# ON MODELING INSURANCE CLAIMS USING COPULAS

FILIP ERNTELL

Master's thesis
2013:E55

LUND UNIVERSITY

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

# ON MODELING INSURANCE CLAIMS USING COPULAS

## Abstract

In this master's thesis, a copula approach is used to model the number of claims made by a customer holding three insurances. It is important for insurance companies to have good models for the risk profiles of their customers, and the number of claims is a key element in calculating the expected cost for the company. Using copulas, multivariate distribution functions are allowed to have any desired marginal distributions and many different dependence structures, as these can be chosen separately.

The data used consists of the number of claims made by 74 770 unique customers during one year. Different count data distributions are considered for the one-dimensional marginal distributions, while four Archimedean copulas are tested as models for the dependence structure. To estimate the parameters of the final model, full maximum likelihood is used, for which new implementations adapted to discrete data were created.

$\chi^2$-tests and likelihood ratio tests determined that negative binomial distribution and zero-inflated Delaporte distribution were the best distributions for the one-dimensional marginals, while Cramér-von Mises method and Kendall's Cramér-von Mises method, using a parametric bootstrap, together with Akaike's Information Criterion, suggested Clayton copula to be the most suitable.

The obtained model is compared to the empirical values and to investigate how well the model fits for different years, it is also fitted to the corresponding data from the following year. The model provides a good fit both compared to the empirical values for the year used for inference as well as for the year used for validation. However, the fit is strongly influenced by the values in the lower tail.

*Keywords:* Insurances, Copulas, Count data, Negative binomial distribution, Delaporte distribution, Full maximum likelihood, Goodness of fit.

## Acknowledgments

I would like to thank my supervisor Nader Tajvidi, for helpful opinions, comments and support in the making of this thesis, Fredrik Thuring, for providing the data and for interesting discussions about possible topics to investigate, and my colleagues in the thesis room for good company the time I spent there. Last but not least, I would also like to thank my friends and family for unconditional support and help.

# CONTENTS

Chapter 1

# INTRODUCTION

## 1.1 Background

In the past twenty years, there has been a growing interest in copulas and their applications. In short, copulas are multivariate distribution functions with uniform one-dimensional margins and are used to join arbitrary multivariate distribution functions to their one-dimensional margins. Working with copulas, multivariate distribution functions are allowed to have any desired marginal distributions, as margins and dependence structure are treated separately [Nelsen, 2006]. Due to the vast number of copula families available, it is possible to capture different dependence structures in a model, while for instance the multivariate normal distribution is limited to linear dependence.

In contrast to general products, insurances are somewhat special. The difference lies in the fact that for most other products, the cost for the company is mainly known at the occasion of the sale and determined by for instance manufacturing costs, wages and so on, making it possible to set the price to be higher than the costs. For an insurance however, the income consists of the yearly fee from the customer while the loss depends on the customer's behavior and the cost of each reported claim, making the profit of each insurance contract stochastic. Because of this, it is of great interest for the insurance company both to be able to set suitable fees based on the risk profile of the customer and to keep the customers with low risk profile that give higher profits and even sell complementary products to them.

These issues make it crucial for the insurance companies to have good models for the number of insurance claims a customer will make, as well as the dependence between the number of claims in different products. This can be used to model risk as well as identify high risk customers and selecting which customers to contact for cross-selling attempts [Thuring, 2012].

## 1.2 Purpose and Structure

In this master's thesis, a copula approach is used to model the dependence between the number of insurance claims during one year made by a customer holding three different insurances. The data consists of 74 770 different customers to a Danish insurance company and has previously been modeled using multivariate credibility theory in [Thuring, 2012], while the number of claims made in two of the insurances, during a different year than the one considered in this thesis, have been modeled using copulas in [Hage, 2013], where the copula parameters was estimated using method of moments. In this project, the parameters of the copula and the one-dimensional marginal distributions are estimated using full maximum likelihood, with new routines created to be able to handle discrete marginal distributions. The obtained model is compared to the empirical values and to investigate how well the model fits for different years, it was also fitted to a validation data set consisting of corresponding data from the following year.

In Chapter 2, an overview of the theoretical background to copulas and the other used concepts is given, followed by a description of the procedure and methods used and a presentation of the results in Chapter 3. In Chapter 4, the results are summarized and discussed.

Chapter 2

# THEORY

## 2.1 Copulas

### 2.1.1 Definition

Assume that $X_1, X_2, \dots, X_d$ are one-dimensional stochastic variables and that a multivariate model for these is to be found. The traditional way involves a $d$-dimensional distribution from a certain family, with joint cumulative distribution function (CDF) $F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}\left(X_1 \leq x_1, \dots, X_d \leq x_d\right)$, which determines both the marginal distributions and the dependence between the different variables. One commonly used example is the multivariate normal distribution. This distribution has normal margins and the dependence is determined by the covariance matrix.

To allow for more freedom in multivariate modeling, one can use copulas. Assume that $X_1, \dots, X_d$ have CDFs $F_1(x) = \mathbb{P}\left(X_1 \leq x\right), \dots, F_d(x) = \mathbb{P}\left(X_d \leq x_d\right)$, respectively. Note that these marginal CDFs all are functions $F : \mathbb{R} \to [0,1]$ and the joint CDF is a function $F_{\boldsymbol{X}} : \mathbb{R}^d \to [0,1]$. This means that each real $d$-dimensional vector $\boldsymbol{x} = (x_1, \dots, x_d)$ leads to a point $\big(F_1(x_1), \dots, F_d(x_d)\big)$ in the $d$-dimensional unit hypercube $[0,1]^d$ and that this vector in turn corresponds to a number $F_{\boldsymbol{X}}(\boldsymbol{x})$ in the interval $[0,1]$. The copula $C$ of $\boldsymbol{X}$ is defined as the function which assigns this value to each point. [Nelsen, 2006]

The essentials of the definition are summarized in Sklar's theorem [Nelsen, 2006, Theorem 2.10.9]:

**Theorem 1** (Sklar's theorem)**.** *Let $F_{\boldsymbol{X}}$ be an $d$-dimensional distribution function with margins $F_1, F_2, \dots, F_d$. Then there exists a $d$-copula $C$ such that for all $\boldsymbol{x}$ in $\mathbb{R}^d$,*

$$F_{\boldsymbol{X}}(x_1, x_2, \dots, x_d) = C\big(F_1(x_1), F_2(x_2), \dots, F_d(x_d)\big). \tag{2.1.1}$$

*If $F_1, F_2, \dots, F_d$ are all continuous, then $C$ is unique; otherwise, $C$ is uniquely determined on* $\operatorname{Ran} F_1 \times \operatorname{Ran} F_2 \times \cdots \times \operatorname{Ran} F_d$. *Conversely, if $C$ is a $d$-copula and $F_1, F_2, \dots, F_d$ are distribution functions, then the function $F_{\boldsymbol{X}}$ defined by* (2.1.1) *is a $d$-dimensional distribution function with margins $F_1, F_2, \dots, F_d$.*

*Remark.* Another way to express (2.1.1) is

$$C(\boldsymbol{U}) = C(u_1, u_2, \dots, u_d) = F_{\boldsymbol{X}}\big(F_1^{-1}(u_1), F_2^{-1}(u_2), \dots, F_d^{-1}(u_d)\big), \tag{2.1.2}$$

where $u_1, u_2, \ldots, u_d \in [0,1]$. [Nelsen, 2006, Corollary 2.10.10]

### 2.1.2 Density Function and Probability Mass Function

For a multivariate distribution with continuously differentiable one-dimensional margins $F_1, \ldots, F_d$ and copula $C$, the joint density function $f_{\boldsymbol{X}}$ is equal to the derivative of the joint CDF. Let

$$\frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \cdots \partial u_d} \triangleq c(u_1, \ldots, u_d), \tag{2.1.3}$$

and let $f_1, \ldots, f_d$ be the marginal density functions, then

$$f_{\boldsymbol{X}}(x_1, \ldots, x_d) = \frac{\partial^d C\big(F_1(x_1), \ldots, F_d(x_d)\big)}{\partial x_1 \cdots \partial x_d} = c\big(F_1(x_1), \ldots, F_d(x_d)\big) \prod_{i=1}^{d} f_i(x_i). \tag{2.1.4}$$

However, as discrete margins are not differentiable, a multivariate distribution with discrete margins does not have a density function. Instead, it has a probability mass function (PMF),

$$p_{\boldsymbol{X}}(x_1, \ldots, x_d) = \mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x}\right). \tag{2.1.5}$$

This can be expressed using the CDF as well, and therefore also the copula:

$$\begin{aligned} p_{\boldsymbol{X}}(x_1, \ldots, x_d) &= \sum_{i_1=0,1} \cdots \sum_{i_d=0,1} (-1)^{i_1+\cdots+i_d} \mathbb{P}\left(X_1 \leq x_1 - i_1, \ldots, X_d \leq x_d - i_d\right) = \\ &= \sum_{i_1=0,1} \cdots \sum_{i_d=0,1} (-1)^{i_1+\cdots+i_d} C\big(F_1(x_1 - i_1), \ldots, F_d(x_d - i_d)\big) \end{aligned} \tag{2.1.6}$$

[Panagiotelis et al., 2012]. Note that when implementing this for marginals that can not take negative values, it might be necessary to check that $x_k - i_k \geq 0$ for $k = 1, \ldots, d$ to avoid errors, depending on how the marginal CDFs are implemented.

### 2.1.3 Conditions for Two-Dimensional Copulas

For a function $C : [0,1]^2 \to [0,1]$ to be a copula, the following conditions need to be satisfied:

1. for every $u$ and $v$ in $[0,1]$,

$$C(u, 0) = C(0, v) = 0 \tag{2.1.7}$$

   and

$$C(u, 1) = u \text{ and } C(1, v) = v; \tag{2.1.8}$$

2. for every $u_1$, $u_2$, $v_1$ and $v_2$ in $[0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0. \tag{2.1.9}$$

[Nelsen, 2006]

*Remark.* Note that (2.1.8) implies that all margins of $C$ are uniform on $[0, 1]$. This holds for all one-dimensional margins for a $d$-dimensional copula as well, which follows from (2.1.2), using that if $F_X$ and $F_{X,Y}$ are a one-dimensional and a two-dimensional CDF respectively, then

$$\lim_{y\to\infty} F_{X,Y}(x, y) = F_X(x) \text{ and } \lim_{x\to\infty} F_X(x) = 1. \tag{2.1.10}$$

Further, if $C$ has second order derivatives, (2.1.9) is equivalent to

$$\frac{\partial^2 C}{\partial u \partial v} \geq 0, \tag{2.1.11}$$

which, due to (2.1.4), is equal to that the density function is non-negative.

### 2.1.4 Fréchet-Hoeffding Bounds

From the conditions in Section 2.1.3, the following theorem can be obtained [Nelsen, 2006, Section 2.2].

**Theorem 2** (Fréchet-Hoeffding bounds, 2 dimensions)**.** *For every two-dimensional copula $C$ and every $u$ and $v$ in [0,1],*

$$\max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v). \tag{2.1.12}$$

*Remark.* The bounds are copulas as well, commonly denoted $M(u, v) = \min(u, v)$ and $W(u, v) = \max(u + v - 1, 0)$ and referred to as *Fréchet-Hoeffding upper and lower bound*, respectively.

Theorem 2 can be generalized to $d$ dimensions [Nelsen, 2006, Theorem 2.10.12].

**Theorem 3** (Fréchet-Hoeffding bounds, $d$ dimensions)**.** *If $C$ is any $d$-dimensional copula, then for every $\boldsymbol{u} \in [0, 1]^d$,*

$$\max(u_1 + \cdots + u_d - d + 1, 0) \leq C(\boldsymbol{u}) \leq \min(u_1, \ldots, u_d). \tag{2.1.13}$$

*Remark.* To clarify that the bound is for the $d$-dimensional case, a superscript $d$ can be added. Note that for $d > 2$, the lower bound is not a copula. [Nelsen, 2006, Section 2.10]

Another important copula is the *independent copula*

$$\Pi^d(\boldsymbol{U}) = U_1 U_2 \cdots U_d. \tag{2.1.14}$$

$M^{(2)}$, $W^{(2)}$ and $\Pi^{(2)}$ all have important interpretations. They represent the cases of increasing monotone dependence, decreasing monotone dependence (see Section 2.3.3) and independence, respectively.

## 2.2 Archimedean Copulas

Among the numerous different copula classes, one of the most important is the Archimedean copulas. The popularity of the Archimedean copulas is due to the multitude of nice properties linked to the members of the class and the simple way in which they are constructed, which have given rise to a vast number of families in the class. [Nelsen, 2006]

To be able to define the Archimedean copulas, we first need to define the concept of *pseudo-inverses*. Let $\varphi$ be a continuous, strictly decreasing function $\varphi : [0,1] \to [0,\infty]$ such that $\varphi(1) = 0$. Then the pseudo-inverse $\varphi^{[-1]}$ of $\varphi$ is the function given by [Nelsen, 2006, Definition 4.1.1],

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 \le t \le \varphi(0), \\ 0, & \varphi(0) \le t \le \infty. \end{cases} \tag{2.2.1}$$

Now, a $d$-dimensional Archimedean copula is defined by

$$C(\boldsymbol{U}) = \varphi^{[-1]}\big(\varphi(u_1) + \cdots + \varphi(u_d)\big). \tag{2.2.2}$$

$\varphi$ is called the *generator function* and if $\varphi(0) = \infty$, that is if $\varphi^{[-1]} = \varphi^{-1}$, $\varphi$ is a *strict* generator. For $C$ to be a copula if $d > 2$, $\varphi$ needs to be strict and $\varphi^{[-1]} = \varphi^{-1}$ needs to be *completely monotonic*, that is

$$(-1)^k \frac{d^k}{dt^k} \varphi^{-1}(t) \ge 0 \tag{2.2.3}$$

for all $t \in (0, \infty)$ and $k = 0, 1, 2, \ldots$ [Nelsen, 2006]. If the inverse of a strict generator of an Archimedean copula is completely monotonic, the copula is positively lower orthant dependent (see Section 2.3.5). This implies that all Archimedean copulas of dimension $d > 2$ are positively lower orthant dependent [Nelsen, 2006, Corollary 4.6.3].

In Table 2.1, some properties of four Archimedean copulas are presented.

**Table 2.1.** Four Archimedean copulas, namely [1] Clayton, [2] Frank, [3] Gumbel and [4] Joe copula. [Nelsen, 2006, Table 4.1]

| $C_\theta$ | $\varphi_\theta(t)$ | $\theta \in$ | Strict | Limiting and special cases, $d = 2$ |
|---|---|---|---|---|
| [1] | $\frac{1}{\theta}\left(t^{-\theta} - 1\right)$ | $[-1, \infty) \setminus \{0\}$ | $\theta \ge 0$ | $C_{-1} = W$, $C_0 = \Pi$, $C_\infty = M$ |
| [2] | $-\log\left(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right)$ | $(-\infty, \infty) \setminus \{0\}$ | Yes | $C_{-\infty} = W$, $C_0 = \Pi$, $C_\infty = M$ |
| [3] | $(-\log t)^\theta$ | $[1, \infty)$ | Yes | $C_1 = \Pi$, $C_\infty = M$ |
| [4] | $-\log\big(1 - (1-t)^\theta\big)$ | $[1, \infty)$ | Yes | $C_1 = \Pi$, $C_\infty = M$ |

## 2.3 Dependence Measures

Dependence between stochastic variables can be measured in a number of ways. One of the most frequently used is *Pearson's correlation coefficient*, commonly called *linear correlation*, defined by

$$\rho_P(X,Y) = \frac{\operatorname{cov}[X,Y]}{\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}}. \tag{2.3.1}$$

Two problems with Pearson's correlation are that it only measures linear dependence and that it is not invariant to strictly increasing transformations.

A different way to measure dependence is to use concordance. Let $(X,Y)$ be a stochastic vector and let $(x_i,y_i)$ and $(x_j,y_j)$ denote two observations. Then $(x_i,y_i)$ and $(x_j,y_j)$ are *concordant* if $(x_i-x_j)(y_i-y_j)>0$ and *discordant* if $(x_i-x_j)(y_i-y_j)<0$ [Nelsen, 2006]. Two dependence measures that are based on concordance are presented below.

### 2.3.1 Kendall's Tau

Let $(x_1,y_1),\ldots,(x_n,y_n)$ denote a random sample of $n$ observations from the random vector $(X,Y)$. Consuider all $\binom{n}{2}$ pairs and let $c$ denote the number of concordant pairs and let $d$ denote the number of discordant pairs. Then *Kendall's tau*, $\tau_K$, for the sample is defined as

$$\tau_K = \frac{c-d}{c+d} = \frac{c-d}{\binom{n}{2}}. \tag{2.3.2}$$

This can be interpreted as the probability of concordance minus the probability of discordance. Formally, let $(X_1,Y_1)$ and $(X_2,Y_2)$ be independent and identically distributed stochastic vectors. Then the probabilistic definition of Kendall's tau is

$$\tau_{X,Y} = \mathbb{P}\left((X_1-X_2)(Y_1-Y_2)>0\right) - \mathbb{P}\left((X_1-X_2)(Y_1-Y_2)<0\right). \tag{2.3.3}$$

Assuming that $X$ and $Y$ has the copula $C$, Kendall's tau can also be defined analytically as

$$\tau_C = 4\iint_{[0,1]^2} C(u,v)\,dC(u,v) - 1. \tag{2.3.4}$$

If $X$ and $Y$ are continuous, then $\tau_{X,Y}=\tau_C$. [Nelsen, 2006, Section 5.1.1]

### 2.3.2 Spearman's Rho

Let $(X_1,Y_1)$, $(X_2,Y_2)$ and $(X_3,Y_3)$ be independent and identically distributed stochastic vectors with copula $C$. Then the probabilistic definition of *Spearman's rho*, $\rho_{X,Y}$, is

$$\rho_{X,Y} = 3\big(\mathbb{P}\left((X_1-X_2)(Y_1-Y_3)>0\right) - \mathbb{P}\left((X_1-X_2)(Y_1-Y_3)<0\right)\big). \tag{2.3.5}$$

The analytical definition is

$$\rho_C = 12 \iint_{[0,1]^2} C(u,v)\, dudv - 3 \tag{2.3.6}$$

Similarly to the case of Kendall's tau, the two definitions are equivalent if $X$ and $Y$ are continuous. [Nelsen, 2006, Section 5.1.2]

### 2.3.3 Monotone Dependence

Two continuous stochastic variables $X$ and $Y$ are *monotone dependent* if there exists a monotone function $g$ for which

$$\mathbb{P}\left(g(Y) = X\right) = 1. \tag{2.3.7}$$

If $g$ is increasing, $X$ and $Y$ are said to be *increasing dependent* and if $g$ is decreasing, $X$ and $Y$ are said to be *decreasing dependent*. Further, a necessary and sufficient condition for $X$ and $Y$ to be increasing (decreasing) dependent is that their copula $C$ is equal to the Fréchet-Hoeffding upper (lower) bound. [Kimeldorf and Sampson, 1978]

### 2.3.4 Kendall's Tau and Spearman's Rho in Some Special Cases

Using the concepts in Section 2.3.3 together with Theorem 5.1.8 and 5.1.9 in [Nelsen, 2006], it follows that both Kendall's tau and Spearman's rho for the Fréchet-Hoeffding upper and lower bound are 1 and $-1$, respectively. Using the notation from Section 2.1.4, this can be written as

$$\tau_M = \rho_M = 1 \text{ and } \tau_W = \rho_W = -1. \tag{2.3.8}$$

Further, (2.3.4) and (2.3.6) gives that

$$\tau_\Pi = 4 \iint_{[0,1]^2} C(u,v)\, dC(u,v) - 1 = 4 \iint_{[0,1]^2} uv\, dudv - 1 = 4 \left( \int_0^1 u\, du \right)^2 - 1 = 0 \tag{2.3.9}$$

and

$$\rho_\Pi = 12 \iint_{[0,1]^2} uv\, dudv - 3 = 12 \left( \int_0^1 u\, du \right)^2 - 3 = 0. \tag{2.3.10}$$

### 2.3.5 Multivariate Dependence Measures

The observant readers might have noticed that all dependence measures above are bivariate. However, it is possible to generalize them to the multivariate case as well, though these generalizations will not be covered here. The interested can find details in for instance [Joe, 1990], where generalizations of Kendall's tau and Spearman's rho are described, [Schmid and Schmidt, 2007], which investigates multivariate versions of Spearman's rho and non-parametric estimation of them, or [Mesfioui and Quessy, 2010], where multivariate non-continuous versions are considered.

One multivariate concept will briefly be mentioned here, that of *orthant dependence* [Nelsen, 2006, Definition 5.7.1]. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a $d$-dimensional random vector. $\boldsymbol{X}$ is *positively lower orthant dependent* if for all $\boldsymbol{x} \in \mathbb{R}^d$

$$\mathbb{P}\left(\boldsymbol{X} \leq \boldsymbol{x}\right) \geq \prod_{i=1}^{d} \mathbb{P}\left(X_i \leq x_i\right) \tag{2.3.11}$$

and *positively upper orthant dependent* if for all $\boldsymbol{x} \in \mathbb{R}^d$

$$\mathbb{P}\left(\boldsymbol{X} > \boldsymbol{x}\right) \geq \prod_{i=1}^{d} \mathbb{P}\left(X_i > x_i\right). \tag{2.3.12}$$

If both (2.3.11) and (2.3.12) holds, $\boldsymbol{X}$ is *positively orthant dependent.* The negative counterparts are defined in the same way, but with reversed inequalities. If $d = 2$, this is equivalent with *quadrant dependence.*

## 2.4 Discrete Marginal Distributions

When at least one of the marginal distribution functions $F_1, F_2, \ldots, F_d$ is discrete, there still exists a copula, but, as mentioned in Theorem 1, this is only uniquely determined on $\operatorname{Ran} F_1 \times \operatorname{Ran} F_2 \times \cdots \times \operatorname{Ran} F_d$. This creates some issues that are needed to keep in mind [Genest and Nešlehová, 2007]:

1. the dependence is not characterized by the copula alone;
2. concordance measures depend on the marginal distributions as well as the copula;
3. the probabilistic and the analytical definitions of $\tau_K$ and $\rho_S$ are not equal;
4. monotone dependence does not imply $|\tau_K| = |\rho_S| = 1$.

However, a multivariate distribution still often inherits dependence properties from the copula, and the parameters of the copulas can still be interpreted as dependence parameters [Genest and Nešlehová, 2007]. Thus, the use of copulas defined as above still makes sense, despite the issues. Furthermore, in [Faugeras, 2012], alternate copula definitions that overcome the listed problems are investigated.

## 2.5 Distributions for Count Data

Count data is a term for data which theoretically can take values at $0, 1, 2, \ldots, \infty$. The Poisson distribution is a common choice when modeling count data, for example it was used in [Thuring, 2012] for modeling the number of insurance claims.

However, as its variance is equal to its expectation and as the distribution is a one parameter distribution, the Poisson distribution lacks somewhat in adaptability, which calls for distributions that can be better tuned to data. Two generalizations are the Negative binomial distribution and the Delaporte distribution, which both are Poisson-mixtures. All three distributions can in turn be generalized by zero-inflation. Below, properties and definitions of these distributions and concepts follow.

### 2.5.1 Poisson Distribution (PO)

The Poisson distribution arises as the number of occurred events during a time interval, when the events occurs with constant intensity. It is also the limit of a binomial distribution, when the number of trials tends to infinity as the success probability approaches zero, while the expected value still is finite [Krishnamoorthy, 2006, Chapter 5].

Let $X$ be Poisson distributed with parameter $\lambda > 0$. Then the PMF is

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, 2, \dots \tag{2.5.1}$$

The expected value is

$$\mathbb{E}\left[X\right] = \lambda \tag{2.5.2}$$

and the variance is

$$\mathrm{Var}\left[X\right] = \lambda. \tag{2.5.3}$$

[Krishnamoorthy, 2006, Chapter 5]

### 2.5.2 Negative Binomial Distribution (NB)

Consider a number of Bernoulli trials with success probability $p$ and let $X$ be the number of failures until the $r$:th success. Then $X$ has a negative binomial distribution, $X \sim \mathrm{NB}^{\mathrm{I}}(r, p)$. The distribution can be extended to $r \in \mathbb{R}^+$. Also, if $X|\Theta = \theta \sim Po(\theta)$ with $\Theta \sim \Gamma(\alpha, \beta)$, then $X \sim \mathrm{NB}^{\mathrm{I}}(\alpha, \frac{\beta}{1+\beta})$. This is proved in Section A.1 in the appendices.

The PMF of $X \sim \mathrm{NB}^{\mathrm{I}}(r, p)$ is

$$p_X(k) = \binom{k+r-1}{r-1} p^r(1-p)^k = \frac{\Gamma(k+r)}{k!\Gamma(r)} p^r(1-p)^k, \tag{2.5.4}$$

for $k = 0, 1, 2, \dots,$ where $r > 0$ and $0 < p < 1$. Note that the first expression does not hold for the extension to $r \in \mathbb{R}^+$, then the second expression is used. The expectation and variance are

$$\mathbb{E}\left[X\right] = \frac{r(1-p)}{p} \tag{2.5.5}$$

and

$$\mathrm{Var}\left[X\right] = \frac{r(1-p)}{p^2} \tag{2.5.6}$$

[Krishnamoorthy, 2006, Chapter 7]

An alternative parametrization sometimes used is

$$\begin{cases} \mu = \mathbb{E}\left[X\right] = \frac{r(1-p)}{p} \\ \sigma = \frac{1}{r} \end{cases} \Leftrightarrow \begin{cases} p = \frac{1}{1+\sigma\mu} \\ r = \frac{1}{\sigma} \end{cases} \tag{2.5.7}$$

We denote this parametrization with $X \sim \mathrm{NB}^{\mathrm{II}}(\mu, \sigma) = \mathrm{NB}^{\mathrm{I}}\left(\frac{1}{\sigma}, \frac{1}{1+\sigma\mu}\right)$. The PMF now is

$$p_X(k) = \frac{\Gamma(k+1/\sigma)}{k!\Gamma(1/\sigma)} \left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^k \tag{2.5.8}$$

for $k = 0, 1, 2, \ldots$, with $\mu > 0$ and $\sigma > 0$ and the expectation and variance are

$$\mathbb{E}\left[X\right] = \mu \tag{2.5.9}$$

and

$$\mathrm{Var}\left[X\right] = \mu + \mu^2\sigma. \tag{2.5.10}$$

[Stasinopoulos et al., 2008, Section A.10]

### 2.5.3 Delaporte Distribution (DEL)

If $X|\Theta = \theta \sim Po(\theta)$ and $\Theta = \lambda + \gamma$ where $\lambda \in \mathbb{R}^+$ and $\gamma \sim \Gamma(\alpha, \beta)$, then $X$ has a Delaporte distribution, $X \sim \mathrm{Del}^{\mathrm{I}}(\lambda, \alpha, \beta)$, with

$$p_X(k) = \begin{cases} \left(\frac{\beta}{1+\beta}\right)^{\alpha} e^{-\lambda}, & \text{if } k = 0, \\ \sum_{i=0}^{k} \frac{\Gamma(\alpha+i)\beta^{\alpha}e^{\lambda}\lambda^{k-i}}{\Gamma(\alpha)i!(k-i)!(1+\beta)^{\alpha+i}}, & \text{if } k = 1, 2, \ldots, \end{cases} \tag{2.5.11}$$

for $\lambda > 0$, $\alpha > 0$ and $\beta > 0$ [Vose, 2008, with parameter $\beta^* = 1/\beta$]. For details of the derivation of the PMF, see Section A.2 in the appendices. The expectation of $X$ is

$$\mathbb{E}\left[X\right] = \lambda + \frac{\alpha}{\beta} \tag{2.5.12}$$

and the variance

$$\mathrm{Var}\left[X\right] = \lambda + \frac{\alpha}{\beta}\left(\frac{1}{\beta} + 1\right) \tag{2.5.13}$$

[Vose, 2008]. This distribution has also got an alternative parametrization, denoted $X \sim \mathrm{Del}^{\mathrm{II}}(\nu, \mu, \sigma) = \mathrm{Del}^{\mathrm{I}}(\mu\nu, \frac{1}{\sigma}, \frac{1}{\mu\sigma(1-\nu)})$. See Section A.2 for detailed derivation. The expectation and variance using this parametrization are

$$\mathbb{E}\left[X\right] = \mu \tag{2.5.14}$$

and

$$\mathrm{Var}\left[X\right] = \mu + \mu^2\sigma(1-\nu)^2 \tag{2.5.15}$$

[Stasinopoulos et al., 2008, Section A.10]

### 2.5.4 Zero-Inflated Distributions

Sometimes, the probability for getting a zero needs to be increased for a distribution to fit data. The idea with *zero-inflation* is to let $X$ be the product between a random variable $Y$ and an independent Bernoulli variable $I$ with success probability $1-\varphi$. By choosing $\varphi$ wisely, the right amount of zeros can be obtained. For a zero-inflated distribution, the PMF becomes

$$p_X(k) = \begin{cases} \varphi + (1-\varphi)p_X(0), & \text{if } k = 0, \\ (1-\varphi)p_X(k), & \text{if } k = 1, 2, \dots, \end{cases} \tag{2.5.16}$$

[Johnson et al., 2005, Section 8.2.3]. As $\mathbb{E}[AB] = \mathbb{E}[A]\,\mathbb{E}[B]$ and $\mathrm{Var}[AB] = \mathrm{Var}[A]\,\mathbb{E}^2[B] + \mathbb{E}\left[A^2\right]\mathrm{Var}[B]$ if $A$ and $B$ are independent, the expectation becomes

$$\mathbb{E}[X] = \mathbb{E}[I]\,\mathbb{E}[Y] = (1-\varphi)\mathbb{E}[Y] \tag{2.5.17}$$

and the variance is

$$\mathrm{Var}[X] = (1-\varphi)\mathrm{Var}[Y] - \varphi(1-\varphi)\mathbb{E}^2[Y] = (1-\varphi)(\mathrm{Var}[Y] - \varphi\mathbb{E}^2[Y]). \tag{2.5.18}$$

#### Zero-Inflated Poisson Distribution (ZIP)

The PMF for a zero-inflated Poisson distribution is

$$p_X(k) = \begin{cases} \varphi + (1-\varphi)e^{-\lambda}, & \text{if } k = 0, \\ (1-\varphi)e^{-\lambda}\frac{\lambda^k}{k!}, & \text{if } k = 1, 2, \dots, \end{cases} \tag{2.5.19}$$

for $\lambda > 0$ and $0 \leq \varphi \leq 1$ [Johnson et al., 2005, Section 8.2.4]. Using (2.5.17) and (2.5.18), the expectation and variance becomes

$$\mathbb{E}[X] = (1-\varphi)\lambda \tag{2.5.20}$$

and

$$\mathrm{Var}[X] = (1-\varphi)(\lambda + \varphi\lambda^2) \tag{2.5.21}$$

respectively.

#### Zero-Inflated Negative Binomial Distribution (ZINB)

The PMF for a zero-inflated negative binomial distribution using parametrization $\mathrm{NB}^{\mathrm{I}}$ is

$$p_X(k) = \begin{cases} \varphi + (1-\varphi)p^r, & \text{if } k = 0, \\ (1-\varphi)\frac{\Gamma(k+r)}{k!\Gamma(r)}p^r(1-p)^k, & \text{if } k = 1, 2, \dots, \end{cases} \tag{2.5.22}$$

for $r > 0$, $0 \leq p \leq 1$ and $0 \leq \varphi \leq 1$ and the expectation and variance are

$$\mathbb{E}[X] = (1-\varphi)\frac{r(1-p)}{p} \tag{2.5.23}$$

and

$$\mathrm{Var}\left[X\right] = (1-\varphi)\frac{r(1-p)}{p^2}\big(1+\varphi r(1-p)\big) \tag{2.5.24}$$

respectively. Using the parametrization $\mathrm{NB}^{\mathrm{II}}$, we get

$$p_X(k) = \begin{cases} \varphi + (1-\varphi)\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}, & \text{if } k = 0, \\ (1-\varphi)\frac{\Gamma(k+1/\sigma)}{k!\Gamma(1/\sigma)}\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^{k}, & \text{if } k = 1,2,\ldots, \end{cases} \tag{2.5.25}$$

for $\mu > 0$, $\sigma > 0$ and $0 \leq \varphi \leq 1$. The expectation and variance then are

$$\mathbb{E}\left[X\right] = (1-\varphi)\mu \tag{2.5.26}$$

and

$$\mathrm{Var}\left[X\right] = \mu(1-\varphi)\big(1+\mu(\varphi+\sigma)\big). \tag{2.5.27}$$

### Zero-Inflated Delaporte Distribution (ZIDEL)

The PMF for a zero-inflated Delaporte distribution using parametrization $\mathrm{Del}^{\mathrm{I}}$ is

$$p_X(k) = \begin{cases} \varphi + (1-\varphi)\left(\frac{\beta}{1+\beta}\right)^{\alpha} e^{-\lambda}, & \text{if } k = 0, \\ (1-\varphi)\sum_{i=0}^{k}\frac{\Gamma(\alpha+i)\beta^{\alpha}e^{\lambda}\lambda^{k-i}}{\Gamma(\alpha)i!(k-i)!(1+\beta)^{\alpha+i}}, & \text{if } k = 1,2,\ldots, \end{cases} \tag{2.5.28}$$

for $\lambda > 0$, $\alpha > 0$, $\beta > 0$ and $0 \leq \varphi \leq 1$. The expectation and variance are

$$\mathbb{E}\left[X\right] = (1-\varphi)\left(\lambda + \frac{\alpha}{\beta}\right) \tag{2.5.29}$$

and

$$\mathrm{Var}\left[X\right] = (1-\varphi)\left(\lambda + \frac{\alpha}{\beta}\left(\frac{1}{\beta}+1\right) + \varphi\left(\lambda+\frac{\alpha}{\beta}\right)^2\right). \tag{2.5.30}$$

Using the parametrization $\mathrm{Del}^{\mathrm{II}}$, expectation and variance are

$$\mathbb{E}\left[X\right] = (1-\varphi)\mu \tag{2.5.31}$$

and

$$\mathrm{Var}\left[X\right] = \mu(1-\varphi)\big(1+\mu(\varphi+\sigma(1-\nu)^2)\big). \tag{2.5.32}$$

### Nested Models

Note that negative binomial distribution, zero-inflated negative binomial distribution, Delaporte distribution and zero-inflated Delaporte distribution are related and can be seen as nested models. This means that two models can be tested using simple hypotheses. In Figure 2.1, the relations are illustrated.

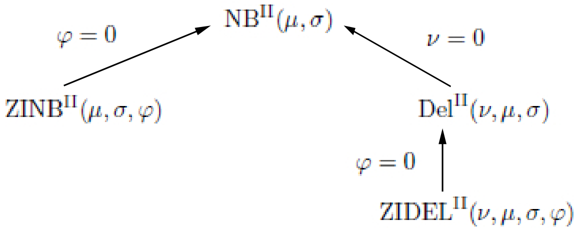

$\mathrm{NB}^{\mathrm{II}}(\mu,\sigma)$
$\varphi = 0$
$\nu = 0$
$\mathrm{ZINB}^{\mathrm{II}}(\mu,\sigma,\varphi)$
$\mathrm{Del}^{\mathrm{II}}(\nu,\mu,\sigma)$
$\varphi = 0$
$\mathrm{ZIDEL}^{\mathrm{II}}(\nu,\mu,\sigma,\varphi)$

**Figure 2.1.** Relationship between the distributions.

## 2.6 Parameter Estimation

When making a multivariate model using copulas, one has two possible choices when estimating the parameters. Either, the parameters of the one-dimensional marginal distributions and the copula are estimated separately or they are treated together all at once. When maximum likelihood is used to estimate both copula parameters and marginal parameters, the first method is called *inference for margins* [Joe and Xu, 1996]. To use maximum likelihood to estimate all parameters at once is called *full maximum likelihood*. A way to estimate the copula parameters without maximum likelihood is described next, followed by a short description of the full maximum likelihood method.

### 2.6.1 Method of Moments

The *method of moments* is based on the fact that the different moments of a stochastic variable often depend on the parameters of its distribution. If the inverse to these relations is available and the moments can be estimated, this gives an estimate of the parameters.

If the analytical expressions for Kendall's tau and Spearman's rho are known for a bivariate distribution, these can be used to estimate the copula parameters. Let for example $(X, Y)$ be a stochastic vector with copula $C$ with parameter $\theta$ and let $\tau_K(\theta) = \tau_C$ and $\rho_S(\theta) = \rho_C$ be Kendall's tau and Spearman's rho, respectively. If we now have estimates $\hat{\tau}_C$ and $\hat{\rho}_C$, we can estimate $\theta$ as

$$\hat{\theta} = \tau_K^{-1}(\hat{\tau}_C) \tag{2.6.1}$$

or

$$\hat{\theta} = \rho_S^{-1}(\hat{\rho}_C). \tag{2.6.2}$$

This is a quite common method, and is for instance used in [Hage, 2013]. However, for copulas with discrete one-dimensional marginals, the estimate based on Kendall's tau might be biased [Genest and Nešlehová, 2007, Section 6.1] and for copulas with dimension $d > 2$, a definition of Kendall's tau and Spearman's rho which depends on the copula parameters is needed.

### 2.6.2 Full Maximum Likelihood

In the full maximum likelihood method, both marginal parameters and copula parameters are estimated at the same time using maximum likelihood.

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a $d$-dimensional stochastic vector with continuous margins, use the notations in Section 2.1.2 and let $F_1, \ldots, F_d$ have parameter vectors $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_d$, respectively, and $C$ parameter vector $\boldsymbol{\theta}$. Now, define the parameter vector $\boldsymbol{\eta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_d, \boldsymbol{\theta})$. Assume $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ are $n$ observation vectors of $\boldsymbol{X}$ independent of each other. The likelihood function is then defined as

$$\begin{aligned}\mathcal{L}\Big(\boldsymbol{\eta}; \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\Big) &\overset{\triangle}{=} \prod_{i=1}^{n} f_{\boldsymbol{X}}(\boldsymbol{x}^{(i)}; \boldsymbol{\eta}) = \\ &= \prod_{i=1}^{n} c\big(F_1(x_1^{(i)}; \boldsymbol{\vartheta}_1), \ldots, F_d(x_d^{(i)}; \boldsymbol{\vartheta}_d); \boldsymbol{\theta}\big) \prod_{j=1}^{d} f_j\big(x_j^{(i)}; \boldsymbol{\vartheta}_j\big),\end{aligned} \tag{2.6.3}$$

using (2.1.4). The log likelihood function is

$$\begin{aligned}\ell\Big(\boldsymbol{\eta}; \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\Big) &\overset{\triangle}{=} \log\Big(\mathcal{L}\Big(\boldsymbol{\eta}; \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\Big)\Big) = \\ &= \sum_{i=1}^{n} \log\Big(c\big(F_1(x_1^{(i)}; \boldsymbol{\vartheta}_1), \ldots, F_d(x_d^{(i)}; \boldsymbol{\vartheta}_d); \boldsymbol{\theta}\big)\Big) + \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{d} \log\Big(f_j\big(x_j^{(i)}; \boldsymbol{\vartheta}_j\big)\Big),\end{aligned} \tag{2.6.4}$$

and the *full maximum likelihood estimate* is defined as

$$\hat{\boldsymbol{\eta}} \overset{\triangle}{=} \arg\max_{\boldsymbol{\eta}} \mathcal{L}\Big(\boldsymbol{\eta}; \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\Big) = \arg\max_{\boldsymbol{\eta}} \ell\Big(\boldsymbol{\eta}; \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\Big). \tag{2.6.5}$$

#### Discrete Case

With the same assumptions as above, but with discrete one-dimensional marginals, the likelihood function is defined, using (2.1.6), as

$$\begin{aligned}&\mathcal{L}\Big(\boldsymbol{\eta}; \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\Big) \overset{\triangle}{=} \prod_{i=1}^{n} p_{\boldsymbol{X}}(\boldsymbol{x}^{(i)}; \boldsymbol{\eta}) = \\ &= \prod_{i=1}^{n} \sum_{j_1=0,1} \cdots \sum_{j_d=0,1} (-1)^{j_1+\cdots+j_d} C\big(F_1(x_1^{(i)} - j_1; \boldsymbol{\vartheta}_1), \ldots, F_d(x_d^{(i)} - j_d; \boldsymbol{\vartheta}_d); \boldsymbol{\theta}\big).\end{aligned} \tag{2.6.6}$$

The full maximum likelihood estimate is still defined as in (2.6.5).

## 2.7 Model Selection and Goodness of Fit

When trying different models, it is important to be able to evaluate how well the model fits the data and which model is the best. It is not only important how well the model fits, but it is also desirable to have a model that is as simple as possible with as few parameters as possible. This is due to the fact that a model can come arbitrarily close to the observed data if the number of parameters is sufficiently high and that the more parameters, the harder it is to estimate them. Using different goodness of fit tests, measures for how well the models meet these desired properties are obtained.

### 2.7.1 Marginal Distributions

#### $\chi^2$-test

Let $X$ be a discrete stochastic variable with support $\{\xi_1, \dots, \xi_m\}$ and let $x_1, \dots, x_n$ be $n$ observations of $X$. Consider the hypothesis that the sample is from a particular discrete distribution with PMF $p_X(k; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $d$-dimensional vector of unknown parameters and let $\hat{\boldsymbol{\theta}}$ be an estimate of $\boldsymbol{\theta}$ based on the sample. The $\chi^2$-test is a test of this hypothesis, where the probabilities from the supposed distribution is compared to the observations. If the squared difference is too large, the hypothesis is rejected. Below, a way to perform this test follows.

1. Find the observed frequencies, that is the number $O_j$ of data points equal to $\xi_j$, $j = 1, 2, \dots, m$.

2. Compute the probabilities $p_j = p_X(\xi_j; \hat{\boldsymbol{\theta}})$ for $j = 1, 2, \dots, m-1$ and $p_m = 1 - \sum_{j=1}^{m-1} p_j$.

3. Compute the expected frequencies $E_j = p_j n$, $j = 1, 2, \dots, m$.

4. Evaluate the $\chi^2$-statistic

$$\chi^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \tag{2.7.1}$$

5. Compare $\chi^2$ with the $(1-\alpha)$th quantile $q_{1-\alpha}$ of a $\chi^2(m-d-1)$-distribution.

6. If $\chi^2 > q_{1-\alpha}$, the hypothesis is rejected.

[Krishnamoorthy, 2006, Chapter 1]

#### Likelihood Ratio Test

Unlike the $\chi^2$-test, the likelihood ratio test does not give information of the general fit of a model. Instead it is a way to select the most appropriate model when comparing a restricted model to an unrestricted counterpart. For instance, nested models can be tested. Consider the hypotheses

$$\begin{cases} H_0 : \ \theta \in \Omega_0 \\ H_1 : \ \theta \in \Omega \setminus \Omega_0, \end{cases}$$

where the dimension of $\Omega_0$ is $r$ and the dimension of $\Omega$ is $m$. Then the likelihood ratio is defined as

$$\lambda = \frac{\sup_{\theta \in \Omega_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Omega} \mathcal{L}(\theta)} \tag{2.7.2}$$

and it can be shown that $-2\log(\lambda) \to \chi^2(m-r)$ under $H_0$. [Madsen, 2008, Section 6.5]

### Akaike's Information Criterion

Just like the likelihood ratio test, Akaike's Information Criterion (AIC) does not give information of the general fit of a model, but is a way to compare different models to each other. It is based on the log likelihood value $\ell(\theta)$ and the number of parameters, where a larger number of parameters is penalized. It is defined as

$$\text{AIC} = 2k - 2\ell(\theta), \tag{2.7.3}$$

where $k$ is the number of estimated parameters in the model and $\ell(\theta)$ is defined as in (2.6.4) or in the corresponding way in the discrete case. The best model according to AIC is the model with the smallest AIC value. [Akaike, 1974] Note that AIC is not restricted to be used for one-dimensional models.

## 2.7.2 Copula

Assume that we have data $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, and that we want to test if the dependence structure is well represented by some copula family $\mathcal{C}_0$. Particularly, we want to test the null hypothesis $H_0 : C \in \mathcal{C}_0$.

### Cramér-von Mises Method

Cramér-von Mises method is based on the empirical copula

$$C_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(F_1\big(x_1^{(i)}\big) \le u_1, \ldots, F_d\big(x_d^{(i)}\big) \le u_d\big), \ \boldsymbol{u} = (u_1, \ldots, u_d) \in [0,1]^d. \tag{2.7.4}$$

The test statistic is

$$S_n \triangleq \int_{[0,1]^d} n\big(C_n(\boldsymbol{u}) - C_{\hat{\theta}}(\boldsymbol{u})\big)^2 \, dC_n(\boldsymbol{u}) \tag{2.7.5}$$

where $C_{\hat{\theta}} \in \mathcal{C}_0$ is the estimated copula [Genest et al., 2009]. Now, assume that we have data $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ and that this has some multivariate distribution, where we know the one-dimensional marginals $F_1, \dots, F_d$. Assume further that we want to use the Cramér-von Mises method to test if $C_{\hat{\theta}}$ is a reasonable description of the associated copula. This can be done as follows.

First, the test statistic defined in (2.7.5) can for numerical purposes be calculated using the Riemann sum approximation as

$$S_n = \sum_{i=1}^{n} \big(C_n(\boldsymbol{u}_i) - C_{\hat{\theta}}(\boldsymbol{u}_i)\big)^2 \tag{2.7.6}$$

[Genest et al., 2009]. A parametric bootstrap for the $p$-value is performed using the following algorithm from [Genest et al., 2009, Appendix A].

1. Compute the *pseudo-observations*

$$\boldsymbol{u}^{(i)} = \big(u_1^{(i)}, \dots, u_d^{(i)}\big) = \big(F_1(x_1^{(i)}), \dots, F_d(x_d^{(i)})\big) \tag{2.7.7}$$

   for $i = 1, \dots, n$. If the marginals are not known, one can use $u_j^{(i)} = \frac{R_{ij}}{n+1} = \frac{n\hat{F}_j\left(x_j^{(i)}\right)}{n+1}$, where $R_{1j}, \dots, R_{nj}$ are the ranks of the $j$th elements in each observation vector.

2. Compute the empirical copula $C_n(\boldsymbol{u})$ according to (2.7.4) and estimate $\theta$ with some estimator $\hat{\theta} = \psi\big(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}\big)$.

3. If there is an analytical expression for $C_\theta$, compute $S_n$ according to (2.7.6), otherwise, do a Monte Carlo approximation.

4. For some large integer $N$, repeat the following steps for every $k = 1, \dots, N$:

   (a) Generate a random sample $\boldsymbol{y}^{(1,k)}, \dots, \boldsymbol{y}^{(n,k)}$ from the multivariate distribution, now with copula $C_{\hat{\theta}}$, and compute their pseudo-observations $\boldsymbol{u}^{*(i,k)} = (u_1^{*(i,k)}, \dots, u_d^{*(i,k)}) = \big(F_1(y_1^{(i,k)}), \dots, F_d(y_d^{(i,k)})\big)$ for $i = 1, \dots, n$.

   (b) Compute the empirical copula

$$C^*_{n,k}(\boldsymbol{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(u_1^{*(i,k)} \leq u_1, \dots, u_d^{*(i,k)} \leq u_d) \tag{2.7.8}$$

   and calculate the estimate $\hat{\theta}_k = \psi\big(\boldsymbol{y}^{(1,k)}, \dots, \boldsymbol{y}^{(n,k)}\big)$.

   (c) If there is an analytical expression for $C_\theta$, compute

$$S^*_{n,k} = \sum_{i=1}^{n} \big(C^*_{n,k}(\boldsymbol{u}^{*(i,k)}) - C_{\hat{\theta}_k}(\boldsymbol{u}^{*(i,k)})\big)^2, \tag{2.7.9}$$

   otherwise, do a Monte Carlo approximation.

An approximative $p$-value for the test is given by $\frac{1}{N} \sum_{k=1}^{N} \mathbb{1}(S^*_{n,k} > S_n)$.

### Kendall's Cramér-von Mises Method

Kendall's Cramér-von Mises method is based on the transformation $V = F_{\boldsymbol{X}}(\boldsymbol{X}) = C(F_1(X_1), \dots, F_d(X_d))$, where $C$ is copula associated with $\boldsymbol{X}$. This transformation is called Kendall's transform. The test statistic is

$$S_n^K \triangleq \int_{[0,1]} n\big(K_n(v) - K_{\hat{\theta}}(v)\big)^2 \, dK_{\hat{\theta}}(v), \tag{2.7.10}$$

where

$$K_{\hat{\theta}}(t) \triangleq \int_{[0,1]^d} \mathbb{1}_{C_\theta(\boldsymbol{u}) \le t} \, dC_\theta(\boldsymbol{u}) \tag{2.7.11}$$

and

$$K_n(\nu) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(v_i \le \nu), \; \nu \in [0,1] \tag{2.7.12}$$

[Genest et al., 2009].

Now, Assume that we have data $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$ and that this has some multivariate distribution, where we know the one-dimensional marginals $F_1, \dots, F_d$. Assume further that we want to use Kendall's Cramér-von Mises method to test if $C_{\hat{\theta}}$ is a reasonable description of the associated copula. A parametric bootstrap for the $p$-value is performed using the following algorithm from [Genest et al., 2009, Appendix B].

1. Compute the pseudo-observations as in (2.7.7) as well as the *rescaled pseudo-observations* using Kendall's transform, $v_1 = C_n\big(\boldsymbol{u}^{(1)}\big), \dots, v_n = C_n\big(\boldsymbol{u}^{(n)}\big)$.

2. Compute $K_n$ as in (2.7.12) and estimate the parameters $\theta$ with some estimator $\hat{\theta} = \psi\big(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}\big)$.

3. If there is an analytical expression for $K_\theta$, compute $S_n^{(K)}$ according to (2.7.10). Otherwise, proceed by Monte Carlo approximation by choosing $m \ge n$ and doing the following extra steps:

   (a) Generate a random sample $\boldsymbol{u}^{*(1)}, \dots, \boldsymbol{u}^{*(m)}$ from the distribution $C_{\hat{\theta}}$.

   (b) Approximate $K_{\hat{\theta}}$ by

   $$B_m^*(t) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(v_i^* \le t), \; t \in [0,1] \tag{2.7.13}$$

   where

   $$v_i^* = C_m\big(\boldsymbol{u}^{*(i)}\big) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}(\boldsymbol{u}^{*(j)} \le \boldsymbol{u}^{*(i)}), \; i \in [1, \dots, m] \tag{2.7.14}$$

   and note that $mB_m^*(v_i^*)$ is the rank of $v_i^*$ among $v_1^*, \dots, v_m^*$.

(c) Approximate $S_n^{(K)}$ by

$$S_n^{(K)} = \frac{n}{m} \sum_{i=1}^{m} \big(K_n(v_i^*) - B_m^*(v_i^*)\big)^2. \tag{2.7.15}$$

4. For some large integer $N$, repeat the following steps for every $k = 1, \dots, N$:

(a) Generate a random sample $\boldsymbol{y}^{(1,k)}, \dots, \boldsymbol{y}^{(n,k)}$ from the multivariate distribution, now with copula $C_{\hat{\theta}}$. Compute the pseudo-observations $\boldsymbol{u}^{*(i,k)} = (u_1^{*(i,k)}, \dots, u_d^{*(i,k)}) = \big(F_1(y_1^{(i,k)}), \dots, F_d(y_d^{(i,k)})\big)$ and the rescaled pseudo-observations $v_{i,k}^* = C_n\big(\boldsymbol{u}^{*(i,k)}\big)$ for $i \in [1, \dots, n]$.

(b) Compute

$$K_{n,k}^*(\nu) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(v_{i,k}^* \leq \nu),\ \nu \in [0,1] \tag{2.7.16}$$

and calculate the estimate $\hat{\theta}_k = \psi\big(\boldsymbol{y}^{(1,k)}, \dots, \boldsymbol{y}^{(n,k)}\big)$.

(c) If there is an analytical expression for $K_\theta$, let

$$S_{n,k}^{(K)*} = \int_0^1 \big(K_{n,k}^*(t) - K_{\hat{\theta}_k}(t)\big)^2 \, dK_{\hat{\theta}_k}(t). \tag{2.7.17}$$

Otherwise, proceed by Monte Carlo approximation as above.

An approximative $p$-value for the test is given by $\frac{1}{N} \sum_{k=1}^{N} \mathbb{1}(S_{n,k}^{(K)*} > S_n^{(K)})$.

Chapter 3

# PROCEDURE AND RESULTS

The implementations used for this chapter were done in R. For most of the distributions, slightly modified versions of distributions found in *gamlss*-package [Rigby and Stasinopoulos, 2005] were used and *fitdistrplus* [Delignette-Muller et al., 2013] was used for the maximum likelihood estimation of the marginal parameters. The copulas were implemented using *copula*-package [Yan, 2007], though a full maximum algorithm suitable for discrete data, as well as the methods for copula goodness of fit, are not included in this package and were therefore created.

## 3.1 Data

The data comes from a Danish insurance company and consists of the number of insurance claims in three different insurance products; building, car and content insurance. All in all there are 95 668 unique customers and the different customers have data collected from different number of years, ranging from one to five. In each year each customer has all three types of insurance and in total, there are 306 196 such observed vectors.

The customers with data from only one year have been removed from the set, and for the remaining 74 770 customers, the second last year is used for inference and the last year is saved for validation later on.

**Table 3.1.** Mean and variance for the marginal distributions

| Marginal | Mean | Variance |
|---|---|---|
| Building | 0.1301 | 0.1481 |
| Car | 0.2153 | 0.3166 |
| Content | 0.1253 | 0.1464 |

The mean and variance for each product are presented in Table 3.1. In Figure 3.1 to 3.6, histograms for both the data set used for inference and the validation data set are shown. Notice the high number of zeros, amounting more than 80% of the data. This suggests that zero-inflated models might be of intereset. It is also worth

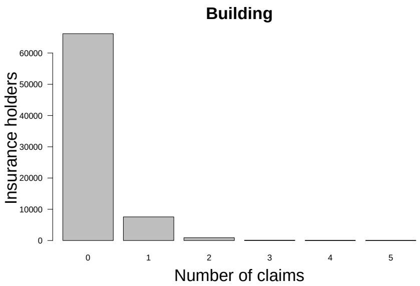


**Figure 3.1.** Histogram for building insurance data.

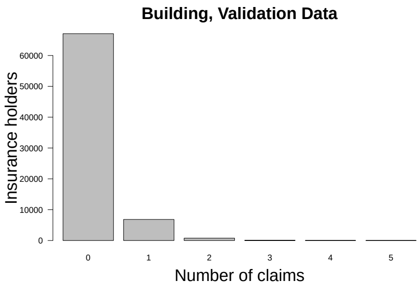


**Figure 3.2.** Histogram for building insurance data, validation data set.

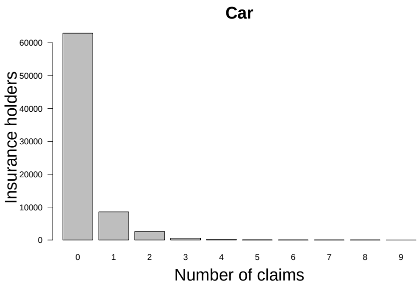


**Figure 3.3.** Histogram for car insurance data.

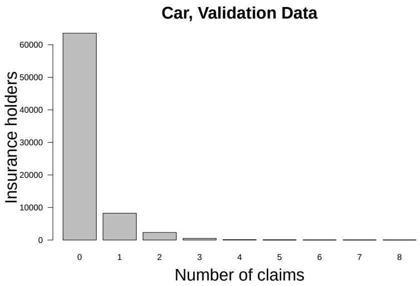


**Figure 3.4.** Histogram for car insurance data, validation data set.

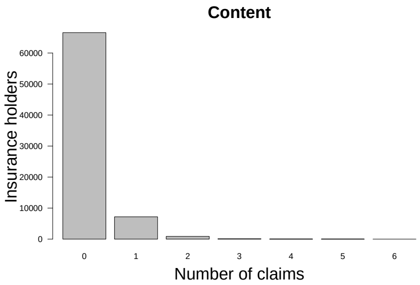


**Figure 3.5.** Histogram for content insurance data.

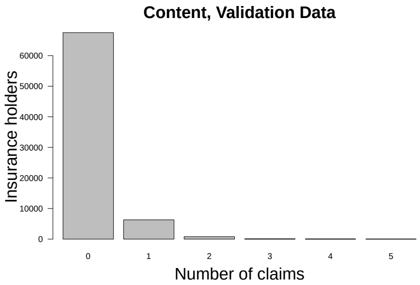


**Figure 3.6.** Histogram for content insurance data, validation data set.

noting that the variance is greater than the mean for all data sets, hinting that the Poisson distribution might not be a sufficient model for the margins. Further, the values for building insurance and content insurance are of the same magnitude, while the mean and variance for car insurance claims are greater.

## 3.2 Modeling Marginal Distributions

The distributions described in Section 2.4 were all tested for each individual marginal and maximum likelihood was used to estimate the parameters. The $\chi^2$-test was used to evaluate if the distribution fits at all while the likelihood ratio test and the AIC values were used to identify the best model. In Tables 3.2, 3.4 and 3.6, the parameter estimates and the goodness of fit statistics are presented for each margin and each model, while the results from the likelihood ratio tests are shown in Table 3.3 and 3.5. For clarity, note that a $p$-value lower than $\alpha = 0.05$ suggests that we should reject $H_0$. The best models are highlighted by italics.

Poisson and zero-inflated Poisson distribution do not fit sufficiently good for any of the data sets. This observation is consistent with the note on the difference in mean and variance in the previous section. For building insurance and content insurance, there are several possible models, but the negative binomial distribution both have the lowest AIC and is the best model according to the likelihood ratio tests and is therefore the chosen model, while for car insurance, zero-inflated Delaporte distribution is the only distribution that is good enough.

**Table 3.2.** Parametrical marginal distribution for car insurance data

| Distr. | Param. est. | Std. error | llh | AIC | $\chi^2$-statistic | $p$-value |
|---|---|---|---|---|---|---|
| Po | $\lambda = 0.2153$ | 0.0017 | $-44\,227.8$ | 88 457.7 | 6 677 568 | 0 |
| ZIP | $\lambda = 0.6468$ | 0.0095 | $-41\,989.1$ | 83 982.1 | 1 901.137 | 0 |
| | $\varphi = 0.6671$ | 0.0045 | | | | |
| NB | $\mu = 0.2153$ | 0.0021 | $-42\,009.1$ | 84 022.2 | 128.2689 | 0 |
| | $\sigma = 2.4088$ | 0.0610 | | | | |
| ZINB | $\mu = 0.5052$ | 0.0226 | $-41\,954.8$ | 83 915.7 | 54.6991 | $8.797 * 10^{-8}$ |
| | $\sigma = 0.3496$ | 0.0662 | | | | |
| | $\varphi = 0.5739$ | 0.0188 | | | | |
| DEL | $\mu = 0.2124$ | 0.0021 | $-42\,011.5$ | 84 028.9 | 135.5536 | 0 |
| | $\sigma = 2.4469$ | 0.1894 | | | | |
| | $\nu = 0.0055$ | 0.0341 | | | | |
| *ZIDEL* | $\mu = 0.5733$ | 0.0163 | $-41\,949.1$ | 83 906.2 | 14.6027 | 0.3328 |
| | $\sigma = 2.8917$ | 1.4164 | | | | |
| | $\nu = 0.7425$ | 0.0778 | | | | |
| | $\varphi = 0.6246$ | 0.0103 | | | | |

**Table 3.3.** Likelihood ratio tests for building insurance data

| $H_0$ | $H_1$ | Degrees of freedom | $p$-value |
|---|---|---|---|
| $NB$ | ZINB | 1 | 1 |
| $NB$ | Del | 1 | 0.8664 |
| $NB$ | ZIDEL | 2 | 1 |

**Table 3.4.** Parametrical marginal distribution for building insurance data

| Distr. | Param. est. | Std. error | llh | AIC | $\chi^2$-statistic | $p$-value |
|---|---|---|---|---|---|---|
| Po | $\lambda = 0.1301$ | 0.0013 | $-30\,420.1$ | 60 842.1 | 985.0602 | 0 |
| ZIP | $\lambda = 0.2572$ | 0.0074 | $-30\,179.8$ | 60 363.7 | 35.4758 | $3.485 * 10^{-6}$ |
| | $\varphi = 0.4943$ | 0.0139 | | | | |
| $NB$ | $\mu = 0.1301$ | 0.0014 | $-30\,167.4$ | 60 338.7 | 0.3712 | 1 |
| | $\sigma = 1.0683$ | 0.0633 | | | | |
| ZINB | $\mu = 0.1306$ | 0.0314 | $-30\,167.4$ | 60 340.7 | 0.3809 | 1 |
| | $\sigma = 1.0565$ | 0.5104 | | | | |
| | $\varphi = 0.0039$ | 0.2390 | | | | |
| DEL | $\mu = 0.1301$ | 0.0014 | $-30\,167.4$ | 60 340.7 | 0.3247 | 1 |
| | $\sigma = 1.1486$ | 0.5422 | | | | |
| | $\nu = 0.0354$ | 0.2192 | | | | |
| ZIDEL | $\mu = 0.1908$ | 0.04702 | $-30\,167.5$ | 60 343.0 | 0.4653 | 1 |
| | $\sigma = 2.9357$ | 4.96516 | | | | |
| | $\nu = 0.6256$ | 0.46527 | | | | |
| | $\varphi = 0.3185$ | 0.16780 | | | | |

## 3.3 Modeling Copula

To model the dependence between the marginal distributions, the four Archimedean copulas tabulated in Table 2.1 in Section 2.2 were considered. Full maximum likelihood was used to get parameter estimates, which implies that new parameter estimations were obtained for the marginals as well.

The goodness of fit was measured using Cramér-von Mises method and Kendall's Cramér-von Mises method, described in Section 2.7. The $p$-values were computed from 1 000 bootstrap values and for Kendall's Cramér-von Mises method, 200 000 Monte Carlo steps were used to estimate $B_m^*$.

It is worth noticing that the value of $S_{n,k}^{(K)}$ in Kendall's Cramér-von Mises method might sometimes be very high. We see that $B_m^*(v_i^*)$ in (2.7.13) compares $v_i^*$ to $v_j^*$.

**Table 3.5.** Likelihood ratio tests for content insurance data

| $H_0$ | $H_1$ | Degrees of freedom | $p$-value |
|---|---|---|---|
| $NB$ | ZINB | 1 | 1 |
| $NB$ | Del | 1 | 0.2492 |
| $NB$ | ZIDEL | 2 | 0.3948 |

**Table 3.6.** Parametrical marginal distribution for content insurance data

| Distr. | Param. est. | Std. error | llh | AIC | $\chi^2$-statistic | $p$-value |
|---|---|---|---|---|---|---|
| Po | $\lambda = 0.1253$ | 0.0013 | $-29\,719.8$ | 59 441.6 | 5341.546 | 0 |
| ZIP | $\lambda = 0.2745$ | 0.0078 | $-29\,404.4$ | 58 812.8 | 157.3674 | 0 |
| | $\varphi = 0.5435$ | 0.0123 | | | | |
| $NB$ | $\mu = 0.1253$ | 0.0014 | $-29\,381.2$ | 58 766.4 | 7.0059 | 0.536 |
| | $\sigma = 1.3129$ | 0.0703 | | | | |
| ZINB | $\mu = 0.1272$ | 0.0252 | $-29\,381.3$ | 58 768.6 | 7.3558 | 0.4988 |
| | $\sigma = 1.2695$ | 0.4626 | | | | |
| | $\varphi = 0.0149$ | 0.1949 | | | | |
| DEL | $\mu = 0.1253$ | 0.0014 | $-29\,380.5$ | 58 767.1 | 3.9395 | 0.9153 |
| | $\sigma = 2.2538$ | 0.7333 | | | | |
| | $\nu = 0.2277$ | 0.1194 | | | | |
| ZIDEL | $\mu = 0.1795$ | 0.0506 | $-29\,380.3$ | 58 768.6 | 3.3408 | 0.9492 |
| | $\sigma = 3.0673$ | 3.0807 | | | | |
| | $\nu = 0.5463$ | 0.3816 | | | | |
| | $\varphi = 0.3021$ | 0.1968 | | | | |

Let $v_\ell^*$ be the lowest value among $v_i^*$, then the lowest possible value of $B_m^*(v_\ell^*)$ is equal to the proportion of observations in the point $(0,0,0)$, as we get a tie in the indicator function for these values. Meanwhile, $K_n(v_i^*)$ in (2.7.15) compares $v_i^*$ to $v_j$, which gives a possibility to get the value $K_n(v_\ell^*) = 0$ if $v_\ell^*$ is less than the lowest value of $v_i$. This possible difference between $B_m^*(v_\ell^*)$ and $K_n(v_\ell^*)$ is furthermore amplified, as $(0,0,0)$ is both the point corresponding to $v_\ell^*$ as well as the by far most common point.

The estimated copula parameters, the log likelihood value as well as the AIC and the goodness of fit measures for the different copulas are presented in Table 3.7. The parameter estimates for the entire models, including the new marginal parameter estimates, are shown in Tables 3.9 to 3.12.

There is a difference in the result of the two goodness of fit measures. According to Cramér-von Mises, Frank and Clayton copula are only significant on the $\alpha = 0.01$ level, while the other two copulas are not suitable at all. However, for Kendall's Cramér-von Mises, all four copula models are significant on the $\alpha = 0.05$ level. As Clayton copula has the lowest AIC and is significant on some level for both goodness of fit methods, it was chosen as the copula for the final model. In Table 3.8, the mean and variance for the marginals in the Clayton copula model have been calculated using the estimated parameters and the formulas shown in Section 2.5.

**Table 3.7.** Summary statistics for copula modeling and copula parameter estimates

| Copula | $\hat{\theta}$ | Std. error | llh | AIC | $p$-val. (KCvM) | $p$-val. (CvM) |
|---|---|---|---|---|---|---|
| *Clayton* | 0.8229 | 0.0256 | −100 667 | 201 352 | 0.258 | 0.026 |
| Frank | 1.5478 | 0.0426 | −100 688.1 | 201 394 | 0.125 | 0.041 |
| Gumbel | 1.0623 | 0.0030 | −101 001.8 | 202 022 | 0.169 | 0 |
| Joe | 1.0661 | 0.0035 | −101 076 | 202 170 | 0.139 | 0 |

**Table 3.8.** Mean and variance for the marginal distributions, calculated using the parameters in the Clayton copula model.

| Marginal | Mean | Variance |
|---|---|---|
| Building | 0.1302 | 0.1481 |
| Car | 0.2145 | 0.3152 |
| Content | 0.1256 | 0.1460 |

**Table 3.9.** Parameter estimates, Clayton copula.

| Marginal | Param. est. | Std. error |
|---|---|---|
| Copula | $\theta = 0.8229$ | 0.0256 |
| Building - NB | $\mu = 0.1302$ | 0.0014 |
| | $\sigma = 1.0560$ | 0.0628 |
| Car - ZIDEL | $\mu = 0.5757$ | 0.0154 |
| | $\sigma = 2.8136$ | 1.1966 |
| | $\nu = 0.7414$ | 0.0680 |
| | $\varphi = 0.6274$ | 0.0096 |
| Content - NB | $\mu = 0.1256$ | 0.0014 |
| | $\sigma = 1.2894$ | 0.0695 |

**Table 3.10.** Parameter estimates, Frank copula.

| Marginal | Param. est. | Std. error |
|---|---|---|
| Copula | $\theta = 1.5478$ | 0.0426 |
| Building - NB | $\mu = 0.1303$ | 0.0014 |
| | $\sigma = 1.0638$ | 0.0631 |
| Car - ZIDEL | $\mu = 0.5775$ | 0.0569 |
| | $\sigma = 3.1839$ | 8.2609 |
| | $\nu = 0.7539$ | 0.3893 |
| | $\varphi = 0.6291$ | 0.0364 |
| Content - NB | $\mu = 0.1257$ | 0.0014 |
| | $\sigma = 1.3033$ | 0.0700 |

**Table 3.11.** Parameter estimates, Gumbel copula.

| Marginal | Param. est. | Std. error |
|---|---|---|
| Copula | $\theta = 1.0623$ | 0.0030 |
| Building - NB | $\mu = 0.1325$ | 0.0014 |
| | $\sigma = 1.1942$ | 0.0656 |
| Car - ZIDEL | $\mu = 0.5992$ | 0.0194 |
| | $\sigma = 6.1547$ | 4.2014 |
| | $\nu = 0.8154$ | 0.0758 |
| | $\varphi = 0.6379$ | 0.0115 |
| Content - NB | $\mu = 0.1280$ | 0.0014 |
| | $\sigma = 1.4436$ | 0.0725 |

**Table 3.12.** Parameter estimates, Joe copula.

| Marginal | Param. est. | Std. error |
|---|---|---|
| Copula | $\theta = 1.0661$ | 0.0035 |
| Building - NB | $\mu = 0.1322$ | 0.0014 |
| | $\sigma = 1.2083$ | 0.0661 |
| Car - ZIDEL | $\mu = 0.5878$ | 0.0157 |
| | $\sigma = 3.3501$ | 1.2708 |
| | $\nu = 0.7408$ | 0.0602 |
| | $\varphi = 0.6318$ | 0.0095 |
| Content - NB | $\mu = 0.1277$ | 0.0014 |
| | $\sigma = 1.4713$ | 0.0735 |

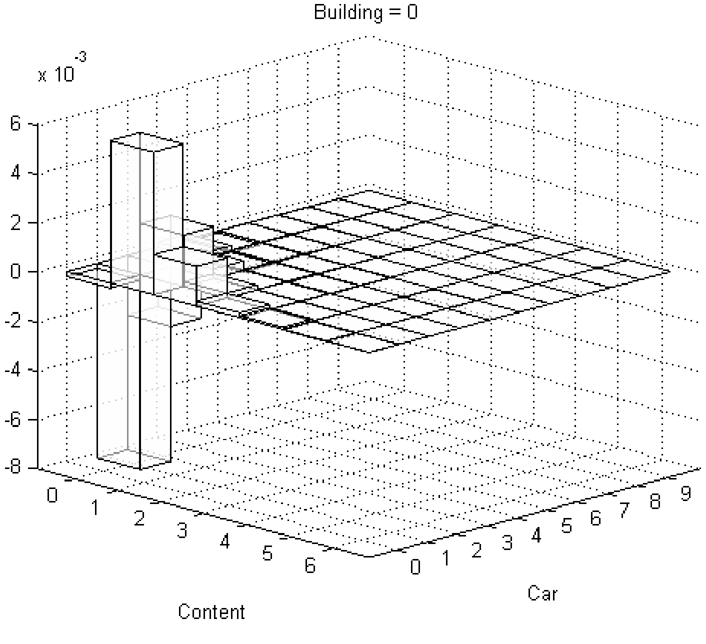


**Figure 3.7.** The difference between the PMF of the model and the empirical PMF, $\hat{p}_{mod}(\boldsymbol{x}) - \hat{p}_{emp}(\boldsymbol{x})$. Building is fixed to 0.

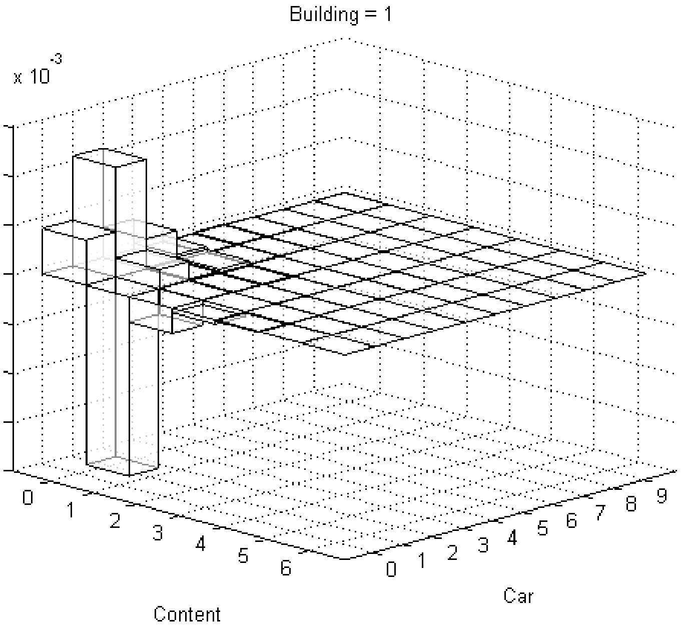


**Figure 3.8.** The difference between the PMF of the model and the empirical PMF, $\hat{p}_{mod}(\boldsymbol{x}) - \hat{p}_{emp}(\boldsymbol{x})$. Building is fixed to 1.

## 3.4 Validation

In Appendix B, contingency tables of the empirical PMF and the PMF for the estimated model are shown to examine how well the model fits to the data and in Figure 3.7 and Figure 3.8, the difference between the PMF of the model and the empirical PMF, $\hat{p}_{mod}(\boldsymbol{x}) - \hat{p}_{emp}(\boldsymbol{x})$, is plotted, for building equal to 0 and 1, respectively. Note that it is not the conditional PMF that is plotted.

To investigate how well the model fits for different years, the model obtained above was fitted to the data from the validation year as well. The marginal parameters were first estimated using maximum likelihood and evaluated using the $\chi^2$-test as in Section 3.2. Then the entire model was estimated using full maximum likelihood and evaluated using Cramér-von Mises method and Kendall's Cramér-von Mises method, as in Section 3.3.

### 3.4.1 Marginal Distributions

All the three marginal distributions fit significantly well according to the $\chi^2$-test. The results are presented in Table 3.13.

**Table 3.13.** Parametrical marginal distributions for validation data

| Marginal | Param. est. | Std. error | llh | $\chi^2$-statistic | $p$-value |
|---|---|---|---|---|---|
| Building (NB) | $\mu = 0.1158$ | 0.0013 | $-27\,805.9$ | 4.3514 | 0.8241 |
| | $\sigma = 1.1826$ | 0.0728 | | | |
| Car (ZIDEL) | $\mu = 0.5540$ | 0.0162 | $-40\,178.4$ | 6.7246 | 0.9782 |
| | $\sigma = 7.9100$ | 3.9491 | | | |
| | $\nu = 0.8315$ | 0.0513 | | | |
| | $\varphi = 0.6359$ | 0.0103 | | | |
| Content (NB) | $\mu = 0.1118$ | 0.0013 | $-27\,070.3$ | 14.1744 | 0.1163 |
| | $\sigma = 1.8108$ | 0.0887 | | | |

### 3.4.2 Copula

For the validation data, both Cramér-von Mises method and Kendall's Cramér-von Mises method support the model. The numerical values are summarized in Table 3.14. The parameter estimates are found in Table 3.15.

**Table 3.14.** Summary statistics for Clayton copula and validation data

| llh | $p$-val. (KCvM) | $p$-val. (CvM) |
|---|---|---|
| $-94\,350$ | 0.647 | 0.094 |

**Table 3.15.** Parameter estimates, Clayton copula.

| Marginal | Param. est. | Std. error |
|---|---|---|
| Copula | $\theta = 0.8244$ | 0.0278 |
| Building - NB | $\mu = 0.1159$ | 0.0013 |
| | $\sigma = 1.1790$ | 0.0727 |
| Car - ZIDEL | $\mu = 0.5547$ | 0.0179 |
| | $\sigma = 7.9946$ | 4.6639 |
| | $\nu = 0.8313$ | 0.0600 |
| | $\varphi = 0.6370$ | 0.0114 |
| Content - NB | $\mu = 0.1120$ | 0.0013 |
| | $\sigma = 1.7976$ | 0.0883 |

Chapter 4

# DISCUSSION

## 4.1 Summary

The purpose with this master's thesis was to model the number of insurance claims in three different insurance types during one year using a three dimensional copula. Further more, maximum likelihood routines suited to models with discrete marginals were to be created and used for inference.

The marginals were modeled using count data distributions able to model data with a lot of zeros. The parameters were estimated using maximum likelihood, the $\chi^2$-test was used for goodness of fit and likelihood ratio tests and the AIC-values were used to choose the model fitting best. The negative binomial distribution proved to be the best distribution to model the building insurance claims as well as the content insurance claims, while the zero-inflated Delaporte distribution provided the best fit for the car insurance claims.

Full maximum likelihood was used to estimate the full models with both one-dimensional marginal distribution and copula parameters and Cramér-von Mises method and Kendall's Cramér-von Mises method with parametric bootstraps were used for goodness of fit testing. The AIC-value was used to choose the best model.

To investigate how well the chosen model performed an other year, it was tested on data for the next year. Neither the $\chi^2$-tests for the marginals or Cramér-von Mises method or Kendall's Cramér-von Mises method gave any reasons to reject the model.

## 4.2 Discussion

All in all, the results seem quite good. The model fits both the empirical values and the validation data from the following year rather well. The goodness of fit tests provides support for it and, as seen in Figure 3.7, the absolute value of greatest difference between the probability mass function of the model and the empirical counterpart is lower than 0.01. Further more, the mean and variance calculated using the models is close to the empirical values, as can be seen in Table 3.1 and Table 3.8.

A part of the reason why several of the Archimedean copulas tested were significantly suitable for model the data might be the fact that all Archimedean copulas are positively lower orthant dependent, that is the probability for all values being low are greater when considering the entire model than when viewing the one-dimensional marginals as independent entities. Most of the observations are low values around zero, why this is a suitable property. Moreover, the probability mass along the line $X_1 = X_2 = X_3$ is not prominent for higher values.

However, as can be seen in Figure 3.7, the error in the point $(0, 0, 0)$ is rather low, suggesting that the error in this point might be very influential for how good the model is considered to be. This is because of the huge number of observations in this point. As a consequence, the error is greater for the values right next to $(0, 0, 0)$. Even though the magnitude of the greatest error is lower than 0.01, this is a problem, especially since this point might not be the most interesting point from a risk perspective. It might be feasible to consider a model estimated conditional on not having a zero to avoid this.

There are some differences in the reached results and the results in [Hage, 2013], where gumbel copula was the only copula that could not be rejected. There are some possible explanations for this. For once, the properties of full maximum likelihood and method of moments might be very different. As the probabilistic and analytical definition of Kendall's tau are not equal, the method of moments might give biased estimates of the copula parameter. The full maximum likelihood routines were investigated using some minor simulation studies during creation, and these gave good estimations even though discrete marginals were used. It is worth mentioning that data from different years were used and that one more dimension was added, something that might affect the results as well.

One source of concern is the difference in results from Cramér-von Mises method and Kendall's Cramér-von Mises method. It seems like Kendall's Cramér-von Mises method is more forgiving than Cramér-von Mises method, but which one that is most reliable is an open question. Taking the huge differences in the values of $S_n^{(K)}$ described and explained in Section 3.3 into account, Cramér-von Mises method is more trustworthy, but as the full maximum likelihood routines, Kendall's Cramér-von Mises method was also investigated using minor simulation studies, with results suggesting that the method performed well for discrete data as well.

The observant reader might have seen that in a few cases, the likelihood for a distribution that is a special case of another distribution is higher than the likelihood for the general distribution, which should be impossible. One explanation for this bewildering fact is likely that different optimization routines for the maximization of the likelihood were used in a few cases to ensure convergence. The mentioned special cases were among these.

Working with discrete copulas creates some difficulties in itself. As mentioned in Section 2.4, a lot of issues arise that need to be taken into account and many things are not clear if they can be done at all. Further more, the copula package in R is implemented for continuous data and, for the investigation to work, several of the functions have been adapted to suit discrete data. Some bugs have been encountered as well, when working with copula package. For instance, elliptical copulas were disregarded from the analysis when it turned out that the values obtained from the function for the CDFs varied even though the input was not altered.

## 4.3 Further Research

Now that both full maximum likelihood and method of moments have been used to model a similar data set, it would be interesting to compare the two methods. Further more, it might be possible to extend the method of moments to dimensions higher than 2 using the generalizations of Kendall's tau and Spearman's rho mentioned in Section 2.3.5. It would also be interesting to compare the copula models with multivariate credibility theory. A way to examine the two methods can be to see how well the methods estimate the conditional expectation of the number of claims in one product given the number of claims in the other products. An extension to this would be to use the data from one year to make predictions for the next year.

As was hinted in the previous section, there are several fields regarding copula models with discrete marginal distributions that can be investigated theoretically, for instance the properties of the different methods of estimation and goodness of fit. The alternate copula definition mentioned in Section 2.4 is interesting and to implement this for the methods used in this thesis would be an interesting path of research. These methods probably need to use some kind of deterministic adaption of the alternate definition to ensure convergence. However, care must be taken so that no new dependence is induced. Some attempts for a maximum likelihood algorithm using a similar method were actually made in this project but was abandoned because of shortness of time.

A possible continuation would be to model the data conditionally of not having any zeros, as mentioned in the previous section. This to avoid letting the vast number of zeros have too much weight. Another way to model is to use pair copula construction, discussed in [Panagiotelis et al., 2012].

Appendix A

# CALCULATIONS

## A.1 Negative Binomial Distribution

Let $\Theta \sim \Gamma(\alpha, \beta)$, with density function

$$f_\Theta(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ x > 0,\ \alpha > 0,\ \text{and}\ \beta > 0 \tag{A.1.1}$$

[Krishnamoorthy, 2006, Chapter 15, with parameter $\beta = 1/b$]. If $X|\Theta = \theta \sim \mathrm{Po}(\theta)$, then $X$ has a negative binomial distribution, $X \sim \mathrm{NB}^{\mathrm{I}}(\alpha, \frac{\beta}{1+\beta})$. The PMF is

$$\begin{aligned} p_X(k) &= \int_0^\infty p_{X|\Theta=x}(k) f_\Theta(x)\,dx = \int_0^\infty \frac{e^{-x}x^k}{k!} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\,dx = \\ &= \frac{\beta^\alpha}{k!\Gamma(\alpha)} \int_0^\infty x^{k+\alpha-1} e^{-x(1+\beta)}\,dx = \frac{\beta^\alpha}{k!\Gamma(\alpha)} \int_0^\infty \left(\frac{y}{1+\beta}\right)^{k+\alpha-1} e^{-y} \frac{dy}{1+\beta} = \\ &= \frac{\beta^\alpha}{k!\Gamma(\alpha)} \left(\frac{1}{1+\beta}\right)^{k+\alpha} \int_0^\infty y^{k+\alpha-1} e^{-y}\,dy = \frac{\Gamma(k+\alpha)}{k!\Gamma(\alpha)} \left(\frac{\beta}{1+\beta}\right)^\alpha \left(\frac{1}{1+\beta}\right)^k, \end{aligned} \tag{A.1.2}$$

where the definition of the gamma function, $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$ is used in the last equality. Now, let $r = \alpha$ and $p = \frac{\beta}{1+\beta}$ and we get the parametrization $X \sim \mathrm{NB}^{\mathrm{I}}(r, p)$.

## A.2 Delaporte Distribution

Let $\Theta = \lambda + G$ where where $\lambda \in \mathbb{R}^+$ and $G \sim \Gamma(\alpha, \beta)$. Then $\Theta$ has a *shifted gamma distribution*, $\Theta \sim \mathrm{SG}^{\mathrm{I}}(\lambda, \alpha, \beta)$. Further, let $X|\Theta = \theta \sim \mathrm{Po}(\theta)$, then $X$ has a *Delaporte distribution*, $X \sim \mathrm{Del}^{\mathrm{I}}(\lambda, \alpha, \beta)$. [Vose, 2008, with parameter $\beta^* = 1/\beta$]

To get the probability mass function for $X$, we first of all establish that $\Theta$ has the density function

$$f_\Theta(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} (x-\lambda)^{\alpha-1} e^{-\beta(x-\lambda)},\ x > \lambda,\ \lambda > 0,\ , \alpha > 0 \text{ and } \beta > 0 \tag{A.2.1}$$

[Krishnamoorthy, 2006, Chapter 15]. The PMF for $X$ then becomes

$$\begin{aligned}
p_X(k) = &\int_\lambda^\infty p_{X|\Theta=x}(k) f_\Theta(x)\,dx = \int_\lambda^\infty \frac{e^{-x}x^k}{k!}\frac{\beta^\alpha}{\Gamma(\alpha)}(x-\lambda)^{\alpha-1}e^{-\beta(x-\lambda)}\,dx = \\
=&\frac{\beta^\alpha}{k!\Gamma(\alpha)}\int_0^\infty e^{-y-\lambda\beta y}(y+\lambda)^k y^{\alpha-1}\,dy = \\
=&\frac{\beta^\alpha e^{-\lambda}}{k!\Gamma(\alpha)}\int_0^\infty e^{-y(1+\beta)}y^{\alpha-1}\sum_{i=0}^{k}\binom{k}{i}y^i\lambda^{k-i}\,dy = \\
=&\frac{\beta^\alpha e^{-\lambda}}{\Gamma(\alpha)}\sum_{i=0}^{k}\frac{\lambda^{k-i}}{i!(k-i)!}\int_0^\infty e^{-y(1+\beta)}y^{\alpha+i-1}\,dy = \\
=&\frac{\beta^\alpha e^{-\lambda}}{\Gamma(\alpha)}\sum_{i=0}^{k}\frac{\lambda^{k-i}}{i!(k-i)!}\int_0^\infty e^{-x}\left(\frac{x}{1+\beta}\right)^{\alpha+i-1}\frac{1}{1+\beta}\,dx = \\
=&\sum_{i=0}^{k}\frac{\beta^\alpha e^{-\lambda}\lambda^{k-i}}{\Gamma(\alpha)i!(k-i)!(1+\beta)^{\alpha+i}}\int_0^\infty e^{-x}x^{\alpha+i-1}\,dx = \\
=&\sum_{i=0}^{k}\frac{\Gamma(\alpha+i)\beta^\alpha e^{\lambda}\lambda^{k-i}}{\Gamma(\alpha)i!(k-i)!(1+\beta)^{\alpha+i}}, \qquad \text{(A.2.2)}
\end{aligned}$$

where the definition of the gamma function and the binomial theorem are put to use, as well as two simple changes of variables. Note that $k=0$ gives $p_X(0) = \left(\frac{\beta}{1+\beta}\right)^\alpha e^{-\lambda}$.

In gamlss package in R, another parametrization is used. Let $\Theta = G(1-\nu)+\nu$, where $0<\nu<1$ and $G \sim \Gamma(\frac{1}{\tilde\sigma^2}, \frac{1}{\tilde\mu\tilde\sigma^2})$. Then, $\Theta$ has a re-parameterized shifted gamma distribution, $\mathrm{SG}^{\mathrm{II}}(\tilde\mu,\tilde\sigma,\tilde\nu) = \mathrm{SG}^{\mathrm{I}}(\tilde\nu, \frac{1}{\tilde\sigma^2}, \frac{1}{\tilde\mu\tilde\sigma^2})$ with density function

$$f_\Theta(x) = f_G\left(\frac{x-\tilde\nu}{1-\tilde\nu}\right)\frac{1}{1-\tilde\nu} = \frac{(x-\tilde\nu)^{1/\tilde\sigma^2-1}e^{-\frac{x-\tilde\nu}{\tilde\sigma^2\tilde\mu(1-\tilde\nu)}}}{(\tilde\sigma^2\tilde\mu(1-\tilde\nu))^{1/\tilde\sigma^2}\Gamma(1/\tilde\sigma^2)}, \qquad \text{(A.2.3)}$$

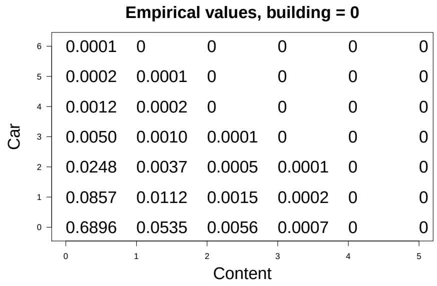
for $x \geq \tilde\nu$, $0<\tilde\nu<1$, $\tilde\mu>0$ and $\tilde\sigma>0$. Now, let $\Theta \sim \mathrm{SG}^{\mathrm{II}}(1,\sigma^{1/2},\nu)$ and consider $\mu\Theta$. This has density function

$$f_{\mu\Theta}(x) = \frac{(x-\mu\nu)^{1/\sigma-1}e^{-\frac{x-\mu\nu}{\mu\sigma(1-\nu)}}}{(\mu\sigma(1-\nu))^{1/\sigma}\Gamma(1/\sigma)}. \qquad \text{(A.2.4)}$$

Let $X|\Theta=\theta \sim \mathrm{Po}(\mu\theta)$, then $X$ has a re-parameterized Delaporte distribution, $X \sim Del^{\mathrm{II}}(\nu,\mu,\sigma) = Del^{\mathrm{I}}(\mu\nu, \frac{1}{\sigma}, \frac{1}{\mu\sigma(1-\nu)})$.

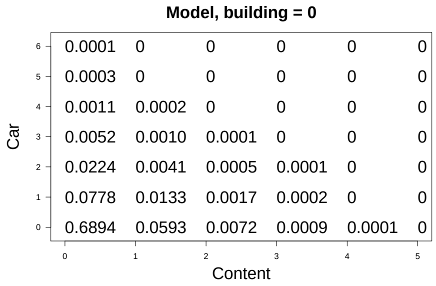
## Appendix B

# CONTINGENCY TABLES


Empirical values, building = 0
Car
Content

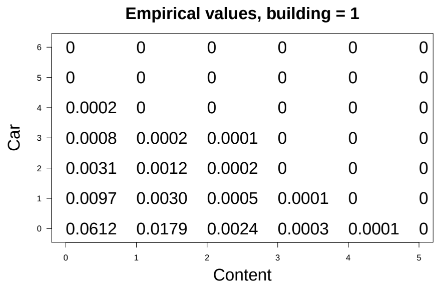
| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0.0001 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.0002 | 0.0001 | 0 | 0 | 0 | 0 |
| 4 | 0.0012 | 0.0002 | 0 | 0 | 0 | 0 |
| 3 | 0.0050 | 0.0010 | 0.0001 | 0 | 0 | 0 |
| 2 | 0.0248 | 0.0037 | 0.0005 | 0.0001 | 0 | 0 |
| 1 | 0.0857 | 0.0112 | 0.0015 | 0.0002 | 0 | 0 |
| 0 | 0.6896 | 0.0535 | 0.0056 | 0.0007 | 0 | 0 |

**Figure B.1.** Contingency table of the empirical PMF for when building = 0.


Model, building = 0
Car
Content

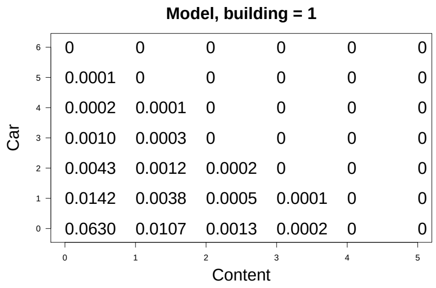
| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0.0001 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.0003 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.0011 | 0.0002 | 0 | 0 | 0 | 0 |
| 3 | 0.0052 | 0.0010 | 0.0001 | 0 | 0 | 0 |
| 2 | 0.0224 | 0.0041 | 0.0005 | 0.0001 | 0 | 0 |
| 1 | 0.0778 | 0.0133 | 0.0017 | 0.0002 | 0 | 0 |
| 0 | 0.6894 | 0.0593 | 0.0072 | 0.0009 | 0.0001 | 0 |

**Figure B.2.** Contingency table of the PMF from the estimated model for when building = 0.


Empirical values, building = 1
Car
Content

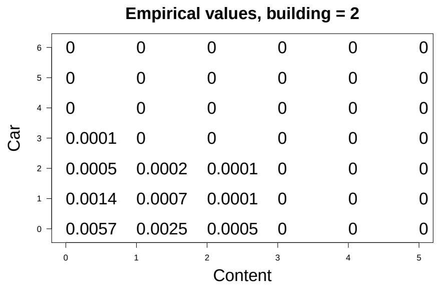
| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.0002 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.0008 | 0.0002 | 0.0001 | 0 | 0 | 0 |
| 2 | 0.0031 | 0.0012 | 0.0002 | 0 | 0 | 0 |
| 1 | 0.0097 | 0.0030 | 0.0005 | 0.0001 | 0 | 0 |
| 0 | 0.0612 | 0.0179 | 0.0024 | 0.0003 | 0.0001 | 0 |

**Figure B.3.** Contingency table of the empirical PMF for when building = 1.


Model, building = 1
Car
Content

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.0001 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.0002 | 0.0001 | 0 | 0 | 0 | 0 |
| 3 | 0.0010 | 0.0003 | 0 | 0 | 0 | 0 |
| 2 | 0.0043 | 0.0012 | 0.0002 | 0 | 0 | 0 |
| 1 | 0.0142 | 0.0038 | 0.0005 | 0.0001 | 0 | 0 |
| 0 | 0.0630 | 0.0107 | 0.0013 | 0.0002 | 0 | 0 |

**Figure B.4.** Contingency table of the PMF from the estimated model for when building = 1.

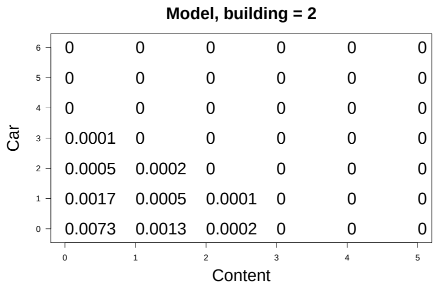


Empirical values, building = 2

| Car | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.0001 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.0005 | 0.0002 | 0.0001 | 0 | 0 | 0 |
| 1 | 0.0014 | 0.0007 | 0.0001 | 0 | 0 | 0 |
| 0 | 0.0057 | 0.0025 | 0.0005 | 0 | 0 | 0 |

Content

**Figure B.5.** Contingency table of the empirical PMF for when building = 2.

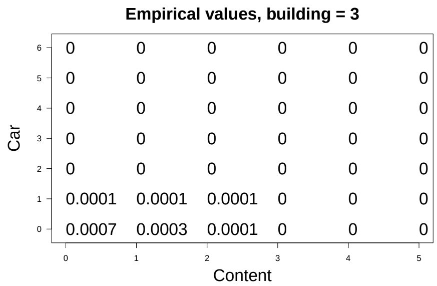


Model, building = 2

| Car | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.0001 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.0005 | 0.0002 | 0 | 0 | 0 | 0 |
| 1 | 0.0017 | 0.0005 | 0.0001 | 0 | 0 | 0 |
| 0 | 0.0073 | 0.0013 | 0.0002 | 0 | 0 | 0 |

Content

**Figure B.6.** Contingency table of the PMF from the estimated model for when building = 2.

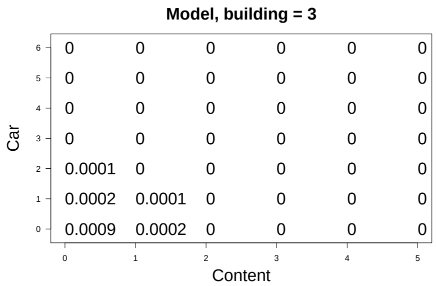


Empirical values, building = 3

| Car | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0.0001 | 0.0001 | 0.0001 | 0 | 0 | 0 |
| 0 | 0.0007 | 0.0003 | 0.0001 | 0 | 0 | 0 |

Content

**Figure B.7.** Contingency table of the empirical PMF for when building = 3.

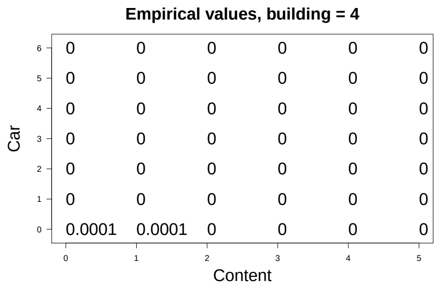


Model, building = 3

| Car | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.0001 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0.0002 | 0.0001 | 0 | 0 | 0 | 0 |
| 0 | 0.0009 | 0.0002 | 0 | 0 | 0 | 0 |

Content

**Figure B.8.** Contingency table of the PMF from the estimated model for when building = 3.

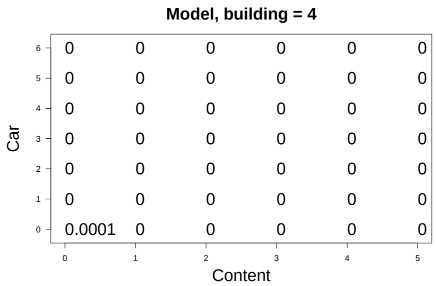


Empirical values, building = 4

| Car | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0.0001 | 0.0001 | 0 | 0 | 0 | 0 |

Content

**Figure B.9.** Contingency table of the empirical PMF for when building = 4.



Model, building = 4

| Car | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0.0001 | 0 | 0 | 0 | 0 | 0 |

Content

**Figure B.10.** Contingency table of the PMF from the estimated model for when building = 4.

# BIBLIOGRAPHY

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

[Delignette-Muller et al., 2013] Delignette-Muller, M. L., Pouillot, R., Denis, J.-B., and Dutang, C. (2013). *fitdistrplus: help to fit of a parametric distribution to non-censored or censored data.* R package version 1.0-1.

[Faugeras, 2012] Faugeras, O. P. (2012). Probabilistic constructions for discrete copulas. Submitted, available on http://hal.archives-ouvertes.fr/hal-00751393/. Retrieved 21 August 2013.

[Genest and Nešlehová, 2007] Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *Astin Bulletin*, 37:475–515.

[Genest et al., 2009] Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199 – 213.

[Hage, 2013] Hage, M. (2013). A copula approach to modeling insurance claims. Master's thesis, Lund University.

[Joe, 1990] Joe, H. (1990). Multivariate concordance. *Journal of Multivariate Analysis*, 35(1):12 – 30.

[Joe and Xu, 1996] Joe, H. and Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia.

[Johnson et al., 2005] Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate Discrete Distributions.* John Wiley & Sons, Inc., Hoboken, New Jersey, 3rd edition.

[Kimeldorf and Sampson, 1978] Kimeldorf, G. and Sampson, A. R. (1978). Monotone dependence. *The Annals of Statistics*, 6:895–903.

[Krishnamoorthy, 2006] Krishnamoorthy, K. (2006). *Handbook of Statistical Distributions with Applications.* Chapman and Hall/CRC, Boca Raton, Florida.

[Madsen, 2008] Madsen, H. (2008). *Time Series Analysis*. Chapman and Hall/CRC, Boca Raton, Florida.

[Mesfioui and Quessy, 2010] Mesfioui, M. and Quessy, J.-F. (2010). Concordance measures for multivariate non-continuous random vectors. *Journal of Multivariate Analysis*, 101(10):2398 – 2410.

[Nelsen, 2006] Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Science+Business Media, New York, NY, 2nd edition.

[Panagiotelis et al., 2012] Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072.

[Rigby and Stasinopoulos, 2005] Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

[Schmid and Schmidt, 2007] Schmid, F. and Schmidt, R. (2007). Multivariate extensions of spearman's rho and related statistics. *Statistics & Probability Letters*, 77(4):407 – 416.

[Stasinopoulos et al., 2008] Stasinopoulos, M., Rigby, B., and Akantziliotou, C. (2008). *Instructions on how to use the gamlss package in R*, 2nd edition. Available on http://www.gamlss.org/. Retrieved 2 September 2013.

[Thuring, 2012] Thuring, F. (2012). A credibility method for profitable cross-selling of insurance products. *Annals of Actuarial Science*, 6:65–75.

[Vose, 2008] Vose, D. (2008). *Risk Analysis: A Quantitative Guide*. John Wiley & Sons, Inc., Hoboken, New Jersey, 3rd edition.

[Yan, 2007] Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21(4):1–21.

# POPULÄRVETENSKAPLIG ARTIKEL

## Copulamodellering av skadeanmälningar i försäkringar

Försäkringar är ganska annorlunda jämfört med de flesta andra produkter man stöter på i sin vardag. För att tjäna pengar på en vara kan ett företag oftast anpassa priset efter tillverkningskostnaden. Ett försäkringsbolag har däremot problemet att kostnaderna inte kommer förrän kunden redan har köpt produkten. När kostnaden väl kommer varierar dessutom både antalet utbetalningar och storleken på dem. Det här kräver att försäkringsbolagen kan göra något slags uppskattning av hur stor risken är att en viss kund ska råka ut för olycksfall, för att kunna anpassa försäkringspremien. En annan sak som är bra för ett försäkringsbolag med flera olika typer av försäkringar, är om det kan välja ut sina bästa, minst riskbenägna kunder för att försöka sälja tilläggsförsäkringar. På det sättet ökar försäkringsbolaget andelen kunder med låg risk för att råka ut för olyckor. För att göra den här typen av anpassningar krävs bra modeller att räkna med. Bland annat måste man kunna beskriva hur olika saker är relaterade till varandra, till exempel sambandet mellan hur många skadeanmälningar en kund gör i sina olika försäkringar.

Ett hjälpmedel som blivit väldigt populärt de senaste åren är något som kallas för copulas. De används för att beskriva beroende mellan olika tal. Idn är att istället för att göra en enda stor modell på en gång, där varje sak som ingår behandlas på samma sätt, modellerar man först en sak i taget. Sedan slås de färdiga modellerna ihop med hjälp av en gemensam modell, copulan. Copulan beskriver bara sambandet mellan de olika småmodellerna. Det fina med den här metoden är dels att man kan välja olika modeller helt fritt för var och en av delmodellerna, dels att många olika slags samband kan beskrivas, beroende på vilken copula man väljer att använda. När man väl valt copula kan man dessutom finjustera inställningarna för att den ska passa så bra som möjligt.

Det är viktigt att välja rätt copula, eftersom egenskaperna varierar mycket beroende på vilken man väljer. Ett exempel på när valet inte blev så lyckat är finanskrisen under 2007 och 2008. Copulas hade blivit populära i modeller för konkursrisk i början av 2000-talet, framför allt tack vare att de har goda egenskaper och att metoden för att göra modeller med dem är intuitiv och lätt att förstå sig på. Bankerna använde sig oftast av den så kallade Gaussiska copulan, som är smidig att använda, men som har en egenskap som gör den olämplig i sammanhanget. Problemet är att extrema händelser sker oberoende av varandra enligt den Gaussiska copulan. Det betyder att om man använder den för att göra en modell för konkursrisken i två företag och vet att det är stor risk att ett av dem ska gå omkull, säger det ingenting om risken för konkurs i det andra företaget. Om något händer i samhället som gör att många företag får ökad konkursrisk samtidigt, missar modellen det. Modellerna började efter ett tag användas för subprimelån också, något de inte var anpassade till. När huspriserna började sjunka kunde många hushåll inte länge betala sina lån, vilket modellerna missade och detta bidrog till finanskrisen.

Det ska kanske nämnas att det inte var modellerna i sig som skapade en finanskris, utan sättet och typerna av finansiella instrument man använde dem på. I vilket fall som helst är det viktigt att öka förståelsen om modellerna så mycket som möjligt. Om man ska använda copulas för att hitta en modell för antalet skadeanmälningar i en personers olika försäkringar stöter man på nya problem. Det centrala problemet är att teorin för copulas främst är anpassad för att fungera för kontinuerlig data, det vill säga tal som kan ha vilka och hur många decimaler som helst, medan antalet skadeanmälningar är heltal. Det är inget problem när man gör modeller för antalet skadeanmälningar i en försäkring i taget, men när man ska slå ihop modellerna behöver man vara försiktig. En annan svårighet är att de flesta programpaket som används för att arbeta med copulas är anpassade för kontinuerlig data. Om man ska arbeta med heltalsdata måste man skriva egna program.

I den här undersökningen studerades data från 74 770 försäkringsinnehavare som var och en hade tre försäkringstyper. Datan bestod av antalet skadeanmälningar per år och målet var att hitta en lämplig copulamodell som passade de aktuella värdena. Först behandlades varje försäkringstyp för sig och sedan valdes den mest lämpliga av fyra olika testade copulas till att slå ihop modellen. Den slutgiltiga modellen anpassades så att den skulle passa datan så bra som möjligt. När den sedan testades på data från året därpå, visade det sig att modellen stämde bra även för det året.

Den här modellen kan alltså användas för att beräkna det förväntade antalet skadeanmälningar för en kund. Då kan man också beräkna den förväntade kostnaden, om man har bra modeller för storleken på utbetalningarna. Med hjälp av detta kan man sätta lagom premier på försäkringarna. Har man bara koll på vad man gör kan man dessutom undvika att skapa en finanskris på kuppen.

Master's Theses in Mathematical Sciences 2013:E55
ISSN 1404-6342

LUNFMS-3050-2013

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/