the total sum insured. Notice that the relation between (X, T) and $\log z$ is linear since

$$X_i = \log \frac{X_i^{'}}{z_i} = \log X_i^{'} - \log z_i \quad \text{and} \quad T_i = \log T_i^{'} - \log z_i, \quad 1 \leq i \leq n.$$

Indeed, as seen in table 6.7, the statistics R and T are greatly reduced when we use the log transformed original data. But we still reject H_R and H_T for the datasets Cargo, Bulk and Container. However, by omitting the standardisation we loose the assumed i.i.d. property. Consequently, these test will no longer be valid, so we need to approach this investigation in a different way.

Table 6.7: Test of quasi independence between $\log X^{'}$ and $\log T^{'}$.

	R	T	5% critical value	p-value(H_R)	p-value(H_T)
Cargo	2.0538	2.9197	1.6448	0.0199	0.0017
Bulk	2.1888	2.7060	1.6448	0.0143	0.0034
Container	3.6013	3.6002	1.6448	0.0002	0.0002
Tank	0.0588	1.3317	1.6448	0.4765	0.0914

Let us still consider the log transformed standardised data $(X_1, T_1), \ldots, (X_n, T_n)$, but assume now that they are independently $TN_2(\mu_{xi}, \mu_{ti}, \sigma_x^2, \sigma_t^2, \rho)$ where

$$\mu_{xi} = \alpha_x + \beta_x \log z_i \,, \quad \mu_{ti} = \alpha_t + \beta_t \log z_i \,.$$

If the linear relation to $\log z$ is the only reason why X and T are dependent we expect ρ in this model to be close to zero. The parameters $(\alpha_x, \alpha_t, \beta_x, \beta_t, \sigma_x^2, \sigma_t^2, \rho)$ can be found by maximising the corresponding log likelihood, as described in section 5.2 on page 51. The estimated parameters for our datasets are displayed in table 6.8 on the facing page. As can be seen in figure 6.5 on the next page, the estimate of $((\alpha_x, \alpha_t, \beta_x, \beta_t)$ fits quite well. Compared to the estimates of ρ done under the TBN model in section 6.3 on page 72, the estimates of ρ under this model was smaller. Still, the test statistic Z rejects the hypothesis $\rho = 0$ for the datasets Cargo and Container.

Table 6.8: Estimated parameters in the normal linear model

	α_x	α_t	β_x	β_t	σ_x^2	σ_t^2	ρ
Cargo	11.66	10.53	-0.9580	-0.9267	0.7275	0.3954	0.2687
Bulk	14.59	9.53	-1.1369	-0.8755	0.9217	0.2811	0.1601
Container	11.28	8.79	-0.9322	-0.8459	0.6066	0.3182	0.4916
Tank	6.4016	11.36	-0.6522	-0.9773	1.2725	0.1999	-0.0635

Table 6.9: Test results of the hypothesis $H_Z : \rho = 0$ under the normal linear model.

	Z	5% critical value	p-value(Z)
Cargo	2.1102	1.6448	0.0174
Bulk	1.4146	1.6448	0.0785
Container	6.0940	1.6448	5.50e-10
Tank	-0.6304	1.6448	0.7357

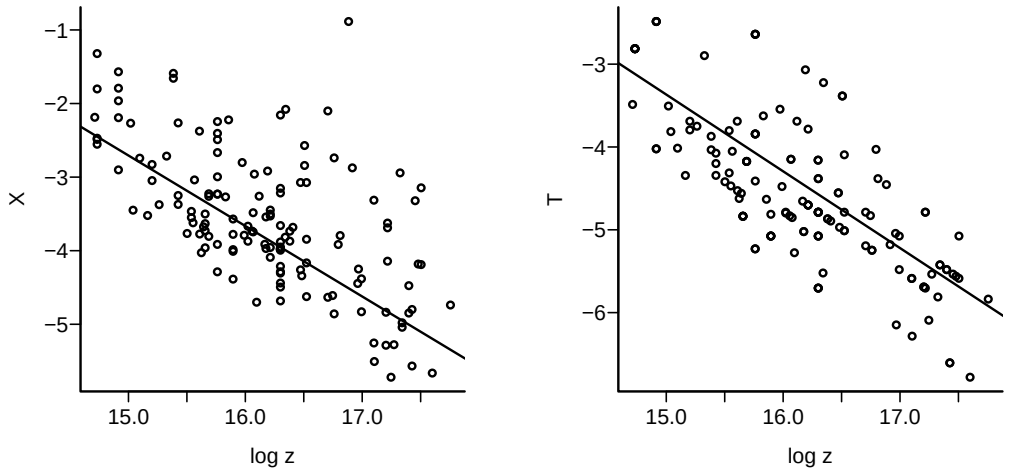

Figure 6.5: The lines $\hat{\alpha}_x + \hat{\beta}_x \log z$ and $\hat{\alpha}_t + \hat{\beta}_t \log z$ plotted together with the observed values of $(\log z, X)$ and $(\log z, T)$ for the dataset Cargo.

6.8 Conclusion and final remarks

The analysis of the dependence between the deductible and claim size given in this chapter was motivated by the following

- A significant association between the claim size and the deductible implies that models which aim to estimate claim sizes based on covariates could benefit on including the deductible as an additional covariate.
- A reconstruction of the unconditional joint distribution of the claim size and deductible provides an estimate of the number of unreported claims.

After testing the assumption of quasi independence in section 6.2 on page 71, we concluded that this assumption fails to hold in three out of the four considered dataset. We also concluded that the assumption of quasi independence was questionable for the fourth dataset. However, we pointed out that one possible explanation of this association could be the standardisation done with the total sum insured.

Amongst the different methods of reconstructing the unconditional joint distribution, we chose to focus on the pure parametric models TBN, Copula1, Copula2 and Copula3. The validity of these models relies on the untestable assumption that the deductible and claim size follows a certain unconditional distribution. In section 6.5 on page 79, we argued that a good fit of the corresponding conditional distribution indicates that the unconditional distribution may represent the unobserved data as well. However, this argument can be quite misleading, specially if the truncated proportion is large. This remains one of the fundamental problems when reconstructing the unconditional distribution using data subject to a dependent truncation.

In section 6.7 on page 86, we addressed the assertion that the dependence between the deductible and claim size can be explained by the standardisation done using the total sum insured. This was done by modelling the mean of the deductible and claim size as linear functions of the log transformed total sum insured. The results clearly indicated that much of the association between the deductible and claim size is explained by this standardisation. However, two of the datasets still seemed to be subject to a dependent truncation.

A

Some proofs

A.1 Quasi independence imply $\rho_c = 0$

Proof: We now give a proof of theorem 2.2.2 on page 10 in the continuous case. We divide the set $A = \{\max(T_1, T_2) < \min(X_1, X_2)\}$ into the following subsets:

$$
\begin{aligned}
A_1 &= \{T_2 < T_1 < X_2 < X_1\}, \\
A_2 &= \{T_1 < T_2 < X_2 < X_1\}, \\
A_3 &= \{T_2 < T_1 < X_1 < X_2\}, \\
A_4 &= \{T_1 < T_2 < X_1 < X_2\}.
\end{aligned}
$$

In this way $A = A_1 \cup A_2 \cup A_3 \cup A_4$
As before, we denote the density of (X, T), X^* and T^* by $h^c(x, t)$, $f(x)$ and $g(t)$, respectively. Note that A_1, A_2, A_3 and A_4 are disjoint sets and that $P(A_1) = P(A_4)$, $P(A_2) = P(A_3)$. Thus, the numerator of ρ_c given in equation 2.4 on page 10 can be rewritten in the following manner:

$$
\begin{aligned}
\sigma_{XT} &= E(X_1 - X_2)(T_1 - T_2)I_A = E(X_1 - X_2)(T_1 - T_2)(I_{A_1} + I_{A_2} + I_{A_3} + I_{A_4}) \\
&= 2E(X_1 - X_2)(T_1 - T_2)(I_{A_1} + I_{A_2}) \\
&= 2\Big\{ \int\int\int\int_{t_2<t_1<x_2<x_1} (x_1 - x_2)(t_1 - t_2)h^c(x_1, t_1)h^c(x_2, t_2)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}t_1\,\mathrm{d}t_2 \\
&\quad + \int\int\int\int_{t_1<t_2<x_2<x_1} (x_1 - x_2)(t_1 - t_2)h^c(x_1, t_1)h^c(x_2, t_2)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}t_1\,\mathrm{d}t_2 \Big\}.
\end{aligned}
$$

For the last integral we make the change of variable

$$
\begin{pmatrix} u_1 \\ u_2 \\ x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ x_1 \\ x_2 \end{pmatrix},
$$

where the Jacobian of the transformation is $J = -1$. With this change of variable, we obtain

$$
\begin{aligned}
\sigma_{XT} &= 2\int\int\int\int_{u_2<u_1<x_2<x_1} (x_1 - x_2)(u_1 - u_2) \\
&\quad \times \{h^c(x_1, u_1)h^c(x_2, u_2) - h^c(x_2, u_1)h^c(x_1, u_2)\}|-1|\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}u_1\,\mathrm{d}u_2.
\end{aligned}
$$

If we now assume quasi independence:

$$
H_0' : \quad h^c(x,t) = \begin{cases} f(x)g(t)/\alpha_0, & x > t, \\ 0, & \text{otherwise}, \end{cases}
$$

we get

$$
\begin{aligned}
\sigma_{XT} &= \frac{2}{\alpha_0^2}\int\int\int\int_{u_2<u_1<x_2<x_1} (x_1 - x_2)(u_1 - u_2) \\
&\quad \times \{\underbrace{f(x_1)g(u_1)f(x_2)g(u_2) - f(x_2)g(u_1)f(x_1)g(u_2)}_{0}\}\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}u_1\,\mathrm{d}u_2 = 0.
\end{aligned}
$$

Hence $\rho_c = 0$ and the proof is complete.

A.2 Proof of the asymptotic properties of r_c

To establish the asymptotic properties of r_c some basic knowledge about the properties of U-statistics is needed, hence the following definition and theorem are given.

A.2.1 U-Statistics

Definition A.2.1: U-statistics

Let $X_1, ..., X_n$ be an i.i.d. random sample from an unknown distribution. Let h be a permutation symmetric function and consider the estimation of $\theta = Eh(X_1, ..., X_r)$. A U-statistic with kernel h will be an unbiased estimator of θ and is defined as

$$U = \frac{1}{\binom{n}{r}} \sum_{\beta} h(X_{\beta_1}, ..., X_{\beta_r})$$

The set β is all unordered subset of r, where the integers can be taken from $\{1, ..., n\}$.

Note that the elements in a U-statistic are in general dependent. Hence, we can't derive the asymptotic behaviour of this statistic by direct application of LLN(Law of Large Numbers) and CLT(Central limit theorem). The solution of this problem is to approximate the original U-statistic by a sum of i.i.d random quantities called projections, which asymptotically have the same distribution as the U-statistic. We can then obtain the asymptotic properties of the underlying U-statistic by applying LLN and CLT on this sum. The details of this procedure is formulated in the following theorem.

Theorem A.2.2: Asymptotic properties of U-statistics

Let $\hat{U}$ be the projection of $U - \theta$ onto the set of all statistics of the form $\sum_{i=1}^{n} g_i(X_i)$ and let

$$h_1(x) = Eh(X_1, X_2, \ldots, X_r | X_1 = x) - \theta.$$

If $Eh^2(X_1, ..., X_r) < \infty$ then $\sqrt{n}(U - \theta - \hat{U}) \xrightarrow{p} 0$. Consequently, the sequence $\sqrt{n}(U - \theta)$ is asymptotically normal with mean 0 and variance

$r^2\zeta$ where, with $X_1, ..., X_r$ and $X_1', ..., X_r'$ denoting i.i.d. variables,

$$\zeta = \text{cov}\{h(X_1, \ldots, X_r), h(X_1, X_2' \ldots, X_r')\},$$
$$\hat{U} = \sum_{i=1}^{n} E(U - \theta \mid X_i) = \frac{r}{n} \sum_{i=1}^{n} h_1(X_i).$$

Proof: Note that the theorem describes the centred projection. A full proof can be found in (van der Vaart, 1998, page 162), but the main idea is the following decomposition:

$$\sqrt{n}\,(U - \theta) = \sqrt{n}\,(\hat{U}) + \sqrt{n}\,(U - \theta - \hat{U}).$$

It can be proved that the last term converge in probability to zero. The normality is then obtained by applying *CLT* on the first term since the elements in $\hat{U}$ are independent. If we can establish that we are dealing with a U-statistic we know the following:

$$U \xrightarrow{P} \theta, \qquad \sqrt{n}\,(U - \theta) \xrightarrow{d} N(0, r^2\zeta).$$

A.2.2 Consistency of r_c

With this knowledge at hand, we now proceed to the proof of equation 2.8 on page 15. The first step is to show that $\frac{1}{n^2} S_{XT}$ is a consistent estimator of σ_{XT}, that is:

$$\frac{1}{n^2} S_{XT} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)(T_i - T_j) I_{ij} \xrightarrow{P} E\left[(X_1 - X_2)(T_1 - T_t) I_A\right].$$

Observe that every element in S_{XT} is repeated once since $(X_i - X_j)(T_i - T_j)I_{ij} = (X_j - X_i)(T_j - T_i)I_{ji}$. This means that:

$$\frac{1}{n^2} S_{XT} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)(T_i - T_j) I_{ij} = \frac{2}{n^2} \sum\sum_{i<j} (X_i - X_j)(T_i - T_j) I_{ij},$$

which asymptotically is the same as:

$$\frac{2}{n(n-1)} \sum\sum_{i<j} (X_i - X_j)(T_i - T_j) I_{ij} = \frac{1}{\binom{n}{2}} \sum\sum_{i<j} (X_i - X_j)(T_i - T_j) I_{ij} = U_1.$$

This is a U-statistic of order $r = 2$ for the kernel:

$$h\left(\begin{pmatrix} x_1 \\ t_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ t_2 \end{pmatrix}\right) = (x_1 - x_2)(t_1 - t_2)1\{\max(t_1, t_2) \leq \min(x_1, x_2)\},$$

estimating $\theta_1 = E(X_1 - X_2)(T_1 - T_2)I_A = \sigma_{XT}$. From theorem A.2.2 on page 91 it now follows that $n^{-2}S_{XT}$ is a consistent estimator of σ_{XT}. Similar arguments will also establish $U_2 = n^{-2}S_{XX} \xrightarrow{P} \sigma_{XX}$ and $U_3 = n^{-2}S_{TT} \xrightarrow{P} \sigma_{TT}$. Consider the function

$$f(x, y, z) = \frac{x}{(yz)^{\frac{1}{2}}} \tag{A.1}$$

We know that $f : \mathbb{R}^3 \to \mathbb{R}$ is a function continuous for all $(\sigma_{XT}, \sigma_{XX}, \sigma_{TT})$ when $\sigma_{XX} > 0$ and $\sigma_{TT} > 0$. It then follows from the *continuous mapping theorem*[1] that:

$$r_c = f(U_1, U_2, U_3) = \frac{U_1}{(U_2U_3)^{\frac{1}{2}}} \xrightarrow{P} f(\sigma_{XT}, \sigma_{XX}, \sigma_{TT}) = \frac{\sigma_{XT}}{(\sigma_{XX}\sigma_{TT})^{\frac{1}{2}}} = \rho_c,$$

and the proof is complete. Note that this proof holds under the conditions $E(X_1 - X_2)^2(T_1 - T_2)^2I_A < \infty$, $E(X_1 - X_2)^4I_A < \infty$ and $E(T_1 - T_2)^4I_A < \infty$ though the result itself may hold under weaker conditions.

A.2.3 Normality of r_c

We now want to establish that r_c is asymptotically normal distributed. By theorem A.2.2 on page 91 U_1, U_2 and U_3 are asymptotically normal distributed. However, if we want to apply the Delta method, we need the joint asymptotic distribution of $\mathbf{U} = (U_1, U_2, U_3)^T$. We therefore proceed with a formal definition of a multivariate U-statistic.

[1] Let X be a random vector in $\mathbb{R}^k$ and let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $P(X \in C) = 1$. Then
(i) If $X_n \xrightarrow{d} X$ then $g(X_n) \xrightarrow{d} g(X)$
(ii) If $X_n \xrightarrow{P} X$ then $g(X_n) \xrightarrow{P} g(X)$
(iii) If $X_n \xrightarrow{a.s.} X$ then $g(X_n) \xrightarrow{a.s.} g(X)$.

Definition A.2.3: One-sample Order-r Multivariate U-statistics

Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution. Let $\mathbf{h}(X_1, \ldots, X_r) = (h_1(X_1, \ldots, X_r), \ldots, h_v(X_1, \ldots, X_r))^T$ be a vector consisting of permutation symmetric functions and consider the estimation of $\boldsymbol{\theta} = (Eh_1(X_1, ..., X_r), \ldots, Eh_v(X_1, ..., X_r))^T$. A U-statistic with kernel vector $\mathbf{h}$ will be an unbiased estimator of $\boldsymbol{\theta}$ and is defined as

$$\mathbf{U} = \frac{1}{\binom{n}{r}} \sum_{\beta} \mathbf{h}(X_{\beta_1}, ..., X_{\beta_r})$$

where, as before, the set β is all unordered subset of r, where the integers can be taken from $\{1, ..., n\}$.

The asymptotic properties of *Order-r Multivariate U-statistics* are derived in Kowalski and Tu (2008) and the result is analogous to that of Theorem A.2.2. However, there is a different representation for ζ, which we will adopt in the next theorem. For the first element in $\mathbf{h}$ we have that

$$\zeta = \text{Var}[h_1^*(X_1)], \quad h_1^*(x) = E[h_1(X_1, \ldots, X_r)|X_1 = x].$$

Theorem A.2.4

Let $\mathbf{h}^*(X_1) = (h_1^*(X_1), \ldots, h_v^*(X_1))^T$ be the natural extension of h_1^* above. Then, under mild regularity conditions

$$\sqrt{n}\,(\mathbf{U} - \boldsymbol{\theta}) \xrightarrow{d} N_v(0, r^2\Sigma_h),$$

where $\Sigma_h = \text{Var}(\mathbf{h}^*(X_1)) = E[(\mathbf{h}^*(X_1) - \boldsymbol{\theta})(\mathbf{h}^*(X_1) - \boldsymbol{\theta})^T]$

Proof: A proof can be found in (Kowalski and Tu, 2008, page 255).

This theorem is applicable to our case since U_1, U_2 and U_3 all are U-statistics of order $r = 2$ and are formed by the same sample. This means, with $\mathbf{U} = (U_1 = n^{-2}S_{XT}, U_2 = n^{-2}S_{XX}, U_3 = n^{-2}S_{TT})^T$ and $\boldsymbol{\theta} = (\sigma_{XT}, \sigma_{XX}, \sigma_{TT})^T$,

$$\sqrt{n}\,(\mathbf{U} - \boldsymbol{\theta}) \xrightarrow{d} N_v(0, 4\Sigma_h),$$

where the covariance matrix is

$$\Sigma_h = E\left[\left(\mathbf{h}^*(\begin{pmatrix}X_1\\T_1\end{pmatrix}) - \boldsymbol{\theta}\right)\left(\mathbf{h}^*(\begin{pmatrix}X_1\\T_1\end{pmatrix}) - \boldsymbol{\theta}\right)^T\right],$$

and the elements of $\mathbf{h}^*$ are

$$\begin{aligned}
h_1^*\left(\begin{pmatrix}X_1\\T_1\end{pmatrix}\right) &= E\left[(X_1 - X_2)(T_1 - T_2)I_A \middle| \begin{pmatrix}X_1\\T_1\end{pmatrix}\right],\\
h_2^*\left(\begin{pmatrix}X_1\\T_1\end{pmatrix}\right) &= E\left[(X_1 - X_2)^2 I_A \middle| \begin{pmatrix}X_1\\T_1\end{pmatrix}\right],\\
h_3^*\left(\begin{pmatrix}X_1\\T_1\end{pmatrix}\right) &= E\left[(T_1 - T_2)^2 I_A \middle| \begin{pmatrix}X_1\\T_1\end{pmatrix}\right].
\end{aligned}$$

Next, let $D(\boldsymbol{\theta}) = \nabla f(\sigma_{XT}, \sigma_{XX}, \sigma_{TT})$, where f is given by equation A.1 on page 93. Thus

$$D(\boldsymbol{\theta}) = \left(\frac{\partial f}{\partial \sigma_{XT}}, \frac{\partial f}{\partial \sigma_{XX}}, \frac{\partial f}{\partial \sigma_{TT}}\right)^T = \left(\frac{1}{(\sigma_{XX}\sigma_{TT})^{\frac{1}{2}}}, \frac{-\sigma_{XT}}{2(\sigma_{XX}^3\sigma_{TT})^{\frac{1}{2}}}, \frac{-\sigma_{XT}}{2(\sigma_{XX}\sigma_{TT}^3)^{\frac{1}{2}}}\right)^T.$$

If we now apply the Delta method, we obtain the asymptotic distribution of r_c:

$$\sqrt{n}\,(r_c - \rho_c) = \sqrt{n}\,(f(\mathbf{U}) - f(\boldsymbol{\theta})) \xrightarrow{d} N\left(0, \sigma^2 = 4D(\boldsymbol{\theta})^T\Sigma_h D(\boldsymbol{\theta})\right),$$

and the proof is complete.

A.2.4 Consistency of the estimated asymptotic variance

A procedure how to estimate $\sigma^2 = 4D(\boldsymbol{\theta})^T\Sigma_h D(\boldsymbol{\theta})$ is described in (Kowalski and Tu, 2008). However, this procedure require rather "nice" expressions of h_1^*, h_2^* and h_3^* and is therefore not applicable in our case. An alternative approach to this problem can be found in appendix A in (Chen *et al.*, 1996).

A.3 Example of Uniform S

We now proceed to show that theorem 2.3.1 on page 22 holds in the simple case when $R_i = 3$ and T is a continuous variable. In general, since $i \in \mathscr{R}_i$ there is always one tie in S_i, therefore $\max(S_i) = r_i - 1$ and $\min(S_i) = -r_i + 1$. This explains the range of S_i. Assume $\mathscr{R}_i = \{i, a, b\}$ where a and b are arbitrary number from the set $\{j | 1 \leq j \leq n, j \neq i\}$. It then follows that

$$R_i = 3 \quad \text{and} \quad S_i \in \{-2, 0, 2\}.$$

As before, let $g^c(t)$ and $G^c(t)$ denote the density and cumulative distribution function of T, respectively. Then

$$\begin{aligned} P(S_i = -2 | R_i = 3) &= P(\operatorname{sgn}(T_a - T_i) = -1 \cap \operatorname{sgn}(T_b - T_i) = -1) \\ &= P(T_a < T_i \cap T_b < T_i) \\ &= \int_{-\infty}^{\infty} P(T_a < T_i \cap T_b < T_i | T_i = t) g^c(t) \, \mathrm{d}t \\ &= \int_{-\infty}^{\infty} P(T_a < t \cap T_b < t) g^c(t) \, \mathrm{d}t. \end{aligned}$$

Under H_0 there is no trend in the selection of T_a and T_b in S_i i.e. they are independent. We can then write the last integral as

$$\int_{-\infty}^{\infty} P(T_a < t) P(T_b < t) g^c(t) \, \mathrm{d}t = \int_{-\infty}^{\infty} G^c(t)^2 g^c(t) \, \mathrm{d}t = E[G^c(T)^2].$$

In general we know that $G^c(T) \sim U[0, 1]$, hence

$$P(S_i = -2 | R_i = 3) = E[G^c(T)^2] = E[U^2] = \int_0^1 u^2 \, \mathrm{d}u = \frac{1}{3}.$$

Similarly we have that

$$\begin{aligned} P(S_i = 0|R_i = 3) &= P(\{T_a < T_i \cap T_b > T_i\} \cup \{T_a > T_i \cap T_b < T_i\}) \\ &= 2\int_{-\infty}^{\infty} P(T_a < t \cap T_b > t) g^c(t)\,\mathrm{d}t \\ &= 2\int_{-\infty}^{\infty} G^c(t)(1 - G^c(t)) g^c(t)\,\mathrm{d}t \\ &= 2\{E[U] - E[U^2]\} \\ &= 2\{\frac{1}{2} - \frac{1}{3}\} = \frac{1}{3}, \end{aligned}$$

and

$$\begin{aligned} P(S_i = 2|R_i = 3) &= P(T_a > T_i \cap T_b > T_i) \\ &= \int_{-\infty}^{\infty} P(T_a > t \cap T_b > t) g^c(t)\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty} (1 - G^c(t))^2 g^c(t)\,\mathrm{d}t \\ &= E[(1 - U)^2] \\ &= E[U_*^2] = \frac{1}{3}, \end{aligned}$$

where the second last equality holds since $1 - U$ also is uniformly distributed on $[0, 1]$. Hence

$$P(S_i = -2|R_i = 3) = P(S_i = 0|R_i = 3) = P(S_i = 2|R_i = 3) = \frac{1}{3}.$$

Thus $S_i|R_i = 3$ is uniformly distributed on $\{-2, 0, 2\}$. Analogous calculations can be done for any value of R_i.

A.4 Alternative representation of the cross-ratio function

In general, when $\pi(x, t)$ is continuous we have that

$$\psi(x, t) = \frac{\pi(x, t) D_1 D_2 \pi(x, t)}{D_1 \pi(x, t) D_2 \pi(x, t)} = \frac{P(X > x, T < t) P(X = x, T = t)}{P(X = x, T < t) P(X > x, T = t)}$$

where the somewhat sloppy notation $X = x$ and $T = t$ should be interpreted as $X \in [x, x + \mathrm{d}x)$ and $T \in [t, t + \mathrm{d}t)$, respectively. Some algebra will establish

the alternative representation of the cross-ratio function:

$$\begin{aligned}
\psi(x,t) &= \frac{P(X > x, T < t)P(X = x, T = t)}{P(X = x, T < t)P(X > x, T = t)} \\
&= \frac{P(X_2 > x, T_2 < t, X_1 = x, T_1 = t)}{P(X_2 = x, T_2 < t, X_1 > x, T_1 = t)} \\
&= \frac{P(X_2 > X_1, T_2 < T_1, \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)}{P(X_1 > X_2, T_2 < T_1, \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)} \\
&= \frac{P(X_2 > X_1, T_2 < T_1, \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)/P(\tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)}{P(X_1 > X_2, T_2 < T_1, \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)/P(\tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)} \\
&= \frac{P(X_2 > X_1, T_2 < T_1 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)}{P(X_1 > X_2, T_2 < T_1 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)},
\end{aligned}$$

where (X_1, T_1) and (X_2, T_2) are independently distributed as in 5.8, $\tilde{X}_{1,2} = \min(X_1, X_2)$ and $\tilde{T}_{1,2} = \max(T_1, T_2)$. Next, note that

$$\begin{aligned}
&P((X_1 - X_2)(T_1 - T_2) < 0 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t) \\
&= 2P(X_2 > X_1, T_2 < T_1 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t),
\end{aligned}$$

since $(X_1 - X_2)(T_1 - T_2) < 0$ if and only if $X_2 > X_1$ and $T_2 < T_1$ or $X_2 < X_1$ and $T_2 > T_1$, and by symmetry these two event have the same probability. Similar we have that

$$\begin{aligned}
&P((X_1 - X_2)(T_1 - T_2) > 0 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t) \\
&= 2P(X_1 > X_2, T_2 < T_1 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t).
\end{aligned}$$

Consequently

$$\begin{aligned}
\psi(x,t) &= \frac{P(X_2 > X_1, T_2 < T_1 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)}{P(X_1 > X_2, T_2 < T_1 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)} \\
&= \frac{2P(X_2 > X_1, T_2 < T_1 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)}{2P(X_1 > X_2, T_2 < T_1 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)} \\
&= \frac{P((X_1 - X_2)(T_1 - T_2) < 0 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)}{P((X_1 - X_2)(T_1 - T_2) > 0 | \tilde{X}_{1,2} = x, \tilde{T}_{1,2} = t)},
\end{aligned}$$

and we have the desired result.

B

Comparison of scatterplots

As a supplement to the goodness of fit test given in section 6.5 on page 79, we made scatterplots of the variables $(X_1, T_1), \ldots, (X_n, T_n)$ simulated from the estimated conditional distributions, where n equal the number of pairs in the original datasets. These scatterplots are displayed in figure B.2, B.3, B.4 and B.4 in the following pages and should be compared with the scatterplot of the original datasets displayed in figure B.1 on the following page.

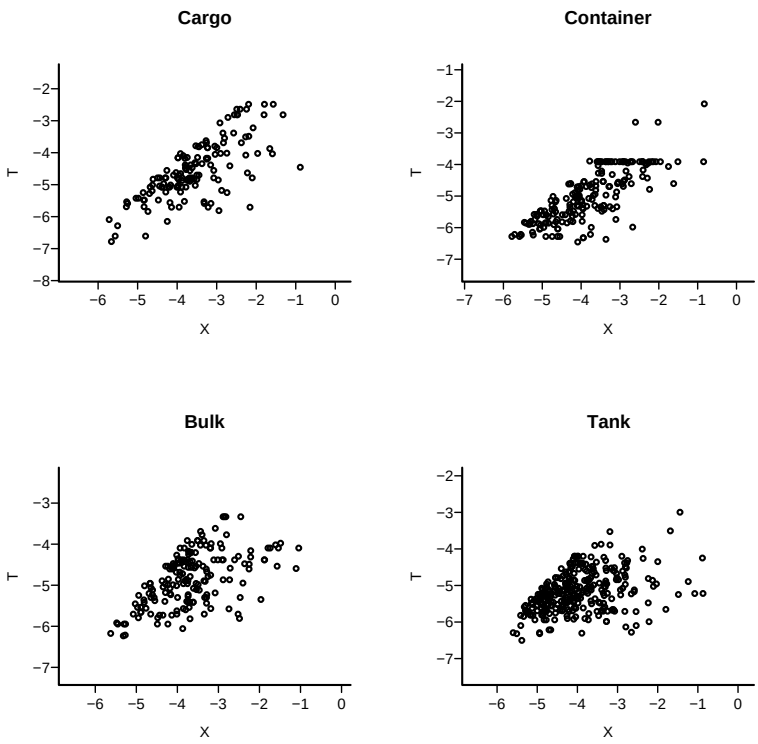

Figure B.1: Scatterplot of the original data

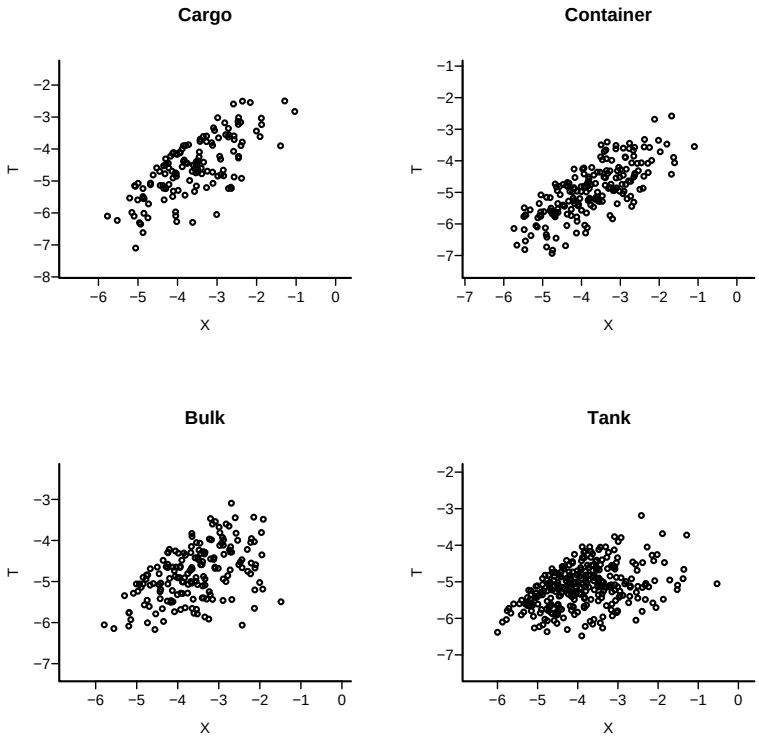

Figure B.2: Scatterplot of randomly drawn vectors from the estimated truncated bivariate normal distribution.

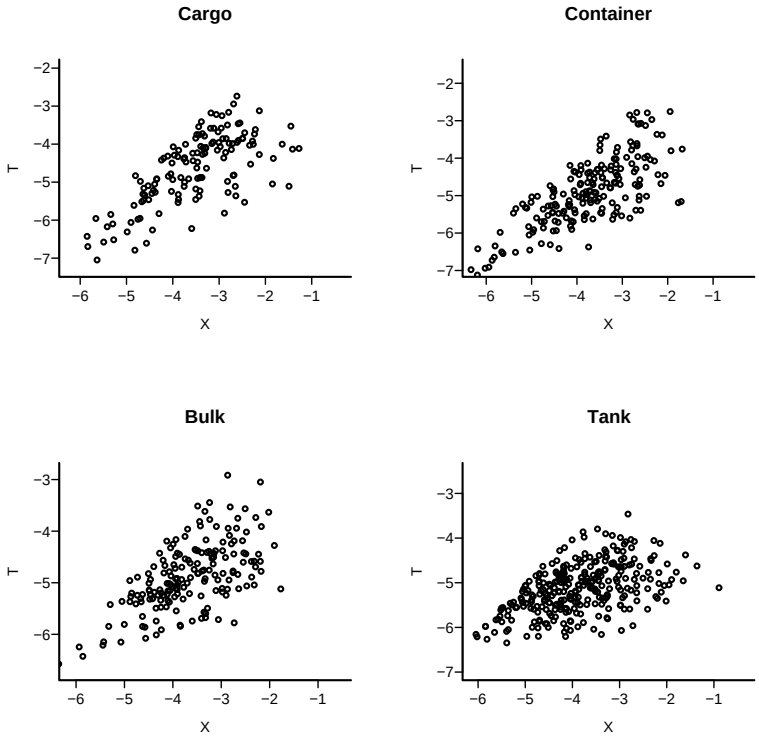

Figure B.3: Scatterplot of randomly drawn vectors from the estimated *Copula1* distribution.

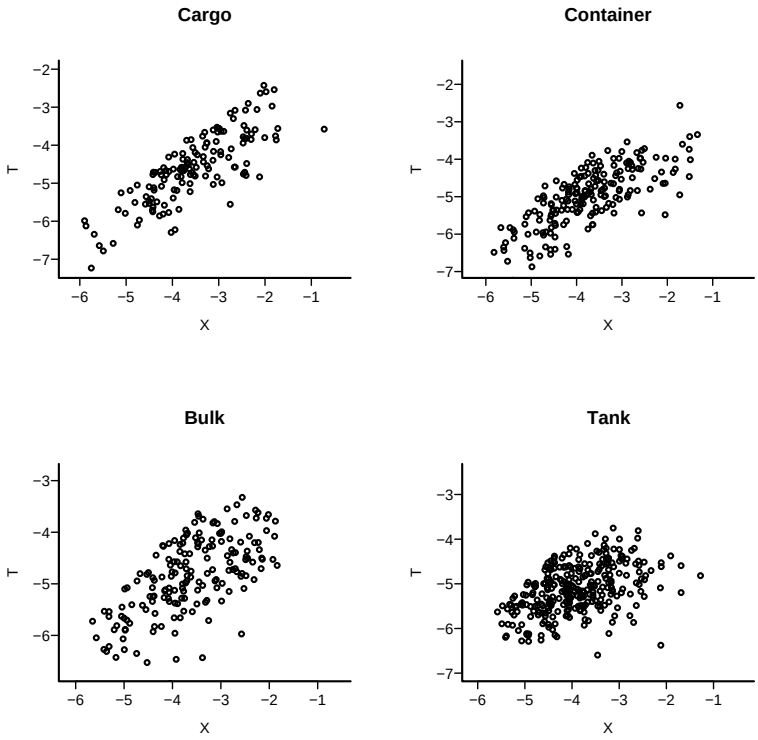

Figure B.4: Scatterplot of randomly drawn vectors from the estimated *Copula2* distribution.

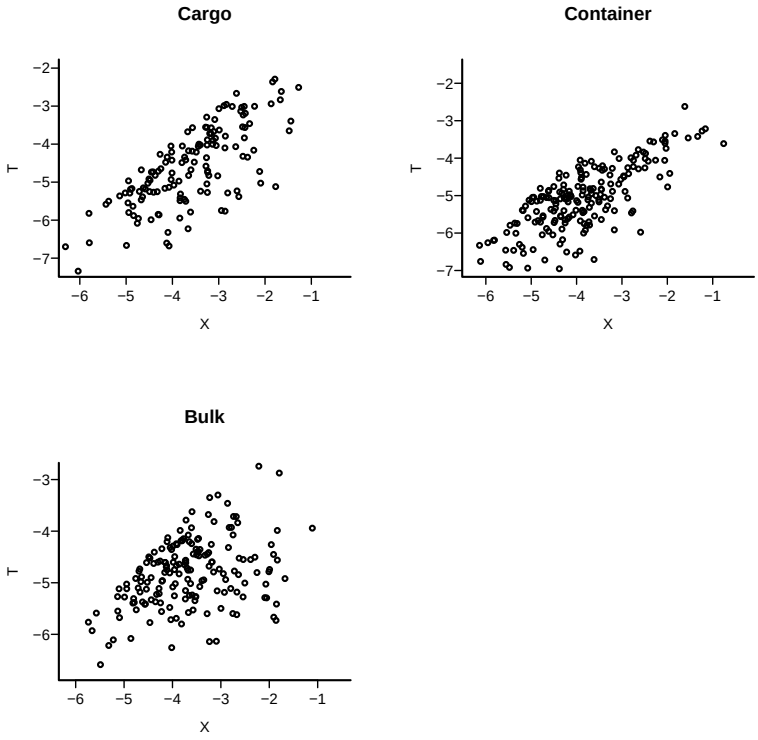

Figure B.5: Scatterplot of randomly drawn vectors from the estimated *Copula3* distribution.

Bibliography

Beaudoin D. and Lakhal-Chaieb L. (2008). 'Archimedean copula model selection under dependent truncation.' *Statistics in medicine*. ISSN 0277-6715. DOI: 10.1002/sim.3316. Cited on page 73.

Casella G. and Berger R.L. (1990). *Statistical inference*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 0-534-11958-1. Cited on page 39.

Chen C.H., Tsai W.Y. and Chao W.H. (1996). 'The product-moment correlation coefficient and linear regression for truncated data'. *Journal of the American Statistical Association*, **volume 91**, no. 435, pages 1181–1186. ISSN 0162-1459. DOI: 10.2307/2291736. Cited on pages 5, 10, 11, 13, 18 and 95.

Kalbfleisch J.D. and Lawless J.F. (1989). 'Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS'. *Journal of the American Statistical Association*, **volume 84**, no. 406, pages 360–372. ISSN 0162-1459. Cited on page 3.

Klein J.P. and Moeschberger M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, second edition. Cited on page 3.

Kowalski J. and Tu X.M. (2008). *Modern applied U-statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons]. ISBN 978-0-471-68227-1. Cited on pages 94 and 95.

Lakhal Chaieb L., Rivest L.P. and Abdous B. (2006). 'Estimating survival under a dependent truncation'. *Biometrika*, **volume 93**, no. 3, pages 655–669. ISSN 0006-3444. DOI: 10.1093/biomet/93.3.655. Cited on pages 54, 55, 56, 62 and 67.

Lynden-Bell D. (1971). 'A method of allowing for known observational selection in small samples applied to 3CR quasars'. *mnras*, **volume 155**, pages 95–+. URL: `http://adsabs.harvard.edu/full/1971MNRAS.155...95L`. Cited on page 27.

McNeil A.J., Frey R. and Embrechts P. (2005). *Quantitative risk management*. Princeton Series in Finance. Princeton University Press. ISBN 0-691-12255-5. Concepts, techniques and tools. Cited on page 47.

Nelsen R.B. (1999). *An introduction to copulas*, volume 139 of *Lecture Notes in Statistics*. Springer-Verlag. ISBN 0-387-98623-5. Cited on pages 48, 49 and 50.

Oakes D. (1989). 'Bivariate survival models induced by frailties'. *Journal of the American Statistical Association*, **volume 84**, no. 406, pages 487–493. ISSN 0162-1459. Cited on pages 59 and 60.

Stute W. and Wang J.L. (2008). 'The central limit theorem under random truncation'. DOI: 10.3150/07-BEJ116. Cited on page 27.

Tsai W.Y. (1990). 'Testing the assumption of independence of truncation time and failure time'. *Biometrika*, **volume 77**, no. 1, pages 169–177. ISSN 0006-3444. URL: `http://biomet.oxfordjournals.org/cgi/content/abstract/77/1/169`. Cited on pages 3, 5, 9, 20, 23 and 31.

van der Vaart A.W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press. ISBN 0-521-49603-9; 0-521-78450-6. Cited on pages 79, 80 and 92.

Wang M.C., Jewell N.P. and Tsai W.Y. (1986). 'Asymptotic properties of the product limit estimate under random truncation'. *The Annals of Statistics*, **volume 14**, no. 4, pages 1597–1605. ISSN 0090-5364. DOI: 10.1214/aos/1176350180. Cited on page 29.

Woodroofe M. (1985). 'Estimating a distribution function with truncated data'. *Ann. Statist.*, **volume 13**, no. 1, pages 163–177. ISSN 0090-5364. DOI: 10.1214/aos/1176346584. Cited on pages 28, 31, 32 and 34.