GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Master Thesis

Multivariate modelling of the dependence 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 April 13, 2020

Processing time of 20 weeks

Confidentiality Clause

Statutory Declaration

Acknowledgments

Contents

1 Introduction

Write in general something like an "abstract", what is the pipeline of the thesis as a whole..

Quick test for user specified compiling and citing:

Lütkepohl and Herwartz [1996]

[R Core Team, 2018]

R Core Team [2018]

1.1 adidas

After World War II, the *"Dassler Brothers Shoe Factory"* (German: *"Gebrüder Dassler Schuhfabrik"*), which was led by *Adolf Dassler* (aka *Adi Dassler*) and his brother Rudolph, was dissolved. The brothers split up and formed their own firms. As a result, the sports shoe factory *"Adi Dassler adidas Sportschuhfabrik"* was founded on August 18th 1949 by Adolf Dassler in Herzogenaurach, a small town in Germany [adidas-group.com].

(a) adidas Performance

(b) adidas Originals

Figure 1.1: Two of the adidas Brand Logos: Performance (left) & Originals (right) [adidas.com media-center]

Today, just over 70 years later, the sportswear designer and manufacturer is known as the *"adidas AG"* (short: *adidas*) and is one of the world's biggest sports and fashion brands. The global headquarters of are located in the birthplace Herzogenaurach and the company is employing over 59,000 people worldwide, with *Kasper Rørsted* leading the brand as CEO since October 1st 2016. In 2019, adidas produced over 1.1 billion sports and sports lifestyle products and is nowadays sponsoring a vast range of athletes, artists and organizations across the globe (e.g. the FIFA World Cup™).

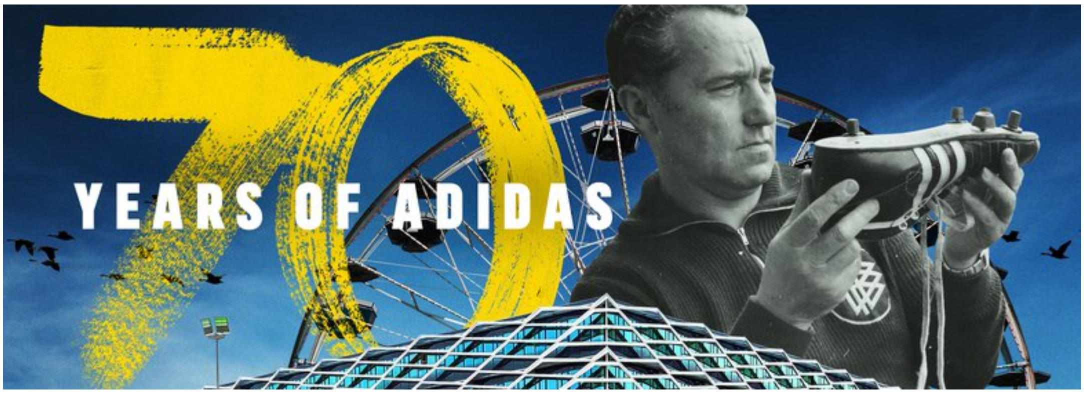

Figure 1.2: adidas celebrates its 70th anniversary and the opening of the ARENA building [adidas 70 years, 2019]

More on DNA, DS&AI, etc...?

1.2 Data Sources

Throughout each season, transactional data are collected from online purchases of the sports brand's e-Commerce 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	Total markdown percentage of the article	Between *0* and *1*
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 e-Commerce 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 dependence structure between net sales of articles.

Besides the transactional data, 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 attribute data

Overall, these are the primary data sources and 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 instances and over 30 variables.

1.3 Motivation

What is the motivation and purpose of this thesis...

Write also towards the end..

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 (such as 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, ...), 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

$$E(y_i) = h(\eta_i) \tag{2.2}$$

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

$$g = h^{-1}, \tag{2.3}$$

is called the *link function* and transforms the mean of the response's distribution to an unbounded continuous scale.

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]. Mathematically speaking, the linear predictor $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$ is extended to

$$\eta_{ij} = \boldsymbol{x'}_{ij}\boldsymbol{\beta} + \boldsymbol{u'}_{ij}\boldsymbol{\gamma}_i \tag{2.4}$$

for individuals $i = 1, \ldots, m$ measured in a longitudinal setting at observed times $t_{i1} < \ldots < t_{ij} < \ldots < t_{in_i}$ or for subjects $j = 1, \ldots, n_i$ in cluster $i = 1, \ldots, m$.
In any case,

- $\boldsymbol{\beta}$ 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*[1] specified as

$$y_{ij} = \boldsymbol{x'}_{ij}\boldsymbol{\beta} + \boldsymbol{u'}_{ij}\boldsymbol{\gamma}_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right) \tag{2.5}$$

or in matrix notation

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i \tag{2.6}$$

with $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$ and $\boldsymbol{\gamma}_i \sim N(\boldsymbol{0}, \mathbf{Q})$ for individuals or clusters $i = 1, \ldots, m$ and positive definite matrix $\mathbf{Q}$. Note that $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_m$ are assumed to be mutually independent.

Similar to GLMs, *Generalized Linear Mixed Models (GLMMs)* relate the linear mixed predictor (Equation 2.4) to the conditional mean $\mu_{ij} = E(y_{ij}|\boldsymbol{\gamma}_i)$ via a

[1] More details on the vector and matrix forms can be found in the appendix.

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

$$y_i = \eta_i^{add} + \varepsilon_i, \tag{2.7}$$

where

$$\eta_i^{add} = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin}, \quad i = 1, \ldots, n. \tag{2.8}$$

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

$$f_j(z_j) = \sum_{l=1}^{d_j} \gamma_{jl} B_l(z_j) \tag{2.9}$$

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-parametric 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 2.7 can be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.10}$$

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 smooth functions $f_j(z_j)$ are centered around zero, such that

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

A convenient trait of additive models is that they also support the incorporation of random effects. Random coefficient terms can straightforwardly be added to the model. Analogously to Section 2.2, we consider data measured in a longitudinal setting with individuals $i = 1, \ldots, m$ observed at times $t_{i1} < \ldots < t_{ij} < \ldots < t_{in_i}$ or clustered data with subjects $j = 1, \ldots, n_i$ in clusters $i = 1, \ldots, m$. Without loss of generality (w.l.o.g.),[2] we can simply add to Equation 2.10 the terms $\boldsymbol{Z}_0\boldsymbol{\gamma}_0$ and $\boldsymbol{Z}_1\boldsymbol{\gamma}_1$ representing the design matrices and coefficients of the random intercepts and random slopes respectively. Explicitly, the coefficients are formulated as $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0i}, \ldots, \gamma_{0m})'$ and $\gamma_1 = (\gamma_{11}, \ldots, \gamma_{1i}, \ldots, \gamma_{1m})'$, whereas the design matrices are expressed as

$$\boldsymbol{Z}_0 = \begin{pmatrix} \mathbf{1}_1 & & & & \mathbf{0} \\ & \ddots & & & \\ & & \mathbf{1}_i & & \\ & & & \ddots & \\ & & & & \mathbf{1}_m \end{pmatrix} \quad \boldsymbol{Z}_1 = \begin{pmatrix} \boldsymbol{x}_1 & & & & \mathbf{0} \\ & \ddots & & & \\ & & \boldsymbol{x}_i & & \\ & & & \ddots & \\ & & & & \boldsymbol{x}_m \end{pmatrix}.$$

More details and technicalities regarding mixed effects in additive models can be found in the appendix.

Extensions of additive models to non-normal responses are consequently called *Generalized Additive Models (GAMs)*, which were first introducd by Hastie and Tibshirani [1986]. If additionally random effects are included, we call them *Generalized Additive Mixed Models (GAMMs)*.

Thus far, we have examined models with main effects and conceivably random effects. Accordingly, these types of effects can likewise be combined with covariate interactions and/or spatial effects. Such models can be described in a unified framework and are titled as (possibly *Generalized*) *Structured Additive Regression Models (STARs)*,

$$y = f_1(\nu_1) + \ldots + f_q(\nu_q) + \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon.$$

The covariates $\nu_1, \ldots, \nu_q$ can be one- or multidimensional and the functions can be of different structure determining the type of effect.

[2]for the indexes

3 Copulas & Dependence 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

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

it holds that

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

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 CDF of a finite set of RVs $X_i, i = 1, \ldots, n$, is equal to the product of the marginals if and only if the RVs X_i are mutually independent, i.e.

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i))$$

$\forall x_1, \ldots, x_n$.
Equally, the exact 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

W.l.o.g., for a random vector $\boldsymbol{X} \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{P})$ and *correlation matrix* $\boldsymbol{P}$, the *Gaussian copula (family)* is given by

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

where Φ is the CDF of $\mathcal{N}(0, \sigma^2)$ and $\Phi_{\boldsymbol{P}}$ is the CDF of $\mathcal{N}_d(\mathbf{0}, \boldsymbol{P})$.

There are special cases to this copula family, namely for $d = 2$ and correlation ρ, the *bivariate Gaussian copula* C_ρ^{Ga} 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

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

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

□

t-Copula

Consider w.l.o.g. $\boldsymbol{X} \sim t_d(\nu, \mathbf{0}, \boldsymbol{P})$ (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

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

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 $C_{\nu,\rho}^t = C_\rho^{Ga}$).

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 $\boldsymbol{x} = \left(t_\nu^{-1}\left(u_1\right), \ldots, t_\nu^{-1}\left(u_d\right)\right)$.

□

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.[3] 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

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

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

then we obtain the *Clayton copula* given by

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

where $\theta \in [-1, \infty)\backslash\{0\}$.
For $\theta > 0$ the generator of the Clayton copula is strict and we arrive at

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

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.

□

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.[4] 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.

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*

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

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.[5]
- 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.

[4] In the bivariate case

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

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 derived.

Spearman's Rho

Consider two RVs X_1 and X_2 with continuous CDFs F_1 and F_2, then the *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[6] 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

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

[6] 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.

□

Both $\rho_S, \rho_\tau \in [-1, 1]$ and any value within this interval is attainable for an arbitrary copula class 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 .

3.3.3 Tail Dependence

Coefficients of tail dependence express the strength of the dependence 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,[7] 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)

[7] Not necessarily true for the other way around

- 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 Structured Additive Conditional Copulas

Modelling of the marginal response distributions along with their dependence structure has been studied so far in a strictly parametric context, not considering any potentially available covariate information. In this section, we will broaden the copula framework by adding conditions given possible covariates for all model parameters, i.e. both for the parameters of the marginals as well as the copula parameter. All involved model parameters will receive *structured additive predictors* (see Section 2.3) to account for possible non-linear or random effects. We will summarily explore *Structured Additive Conditional Copulas* and for extensive literature, we refer to Klein and Kneib [2016] and Vatter and Nagler [2019].

To get started, we define $(Y_1, Y_2)'$ to be independent bivariate responses[8] and $\boldsymbol{\nu}$ being the information contained in covariates. Ergo, Equation 3.4 of Sklar's theorem can be extended to the conditional case

$$F_{1,2}\left(Y_1, Y_2|\boldsymbol{\nu}\right) = C\left(F_1\left(Y_1|\boldsymbol{\nu}\right), F_2\left(Y_2|\boldsymbol{\nu}\right)|\boldsymbol{\nu}\right) \tag{3.32}$$

in conjunction with all facets of Section 3.1 [Patton, 2006].
The marginal CDFs $F_d\left(y_{id}|\boldsymbol{\nu}_i\right)$ for observations $i = 1, \ldots, n$ can also be stated as

$$F_d\left(y_{id}|\vartheta_{i1}^{(d)}, \ldots, \vartheta_{iK_d}^{(d)}\right), \quad d = 1, 2, \tag{3.33}$$

i.e. the distribution F_d has a total of K_d parameters, denoted as $\vartheta_{i1}^{(d)}, \ldots, \vartheta_{iK_d}^{(d)}$. To relate all parameters of the marginals to structured additive predictors $\eta_i^{\vartheta_k^{(d)}}, k = 1, \ldots, K_d$ consisting of the covariates $\boldsymbol{\nu}_i$ (see Section 2.3), we employ strictly increasing response mappings $h_k^{(d)}$ to ensure proper domain allocation, i.e.

$$\vartheta_{ik}^{(d)} = h_k^{(d)}(\eta_i^{\vartheta_k^{(d)}}). \tag{3.34}$$

[8]Continuous responses in the paper, but look at "A note on identification of bivariate copulas for discrete count data" to excuse this when explanatory variables are involved.

Assuming that the parameters of the copula can also depend on covariates $\boldsymbol{\nu}_i$ while Sklar's theorem applies as usual, the left-hand side of Equation 3.32 can equivalently be stated as

$$F_{1,2}(y_{i1}, y_{i2}|v_i) = F_{1,2}(y_{i1}, y_{i2}|\vartheta_{i1}^{(1)}, \ldots, \vartheta_{iK_1}^{(1)}, \vartheta_{i1}^{(2)}, \ldots, \vartheta_{iK_2}^{(2)}, \vartheta_{i1}^{(c)}, \ldots, \vartheta_{iK_c}^{(c)}),$$

where the last share of parameters $\vartheta_{i1}^{(c)}, \ldots, \vartheta_{iK_c}^{(c)}$ belong to the copula. Similar to Equation 3.34, the copula parameters are modelled as $\vartheta_{ik}^{(c)} = h_k^{(c)}(\eta_i^{\vartheta_k^{(c)}})$ with K_c being the number of parameters.

3.5 Vine Copulas

Vine copulas to be written down...

$\boldsymbol{x}_i = (xi1, \ldots, xid)T$

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

GAMM Generalized Additive Mixed Model

STAR Structured Additive Regression Model

References

[adidas 70 years 2019] adidas 70 years: *adidas celebrates its 70th anniversary and the opening of the ARENA building.* 2019. – `https://www.adidas-group.com/en/media/news-archive/press-releases/2019/adidas-celebrates-70th-anniversary`, Last accessed on 2020-04-01

[adidas-group.com] adidas-group.com: *adidas-group*. – `https://www.adidas-group.com/en/`, Last accessed on 2020-04-01

[adidas.com media-center] adidas.com media-center: *adidas media-center. Pictures and Videos*. – `https://www.adidas-group.com/en/media/media-center/`, Last accessed on 2020-04-01

[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

[Hastie and Tibshirani 1986] Hastie, Trevor ; Tibshirani, Robert: Generalized Additive Models. In: *Statist. Sci.* 1 (1986), 08, Nr. 3, S. 297–310. – URL `https://doi.org/10.1214/ss/1177013604`

[Hastie and Tibshirani 1990] Hastie, Trevor J. ; Tibshirani, Robert J.: Generalized additive models, volume 43 of. In: *Monographs on statistics and applied probability* 15 (1990)

[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

[Lütkepohl and Herwartz 1996] Lütkepohl, Helmut ; Herwartz, Helmut: Specification of varying coefficient time series models via generalized flexible least squares. In: *Journal of Econometrics* 70 (1996), Nr. 1, S. 261–290

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

[Patton 2006] Patton, Andrew J.: Modelling asymmetric exchange rate dependence. In: *International economic review* 47 (2006), Nr. 2, S. 527–556

[R Core Team 2018] R Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (Veranst.), 2018. – URL `https://www.R-project.org/`

[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

[Trivedi and Zimmer 2017] Trivedi, Pravin ; Zimmer, David: A note on identification of bivariate copulas for discrete count data. In: *Econometrics* 5 (2017), Nr. 1, S. 10

[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