University of Cologne

Malte Bonart

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

Master’s 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 Abbreviations

cdf	Cumulative distribution function
iid	Identically and independent distributed
SMM	Simulated methods of moments
DGP	Data generating process

1. Introduction

Models with copula functions became increasingly popular since the 1990th (Nelsen 1999, p. 1). The concept was first introduced in the work by Sklar (1959). They 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.

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 (Patton 2009, p. 771).

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. 774 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).

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

Research question: What do I want to analyze?

How similar is the dependency structure of the political communication on social media channels compared to financial markets?

Can we detect structural breaks in the dependency structure of the political communication on social media channels? This could be an indicator of political eruptions such as elections, scandals or political events.

extreme dependence during economic crisis (-> elections)

Relevance: Why do I ask this question? Why is it relevant?

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 implementation details of the software package *factorcopula*, written in the statistical programming language R (Bonart 2018). With the package, factor copulas can be fitted to real data and structural breaks can be detected. The validity of the package and the methods is illustrated by a small simulation study. In the last chapter, the methods are applied to a large dataset of textual social media posts from german politicians and political parties. Here, the goal is to identify temporal dependencies between different topics and to test for changes in the dependence structure due to important political events. The last chapter summarizes the findings and critically discusses the presented methods.

2. Theoretical foundation

2.1. Copula theory

A function of the type $C : [0, 1]^N \to [0, 1]$, with $N \geq 2$ is called a *copula* if

1. $C(u_1, \ldots, u_N) = 0$, if $\exists i \in \{1, \ldots, N\} : u_i = 0$
2. $C(1, \ldots, 1, x_i, 1, \ldots, 1) = x_i$
3. The C-*volume* of every *N-box* is postive

The last property is also called the *N-increasing* property (Sempi 2011, p. 302). From this definition it follows that, in a statistical sense, a copula function is a multivariate distribution $C(u_1, \ldots, u_N) = P(U_1 \leq u_1, \ldots, U_N \leq u_N)$ with uniform marginals $U_i \sim U(0,1) \forall i \in \{1, \ldots, N\}$ (Joe 2015, p. 7).

Sklar (1959) showed, that every d-variate distribution $F(x_1, \ldots, x_n)$ can be expressed in terms of its marginal distributions $F_1(x), \ldots, F_n(x)$ and a copula function $C(u_1, \ldots, u_N)$ such that $F(\mathbf{x}) = C(F_1(x_1), \ldots, F_N(x_N))$.

If F is *continuous* with marginal quantile functions $F_1^{-1}, \ldots, F_N^{-1}$ then $C(\mathbf{u})$ is uniquely determined by $C(\mathbf{u}) = F(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_N))$.

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 be expressed as a function of their copula (Schmid et al. 2010, p. 210). Two widely used measures are Spearman's

$$\rho_{X_1,X_2} = 12 \int\int_{[0,1]^2} u_1 u_2 dC(u_1, u_2) - 3 \qquad (2.1)$$

and Kendal's rank correlation

$$\tau_{X_1,X_2} = 4 \int\int_{[0,1]^2} C(u_1, u_2) dC(u_1, u_2) - 1 = 4E(C(U_1, U_2)) - 1. \qquad (2.2)$$

Bivariate measures can be extended to the multivariate case where there measure the strength of association embedded in the N-dimensional copula of a multivariate vector $\boldsymbol{X}$.

One can also use the average over all bivariate dependency measures.

Multivariate versions of Spearman's rank correlation can be defined in terms of an underlying copula function as ((Schmid et al. 2010, p. 215ff)):

$$\rho = \frac{N+1}{2^N-(N+1)}(2^N \int_{[0,1]^N} C(\boldsymbol{u})d\boldsymbol{u} - 1) \tag{2.3}$$

Analogously, Kendall's rank correlation is defined as

$$\tau = \frac{1}{2^{N-1}-1}(2^N \int_{[0,1]^N} C(\boldsymbol{u})d\boldsymbol{u} - 1). \tag{2.4}$$

Lower and upper tail dependency for two variables X and Y is defined as

$$\begin{aligned} \tau_{XY}^L &= \lim_{q \to 0} \frac{P(X \leq F_X^{-1}(q), Y \leq F_Y^{-1}(q))}{q} \\ \tau_{XY}^U &= \lim_{q \to 1} \frac{P(X > F_X^{-1}(q), Y > F_Y^{-1}(q))}{q} \end{aligned} \tag{2.5}$$

2.2. Copula models for multivariate time series

Conditional copula as presented in (Patton 2006) and (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 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 $\{\boldsymbol{Y}_{t-1}', \boldsymbol{Y}_{t-2}', \dots, \boldsymbol{X}_t', \boldsymbol{X}_{t-1}', \dots\}$. The conditional means and variances of $\boldsymbol{Y}_t|\mathcal{F}_{t-1}$ are estimated parametrically. The observations are then 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.

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:

$$\boldsymbol{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \boldsymbol{\eta}_t, \tag{2.6}$$

with $\boldsymbol{\sigma}_t = \text{diag}(\sigma_{1t}, \dots, \sigma_{Nt})$. The innovations $\boldsymbol{\eta}_t = (\eta_{1t}, \dots, \eta_{Nt})'$ are independent of past information and iid distributed according to some multivariate distribution function $F_\eta(x_1, \dots, x_N)$.

The cdf of the innovations can be expressed in terms of a copula and the marginal distributions such that $F_\eta(x_1, \dots, x_N) = C(F_{\eta_1}(x_1), \dots, F_{\eta_N}(x_N); \theta_t)$.

2.3. Factor copulas

(Oh and Patton 2017)

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

Consider a set of artificial variables $X_i = \sum_{k=1}^{K} \beta_{ik} Z_k + \epsilon_i$ with $i = 1, \dots, N$ the dimension of the observable data $\boldsymbol{Y} = (Y_1, \dots, Y_N)'$ and $k = 1, \dots, K$ the number of latent variables. The latent variables Z_k and the error term e_i follow some parametrized distributions such that $\epsilon_i \overset{iid}{\sim} F_\epsilon(\boldsymbol{\gamma_\epsilon})$ and $Z_k \sim F_{Z_k}(\boldsymbol{\gamma_{Z_k}})$ with $Z_i \perp Z_j \forall i \neq j$, $Z_k \perp \epsilon_i \forall i, k$ and $\boldsymbol{\gamma_\epsilon}$, $\boldsymbol{\gamma_{Z_k}}$ some distribution specific parameter vectors.

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

The factor copula is therefore completely defined via the parameter vector $\boldsymbol{\theta} = (\beta_{11}, \dots, \beta_{i1}, \dots, \beta_{ik}, \boldsymbol{\gamma}'_{\boldsymbol{Z_1}}, \dots, \boldsymbol{\gamma}'_{\boldsymbol{Z_K}}, \boldsymbol{\gamma}'_{\boldsymbol{\epsilon}})'$. The number of latent variables K and the distribution functions $F_{Z_1}, \dots, F_{Z_k}, F_\epsilon$ are hyperparameters of the model which have to be chosen prior to the estimation.[1]

[1] 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.

The latent factor structure is linked to the observable data via the copula function because it holds that $F_{\boldsymbol{Y}}(y_1,\ldots,y_n) = C_\theta(F_{Y_1}(y_1),\ldots,F_{Y_N}(y_N))$. The model can be summarized in the following set of equations:

$$\begin{aligned}
\boldsymbol{Y} &= (Y_1,\ldots,Y_N)' \\
F_{\boldsymbol{Y}} &= C_f(F_{Y_1}(y_1),\ldots,F_{Y_N}(y_N);\boldsymbol{\theta}) \\
\boldsymbol{X} &= (X_1,\ldots,X_N)' = \boldsymbol{\beta Z} + \boldsymbol{\epsilon} \\
F_{\boldsymbol{X}} &= C_f(F_{X_1}(x_1),\ldots,F_{X_N}(x_N);\boldsymbol{\theta})
\end{aligned} \tag{2.7}$$

It is important to note, that the artificial variables $\boldsymbol{X}$ are only used for the construction of the factor copula function $C_f(u_1,\ldots,u_N)$. Once this copula function is determined, the artificial variables and its marginal distributions $F_{X_i}(x)$ are of no interest. Using the copula function together with the marginal distributions of the observable variables $F_{Y_i}(y)$ one can then determine the joint distribution of $\boldsymbol{Y}$.

This approach allows for a two-stage estimation in which first the marginal distributions are estimated flexibly and second the factor structure for the possibly high dimensional copula function is fitted to the data. For the factor copula and the joint distribution of the artificial variables as defined in (2.7) 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 shows four different simulations from a one factor equidependence factor copula model. The marginal distributions are stan-

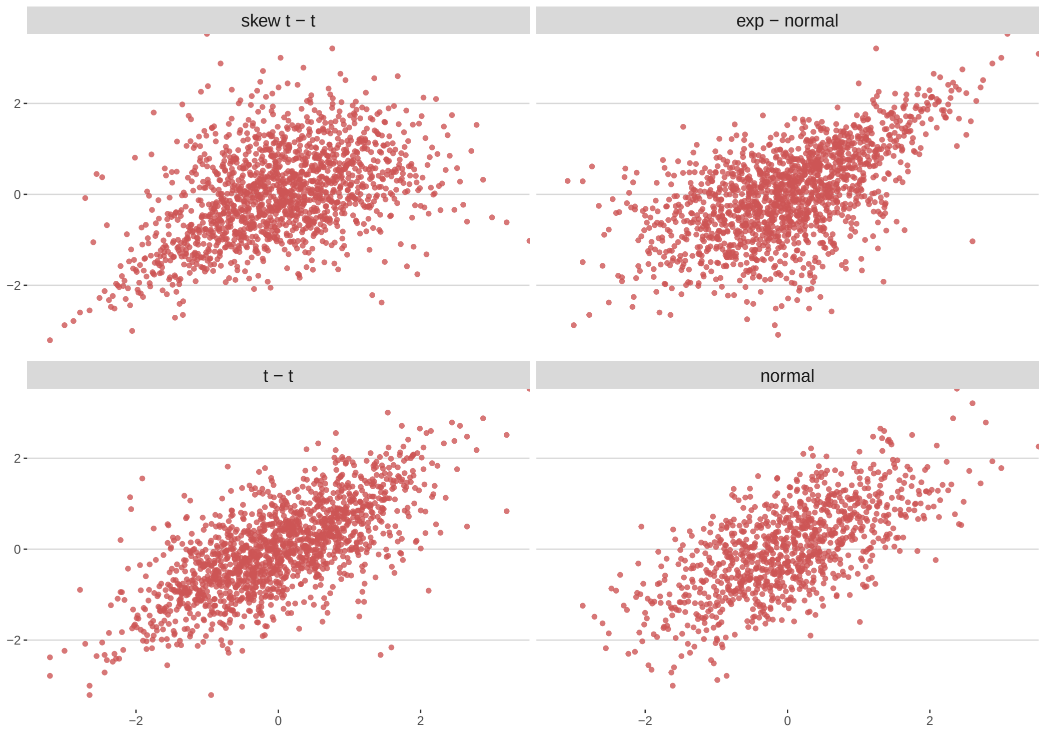


Figure 2.1: Illustration of different equidependence factor copula models with $N = 2$, $\beta = 1.5$, standard normal distributed marginals and different distributions for the latent variable and the error term.

dard normal while the distributions of the latent variable and the error term differ.

The block-equidependence model is less restrictive and is especially suitable for variables which can be naturally partitioned into different groups.[2] 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 model can be summarized as:

$$
\begin{aligned}
\boldsymbol{X} &= (X_1^1, \ldots, X_{s_1}^1, \ldots, X_1^D, \ldots, X_{s_D}^D)' = \boldsymbol{\beta Z} + \boldsymbol{\epsilon} \\
\boldsymbol{Z} &= (Z_0, Z_1, \ldots, Z_D)' \\
\boldsymbol{\beta} &= \begin{pmatrix}
\beta^1 & \beta^{D+1} & 0 & \cdots & 0 \\
\beta^1 & \beta^{D+1} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\beta^1 & \beta^{D+1} & 0 & \cdots & 0 \\
\beta^2 & 0 & \beta^{D+2} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\beta^D & 0 & 0 & \cdots & \beta^{D+D} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\beta^D & 0 & 0 & \cdots & \beta^{D+D}
\end{pmatrix},
\end{aligned}
\tag{2.8}
$$

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

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. In this case, pseudo-maximum likelihood is 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. Here, the population based statistics are replaced by their sample counterparts (inversion method).

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 (2.2) - (2.5) is not available in general. If it was 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

$$\hat{\boldsymbol{\theta}} = \arg\min Q(\boldsymbol{\theta}) = \arg\min \boldsymbol{g}(\boldsymbol{\theta})'\hat{\boldsymbol{W}}\boldsymbol{g}(\boldsymbol{\theta}) \tag{2.9}$$

with

$$\boldsymbol{g}(\boldsymbol{\theta}) = \hat{\boldsymbol{m}} - \tilde{\boldsymbol{m}}(\boldsymbol{\theta}), \tag{2.10}$$

where $\hat{m}$ and $\tilde{m}$ are the vector of dependencies computed with the observable and the simulated data respectively.

Oh and Patton (2013, p. 691ff) showed that under a set of assumptions, the SMM is weakly consistent and asymptotically normal distributed:

$$\frac{1}{\sqrt{1/T + 1/S}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \Sigma_0) \text{ for } T, S \to \infty \tag{2.11}$$

, with covariance matrix $\Sigma_0 = 12$

2.5. Structural break test for factor copulas

Note that we assume that the functional form of the copula is time invariant while the copula's parameters can vary over time (Patton 2006, p. 542).

The model presented in 2.2 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.

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

The test statistics is

$$P = \sup_{s \in [\epsilon, 1]} s^2 T(\boldsymbol{\theta}_{sT,S} - \boldsymbol{\theta}_{T,S})'(\boldsymbol{\theta}_{sT,S} - \boldsymbol{\theta}_{T,S}) \tag{2.12}$$

Under the null hypothesis and given some assumptions the test statistics converges in distribution to

$$P \xrightarrow{d} \sup_{s \in [\epsilon, 1]} (\boldsymbol{A}^*(s) - s\boldsymbol{A}^*(1))'(\boldsymbol{A}^*(s) - s\boldsymbol{A}^*(1)), \tag{2.13}$$

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

In the following a structural break test is presented (Manner, Stark, and Wied 2017).

(Patton 2009)

(Oh and Patton 2013)

(Manner, Stark, and Wied 2017)

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

3.1. Overview and usage

The package consists of a set of main functions which can be used to construct, simulate and fit a factor copula model. The definition of the model is handled by the functions `config_factor`, `config_error` and `config_beta` for difining the distriution of the latent variables, the error term and the factor loadings matrix:

```
library(factorcopula)
k <- c(1, 2, 3)
Z <- config_factor(rst = list(nu = 1/0.25, lambda = -0.8))
eps <- config_error(rt = list(df = 1/0.25))
beta <- config_beta(k, Z = 1)
```

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. Instead of havin fixed distributional parameters one can also specify them as non-standard evaluated model parameters:

```
Z <- config_factor(rst = list(nu = 1/df, lambda = lambda),
                   par = c("df", "lambda"))
```

The central function `fc_create` creates the actual copula model from the specifications. It returns a function which can be used to simulate values from the copula model given a *named* parameter vector θ. During optimazation it is crucial that the random seed is fixed otherwise numerical instabilities can occur (app. to Oh and Patton 2013, p. 12f). This can be achieved by passing a seed integer to the function. The function is optimized in such a way that given a fixed seed, it tries to call the random number generators specified in `config_factor` or `config_error` as less as possible. Simulating new random values is costly and can be avoided if the seed is fixed and only the beta parameters change. Therefore, the function keeps track of the previous state and only updates the random numbers if nescessary. This can improve the performance in the optimizataion process massively, especially when only beta parameters are about to be optimized.

3.2. Optimization strategy

3.3. 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 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:

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

For all variations of N and T, $S = 25000$ was chosen. Each simulation was repeated 500 times. The bias and mean squared error of $\hat{\beta}$ was approximated by using the empirical average and standard deviation of all 500 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.

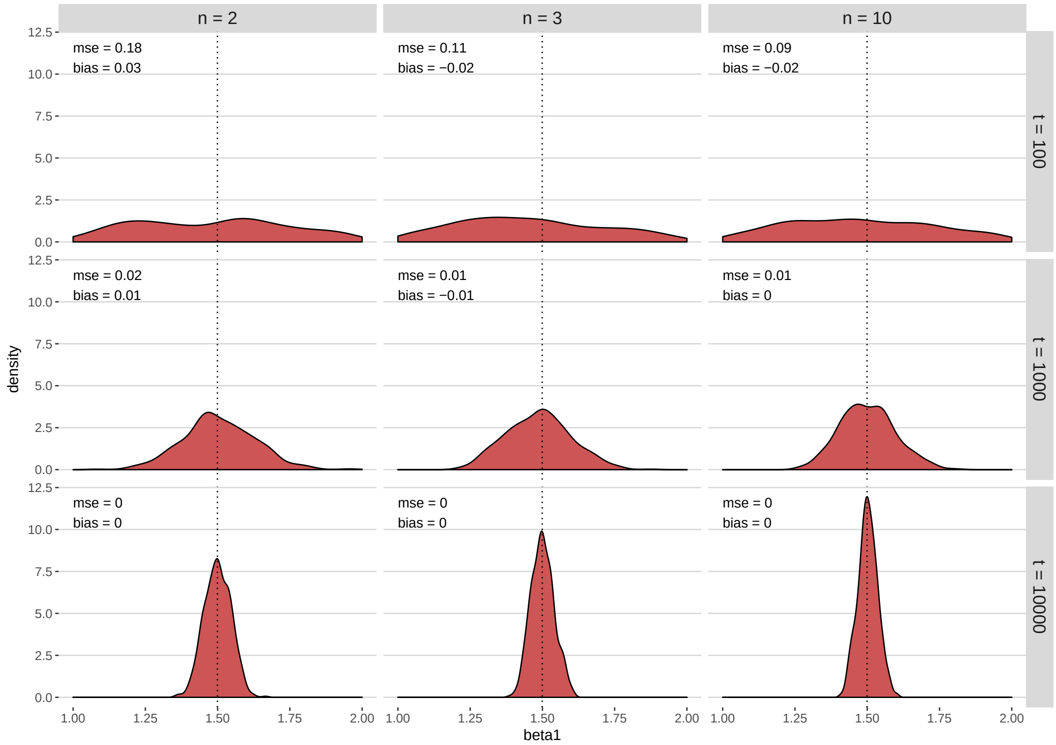


Figure 3.1: Approximated density of $\hat{\theta}$ for an equidependence skew t - t factor copula model with $\beta = 1.5$, $S = 25000$ and standard normal distributed marginals. Each simulation is based on 500 Monte-Carlo replications.

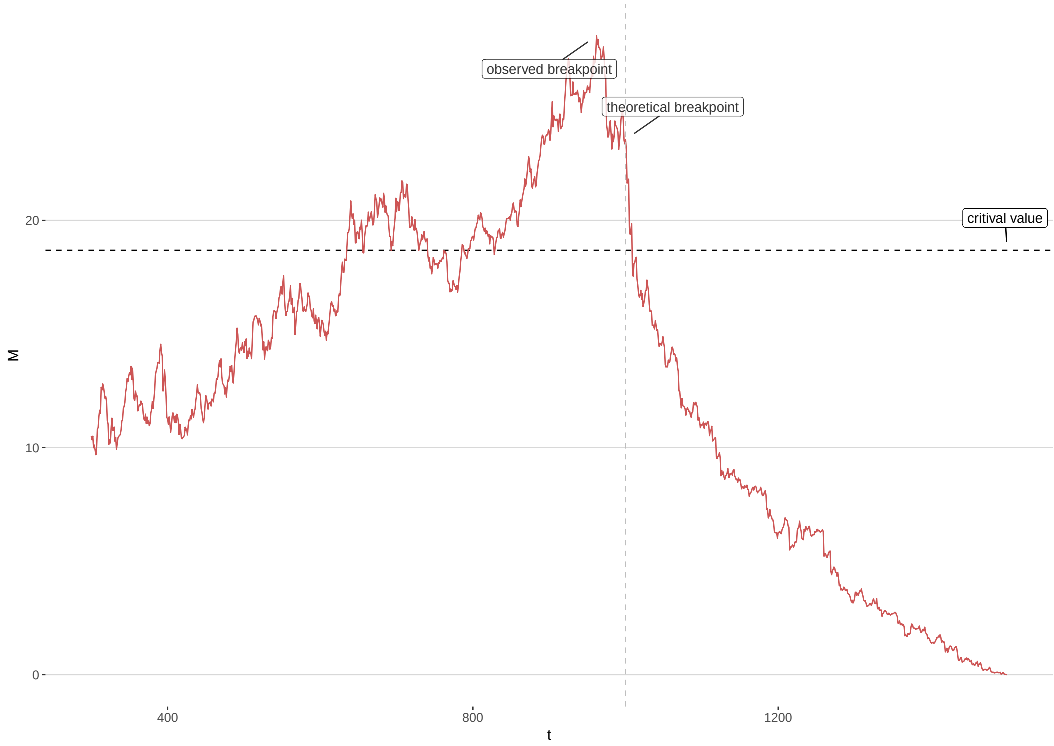
For the second study, a more sophisticated model was chosen to illustrate the effectiveness of the approach even for high dimensional problems and complicated dependence structures. Analogous to the empirical examples in Manner, Stark, and Wied (2017) and Oh and Patton (2017) a *bloc-equidependence* model as described in section 2.3 was chosen. The model was based on the equations:

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

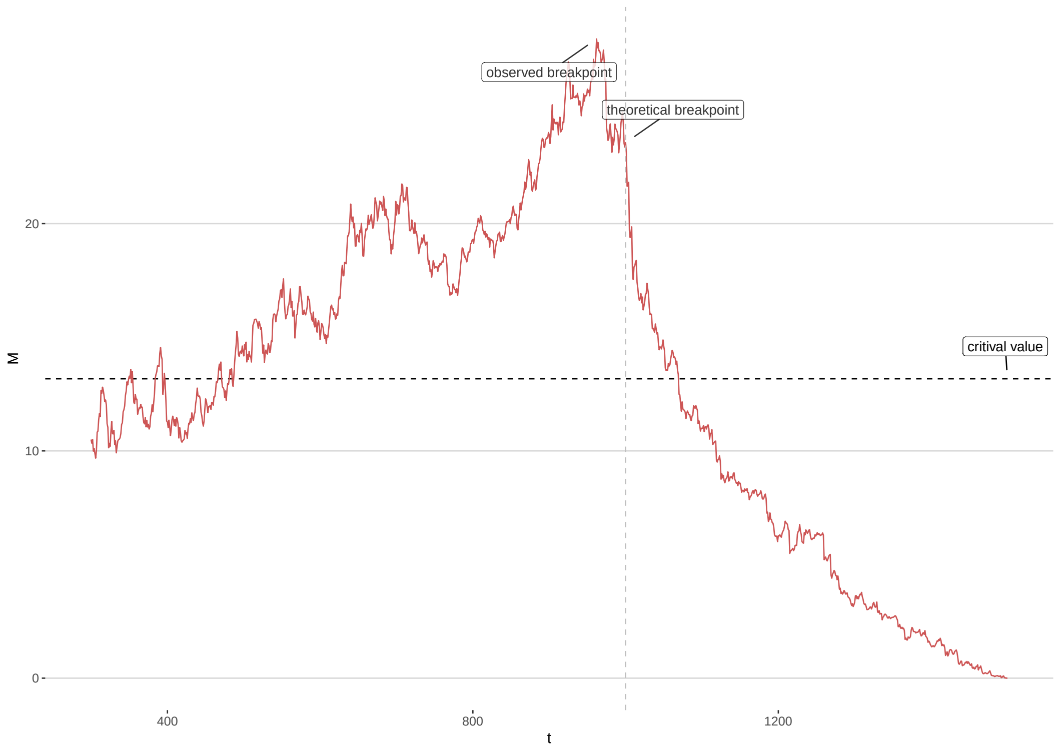
The 21 observable variables were partitioned in 3 groups of equal size.

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 $T = 1500$, $S = 25000$ 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 $t = 300, \ldots, 1500$. Both the moments and copula based versions are shown together with the estimated critical value based on 1000 bootstrap replications.



(a) Test based on moments



(b) Tets based on factor copula

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 $t = 1000$.

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

4. Modelling topic dependencies over time with factor copulas

4.1. The *btw17* dataset

The dataset consists of posts from public pages created on the social media platform *facebook*. The data was collected by a *Java* program which was built using the *restfb* client software to call Facebook's official *Graph API* and the *Mongo DB* client software to store the data in a document-oriented database (Facebook 2018; Allen and Bartels 2018; MongoDB Inc 2018).

The set of public pages was created by manually researching facebook accounts of political candidates which competed in the German federal election held in late September 2017. In addition, official pages from political parties both on the federal and regional level, were included. Candidates and organizations from the seven largest parties (CDU, CSU, SPD, Die Linke, Bündnis 90/ Die Grünen, AfD, FDP) were tracked.

For this analysis the data is restricted on textual posts created between January 2014 and December 2017. This results in 663638 posts from 1217 candidates and parties. 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 1200 active accounts. After the election one can observe a drop down in active accounts.

[insert figure here]

4.2. Data processing and descriptive analysis

4.3. Factor copula estimation

4.4. Structural break test

4.5. Results

A. Appendix

B. References

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.

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.

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.

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 22. März 2018

(Malte Bonart)