Malte Bonart

Testing for Structural Breaks in Factor Copula Models - Implementation and Application in Social Media Topic Analysis

Master thesis

Supervisor: Prof.Dr. Dominik Wied

Submitted for the Master Examination in Economics at the Faculty of Management, Economics and Social Sciences of the University of Cologne in June 2018.

This is my abstract.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

cdf   Cumulative distribution function

iid   Identically and independent distributed

SMM   Simulated methods of moments

EDF   Empirical distribution function

DGP   Data generating process

btw17   Bundestag election 2017

API   Application programming interface

1. Introduction

Statistical models involving copula functions became increasingly popular since the 1990th (Nelsen 1999, p. 1). The concept was first introduced in the work by Sklar (1959). Copulas are mainly used in two ways: First, to model the dependence structure of multivariate distributions independent of their underlying marginal distributions and second, to construct bivariate or multivariate distributions (Sempi 2011, p. 302). This paper focuses on the first application. Up to the year 1999 the there are 38 publication listed for the topics *copula* and *dependence.* From 2000 - 2008 there were 334 publication and from 2009 - 2017 this number increased to 2048 publication.[1]

Skalr's theorem can be used to construct multivariate models first, by specifying the marginal distributions of the random variables involved and second, by specifying the dependence structure among the variables via a copula function Sempi 2011. By doing so, one allows for non-parametrized or semi-parametrized estimation of the marginal distributions together with a parametrized copula. High dimensional problems become traceable since the number of parameters can be drastically reduced (Patton 2009, p. 777).

For time series data copula theory can be used in two ways: First, to describe the cross sectional dependence structure by estimating the conditional copula function of the conditional joint distribution $F(\mathbf{y}|\mathcal{F}_{t-1})$ with $\mathbf{Y}_t = [Y_{1t}, \ldots, Y_{nt}]'$ and past information $\mathcal{F}_{t-1}$. To obtain a valid distribution, the information set must be the same for both the copula and the marginal distributions. Second, copulas can be used to describe the dependence between observations of a univariate time series $[Y_t, Y_{t+1}, \ldots, Y_{t+n}]'$. This is related to the study of Markov processes. (Patton 2009, p. 771 ff). This paper focuses on the first application.

Applications for copula modeling can be found in various disciplines but they became increasingly popular in the field of finance, actuarial science and hydrology Sempi 2011.

Correlation or covariance matrices can be used to model linear dependence especially for multivariate normal or t-distributions. But they lack the ability to model the dependence e.g. in the presence of heavy tails or outliers (Kumar 2011).

---

[1]Numbers are taken from the *Web of Science Core Collection* for the search term `TOPIC: (copula) AND TOPIC: (dependence)`.

Rank correlation matrices such as *Spearman's rho* are invariant under monotonic transformations but they are not moment-based.

Copulas are usefull whenever one wants to draw conclusion about the full joint distribution but other models are more suited if one focuses on predicting the conditional mean (Patton 2009, p. 769)

This thesis summarizes and strucutures the current developement in the field of factor copula models. It contributes to the sicentific disucssion in two ways: First, a software library for a consistent specification, simulation and estimation of factor copula models is presented. With the library, the methods and the strucutral break test are made avalaible to a broader scientific audience and can be used by applied reasearchers. Second, we study novel ways of applying the discussed methods to areas outside of the finance community. The strucutral break test is applied to an aggregated dataset of social media posts from german politicians and political parties.

The thesis is structured in four main chapters: The first chapter lays the theoretical foundation by summarizing important aspects of copula theory and by presenting the factor copula approach, its estimation strategy via simulated methods of moments and a suitable test for time varying dependence structures. The second chapter presents implementational details of the software library *factorcopula*, written in the statistical programming language R (Bonart 2018). To ensure the validity of the package and its performance two simulation studies are performed. In the last chapter, the methods are applied to a large dataset of textual social media posts from german politicians and political parties. The goal is to illustrate and assess novel ways of applying the factor copula approach and the structural break test.

2. Theoretical foundation

In this chapter we present and summarize the theoretical foundation of this thesis. First, the general idea behind copula functions is introduced. Second, a special class of copula models, the so called *factor copula* model is presented. Third, we discuss copula models in the context of time series data and present a specific framework for modelling dependencies of multivariate time series. Fourth, the *simulated methods of moments* is explained. It is the estimation method which is used throughout this work. Finally we summarize the ideas

of a structural break test for the possibly time varying parameters of a factor copula model.

2.1. Copula theory

The joint cumulative distribution function $F_Y(y_1, \ldots, y_N) = P(Y \leq y_1, \ldots, Y \leq y_N)$ for some multivariate random variable Y of dimension N and its corresponding marginal distributions $F_{Y_i}(y_i) = P(Y_i \leq y_i) \ \forall \ i = 1, \ldots, N$ can be used to describe a joint probability distribution. Estimating F_Y is computationally demanding espacialy if N becomes large. Therefore, a copula function is introduced which can be used to link the marginal and the joint distribution function.

A function of the type $C : [0,1]^N \to [0,1]$, with $N \geq 2$ is called a *copula* if first, it shares the common properties of a multivariate distribution function such that $C_U(u_1, \ldots, u_N) = P(U_1 \leq u_1, \ldots, U_N \leq u_N)$ and if second, its marginal distributions are $U_i \sim U(0,1)$, e.g. uniformly distributed (Joe 2015, p. 7). A thereom by Sklar (1959) states, that every d-variate distribution function $F_Y(y_1, \ldots, y_n)$ can be eypressed in terms of its marginal distributions and a copula function such that

$$F(y_1, \ldots, y_N) = C(F_1(y_1), \ldots, F_N(y_N)). \tag{2.1}$$

If F_Y is *continuous* with marginal quantile functions $F_{Y_1}^{-1}, \ldots, F_{Y_N}^{-1}$ then $C(\mathbf{u})$ is uniquely determined by $C(\mathbf{u}) = F(F_{Y_1}^{-1}(u_1), \ldots, F_{Y_N}^{-1}(u_N))$.

Some bivariate measures of dependency share the property of *scale invariance*. Thus, the measures are invariant with respect to the marginal distributions and can therefore expressed as a function of their copula (Schmid et al. 2010, p. 210). The widely used linear correlation coefficient is dependend on both the copula and the marginal distributions (Patton 2009, p. 768). Two scale invariant measures are Spearman's rank correlation

$$\rho_{Y_i, Y_j} = 12 \int \int_{[0,1]^2} u_1 u_2 \, dC(u_i, u_j) - 3, \tag{2.2}$$

and Kendal's rank correlation

$$\tau_{Y_i, Y_j} = 4 \int \int_{[0,1]^2} C(u_i, u_j) dC(u_i, u_j) - 1 = 4E(C(U_i, U_j)) - 1. \qquad (2.3)$$

Lower and upper tail dependency is defined as

$$
\begin{aligned}
\tau^L_{Y_i, Y_j} &= \lim_{q \to 0} \frac{P(Y_i \le F^{-1}_{Y_i}(q), Y_j \le F^{-1}_{Y_j}(q))}{q} \\
\tau^U_{Y_i, Y_j} &= \lim_{q \to 1} \frac{P(Y_i > F^{-1}_{Y_i}(q), Y_j > F^{-1}_{Y_j}(q))}{q}
\end{aligned}
\qquad (2.4)
$$

2.2. Factor copulas

Factor copulas are a family of copula models for which the copula function $C(u_1, \ldots, u_N)$ is based on a latent factor structure as defined in Oh and Patton (2017, p. 140 ff).

Consider a set of artificial variables $X_i, i = 1, \ldots, N$ which linearly depend on some latent factors $Z_k, k = 1, \ldots, K$ and some error e_i such that $X_i = \sum_{k=1}^{K} \beta_{ik} Z_k + \epsilon_i$, with the linear coefficients β_{ik}. The latent variables Z_k and the error term ϵ_i follow some parametrized distributions, e.g. $\epsilon_i \overset{iid}{\sim} F_\epsilon(\boldsymbol{\gamma_\epsilon})$ and $Z_k \sim F_{Z_k}(\boldsymbol{\gamma_{Z_k}})$ with $Z_i \perp Z_j \forall i \ne j$, $Z_k \perp \epsilon_i \forall i, k$ and $\boldsymbol{\gamma_\epsilon}, \boldsymbol{\gamma_{Z_k}}$ some distribution specific parameter vectors.

The joint probability function $F_X(x_1, \ldots, x_N)$ of the artificial variables can then be expressed in terms of its marginal distributions $F_{X_i}(x)$ and a factor copula function $C_U(u_1, \ldots, u_N)$ such that $F_X(x_1, \ldots, x_N) = C_U(F_{X_1}(x_1), \ldots, F_{X_N}(x_N); \boldsymbol{\theta})$.

Thus, the factor copula model is completely defined by specifiying the parameter vector $\boldsymbol{\theta} = (\beta_{11}, \ldots, \beta_{i1}, \ldots, \beta_{ik}, \boldsymbol{\gamma'_{Z_1}}, \ldots, \boldsymbol{\gamma'_{Z_K}}, \boldsymbol{\gamma'_\epsilon})'$. The number of latent variables K and the distribution functions $F_{Z_1}, \ldots, F_{Z_k}, F_\epsilon$ are hyper-parameters of the model which have to be chosen prior to the estimation.[2]

The artifical variables are only used to construct the factor copula function $C_U(u_1, \ldots, u_N)$ which is then used to model the joint distribution of the observable variables such that $F_Y(y_1, \ldots, y_n) = C_U(F_{Y_1}(y_1), \ldots, F_{Y_N}(y_N))$. Once

---

[2]Oh and Patton (2017, p. 143ff) provide a heuristic of finding the number of latent variables by analyzing so called *scree-plots*: Ordered eigenvalues from the sample rank-correlation matrix of the data.

this copula function is determined, the artificial variables and its marginal distributions $F_{X_i}(x)$ are of no interest. Using matrix notation, the model can be summarized in the following set of equations:

$$
\begin{aligned}
\boldsymbol{Y} &= (Y_1, \ldots, Y_N)' \\
\boldsymbol{X} &= (X_1, \ldots, X_N)' = \boldsymbol{\beta Z} + \boldsymbol{\epsilon} \\
F_Y &= C_U(F_{Y_1}(y_1), \ldots, F_{Y_N}(y_N); \boldsymbol{\theta}) \\
F_X &= C_U(F_{X_1}(x_1), \ldots, F_{X_N}(x_N); \boldsymbol{\theta})
\end{aligned}
\tag{2.5}
$$

This approach allows for a two-stage estimation: First, the marginal distributions $\hat{F}_{Y_i}$ are estimated parametrically or non-parametrically, e.g. by using the empirical distribution function. Second, the factor structure for the possibly high dimensional copula function is fitted to the data by finding $\hat{\boldsymbol{\theta}}$. For the factor copula and the joint distribution of the artificial variables as defined in (2.5) a closed form usually does not exist. Therefore, one has to rely on simulation methods as described in section 2.4.

A lower bound for the number of parameters $P = |\boldsymbol{\theta}|$ to be estimated is given by the size of the factor matrix $\boldsymbol{\beta}$ which is $|\boldsymbol{\beta}| = N \times K$. To reduce the number of parameters Oh and Patton (2017) present two restrictions on $\boldsymbol{\beta}$: the *equidependence* and the *block-equidependence* model.

For the first model it is assumed that $K = 1$ and $\boldsymbol{\beta} = (\beta, \ldots, \beta)'$. Thus, the model consists of a single latent factor and a single factor loading β which is the same for all variables. This implies that each pairwise dependency is the same for all observable variables.
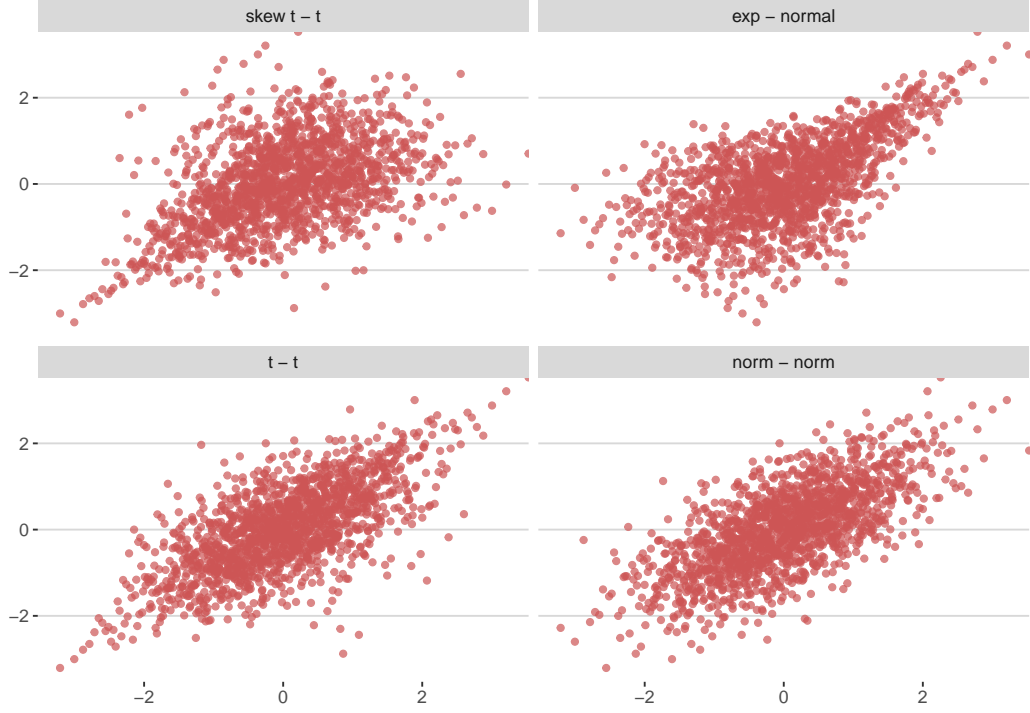
Figure 2.1: Illustration of different equidependence factor copula models with $N = 2$, $\beta = 1.5$, $Y_i \sim N(0,1)$ and different distributions for the latent variable and the error term.

Figure 2.1 shows four different simulations from a two dimensional one factor equidependence factor copula model. The marginal distributions are standard normal, the linear coefficient is fixed at $\beta = 1.5$ but the distributions of the latent variable and the error term differ. All models produce positive dependence and the rank correlation

The block-equidependence model is less restrictive and is especially suitable for variables which can be naturally partitioned into different groups.[3] The model assumes a common factor for all groups and a group specific factor for each group. Thus, each variable is only affected by two factors. For the factor matrix, it is further assumed that all variables in the same group have the same factor loading while variables in different groups can have different loadings. This implies that the pairwise intra-group dependencies are equal while the pairwise inter-group dependencies can vary between the groups.

Formally, consider a partition of $\boldsymbol{X} = (X_1, \ldots, X_N)'$ into D groups X_j^i, where $i = 1, \ldots, D$, $j = 1, \ldots, s_i$ and s_i the number of variables in group i. Then the

---

[3]E.g. this could be stock market prices grouped into different industry sectors.

model can be summarized as:

$$\boldsymbol{X} = (X_1^1, \ldots, X_{s_1}^1, \ldots, X_1^D, \ldots, X_{s_D}^D)' = \boldsymbol{\beta}\boldsymbol{Z} + \boldsymbol{\epsilon}$$

$$\boldsymbol{Z} = (Z_0, Z_1, \ldots, Z_D)'$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta^1 & \beta^{D+1} & 0 & \cdots & 0 \\ \beta^1 & \beta^{D+1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta^1 & \beta^{D+1} & 0 & \cdots & 0 \\ \beta^2 & 0 & \beta^{D+2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta^D & 0 & 0 & \cdots & \beta^{D+D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta^D & 0 & 0 & \cdots & \beta^{D+D} \end{pmatrix}, \tag{2.6}$$

where $\boldsymbol{\beta}$ is of size $N \times (D+1)$, $N = \sum_{i=1}^{D} s_i$ but with only $2D$ different factor loadings.

2.3. Copula models for multivariate time series

Sklar's theorem in (2.1) which shows the link between a copula function and the multivariate distribution can be easily extended to the multivariate time series case: Consider the random vector $Y_t = (Y_{1t}, \ldots, Y_{Nt})$. The goal is to model the conditional multivariate distribution of $\boldsymbol{Y}_t | \mathcal{F}_{t-1}$, where the σ-algebra $\mathcal{F}_{t-1}$ possibly contains past information and information from other exogenous variables $\{\boldsymbol{Y}_{t-1}', \boldsymbol{Y}_{t-2}', \ldots, \boldsymbol{X}_t', \boldsymbol{X}_{t-1}', \ldots\}$.

Using the definition of a conditional copula as in (Patton 2006) one can write

$$F_{Y_t | \mathcal{F}_{t-1}}(y_1, \ldots, y_N) = C_{Y_t | \mathcal{F}_{t-1}}(F_{Y_1 | \mathcal{F}_{t-1}}(y_1), \ldots, F_{Y_N | \mathcal{F}_{t-1}}(y_N)).$$

To have a valid conditional multivariate distribution, the conditioning set must be the same for the marginal distributions and the copula (Patton 2009, p. 772).

For this work we use a semiparametric copula-based multivariate dynamic model as described in Chen and Fan (2006, p. 129 ff). The conditional

means and variances of $\boldsymbol{Y}_t|\mathcal{F}_{t-1}$ are estimated parametrically. If we denote the parametrized conditional mean of a single variable as $\mu_{it} = E(Y_{it}|\mathcal{F}_{t-1}; \boldsymbol{\phi})$ and the parametrized conditional standard deviation as $\sigma_{it} = \sqrt{V(Y_{it}|\mathcal{F}_{t-1}; \boldsymbol{\phi})}$ we can write the multivariate time series as:

$$\boldsymbol{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \boldsymbol{\eta}_t, \tag{2.7}$$

with $\boldsymbol{\sigma}_t = \text{diag}(\sigma_{1t}, \ldots, \sigma_{Nt})$. The innovations $\boldsymbol{\eta}_t = (\eta_{1t}, \ldots, \eta_{Nt})'$ have a standardized conditional distribution which can be modelled with a conditional copula.

For this model, the marginal distributions of the innovations are modelled with the empirical distribution function (EDF) and it is assumend that they are independent of past informtion and iid distributed such that $F_{\boldsymbol{\eta}_t} = F_{\boldsymbol{\eta}} \forall t = 1, \ldots, T$.

In other words, the observations are first filtered by removing serial dependence or volatility clustering such that the leftover standardized innovations are independent of past information. Finally, the innovations are modeled using a parametric copula and nonparametric rank based estimates of the marginal distributions. The cdf of the innovations can be expressed in terms of a copula and the marginal distributions such that $F_\eta(x_1, \ldots, x_N) = C(F_{\eta_1}(x_1), \ldots, F_{\eta_N}(x_N); \theta_t)$.

2.4. Simulated methods of moments estimation for factor copulas

Estimation methods used for copula models depends on the degree of parametrization: For fully parametrized models for the copula and the marginal distributions maximum likelihood or multi-stage maximum likelihood is used. But one can also non parametrically estimate the marginal distributions and combine them with a parametric copula as shown in the previous section.

In this case, pseudo-maximum likelihood estimation can be used. If a closed form functional relation of spearman's rho or kendall's thau to the copula parameters is available, one can also solve the system directly by using a method of moments approach. For the SMM approach, the population based statistics are replaced by their sample counterparts (inversion method).

2. Theoretical foundation

For the factor copula model a closed form one to one mapping of the copula's parameters θ to measures of dependency as defined in (2.3) - (2.4) is not available in general. If it were available, methods of moments or generalized methods of moments (if the number of moment conditions is larger than the number of parameters) could be applied (Oh and Patton 2013, p. 689f).

Instead one can use a set of scale-invariant empirical dependence measures calculated with simulations from the artificial variables $\boldsymbol{X}$ and compare them to the dependence measures obtained from the observable data $\boldsymbol{Y}$. Minimizing the weighted squared difference of the two dependency vectors yields an estimator for θ.

Formally, the estimator is given by

$$\hat{\boldsymbol{\theta}} = \arg\min Q(\boldsymbol{\theta}) = \arg\min \boldsymbol{g}(\boldsymbol{\theta})'\hat{\boldsymbol{W}}\boldsymbol{g}(\boldsymbol{\theta}) \tag{2.8}$$

with

$$\boldsymbol{g}(\boldsymbol{\theta}) = \hat{\boldsymbol{m}} - \tilde{\boldsymbol{m}}(\boldsymbol{\theta}), \tag{2.9}$$

where $\hat{\boldsymbol{m}}$ and $\tilde{\boldsymbol{m}}$ are vectors of empirical dependency measures computed from the observable and the simulated data respectively.

For the dependency measures one uses the empircal counterparts of Spermas Rho as in 2.2 and lower and upper tail dependency as stated in 2.4. Other invariant measures which only dependend on the copula, e.g. Kendall's Tau, could also be used.

$$\hat{\rho}_{ij} = \frac{12}{T}\sum_{t=1}^{T} I(\hat{F}_i(x_i)\hat{F}_j(x_j)) - 3$$

$$\hat{\lambda}_{ij} =$$

Oh and Patton (2013, p. 691ff) showed that given some assumptions the SMM estimator is weakly consistent and asymptotically normal distributed. The assumptions ensure that for both the iid and the time series case the sample dependency measures converge in probability to their theoretical population values. Further it is important to note, that it is assumed that the population

version of the moment conditions (2.9) is differentiable at the true parameter $\boldsymbol{\theta}_0$ while this is not true for the finite sample version.

The convergence can be summarized as:

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta_0}, (\frac{1}{T} + \frac{1}{S})\boldsymbol{\Omega}) \ \text{ for } \ T, S \to \infty, \tag{2.10}$$

with covariance matrix $\boldsymbol{\Omega} = (\boldsymbol{G'WG})^{-1}\boldsymbol{G'W'\Sigma WG}(\boldsymbol{G'WG})^{-1}$, where $\boldsymbol{G} = \frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_0}$ is the Jacobian matrix of the first order derivitives of the moments function $\boldsymbol{g}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}$ is the asymptotic variance of the moments estimator $\hat{\boldsymbol{m}}$.

The matrix $\boldsymbol{W}$ is a positive definite weight matrix and can be choosen to be the identity matrix. If the efficient weight matrix $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$ is used the aymptotic variance simplifies to $\boldsymbol{\Omega} = (\boldsymbol{G'WG})^{-1}$.

It is crucial to note, that the standard error of the copula parameters $\hat{\boldsymbol{\theta}}$ is not affected by the error arriving from estimating $\hat{\boldsymbol{\phi}}$, which is the parameter vector for the marginal models for the conditional mean and variance of $\boldsymbol{Y}_t$ (see (2.7)). The asymptotic results as they are state here are ony valid if the marginal distribution of the residuals is estimated non-parametically by using the empirical distribution function.

The asymptotic variance $\boldsymbol{\Sigma}$ can be estimated using an iid bootstrap procedure:

Input: $\hat{\boldsymbol{\epsilon}}_t$
Output: $\hat{\boldsymbol{\Sigma}}_{T,B}$
for $b \leftarrow 1$ **to** B **do**
    sample with replacement: $\{\hat{\boldsymbol{\epsilon}}_t^{(b)}\}_{t=1}^T$;
    compute bootstrap sample moments: $\hat{\boldsymbol{m}}_T^{(b)}$;
end
compute sample moments: $\hat{\boldsymbol{m}}_T$;
compute difference: $\boldsymbol{g}_{T,B} = \hat{\boldsymbol{m}}_T^{(b)} - \hat{\boldsymbol{m}}_T$;
$\hat{\boldsymbol{\Sigma}}_{T,B} = \frac{T}{B} \sum_{b=1}^B (\boldsymbol{g}_{T,B}\boldsymbol{g}_{T,B}')$;
return $\hat{\boldsymbol{\Sigma}}_{T,B}$;
         Algorithm 1: Estimation of the covariance matrix of $\hat{\boldsymbol{m}}_T$.

The derivative $\boldsymbol{G}$ can be estimated using a numerical approximation arround the parameter estimator $\hat{\boldsymbol{\theta}}$. For the kth column we can write:

$$\hat{\boldsymbol{G}}_{T,S,k} = \frac{\boldsymbol{g}_{T,S}(\hat{\boldsymbol{\theta}} + \boldsymbol{e}_k\epsilon_{T,S}) - \boldsymbol{g}_{T,S}(\hat{\boldsymbol{\theta}} - \boldsymbol{e}_k\epsilon_{T,S})}{2\epsilon_{T,S}},$$

where $\boldsymbol{e}$ is the kth unit vector and $\epsilon_{T,S}$ the step size which is usually set to $\epsilon_{T,S} = 0.1$.

2.5. Structural break test for factor copulas

In the following section a structural break test based on (Manner, Stark, and Wied 2017) for a change of the copula parameters in (2.7) is presented. Note that it is assumed that the functional form of the copula is time invariant while the copula's parameters θ_t for $t = 1, \ldots, T$ can vary over time. The model presented in 2.3 allows for a wide variety of parametrization and copula functions. Here, we focus on the factor copula model and the SMM estimation procedure as presented in the previous sections.

The test is based on recursive estimations of the copula model at each timepoint $t \in \epsilon 1, \ldots, T$. θ_t is the estimated parameter using the data up to point t. The strictly positive trimming parameter ϵ has to be choosen such that the model parameters don't fluctutate too much just because of small sample sizes. The null hyopthesis states that the estimator is time invariant while the alternative indicates at least one significant break point in time.

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_T \qquad H_1 : \theta_t \neq \theta_{t+1} \text{ for some } t = \{1, \ldots, T\}$$

The the test statistics is the maximum of the scaled squared distance between the full model and the recursive estimator:

$$P = \max_{\epsilon T \leq t \leq T} \frac{t}{T}^2 T(\boldsymbol{\theta}_{t,S} - \boldsymbol{\theta}_{T,S})'(\boldsymbol{\theta}_{t,S} - \boldsymbol{\theta}_{T,S}). \tag{2.11}$$

An alternative test statistic is solely based on the moment generating functions and does not involve the estimation of a copula model:

$$M = \max_{\epsilon T \leq t \leq T} \frac{t}{T}^2 T(\hat{\boldsymbol{m}}_t - \hat{\boldsymbol{m}}_T)'(\hat{\boldsymbol{m}}_t - \hat{\boldsymbol{m}}_T). \tag{2.12}$$

If the test statistic is larger than some critical value, the time t where the test statistic occurs can be interpreted as the detected breakpoint.

Under the null hypothesis and given some assumptions which are similar in spirit to the assumptions for the asymptotic distribtion as in (2.10) the test statistics converges in distribution to

$$P \xrightarrow{d} \max_{\epsilon T \leq t \leq T} (\boldsymbol{A}^*(t) - s\boldsymbol{A}^*(T))'(\boldsymbol{A}^*(t) - s\boldsymbol{A}^*(T)), \quad (2.13)$$

with $\boldsymbol{A}^*(t) = (G'WG)^{-1}G'W(A(t) - \frac{s}{\sqrt{k}}A(1))$ and $T, S \to \infty, \frac{S}{T} \to k$ or $\frac{S}{T} \to \infty$.

The distribution under the null hypothesis can be estimated using the following bootstrap procedure:

Input: $\hat{\boldsymbol{\epsilon}}_t; \alpha$
Output: K
compute full sample moments $\hat{\boldsymbol{m}}_T$;
if *copula based test* **then**
$\quad | \quad L \leftarrow (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'$;
else
$\quad | \quad L \leftarrow 1$;
end
for $b \leftarrow 1$ **to** B **do**
$\quad$ generate bootstrap sample $\{\hat{\boldsymbol{\epsilon}}_t^{(b)}\}_{t=1}^T$;
$\quad$ **for** $t \leftarrow T$ **to** ϵT **do**
$\quad\quad$ compute recursive bootstrap sample moments $\hat{\boldsymbol{m}}_t^{(b)}$;
$\quad\quad A_t^{(b)*} \leftarrow L\frac{t}{T}\sqrt{T}(\hat{\boldsymbol{m}}_t^{(b)} - \hat{\boldsymbol{m}}_T)$;
$\quad\quad K_t^{(b)} \leftarrow (A_t^{(b)*} - \frac{t}{T}A_T^{(b)*})'(A_t^{(b)*} - \frac{t}{T}A_T^{(b)*})$;
$\quad$ **end**
$\quad$ calculate the maximum $K^{(b)}$ over all $K_t^{(b)}$;
end
calculate the $1 - \alpha$ sample quantile K from all $K^{(b)}$;
return K;

Algorithm 2: Estimation of the critical value.

3. *factorcopula* - an R package for simulation and estimation of factor copulas

In this chapter we present the implementation of the previous methods, using the programming language R. The functions are bundled in an R-package such that the methods can be easily installed and distributed. The validity of the package is tested in a simulation study.

3. factorcopula - an R package for simulation and estimation of factor copulas

3.1. Overview and usage

The package consists of a set of high level functions which can be used to construct, simulate and fit various factor copula models. The definition of the factor copula model is handled by the functions `config_factor`, `config_error` and `config_beta` which define the distriution of the latent variables $\boldsymbol{Z}$, the error term $\boldsymbol{\epsilon}$ and the matrix of factor loadings $\boldsymbol{\beta}$.

For the specification of the distributions, the function name of any available random number generator can be used. Additional arguments for the function can be declared in a named list. Distributional parameters can either be pre-defined or passed as expressions which are not immediately evaluated. To distinguish model parameters from pre-defined distributional parameters, an additional character vector with the name of the model parameters has to be passed to `config_factor` or `config_error`.

In the following example, a factor copula with two latent variables is defined. The first factor is skewed-t distributed, the second is normal distributed. The error term is t distributed with 4 degrees of freedom. The distributional parameters of the skewed-t distribution are not pre-defined and will be included in the estimation process.

```
library(factorcopula)
Z <- config_factor(rst = list(nu = 1/dfInv, lambda = lambda),
                   rnorm = list(),
                   par = c("dfInv", "lambda"))
eps <- config_error(rt = list(df = 4))
```

For specifying the matrix of factor loadings one can either manually construct a character matrix of parameters or use the function `config_beta` which accepts an input vector `k` and the number of latent variables.

The vector k defines the group for each variable $i = 1, \ldots, N$. Thus, an equidependence model can be specified with $k = (1, 1, \ldots, 1)$, an unrestrictive model with $k = (1, \ldots, N)$ and a bloc-equidependence model with $k = (1, 1, \ldots, 2, 2, \ldots, M, M, \ldots)$, where M is the number of groups.

In this example an equidependence model, with one β-parameter for each latent variable is constructed:

```
k <- c(1, 1, 1)
beta <- config_beta(k, Z = 2)
```

The two functions `fc_create` and `fc_fit` are used to either simulate values from a factor copula or to fit a model to a real dataset.

The function `fc_create` returns itself a function which can be used to simulate values from the copula model. The simulation function accepts a *named* vector θ of parameters, the number of simulations S and an optional random seed. If a random seed is provided the functions always uses the same underlying random numbers.

The simulation of new values is a time consuming part. To avoid unnescessary calls to the random number generators, the function remembers the state of the Z and ϵ matrix over the function calls and only updates the values if nescessary. Thus if neither the seed nor the distributional parameters change, the function directly proceeds to the calculation of $\boldsymbol{\beta Z + epsilon}$.

During optimazation, the random seed is kept fixed to avoid numerical instabilities (app. to Oh and Patton 2013, p. 12f). Therefore, by using the described memory functionability one can greatly save computational costs. This improves the performance of the optimizataion process massively, especially when only beta parameters are about to be optimized. In this case, the random number generators are only called once.

```
copfun <- fc_create(Z, eps, beta)
theta <- c(beta1 = 2, beta2 = 1, dfInv = 0.25, lambda = -0.8)
X <- copfun(theta, S = 10)
```

As its core, the function `fc_fit` uses the `sbplx` function from the `nloptr` package (Johnson n.d.). The function is a re-implementation of the Subplex algorithm by (Rowan 1990) with support for bound constraints. It is based on the well known Nelder-Mead simplex method and solves the objective function on subspaces of the input space. For parallel computations, the function `fc_fit` accepts a cluster object created by the `snow` or `parallel` ackage. The function uses several randomly selected starting values. The number of different starting values can also be specified with the `trials` argument.

3.2. Simulation study

To illustrate the discussed methods and the validity and performance of the package two simulation studies were performed. First, an equidependence factor copula model with varying dimensions was estimated repeatedly to show the consistency of the SMM procedure. Second, both the moments and copula based structural break test was performed for a bloc-equidependence factor copula model.
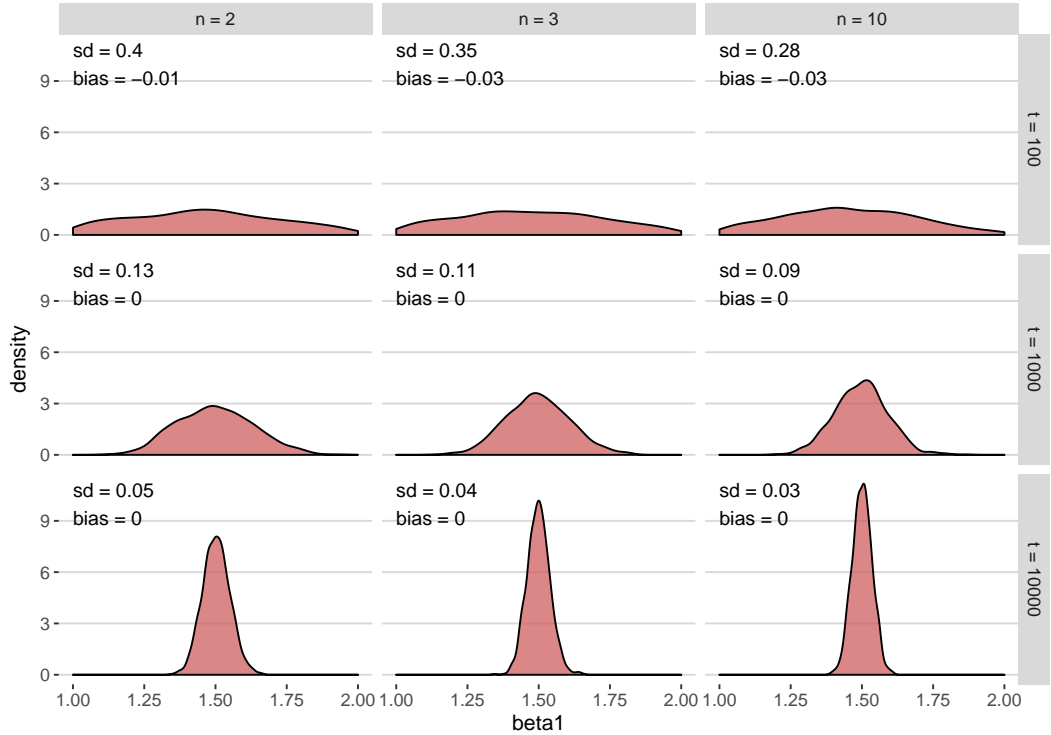

Figure 3.1: Approximated density of $\hat{\theta}$ for an equidependence skew t - t factor copula model with $\beta = 1.5$, $S = 25000$ and standard normal distributed marginals. Each simulation is based on 1000 Monte-Carlo replications.

Figure 3.1 shows the results for the first study. The DGP was based on a simple equidependence model with one skew-t distributed latent variable and a single factor loading $\beta = 1.5$. The error term is t-distributed and the marginal distribution of the observed values is iid standard normal distributed:

$$\boldsymbol{Y} = (Y_1, Y_2)' \sim F_{\boldsymbol{Y}} = C(F_{Y_1}, F_{Y_2})$$
$$(X_1, X_2)' = (\beta, \beta)'Z + \boldsymbol{\epsilon} \tag{3.1}$$
$$Y_1, Y_2 \sim N(0,1), Z \sim skew - t(4, -0.8), \epsilon \sim t(4).$$

15

For all variations of N and T, $S = 25000$ was chosen. Each simulation was repeated 1000 times. The bias and standard deviation of $\hat{\beta}$ was approximated by using the empirical average and standard deviation of all 1000 simulations.

One can clearly see that for larger T the SMM estimator converges to the true parameter. For $t = 10000$ the bias and mean squared error is virtually zero. For larger N one can also note a drop in the mean squared error.

For the second study, a more sophisticated model was chosen to illustrate the effectiveness of the approach even for high dimensional problems and complicated dependence structures. Analogous to the empirical examples in Manner, Stark, and Wied (2017) and Oh and Patton (2017) a *bloc-equidependence* model as described in section 2.2 was chosen. The model was based on the equations:

$$
\begin{aligned}
\boldsymbol{Y} = (Y_1, \dots, Y_{21})' &\sim F_{\boldsymbol{Y}} = C(F_{Y_1}, \dots, F_{Y_{21}}) \\
(X_1, \dots, X_{21})' &= \boldsymbol{\beta Z} + \boldsymbol{\epsilon} \\
Y_i \sim N(0,1), Z_0 &\sim skew - t(4, -0.8), Z_j \sim t(4), \epsilon \sim t(4) \\
i &= 1, \dots, 21, j = 1, \dots, 3 \\
k_1 &= k_2 = k_3 = 7 \\
\theta &= (\beta_1, \dots, \beta_6)'
\end{aligned}
\tag{3.2}
$$

The 21 observable variables were partitioned in 3 groups of equal size. Therefore the parameters to be estimated reduced from $0.5 * N * (N-1) = 210$ to just $2M = 6$. As in the first example the marginal distributed are iid standard normal. We chose $T = 1500$, $S = 25 \times T$ and a breakpoint at $t = 1000$. Before the break, $\theta_0 = (0,1,1,0,1,1)$ and after the break $\theta_1 = (1.5,1,1,1.5,1,1)$. Thus, only the intra- and interdependence for the first group increases from 0 to 1.5.

Figure 3.2 shows the result of the break test for a single recursive run of the simulation for $t = 300, \dots, 1500$. Both the moments and copula based versions are shown together with the estimated critical value based on 2000 bootstrap replications.
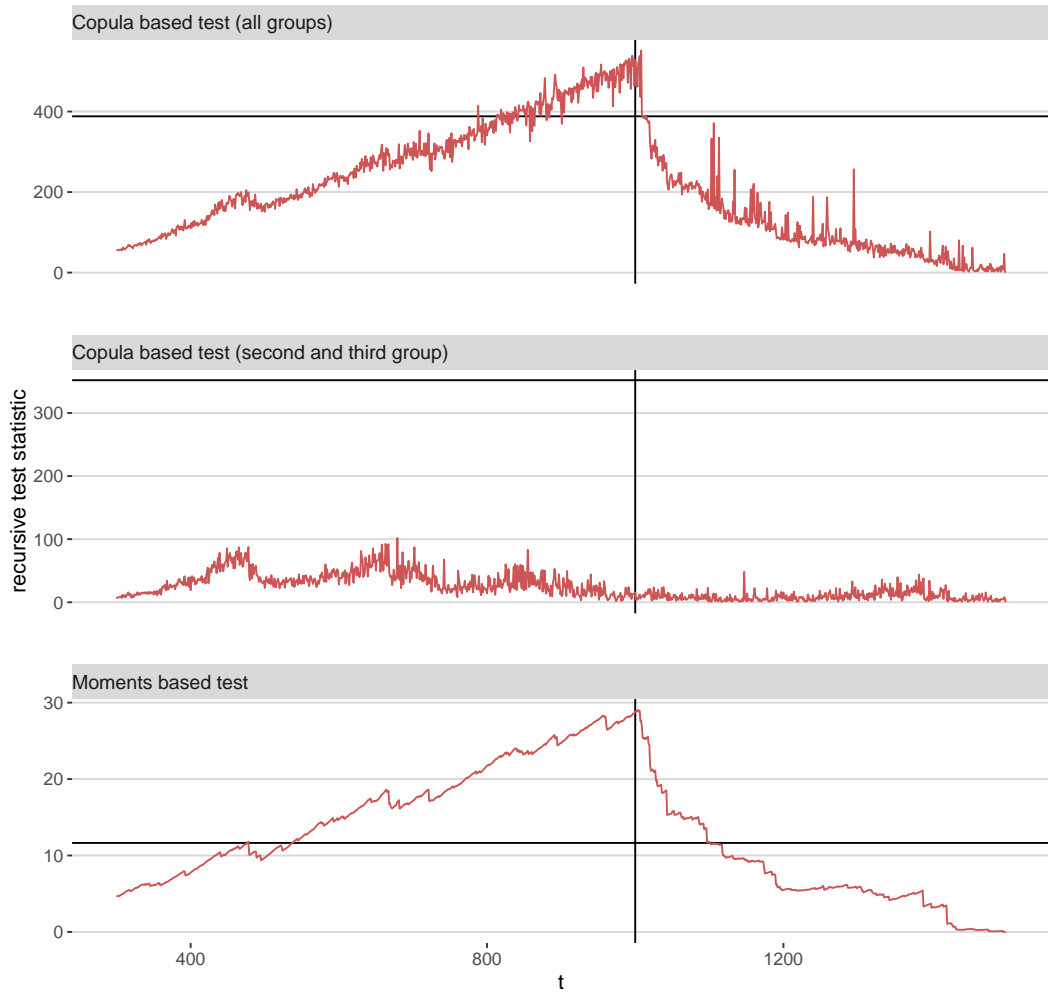
Figure 3.2: Structural break test for a bloc-equidependence model with $N = 21$ and 3 groups of equal size. The theoretical breakpoint is at $t = 1000$ and is modelled as a change of the intra- and interdependenvy of the first group from 0 to 1.5.

The moments based test detects a breakpoint at $t = 962$ which is close to the true value but less precise than the copula based test. The null hypothesis of no break is clearly rejected since the test statistic fluctuates too strong.

coefficient	before ($t \leq 1000$)	after ($t > 1000$)	full model
β_1	0.00 (0.10)	1.96 (0.69)	0.00 (0.13)
β_2	1.17 (0.37)	0.88 (0.29)	1.28 (0.30)
β_3	0.96 (0.27)	0.83 (0.30)	0.99 (0.27)
β_4	0.03 (0.15)	1.55 (0.64)	0.88 (0.15)
β_5	0.92 (0.18)	1.01 (0.32)	0.92 (0.18)
β_6	1.04 (0.20)	1.13 (0.40)	1.09 (0.20)
Q	0.0008	0.0026	0.0025
T	1000	500	1500
S	37500	37500	37500

Table 3.1: Estimation results for the bloc-equidependence factor copula model before and after the breakpoint and for the full dataset. Standard errors in paranthesis (estimated with $B = 2000$ bootstrap samples).

4. Modelling topic dependencies over time with factor copulas

In this chapter we apply the previously discussed methods to a set of relative frequencies over time. The frequencies are derived by counting the relative number of times a specific regular expression appeared in documents of a large text collection of social media posts.

First the dataset is presented. Second, we give a detailed description of the feature generation process and some descriptive overview. Third, the model setup is explaind and estimation results of various factor copula models applied to the residuals of the word frequencies are presented. Finally, we apply both the moments and copula based structural break test.

4.1. The *btw17* social media dataset

The original dataset consists of social media posts published by public pages on *Facebook* between January 2014 and December 2017.

Using the official list of the candidates for the Bundestag election in 2017 (btw17), the public facebook page for each of the politicians was manually

researched. Only candidates from the seven largest parties (CDU, CSU, SPD, Die Linke, Bündnis 90/ Die Grünen, AfD, FDP) were considerd. Arround 83.8% of all 2516 candidates have an account on Facebook (see Stier et al. 2018, p. 16).

Due to API and privacy restrictions only information from public pages can be accessed such that only arround 52.4% of the social media accounts could be considered for the data collection. In addition, 113 official pages from political parties, both on the federal and regional level, were added.

The data collection software used the *restfb* Java client library to call Facebook's official *Graph API*. The posts were stored in a document orientated database on servers in Germany. In addition to the posts's content, they are tagged with a timestamp, the user-id of the author and the number of likes and shares the post has gained uppon collecting it from the API. (Facebook 2018; Allen and Bartels 2018; MongoDB Inc 2018).

For this analysis the data is restricted on textual posts only. This results in nearly 664 thousand posts. Figure 4.1 shows the monthly number of active accounts and the monthly number of posts by party.[4] In early 2014 approximately 500 accounts were active. This number increased steadily to roughly 750 accounts in mid 2016. From then until the election in September 2017 the number increased rapidly to almost 1200 active accounts followed by drop after the election. One can also observe a sharp rise in the number of posts in the month of the election, followed by a decline afterwards.

Table 4.1 shows the overall number of *textual* posts, accounts, likes and shares per party. The likes and shares are based on the aggregated sum of the number of likes and shares for each post from this party. Allthough the right and left wing parties *AfD* and *Linke* have a relatively small number of accounts and posts they generate by far the greatest number of attention in terms of likes and shares. The social democratics party *SPD* has over twive more posts than the *AfD* but roughly generates only half of the likes and only a fifth of the shares.

---

[4]An account was defined active if it has at least one post in the month considered.
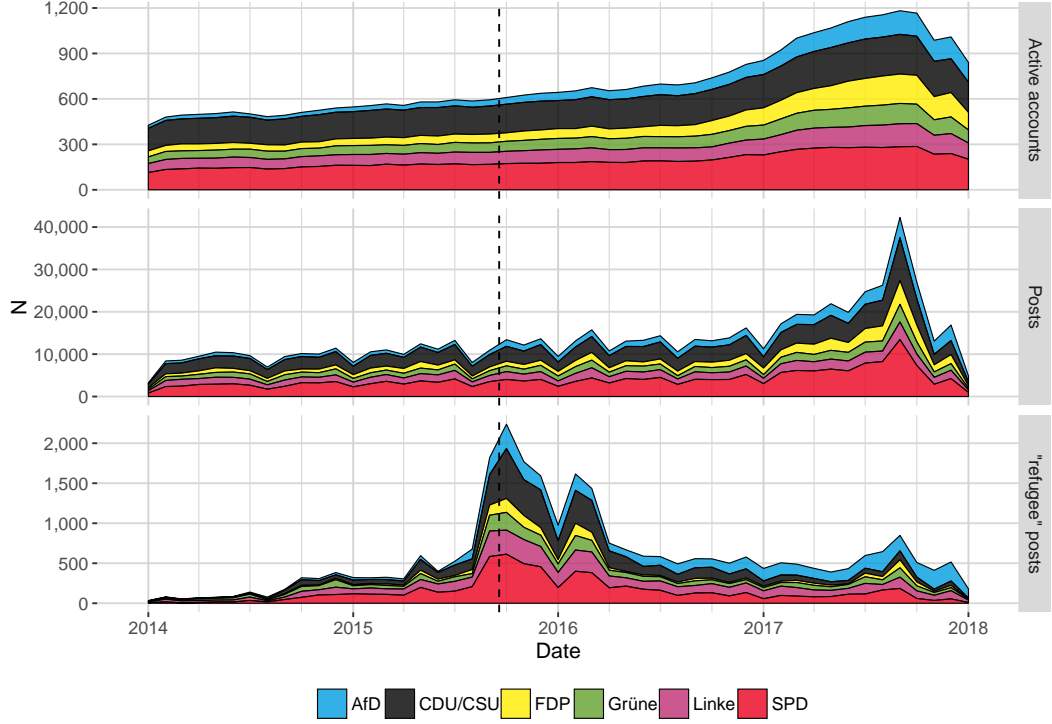
Accounts and posts by party over time

Figure 4.1: Number of active accounts and number of posts per party and month. The bottom panel shows the number of posts matching the regular expression `flucht|fluecht`. The vertical line indicates the breakpoint detected by the moments based test.

4.2. Data processing and descriptive analysis

For this analysis we focus on how the German refugee crisis in 2015 and 2016 is perceived by the political partys in the social media.

First, each post in the dataset is cleaned by removing links and stopwords, transforming umlauts and converting the text to lowercase letters.

Second, a post is considered to be related to the refugee crisis if it matches the regular expression `flucht|fluecht`. For example a match occurs if a post contains the words "flüchtlingskrise", "fluchtursachen", "flüchten" or "flüchtlingsheime".

Third, for each party and day we count the relative number of posts related to the refugee crisis. To account for the importance of a post for the public audience we calculate a weighted average using the weights $log(l_i + 1)$, where l_i is the number of likes of the post. The resulting dataset consists of a columns

party	posts	accounts	likes	shares	share of posts related to "refugees"
			(in Million)		(in %)
AfD	74724	162	18.36	7.59	6.96
CDU/CSU	169115	267	14.72	1.83	3.49
FDP	71083	201	6.32	0.77	2.39
Grüne	67188	139	4.03	1.28	4.62
Linke	84723	158	16.03	4.23	6.23
SPD	196805	290	10.11	1.57	3.66
All parties	663 638	1217	69.57	17.27	4.28

Table 4.1: Overall number of posts, active accounts, likes and shares over the observation period from "2014-01-01" - "2017-12-31". The last two columns show the fraction of posts matching the regular expression `flucht|fluecht`.

for each party, six in total. Each row represents the daily weighted relative frequency of posts matching the regular expression (compare to Figure A.1 in the Appendix). Using the full time span of four years results in $T = 1461$ observations. Thus, the aggregated dataset can be seen as an indicator for how much a party discusses and talks about the topic "refugee crisis" on each day. A positive dependence between all parties would indicate that this topic is relevant for all parties whereas we would see no dependence if the partys talk about this topic independently from each other.

To remove time dependence as discussed in **??**, we estimate an ARIMA model for each of the time series. The best univariate model was choosen by running some model candidates over a range of parameters. The model candidate with the lowest Akaike information criterion was considered. All univariate models used $I = 1$, $AR \in \{0, 1\}$ and $MA \in \{0, 1\}$. In the following the estimated standardized residuals are used.

Party	AfD	CDU/CSU	FDP	Grüne	Linke	SPD
AR	4	1	2	3	2	1
MA	1	2	1	3	2	1

Table 4.2: ARMA Model parameters for the ARIMA-GARCH model used to estimate the standardized residuals.

4.3. Results

Table **??** shows the overall pairwise dependency measures used also for the SMM procedure. The dependencies are very weak to weak.

Pairs	Rank - correlation	Quantile-dependence			
		0.05	0.1	0.90	0.95
AfD-CDU/CSU	0.07	0.14	0.17	0.13	0.10
AfD-FDP	0.08	0.10	0.14	0.13	0.07
AfD-Grüne	0.03	0.07	0.16	0.09	0.10
AfD-Linke	0.08	0.07	0.15	0.19	0.08
AfD-SPD	0.09	0.11	0.16	0.16	0.08
CDU/CSU-FDP	0.17	0.16	0.25	0.18	0.14
CDU/CSU-Grüne	0.13	0.11	0.20	0.14	0.07
CDU/CSU-Linke	0.10	0.10	0.18	0.16	0.14
CDU/CSU-SPD	0.20	0.22	0.22	0.19	0.14
FDP-Grüne	0.10	0.08	0.19	0.15	0.10
FDP-Linke	0.10	0.10	0.16	0.17	0.10
FDP-SPD	0.12	0.19	0.23	0.18	0.07
Grüne-Linke	0.09	0.07	0.12	0.19	0.08
Grüne-SPD	0.12	0.14	0.21	0.16	0.14
Linke-SPD	0.15	0.16	0.16	0.23	0.21
Average	0.11	0.12	0.18	0.16	0.11

First various factor copula models were fitted to the complete residuals.

	Equidependence			Unrestrictive		
	norm-norm	t-t	skewt-t	norm-norm	t-t	skewt-t
β_1	0.43	0.42	0.42	0.23	0.23	0.25
β_2	-	-	-	0.62	0.52	0.63
β_3	-	-	-	0.42	0.44	0.43
β_4	-	-	-	0.32	0.34	0.33
β_5	-	-	-	0.37	0.41	0.35
β_6	-	-	-	0.63	0.60	0.67
df	-	99	96	-	58	38
λ	-	-	-0.29	-	-	-0.36
Q	0.0036	0.0034	0.0031	0.1081	0.0979	0.0903

Table 4.3: Estimation results for different one-factor copula specifications.

	Moments based test		Copula based test
	unrestrictive	restrictive	restrictive
test statistics	6.05	116.95	1.51
95%-CV	1.98	61.90	1.67 (p-value: 0.0685)
breakpoint	2015-09-18	2015-09-18	2015-03-13

Table 4.4: Summary of break point detection methods for the social media dataset.

coefficient	before ($t \leq 626$)	after ($t > 626$)	full model
β_1	0.39 (0.03)	0.41 (0.03)	0.42 (0.03)
λ	-0.55 (0.75)	-0.33 (0.54)	-0.29 (0.37)
Q	0.0029	0.0023	0.0031
T	626	835	1461
S	36525	36525	36525

Table 4.5: Estimation results for the btw17 dataset. A bloc-equidependence factor copula model is estimated before and after the breakpoint at $t = 626$. Standard errors in paranthesis (estimated with $B = 2000$ bootstrap samples). Degrees of freedom were fixed at the full model estimate of $df = 96$.

5. Discussion

A. Appendix

A.1. Notes on data and source code access

An online version of this thesis is publicly available as a git-repository under https://github.com/bonartm/factorcopula-thesis. The repository contains notes on how to install all dependencies. The source code files for the analyses are located online in the `source` folder.

Due to data restrictions by Facebook it is not possible to publish the original dataset of Facebook posts. However, the residuals of the ARIMA-GARCH models which were applied to the aggregated Facebook data are located online at `data/topics_residuals.rds`.

The methods for the simulation and estimation of factor copula models and the break test are not part of this repository. Instead, they are available via the R-package `factorcopula`. The package can be installed from Github. Further notes can be found under: https://github.com/bonartm/factorcopula (See also chapter 3).

For almost all estimation procedures, the HPC cluster of the Universiy of Cologne was utlizied (Achter et al. 2018). To simplify the workflow we wrote the R-package `cheopsr` which allows the execution of job scripts from within the local R environment. The package is available online at https://github. com/bonartm/cheopsr. To run the package, a Unix-like system and accesss rights to the HPC cluster are obligatory.

A. Appendix

A. Appendix

A.2. Additional figures
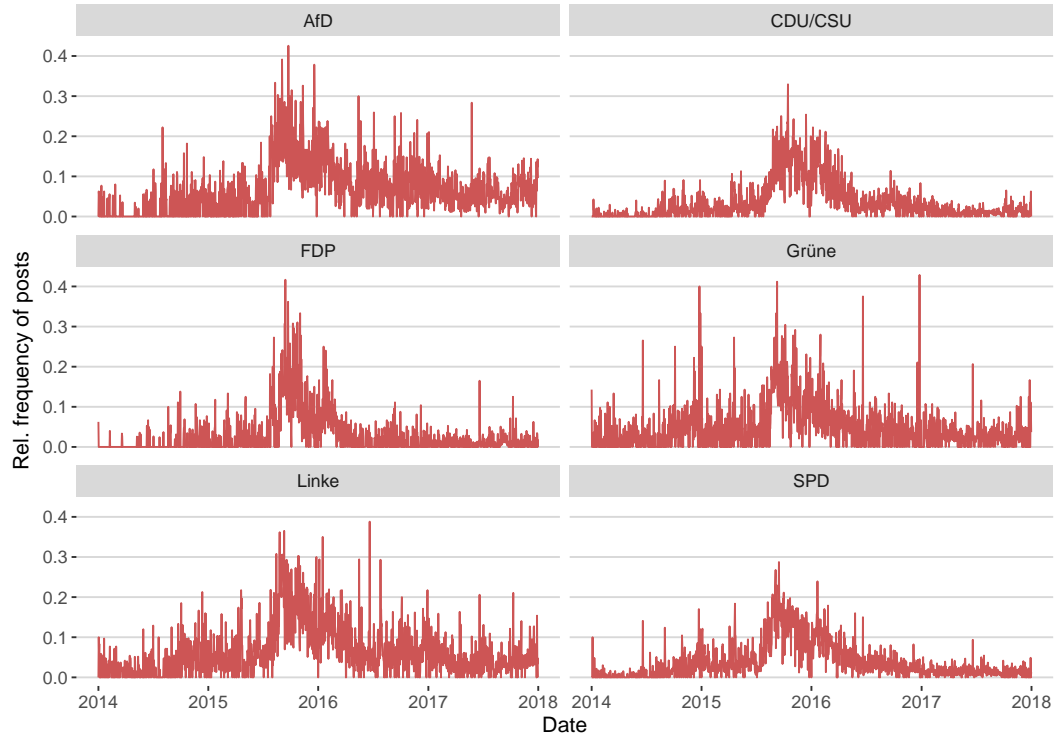
(a) Observed daily relative frequency of posts which match the regular expression.
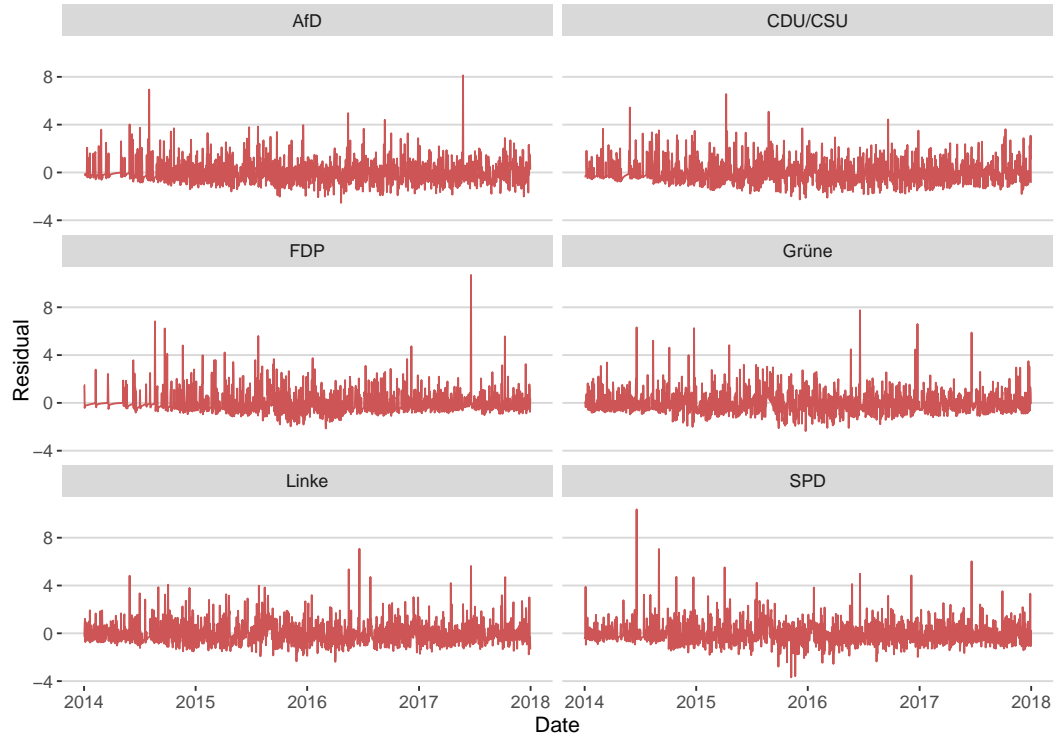
(b) Residuals of ARIMA models applied to the time series.
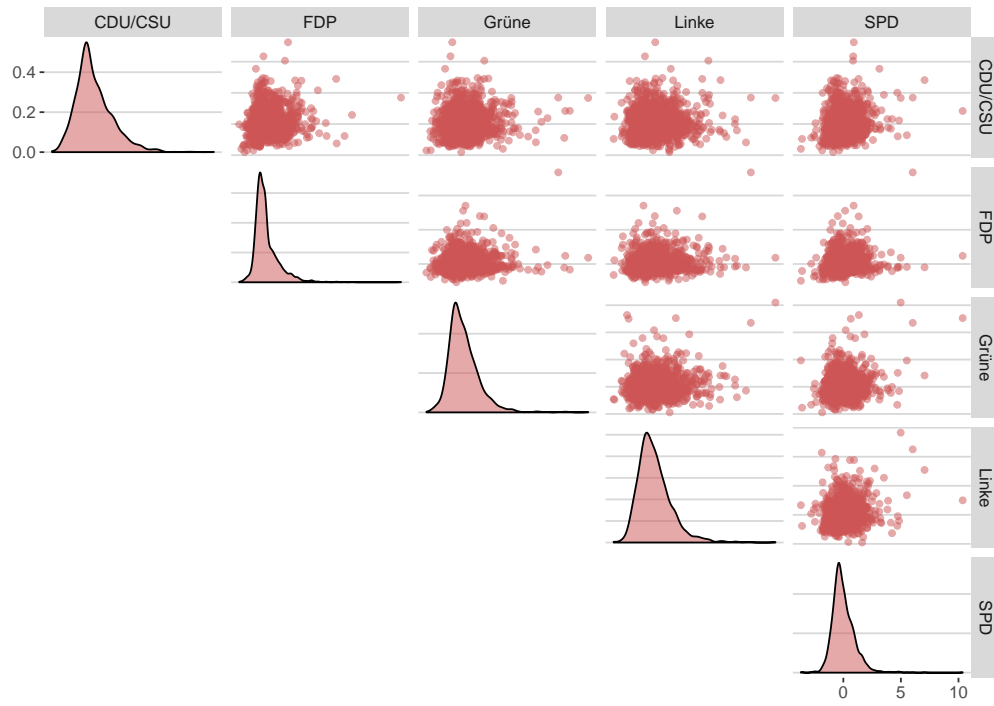
Figure A.1

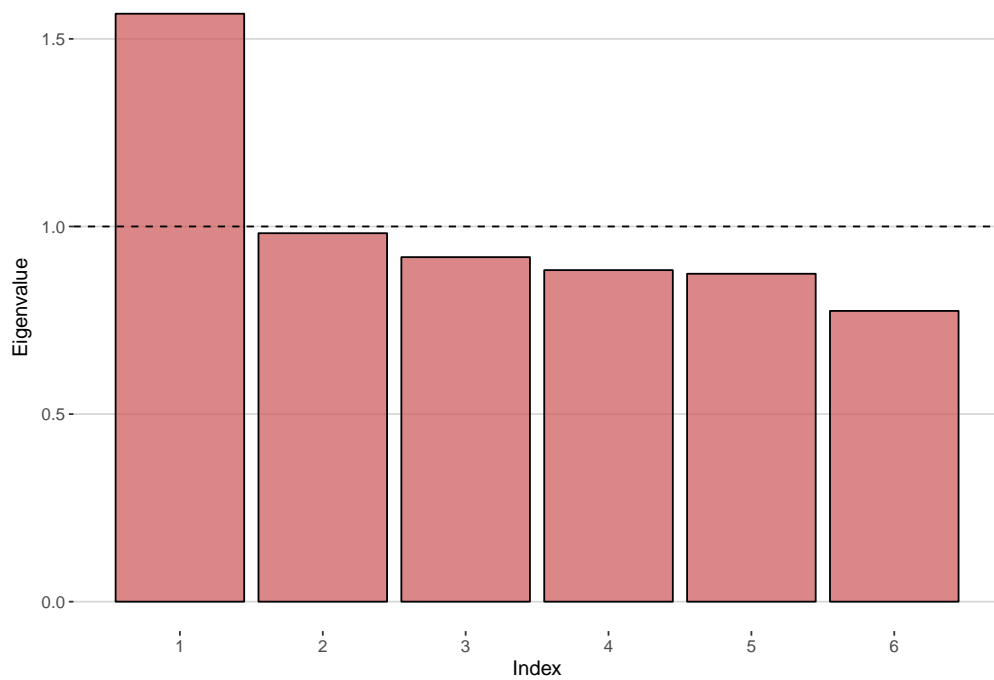Figure A.2: Pairwise scatterplot of the estimated residuals.

Figure A.3: Scree-plot of ranked eigenvalues based on the pairwise rank-correlation matrix.

B. References

Achter, V. et al. (2018). *CHEOPS Cologne High Efficient Operating Platform for Science.* Tech. rep. https://rrzk.uni-koeln.de/sites/rrzk/HPC_Projekte/CHEOPS_Brief_Instructions.pdf. University of Cologne, Regional Computing Centre.

Allen, M. and N. Bartels (2018). *RestFB - a pure Java Facebook Graph API client.* http://restfb.com.

Bonart, M. (2018). *factorcopula - high dimensional depence modelling with factor copulas.* https://github.com/bonartm/factorcopula.

Chen, X. and Y. Fan (2006). "Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification". In: *Journal of Econometrics* 135.1, pp. 125–154.

Facebook (2018). *Facebook's Graph API.* https://developers.facebook.com/docs/graph-api.

Joe, H. (2015). *Dependence modelling with copulas.* Taylor & Francis Group.

Johnson, S. G. *The NLopt nonlinear-optimization package.* URL: http://ab-initio.mit.edu/nlopt.

Kumar, P. (2011). "Copulas: Distribution Functions and Simulation". In: *International Encyclopedia of Statistical Science.* Ed. by M. Lovric. Berlin, Heidelberg: Springer, pp. 309–312.

Manner, H., F. Stark, and D. Wied (2017). "Testing for Structural Breaks in Factor Copula Models". Working Paper.

MongoDB Inc (2018). *Java driver for Mongo DB.* https://mongodb.github.io/mongo-java-driver/.

Nelsen, R. B. (1999). *An Introduction to Copulas.* Springer.

Oh, D. H. and A. J. Patton (2013). "Simulated Method of Moments Estimation for Copula-Based Multivariate Models". In: *Journal of the American Statistical Association* 108.502, pp. 689–700.

– (2017). "Modeling Dependence in High Dimensions With Factor Copulas". In: *Journal of Business & Economic Statistics* 35.1, pp. 139–154.

Patton, A. J. (2006). "MODELLING ASYMMETRIC EXCHANGE RATE DEPENDENCE". In: *International Economic Review* 47.2, pp. 527–556.

– (2009). "Copula-Based Models for Financial Time Series". In: *Handbook of Financial Time Series.* Ed. by T. Andersen, R. A. D. andJ.-P. Kreiß, and T. Mikosch. Berlin, Heidelberg: Springer, pp. 767–785.

B. References

Rowan, T. H. (1990). "Functional Stability Analysis of Numerical Algorithms". UMI Order No. GAX90-31702. PhD thesis. Austin, TX, USA.

Schmid, F. et al. (2010). "Copula-Based Measures of Multivariate Association". In: *Copula Theory and Its Applications.* Ed. by P. Jaworski et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 209–236.

Sempi, C. (2011). "Copulas". In: *International Encyclopedia of Statistical Science.* Ed. by M. Lovric. Berlin, Heidelberg: Springer, pp. 302–305.

Sklar, A. (1959). "Fonctions de répartition à n dimensions et leurs marges". In: *Publications de l'Institut de Statistique de L'Université de Paris* 8, pp. 229–231.

Stier, S. et al. (2018). "Systematically Monitoring Social Media: the case of the German federal election 2017". In: *GESIS papers* 2018/04, p. 25.

C. Statutory Declaration

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Köln, den 7. Mai 2018

(Malte Bonart)