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 12, 2020

Processing time of 20 weeks

# Contents

# 1 Introduction

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

## 1.1 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)<br>(Spring-Summer [SS]: December - May)<br>Fall-Winter [FW]: June - November) | Factors:<br>*SS17, FW17,*<br>*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.

# 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

The *Generalized Linear Model (GLM)* is 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 Generalized Additive Models

## 2.3 Mixed Effects Models

The *Linear Mixed Model (LMM)* is powerful tool 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 $\varepsilon_{ij}$ 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, the *Generalized Linear Mixed Model (GLMM)* relates 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.

# 3 Copulas & Dependence Modelling

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 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 in $[0,1]$.

**Sklar's Theorem**

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

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{3.1}$$

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 3.1 is a $d$-dimensional CDF.

□

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

$$c(\boldsymbol{u}) = \frac{\partial^d C\left(u_1, \ldots, u_d\right)}{\partial u_1 \cdots \partial u_d} \tag{3.2}$$

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

By virtue of Equation 3.1 in Sklar's theorem and given that

$$C(\boldsymbol{u}) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right), \tag{3.3}$$

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

$$c(u_1, \ldots, u_d) = \frac{f\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right)}{\prod_{i=1}^{d} f_i\left(F_i^{-1}(u_i)\right)} \tag{3.4}$$

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

**Invariance Principal**

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

□

**The Fréchet-Hoeffding Bounds**

Let $C(\boldsymbol{u}) = C(u_1, \ldots, u_d)$ be any $d$-dimensional copula.
Then for $W(\boldsymbol{u}) = \max\left\{\sum_{i=1}^{d} u_i - d + 1, 0\right\}$ and $M(\boldsymbol{u}) = \min_{1 \le i \le d}\{u_i\}$,
it holds that

$$W(\boldsymbol{u}) \le C(\boldsymbol{u}) \le M(\boldsymbol{u}), \quad \boldsymbol{u} \in [0,1]^d \tag{3.5}$$

Note that $W$ is a copula if and only if $d = 2$, whereas $M$ is a copula for all $d \ge 2$

□

# 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

# References

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

[Hofert et al., 2019] Hofert, M., Kojadinovic, I., Mächler, M., and Yan, J. (2019). *Elements of copula modeling with R*. Springer.

[Lütkepohl, 2005] Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

[McNeil et al., 2015] McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press.

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

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