GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Master Thesis

Multivariate modelling of the dependency structure between article sales of a sportswear manufacturer

Author

Petros Christanas

from Nuremberg

Matriculation Number

11604278

Applied Statistics M.Sc.

Chair of Statistics and Econometrics

Supervisors

Prof. Dr. Thomas Kneib

Dipl.-Vw. Quant. Fabian H. C. Raters

Submitted March 28, 2020

Processing time of 20 weeks

Confidentiality Clause

Statutory Declaration

Acknowledgments

Contents

1 Introduction

Write introduction here and "upper" subsections here (adidas, Motivation, etc...)

1.1 adidas

1.2 Data Sources

Throughout each season, transactional data are collected from online purchases of the sports brand's eCommerce website. Specifically, we are provided with weekly sales data for western European countries. A short description is depicted in Table 1.1.

Column	Description	Values
week_id	Calender week of a specific year (YYYYWW)	Factors: *201648, ..., 201852*
article_number	Unique article identification number (article ID)	Factors: *10669, 10, ...*
min_date_of_week	Minimum date of the respective week; always a Monday (YYYY-MM-DD)	Dates: *2016-11-28, ..., 2018-12-24*
art_min_price	Minimal recorded price of the article	Non-negative (integer) value
month_id	Calender month of a specific year (YYYYMM)	Factors: *201612, ..., 201812*
season	Season of year (format: SSYY) (Spring-Summer [SS]: December - May) Fall-Winter [FW]: June - November)	Factors: *SS17, FW17,* *SS18, FW18, SS19*
bf_w	Weekly "Black Friday" promotion intensity of the article	Between 0 and 1
ff_w	Weekly "Friends & Family" promotion intensity of the article	Between 0 and 1
ot_w	Weekly article promotion intensity of "Other" type	Between 0 and 1
gross_demand_quantity	Weekly amount of added articles to shopping cart	Non-negative (integer) value
base_price_locf	Retail price of the article without any discounts	Non-negative (integer) value
total_markdown_pct		
day_of_month	Day of the month	Integers: *1 - 31*
month_of_year	Month of the year	Factors: *January, ..., December*
year	Year	Integers: *2016, 2017, 2018*
week_of_year	Week of the year	Integers: *1 - 52*

Table 1.1: Transactional raw data description from online purchases of western European countries

Due to legal regulations of the company, some columns had to undergo anonymization in order for the data to be released. To ensure data protection and confidentiality, numeric variables (with exception of time-indicating columns) were transformed. As a consequence for the analysis part, most integer values were converted to float numbers. This fact should be kept in mind by the reader, since the above table serves as a reminder and reference point for the data

documentation.

Another peculiarity of this setup is to be considered, too. We will often refer to the variable *gross demand quantity* as *sales*, even though it is obviously not exactly the same. In the eCommerce environment, there are several stages before the purchase is complete, e.g. addition to cart, removal from cart, proceeding to checkout & even the return of bought articles. Targeting the articles added to cart, i.e. the (gross) demand quantity, provides the optimal data extraction for analytical purposes and is the closest to adequately model the dependency structure between net sales of articles.

Besides the transactional data, *article master data*, i.e. attributes of the articles, are provided and described in Table 1.2.

Column	Description	Values (all Factors)
article_number	Unique article identification number (article ID)	*10669, 10, ...*
gender	Gender type of the article (Men, Women, Unisex)	*M, W, U*
age_group	Age group of the article (Adult, Infant, Junior, Kids)	*A, I, J, K*
product_division_descr	Product division of the article	*APPAREL, FOOTWEAR, HARDWARE*
product_group_descr	Product group of the article	*BAGS, BALLS, FOOTWEAR ACCESSORIES, SHOES, ...*
color	Consolidated color group of the article	*BEIGE, BLACK, BROWN, ORANGE, PINK, RED, ...*
sports_category_descr	Sports category of the article	encoded: *SC_1, ..., SC_22*
sales_line_descr	Sales line of the article	encoded: *SL_1, ..., SL_379*
business_unit_descr	The article's Business Unit membership	encoded: *BU_1, ..., BU_18*
business_segment_descr	The article's Business Segment membership	encoded: *BS_1, ..., BS_49*
sub_brand_descr	Sub-brand of the article	encoded: *sub-brand_1, ..., sub-brand_4*
item_type	Item type of the article	encoded: *IT_1, ..., IT_171*
brand_element	Brand element of the article	encoded: *BE_1, ..., BE_131*
product_franchise_descr	Product franchise of the article	encoded: *franchise_1, ..., franchise_72*
product_line_descr	Product line of the article	encoded: *PL_1, ..., PL_105*
franchise_bin	Franchise indicator of the article	*FRANCHISE, NON-FRANCHISE*
category	Category of the article	encoded: *category_1, category_2*

Table 1.2: Article master data

Plenty of additional information is stored in the database, but we are neglecting columns omitted in these tables, as they are redundant, already summarized, transformed or simply do not provide any value.
Overall, we will be dealing with data collected over two years, namely the years 2017 and 2018, while some transactions of late 2016 might be attached marginally. In summary, after joining the transactional observations to the article

attributes by the article ID, this translates to a dataset of over 587,000 rows and over 30 variables.

1.3 Motivation

2 Statistical Theory & Methods

This chapter introduces various statistical methods used during the conduction of this thesis. It is assumed that basic understanding and knowledge of the reader regarding mathematical foundations of statistics (like linear algebra, probability theory, etc) already exists.

2.1 Generalized Linear Models

Generalized Linear Models (GLMs) are an extension of the classical *Linear Regression Model (LM)*

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n$$

which in matrix notation can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where the response variable y_i can take values from several probability distributions (e.g. Poisson, Binomial, Gamma and others), which are members of the exponential family [Fahrmeir et al., 2003]. The linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta} \tag{2.1}$$

is passed through a *response function h* (a one-to-one, twice differentiable transformation), such that

$$E(y_i) = h(\eta_i)$$

i.e. the expected value of the response variable belongs to the value range of *h*. The inverse of the response function,

$$g = h^{-1}$$

is called the *link function*.

2.2 Mixed Effects Models

Linear Mixed Models (LMMs) are powerful tools when dealing with clustered data or data with a longitudinal structure (repeated measurements of individuals). As in the classical LM, there are population-specific effects, namely the parameter

vector of *fixed effects* $\boldsymbol{\beta}$, as well as the cluster- or individual-specific effects of such models called *random effects* [Fahrmeir et al., 2003]. In the following, we will refer to our clusters or individuals as "groups" for briefness. Mathematically speaking, the linear predictor $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$ is extended to

$$\eta_{ij} = \boldsymbol{x'}_{ij}\boldsymbol{\beta} + \boldsymbol{u}'_{ij}\boldsymbol{\gamma}_i, \quad j = 1, \ldots, m, \quad i = 1, \ldots, n_i, \tag{2.2}$$

where

- i is the number of groups
- j is the number of observations per group
- $\boldsymbol{\beta}$ is the vector of fixed effects
- $\boldsymbol{\gamma}_i$ is the vector of random effects
- $\boldsymbol{x}'_{ij}$ is the vector of covariates and
- $\boldsymbol{u}'_{ij}$ is a subvector of $\boldsymbol{x}'_{ij}$.

$\boldsymbol{x}'_{ij} = (1, x_{ij1}, \ldots, x_{ijk})$ and $\boldsymbol{u}'_{ij} = (1, u_{ij1}, \ldots, u_{ijk})$ are therefore the design vectors and ε_{ij} are the error terms of the *measurement model*

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{u}'_{ij}\boldsymbol{\gamma}_i + \epsilon_{ij}, \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right) \tag{2.3}$$

or in matrix notation

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i \tag{2.4}$$

for group $i = 1, \ldots, m$ with $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$.

Similar to GLMs, *Generalized Linear Mixed Models (GLMMs)* relate the linear mixed predictor 2.2 to the conditional mean $\mu_{ij} = E(y_{ij}|\boldsymbol{\gamma}_i)$ via a suitable response function h, such that $\mu_{ij} = h(\eta_{ij})$ and thus the conditional density of y_{ij} belongs to the exponential family.

2.3 Additive Models

Additive Models expand models with just a linear predictor

$$\eta_i^{lin} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

(such as the LM) to

$$y_i = \eta_i^{add} + \varepsilon_i, \tag{2.5}$$

where

$$\eta_i^{add} = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin}, \quad i = 1, \ldots, n. \tag{2.6}$$

The functions $f_1(z_1), \ldots, f_q(z_q)$ are non-linear univariate *smooth effects* of the *continuous* covariates $z_1, \ldots, z_q$ and are defined as

$$f_j(z_j) = \sum_{l=1}^{d_j} \gamma_{jl} B_l(z_j) \tag{2.7}$$

with $B_l(z_j)$ being *basis functions* for $j = 1, \ldots, q$ and d_j the number of basis functions for covariate z_j. The regression coefficients of the basis functions $B_l(z_j)$ are labeled as γ_{jl}. There is a wide variety of basis functions which can be used to flexibly model the data in a non- or semiparametric manner. For more content on basis functions we refer to Wood [2017] and Fahrmeir et al. [2003]. The basis functions evaluated at the observed covariate values are summarized in the design matrices $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_q$ and the additive model 2.5 can be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Accordingly, the vector of function values evaluated at the observed covariate values $z_{1j}, \ldots, z_{nj}$ is denoted by $\boldsymbol{f}_j = (f_j(z_{1j}), \ldots, f_j(z_{nj}))'$ and therefore $\boldsymbol{f}_j = \boldsymbol{Z}_j\boldsymbol{\gamma}_j$. To ensure identifiability of the additive model, the functions $f_j(z_j)$ are centered around zero, such that

$$\sum_{i=1}^{n} f_1(z_{i1}) = \ldots = \sum_{i=1}^{n} f_q(z_{iq}) = 0.$$

Extensions of additive models to non-normal responses are consequently called *Generalized Additive Models (GAMs)*.

MAYBE INCLUDE MODELS WITH INTERACTIONS; MIXED EFFECTS

3 Copulas & Dependency Structures

Multivariate distributions consist of the marginal distributions and the dependence structure between those marginals. These components can be specified separately in a single framework with the help of copula functions. This chapter introduces the concept of modelling such dependency structures with copulas, which is the main focus of this thesis.

3.1 Introduction to Copulas

A d-dimensional function $C : [0,1]^d \to [0,1]$ is called a *copula*, if it is a Cumulative Distribution Function (CDF) with uniform margins, i.e.

$$P\left(U_1 \leq u_1, \ldots, U_d \leq u_d\right) = C\left(u_1, \ldots, u_d\right)$$

where $U_i,\ i = 1, \ldots, d$ are uniformly distributed Random Variables (RVs) in $[0,1]$. Since C is a CDF, following properties emerge:

- $C(\boldsymbol{u}) = C\left(u_1, \ldots, u_d\right)$ is increasing in each component $u_i, i = 1, \ldots, d$.
- The i^{th} marginal distribution is obtained by setting $u_j = 1$ for $j \neq i$ and it has to be uniformly distributed

  $$C\left(1, \ldots, 1, u_i, 1, \ldots, 1\right) = u_i$$

- For $a_i \leq b_i$, the probability $P\left(U_1 \in [a_1, b_1], \ldots, U_d \in [a_d, b_d]\right)$ mus be non-negative, so we obtain the *rectangle inequality*

  $$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1 + \cdots + i_d} C\left(u_{1,i_1}, \ldots, u_{d,i_d}\right) \geq 0, \tag{3.1}$$

  where $u_{j,1} = a_j$ and $u_{j,2} = b_j$.

The reverse is also true, i.e. any function C that satisfies the above properties is a copula. Furthermore, $C\left(1, u_1, \ldots, u_{d-1}\right)$ is also a $(d-1)$-dimensional copula and thus all k-dimensional marginals with $2 < k < d$ are copulas.

Generalized Inverse

For a CDF, the *generalized inverse* is defined by

$$F^{\leftarrow}(y) := \inf\{x : F(x) \geq y\}$$

(similar to the definition of a *quantile function*).

□

Probability Transformation

If a RV Y has a continuous CDF F, then

$$F(Y) \sim U[0,1]. \tag{3.2}$$

□

The reverse of the *probability transformation* is the *quantile transformation*.

Quantile Transformation

If $U \sim U[0,1]$ and F be a CDF, then

$$P\left(F^{\leftarrow}(U) \leq x\right) = F(x) \tag{3.3}$$

□

The above two transformations allow us to move back and forth between $\mathbb{R}^d$ and $[0,1]^d$ and are the primary building blocks regarding copulas. Against this backdrop, we introduce *Sklar's theorem* which is considered the foundation of all copula related applications.

Sklar's Theorem Sklar [1959]

Let F be a d-dimensional CDF with marginal distributions $F_i,\ i = 1,\ldots,d$. Then there exists a copula C such that

$$F(x_1,\ldots,x_d) = C(F_1(x_1),\ldots,F_d(x_d)) \tag{3.4}$$

for all $x_i \in \mathbb{R},\ i = 1,\ldots,d$.

The copula C is unique, if $\forall i = 1,\ldots,d$, F_i is continuous. Otherwise C is uniquely determined only on $Ran(F_1) \times \ldots \times Ran(F_d)$, where $Ran(F_i)$ is the range of F_i.

Conversely, if C is a d-dimensional copula and $F_1,\ldots,F_d$ are univariate CDF's, then F as defined in Equation 3.4 is a d-dimensional CDF.

□

If the copula has a Probability Density Function (PDF), then the *copula density* is defined as

$$c(\boldsymbol{u}) = \frac{\partial^d C\left(u_1, \ldots, u_d\right)}{\partial u_1 \cdots \partial u_d} \tag{3.5}$$

for a differentiable copula function C and the realization of a random vector $\boldsymbol{u} = (u_1, \ldots, u_d)$.

By virtue of Equation 3.4 in Sklar's theorem and given that

$$C(\boldsymbol{u}) = F\left(F_1^{\leftarrow}\left(u_1\right), \ldots, F_d^{\leftarrow}\left(u_d\right)\right), \tag{3.6}$$

i.e. inversible CDFs $F_i,\ i = 1, \ldots, d,$ we can rewrite the copula density to

$$c\left(u_1, \ldots, u_d\right) = \frac{f\left(F_1^{\leftarrow}\left(u_1\right), \ldots, F_d^{\leftarrow}\left(u_d\right)\right)}{\prod_{i=1}^{d} f_i\left(F_i^{\leftarrow}\left(u_i\right)\right)} \tag{3.7}$$

for densities f of F and $f_1, \ldots, f_d$ of the corresponding marginals.

Invariance Principal

Suppose the RVs $X_1, \ldots, X_d$ have continuous marginals and copula C. For strictly increasing functions $T_i : \mathbb{R} \to \mathbb{R}, i = 1, \ldots, d$, the RVs $T_1(X_1), \ldots, T_d(X_d)$ also have copula C.

□

Fréchet-Hoeffding Bounds

Let $C(\boldsymbol{u}) = C(u_1, \ldots, u_d)$ be any d-dimensional copula.

Then, for

$$W(\boldsymbol{u}) = \max\left\{\sum_{i=1}^{d} u_i - d + 1,\ 0\right\} \tag{3.8}$$

as well as

$$M(\boldsymbol{u}) = \min_{1 \le i \le d}\left\{u_i\right\}, \tag{3.9}$$

it holds that

$$W(\boldsymbol{u}) \le C(\boldsymbol{u}) \le M(\boldsymbol{u}), \quad \boldsymbol{u} \in [0,1]^d. \tag{3.10}$$

We call W the *lower Fréchet-Hoeffding bound* and M the *upper Fréchet-Hoeffding*

bound.

Note that W is a copula if and only if $d = 2$, whereas M is a copula for all $d \geq 2$ (more on this later in Section 3.2.1).

□

MORE ON COPULA THEORY (NOTES)

3.2 Copula Classes

In this section we will take a look at three very popular *copula classes*, namely *fundamental, elliptical and archimedean copulas*. For each class, we will present a few (parametric) *copula families* which are widely used.

3.2.1 Fundamental Copulas

Fundamental copulas are a basic class of copulas, which emerge directly from the copula framework and do not depend on any parametric components.

Independence Copula

It is well known that the joint probability of a finite set of events $E_i, i = 1, \ldots, k$, is equal to the product of the marginals if and only if the Events E_i are independent.

$$\mathrm{P}\left(\bigcap_{i=1}^{k} E_i\right) = \prod_{i=1}^{k} \mathrm{P}\left(E_i\right).$$

Analogously, the same concept applies when we talk about the *independence copula*

$$\Pi(\boldsymbol{u}) = \prod_{i=1}^{d} u_i. \tag{3.11}$$

As a result of Sklar's theorem the RVs u_i are independent if and only if their copula is the independence copula, i.e.

$$C(\boldsymbol{u}) = \Pi(\boldsymbol{u})$$

and thus the copula density would be

$$c(\boldsymbol{u}) = 1, \quad \boldsymbol{u} \in [0,1]^d.$$

□

From Equation 3.10, it is obvious that the Fréchet-Hoeffding bounds correspond to the extreme cases of perfect dependence between the RVs $X_i, i = 1, \ldots, d$.

Comonotonicity Copula

Consider the RVs $X_1, \ldots, X_d$ and strictly increasing transformations $T_1, \ldots, T_d$ and $X_i = T(X_i)$ for $i = 2, \ldots, d$. Making use of the *invariance principal*, it can be shown that these RVs have as copula the upper Fréchet-Hoeffding bound

$$M(\boldsymbol{u}) = \min\{u_1, \ldots, u_d\}.$$

Since there is perfect positive dependence between those RVs, we call M the *comonotonicity copula*. The number of dimensions d can be any finite number greater than or equal to 2 for M to be a copula, as the minimum remains well defined.

□

Countermonotonicity Copula

Similar to the comonotonic case, it can be shown that if two RVs X_1 and X_2 are perfectly negatively dependent, their copula is the lower Fréchet-Hoeffding bound

$$W(\boldsymbol{u}) = \max\left\{\sum_{i=1}^{d} u_i - d + 1,\ 0\right\}.$$

Therefore, W is known as the *countermonotonicity copula*. Because of the fact that countermonotonicity is not valid for a dimension greater than 2, we end up with the restriction $d = 2$ for W to be indeed a copula.

□

3.2.2 Elliptical Copulas

Copulas which can be derived from known multivariate distributions like for example the *Multivariate Normal (or Gaussian) Distribution* or the *Multivariate Student's t-Distribution* are called *implicit copulas*. *Elliptical copulas* are implicit copulas which arise via Sklar's theorem from elliptical distributions like the mentioned examples.

Gaussian Copula

Without loss of generality (w.l.o.g.), for a random vector $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{P})$ and *correlation matrix* $\boldsymbol{P}$, the *Gaussian copula (family)* is given by

$$C^{Ga}_{\boldsymbol{P}}(\boldsymbol{u}) = \Phi_{\boldsymbol{P}}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right), \tag{3.12}$$

where Φ is the CDF of $\mathcal{N}(0, \sigma^2)$ and $\Phi_{\boldsymbol{P}}$ is the CDF of $\mathcal{N}_d(\boldsymbol{0}, \boldsymbol{P})$.
There are special cases to this copula family, namely for $d = 2$ and correlation ρ, the *bivariate Gaussian copula* C^{Ga}_{ρ} is equivalent to

- the independence copula Π if $\rho = 0$,
- the comonotonicity copula M if $\rho = 1$ and
- the countermonotonicity copula W if $\rho = -1$

The density of the Gaussian copula is given by

$$c^{\mathrm{Ga}}_{\boldsymbol{P}}(\boldsymbol{u}) = \frac{1}{\sqrt{\det \boldsymbol{P}}} \exp\left(-\frac{1}{2}\boldsymbol{x}'\left(\boldsymbol{P}^{-1} - \boldsymbol{I}_d\right)\boldsymbol{x}\right), \tag{3.13}$$

where $\boldsymbol{x} = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d))$.

□

t-Copula

Consider w.l.o.g. $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{0}, \boldsymbol{P})$ (multivariate Student's t-distribution) with ν Degrees of Freedom (d.o.f.) and $\boldsymbol{P}$ a correlation matrix, then the *t-copula (family)* is given by

$$C^{t}_{\nu,\boldsymbol{P}}(\boldsymbol{u}) = t_{\nu,\boldsymbol{P}}\left(t^{-1}_{\nu}(u_1), \ldots, t^{-1}_{\nu}(u_d)\right), \tag{3.14}$$

where t_ν is the CDF of the univariate Student's t-distribution and $t_{\nu,\boldsymbol{P}}$ is the CDF of the multivariate Student's t-distribution (both with ν d.o.f.).
For the *bivariate t-copula* ($d = 2$), the special cases are the same as for the Gaussian copula except that $d = 0$ does not yield the independence copula (unless $\nu \to \infty$ in which case $C^{t}_{\nu,\rho} = C^{Ga}_{\rho}$).
The density of $C^{t}_{\nu,\boldsymbol{P}}$ is given by

$$c^{t}_{\nu,\boldsymbol{P}}(\boldsymbol{u}) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\sqrt{\det \boldsymbol{P}}}\left(\frac{\Gamma(\nu/2)}{\Gamma((\nu+1)/2)}\right)^d \frac{\left(1 + \boldsymbol{x}'\boldsymbol{P}^{-1}\boldsymbol{x}/\nu\right)^{-(\nu+d)/2}}{\prod_{j=1}^{d}\left(1 + x_j^2/\nu\right)^{-(\nu+1)/2}}, \tag{3.15}$$

where $\boldsymbol{x} = (t^{-1}_{\nu}(u_1), \ldots, t^{-1}_{\nu}(u_d))$.

□

3.2.3 Archimedean Copulas

Unlike implicit copulas, *explicit copulas* can be specified directly by taking into account certain constructional principles. The most important aspects of a such explicit copulas, in particular *archimedean copulas*, are showcased in this subsection. Archimedian copulas are of the general form

$$C(\boldsymbol{u}) = \phi^{-1}\left(\phi\left(u_1\right) + \cdots + \phi\left(u_d\right)\right), \tag{3.16}$$

where the function $\phi : [0,1] \to [0,\infty)$ is the *(archimedean) generator* and satisfies the following properties:

- ϕ is strictly decreasing in the entire domain $[0,1]$
- We set $\phi(1) = 0$
- If $\phi(0) = \lim\limits_{u \to 0^-} \phi(u) = \infty$, then ϕ is called *strict*.

Based on Equation 3.16 and according to the form of the generator, we can construct several copula families. Three of the most popular ones are the *Gumbel*, the *Clayton* and the *Frank copula*, which will be discussed.[1] The advantage of such copulas lies in the fact that they interpolate between certain fundamental dependency structures.

Clayton Copula

If the generator takes on the form

$$\phi_{Cl}(u) = \frac{1}{\theta}\left(u^{-\theta} - 1\right) \tag{3.17}$$

then we obtain the *Clayton copula* given by

$$C_\theta^{Cl}\left(u_1, u_2\right) = \left(\max\left\{u_1^{-\theta} + u_2^{-\theta} - 1, 0\right\}\right)^{-\frac{1}{\theta}}, \tag{3.18}$$

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

For $\theta > 0$ the generator of the Clayton copula is strict and we arrive at

$$C_\theta^{Cl}\left(u_1, u_2\right) = \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-\frac{1}{\theta}}. \tag{3.19}$$

Note that for $\theta = -1$, we obtain the lower Fréchet-Hoeffding bound, whereas for the limits $\theta \to 0$ and $\theta \to \infty$ we arrive at the independence copula and the comonotonicity copula respectively.

[1] We will look into these copulas for the bivariate case ($d = 2$) only.

□

Gumbel Copula

If the generator takes on the form

$$\phi_{Gu}(u) = (-\ln u)^{\theta}, \quad \theta \in [1, \infty), \tag{3.20}$$

then we arrive at the *Gumbel copula* given by

$$C_{\theta}^{Gu}(u_1, u_2) = \exp\left[-\left((-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}\right)^{\frac{1}{\theta}}\right]. \tag{3.21}$$

Note that for $\theta = 1$, we obtain the independence copula, while for $\theta \to \infty$ the Gumbel copula converges to the comonotonicity copula.

□

Frank Copula

If the generator takes on the form

$$\phi_{Fr}(u) = \ln\left(e^{-\theta} - 1\right) - \ln\left(e^{-\theta u} - 1\right), \quad \theta \in \mathbb{R}\backslash\{0\}, \tag{3.22}$$

we obtain the *Frank copula* given by

$$C_{\theta}^{Fr}(u_1, u_2) = -\frac{1}{\theta}\ln\left(1 + \frac{\left(e^{-\theta u_1} - 1\right)\cdot\left(e^{-\theta u_2} - 1\right)}{e^{-\theta} - 1}\right) \tag{3.23}$$

□

MAYBE MORE ON THESE WITH SOME PRETTY PLOTS

3.3 Dependence Measures

Dependence measures allow us to summarize a particular kind of dependence into a single number.[2] Recall the Fréchet-Hoeffding bounds (Equation 3.8 and Equation 3.9). They are an example of such kind of dependence measures. After all, they represent perfect negative or positive dependence. In this section, we will take a closer look into three classes of dependence measures along with appropriate association metrics.

[2] In the bivariate case

3.3.1 Linear Correlation

Undoubtedly, the most famous association metric for two RVs X_1 and X_2 is the *Linear or Pearson's correlation coefficient*

$$\rho\left(X_1, X_2\right) = \frac{\mathrm{Cov}\left(X_1, X_2\right)}{\sqrt{\mathrm{Var}(X_1)}\sqrt{\mathrm{Var}(X_2)}} \in [-1, 1]. \tag{3.24}$$

Note that $E(X_1) < \infty$ and $E(X_2) < \infty$ have to hold, i.e. the first two moments have to exist for ρ to be defined.

The Pearson correlation coefficient is interpretable for RVs which have (approximately) a linear relationship, where $\rho = -1$ indicates perfect negative linear correlation, $\rho = 1$ indicates perfect positive linear correlation and $\rho = 0$ indicates no correlation between X_1 and X_2. However, comprehensibility of this measure comes along with some drawbacks:

- A correlation of 0 is in general not equivalent to independence. This property holds only for normally distributed RVs.[3]
- ρ is invariant only under linear transformations, but not under transformations in general.
- Given the marginals and correlation ρ, one is able to construct a joint distribution only for the class of elliptical distributions. (MAYBE PLOT qrm)
- Given the marginals, only for elliptically distributed RVs any $\rho \in [-1, 1]$ is attainable.

3.3.2 Rank Correlation

To compensate some of the drawbacks of linear correlation, we take advantage of correlation measures based on the ranks of data. *Rank correlation coefficients*, like the ones presented below, are always defined and obey to the invariance principal. This means that these coefficients only depend on the underlying copula and they can thereof be directly derive.

Spearman's Rho

Consider two RVs X_1 and X_2 with continuous CDFs F_1 and F_2, then the

[3] e.g. $X_2 = X_1^2$ implies perfect dependence, yet $\rho(X_1, X_2) = 0$. Conversely though, independence always yields $\rho = 0$.

Spearman's rho correlation coefficient is simply the linear correlation between the CDFs

$$\rho_S = \rho\left(F_1\left(X_1\right), F_2\left(X_2\right)\right). \tag{3.25}$$

The reason being is that by applying the CDF to data, naturally a multiple of the ranks of the data are obtained, which essentially is equivalent to

$$\rho_S = \rho(Ran(X_1), Ran(X_2)) \tag{3.26}$$

Due to the invariance principle, we also obtain Spearman's rho directly from the unique copula via

$$\rho_{\mathrm{S}} = 12 \int_0^1 \int_0^1 C\left(u_1, u_2\right) \mathrm{d}u_1 \mathrm{d}u_2 - 3. \tag{3.27}$$

□

Kendall's Tau

Let $X_1 \sim F_1$ and $X_2 \sim F_2$ be two RV and let $(\tilde{X}_1, \tilde{X}_2)$ be an independent copy[4] of (X_1, X_2). Then *Kendall's tau* is defined by

$$\begin{aligned} \rho_\tau &= E\left[\operatorname{sign}\left(\left(X_1 - X_1'\right)\left(X_2 - X_2'\right)\right)\right] \\ &= P\left(\left(X_1 - X_1'\right)\left(X_2 - X_2'\right) > 0\right) - P\left(\left(X_1 - X_1'\right)\left(X_2 - X_2'\right) < 0\right). \end{aligned} \tag{3.28}$$

Similarly to Spearman's rho, using the invariance principal, we can directly derive Kendall's tau from the unique copula by

$$\rho_\tau\left(X_1, X_2\right) = 4 \int_0^1 \int_0^1 C\left(u_1, u_2\right) dC\left(u_1, u_2\right) - 1. \tag{3.29}$$

□

Both $\rho_S, \rho_\tau \in [-1, 1]$, where any value within this interval is attainable in contrast to the Pearson coefficient. If any of these rank correlations is -1 (or 1), we are in the countermonotonic (or comonotonic) case. If ρ_S (or ρ_{tau}) $= 0$, this does not necessarily imply independence between X_1 and X_2, although the opposite direction holds. Furthermore, they are not limited to be invariant just under linear transformations .

[4] An independent copy $\tilde{X}$ of a RV X is a RV that inherits from the same distribution as X and is independent of X.

3.3.3 Tail Dependence

Coefficients of tail dependence express the strength of the dependency in the extremes of distributions, i.e. the joint tails. We distinguish between *lower* and *upper tail dependence* between $X_j \sim F_j, j = 1, 2$ and provided that the below limits exist, they are given by

$$\lambda_l = \lim_{q \to 0^+} P\left(X_2 \leq F_2^{\leftarrow}(q) | X_1 \leq F_1^{\leftarrow}(q)\right) \tag{3.30}$$

and

$$\lambda_u = \lim_{q \to 1^-} P\left(X_2 > F_2^{\leftarrow}(q) | X_1 > F_1^{\leftarrow}(q)\right). \tag{3.31}$$

If λ_l (or λ_u) $= 0$, then we say that X_1 and X_2 are *asymptotically independent* in the lower (or upper) tail,[5] otherwise we have lower (or upper) tail dependence.
For continuous CDFs and by using Bayes' theorem, these expressions can be re-written to

$$\begin{aligned} \lambda_l &= \lim_{q \to 0^+} \frac{P\left(X_2 \leq F_2^{\leftarrow}(q), X_1 \leq F_1^{\leftarrow}(q)\right)}{P\left(X_1 \leq F_1^{*}(q)\right)} \\ &= \lim_{q \to 0^+} \frac{C(q,q)}{q} \end{aligned}$$

and similarly

$$\lambda_u = 2 - \lim_{q \to 1^-} \frac{1 - C(q,q)}{1 - q}.$$

Therefore, tail dependencies can be assessed by means of the copula itself when approaching the points $(0, 0)$ and $(1, 1)$. In addition, for all radially symmetric copulas (e.g. the bivariate Gaussian or the t-copula) we have $\lambda_l = \lambda_u = \lambda$.
Some examples are:

- Clayton: $\lambda_l = 2^{-1/\theta}$, $\lambda_u = 0$ (only lower tail dependence)
- Gumbel: $\lambda_l = 0$, $\lambda_u = 2 - 2^{1/\theta}$ (only upper tail dependence)
- Frank: $\lambda_l = 0$, $\lambda_u = 0$ (no tail dependence)

Following such guidelines, the choice of a practicable copula can be facilitated.

3.4 Conditional Copulas

3.5 Vine Copulas

[5] Not necessarily true for the other way around

4 Data Exploration

5 Modelling

6 Conclusion

Appendix

Include appendix here...

List of Figures

List of Tables

List of Abbreviations

BIC Bayesian Information Criterion

GLM Generalized Linear Model

LM Linear Regression Model

LMM Linear Mixed Model

GLMM Generalized Linear Mixed Model

CDF Cumulative Distribution Function

PDF Probability Density Function

RV Random Variable

w.l.o.g. Without loss of generality

d.o.f. Degrees of Freedom

GAM Generalized Additive Model

References

[Embrechts et al. 2001] Embrechts, Paul ; Lindskog, Filip ; McNeil, Alexander: Modelling dependence with copulas. In: *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich* 14 (2001)

[Fahrmeir et al. 2003] Fahrmeir, L. ; Kneib, T. ; Lang, S. ; Marx, B.: *Regression; Models, Methods and Applications. 2013*. 2003

[Hofert et al. 2019] Hofert, Marius ; Kojadinovic, Ivan ; Mächler, Martin ; Yan, Jun: *Elements of copula modeling with R*. Springer, 2019

[Klein and Kneib 2016] Klein, Nadja ; Kneib, Thomas: Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. In: *Statistics and Computing* 26 (2016), Nr. 4, S. 841–860

[Lütkepohl 2005] Lütkepohl, Helmut: *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005

[McNeil et al. 2015] McNeil, Alexander J. ; Frey, Rüdiger ; Embrechts, Paul: *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015

[Ruppert and Matteson 2015] Ruppert, David ; Matteson, David S.: *Copulas*. S. 183–215. In: *Statistics and Data Analysis for Financial Engineering: with R examples*. New York, NY : Springer New York, 2015. – URL `https://doi.org/10.1007/978-1-4939-2614-5_8`. – ISBN 978-1-4939-2614-5

[Schmidt 2007] Schmidt, Thorsten: Coping with copulas. In: *Copulas-From theory to application in finance* (2007), S. 3–34

[Sklar 1959] Sklar, M.: Fonctions de repartition an dimensions et leurs marges. In: *Publ. inst. statist. univ. Paris* 8 (1959), S. 229–231

[Vatter and Chavez-Demoulin 2015] Vatter, Thibault ; Chavez-Demoulin, Valérie: Generalized additive models for conditional dependence structures. In: *Journal of Multivariate Analysis* 141 (2015), S. 147–167

[Vatter and Nagler 2018] Vatter, Thibault ; Nagler, Thomas: Generalized additive models for pair-copula constructions. In: *Journal of Computational and Graphical Statistics* 27 (2018), Nr. 4, S. 715–727

[Vatter and Nagler 2019] Vatter, Thibault ; Nagler, Thomas: gamCopula-package: Generalized Additive Models for Bivariate Conditional... (2019)

[Wood 2017] Wood, Simon N.: *Generalized additive models: an introduction with R*. CRC press, 2017