Malte Bonart

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

Master thesis

Supervisors: Prof.Dr. Dominik Wied, Florian Stark

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

Abstract

Multivariate statistical models based on copula functions have gained much popularity during the last years. In the field of finance they are used to model complex dependence structures between financial assets. A multivariate distribution, can always be expressed in terms of its marginal distributions and a copula function. In contrast to the linear correlation coefficient, the dependencies described by a copula are invariant to monotone transformations of the marginal distributions. For multivariate time series, copulas allow for a multistage estimation process, in which first the marginal distributions are estimated using standard univariate time series models and second a static copula model is applied to the residuals.

A new class of copula models, the so called factor copulas, are useful for high dimensional problems. Here, the dependence structure is modeled as a linear factor model for which the dependencies are described by a lower dimensional set of latent variables. For estimation, a simulation based technique based on the Generalized Method of Moments can be adapted to this type of model.

This work summarizes and structures the current state of development in the field of factor copula models including the estimation procedure and a test for structural breaks in its parameters. This thesis contributes to the current research by providing an open-source software package for the programming language *R* which implements the methods and makes them available to a broader audience. This can improve future research on this topic.

The validity and functionality of the theory and its implementation is tested in two simulation studies. Further, we investigate how these methods can be used for the detection of breaks in other research areas. Using real text data from an online social network, the case of the German refugee crisis in late 2015 is analyzed: For each major German party we derive the importance of topics related to refugees in the online political discourse over time. The non-parametric structural break test detects a break during the time of the refugee crisis but a copula model estimated before and after the break yields no significant difference in its parameters.

The package is capable of estimating high dimensional factor copula models. However, due to the simulation based estimation technique and the non-regular properties of the objective function the numerical methods are often unstable, inefficient and computationally demanding. The choice of the underlying optimization algorithm strongly effects the estimation results. Future research should therefore concentrate on improving the stability and efficiency of the estimation techniques and the underlying optimization procedures.

Contents

List of Figures iv

List of Tables iv

List of Algorithms iv

List of Abbreviations v

1. Introduction 1

2. Theoretical foundation 4

2.1. Copula theory 4

2.1.1. Definitions 4

2.1.2. Copulas and dependency 6

2.2. Factor copulas 8

2.2.1. General set up 8

2.2.2. Model restrictions 9

2.3. Copula models for multivariate time series 12

2.4. Simulated method of moments estimation for factor copulas . . 14

2.4.1. SMM and the Generalized method of moments 15

2.4.2. SMM for factor copulas 16

2.5. A structural break test for factor copulas 19

3. *factorcopula* - an R package for simulation and estimation of factor copulas 22

3.1. Copula specification and simulation 22

3.2. Optimization strategy 24

3.3. Simulation study 25

3.3.1. Convergence for the equidependence model 25

3.3.2. Structural break in a bloc-equidependence model 27

4. Modelling dependencies in topic salience over time 30

4.1. The *btw17* social media dataset 32

4.2. Data processing and descriptive analysis 33

4.3. Results 38

5. Discussion and summary 41

A. Appendix **45**
A.1. Notes on data and source code access 45
A.2. Code examples 45
A.3. Additional figures 47

B. References **49**

C. Statutory Declaration **52**

List of Figures

2.1. Simulations from different equidependence factor copula models 10
3.1. Monte-Carlo density estimators for the parameter of an equidependence factor copula model 26
3.2. Parametric and non-parametric structural break test for a bloc-equidependence model 28
4.1. Monthly asylum seekers and applications from 2014 - 2017 . . . 31
4.2. Number of active accounts and posts per party and month . . . 34
4.3. Univariate time series and estimated residuals for the btw17 dataset 36
A.1. Pairwise scatter-plot of the estimated residuals for the btw17 dataset 47
A.2. Scree-plot of ranked eigenvalues based on the pairwise rank-correlation matrix 48

List of Tables

3.1. Estimation results for the bloc-equidependence simulation . . . 29
4.1. Overall number of posts, active accounts, likes and shares for the btw17 dataset 35
4.2. Parameters for the ARIMA-GARCH(1,1) model 35
4.3. Pairwise sample dependencies for the daily topic salience of each party 37
4.4. Estimation results for different one-factor copula specifications . 38
4.5. Results for the parametric and non-parametric break point detection tests 39
4.6. Post and pre-break estimation results for the btw17 dataset . . . 40

List of Algorithms

1. Bootstrap procedure for estimating the asymptotic variance $\Sigma_{T,B}$ 18
2. Bootstrap procedure for estimating the distribution of the test statistics under H_0 21

List of Abbreviations

API	Application programming interface
btw17	Bundestag election 2017
cdf	Cumulative distribution function
CV	Critical value
DGP	Data generating process
EDF	Empirical distribution function
GMM	Generalized Methods of Moments
iid	Independent and identically distributed
ML	Maximum Likelihood
SMM	Simulated methods of moments

1. Introduction

Many multivariate statistical models are based on the assumption that the variables follow a multivariate Gaussian distribution. Here, the natural measure of dependence is the covariance or correlation matrix between the variables (Joe 2015, p. 25). But using simple correlations as the only dependence measure can be misleading. In many cases real world data does not follow a Gaussian distribution and shows other distributional features, such as skewness or heavy tails. Correlation requires that the variance of the marginal distribution is finite. This requirement can be problematic when dealing with heavy tailed distributions. Further more, correlation is only a measure of linear dependence and is not invariant under monotonic transformations of the variables (Embrechts, Mcneil, and Straumann 2002, pp. 6-8). For instance applying the strictly increasing logarithmic function to the variables usually results in a different correlation matrix.

Due to this shortcomings of the ordinary linear correlation coefficient, other and more flexible methods for describing the dependence structure between non Gaussian variables have been developed. In this area of research, models based on so called copula functions became increasingly popular since the 1990th (Nelsen 1999, p. 1). Especially in the last decade, an exponentially increasing boom in the research activity can be observed: Up to the year 1999, only 38 publications with the topics *copula* and *dependence* have been published in peer-reviewed journals. From 2000 - 2008 334 copula related publications are listed and for the years 2009 - 2017 the number of publications increased to 2048.[1]

Applications of copula based models can be found in many applied disciplines and one of its first usage was in the field of survival analysis in bio-statistics (Fermanian 2017). But it recent popularity is mainly driven by applications in finance and insurance science, especially in quantitative risk management, e.g. for estimating the Value at Risk measure of a portfolio (Embrechts 2009, p. 644).

At first, any valid multivariate distribution function for which its marginal distributions have a uniform distribution on the interval $[0, 1]$ can be called a copula function (Nelsen 1999, p. 1). Its popularity is based on the results of

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

a theorem by Sklar (1959): It implies that multivariate distributions can be constructed by separately specifying the marginal distributions of the random variables and by defining the dependence structure among the variables. The dependence structure is described with a copula function. It links the multivariate distribution function to its marginal distributions. In case of continuous multivariate distributions the copula function is uniquely defined. In contrast to the linear correlation coefficient, scale-invariant dependency measures can be expressed in terms of an underlying copula function.

The theorem can be applied in both directions: First, to model the dependence structure of multivariate distributions independent of their underlying marginal distributions and second, to construct bivariate or multivariate distributions based on a copula function and marginal distributions (Sempi 2011, pp. 302-304).

The splitting of a multivariate distribution into its marginals and a copula allows for a two-stage estimation process in multivariate models: The marginal distributions and the copula function can be estimated separately. By doing so, semi-parametric techniques can be utilized. For instance, the approach described in the following chapters uses the non-parametric empirical distribution function (EDF) for the marginals while the copula function is estimated parametrically. (Patton 2009, p. 777).

Originally, copulas were formulated for static data, where the observations are independent from each other. But recently, copula theory has been adapted to time series: There, it 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 at some time point given past information. Structural break tests or time series models for the parameters of the copula function can applied here to analyze temporal changes in the dependence structure. Second, copulas are used to describe the dependence between observations of a univariate time series. This is related to the study of Markov processes. (Patton 2009, p. 771 ff).

This thesis focuses on the task of dependence modelling for the multivariate time series case. Especially, we try to summarize and structure the current development in the field of factor copula models, which are a specific class of copulas used for high dimensional applications. For this type of model, the dependence structure is modeled as a linear factor model for which the

dependencies can be described by a lower dimensional set of latent variables. Furthermore, this paper contributes to the scientific discussion in two ways:

First, a software library, written in the language *R* is presented. It allows the consistent specification, simulation and estimation of general factor copula models. By doing so, the methods and the structural break test are made available to a broader scientific audience and can be easily studied and improved by other researchers.

Second, we study a novel way of applying the discussed methods to a non finance related area of research. Specifically, the case of the German and European refugee crisis in late 2015 and its perception by political parties on the social media platform Facebook is analyzed: Using an aggregated dataset of textual social-network posts from German politicians and political parties we derive a dynamic party-specific measurement of topic salience.

Topic or issue salience is a fuzzy concept used in the political sciences to describe the importance or prominence which various actors in a political system such as voters, parties or elites, place on certain issues (Wlezien 2005, p. 557). In this study, the importance of refugee and asylum related issues in its political online discourse is measured separately for each party over time. Using the methods of dependence modelling and a structural break test we ask whether and how the dependencies in topic salience between the parties changed due to the refugee crisis.

In the following, we give an overview over the next chapters: The first chapter deals with the theoretical basics of copula theory, introduces the time series framework and the factor copula model. The dynamic model is based on the assumption of a time-invariant copula given that each marginal time series is filtered by its time dependent conditional means and variances. Further, the estimation method known as Simulated Method of Moment (SMM) is presented. It is based on the comparison of a vector of dependencies measures calculated independently with simulations from the factor model and with the observed residual data. The last theoretical section deals with a structural break test for the possibly time varying parameters of the copula.

The second chapter gives an overview over the software library and discusses the underlying numerical optimization routines. The validity and functionality of the theory and its implementation is tested in two simulation studies.

The third chapter focuses on the empirical application. The social-media dataset, its aggregation and the derivation of the topic salience measurements are described. Then the marginal models for the univariate time series are discussed. Finally, the results of applying various factor copula models and structural break tests to the data are presented. The last chapter discusses the results and gives a critical assessment of the thesis and its discussed methods.

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 (SMM) and its origin in the Generalized methods of moments (GMM) is presented. The SMM is the estimation method which is used throughout this paper. Finally we summarize the ideas of a structural break test for possibly time varying parameters of a factor copula model.

2.1. Copula theory

The joint cumulative distribution function (cdf) $F_Y(y_1, \ldots, y_N) = P(Y_1 \leq y_1, \ldots, Y_N \leq y_N)$ for some multivariate random vector Y of dimension N has the continuous marginal distributions $F_{Y_i}(y_i) = P(Y_i \leq y_i)\ \forall\ i = 1, \ldots, N$. Estimating F_Y is computationally demanding especially if N becomes large. Therefore, a copula function is introduced which can be used to link the marginal and the joint cdf.

2.1.1. Definitions

A function of the type $C : [0,1]^N \to [0,1]$, with $N \geq 2$ is called a *copula* if it is the distribution function of a random vector U such that $C_U(u_1, \ldots, u_N) = P(U_1 \leq u_1, \ldots, U_N \leq u_N)$ and if its marginal distributions are $U_i \sim Unif(0,1)$,

e.g. uniformly distributed (Joe 2015, p. 7).[2] The theorem by Sklar (1959) states, that every d-variate distribution function $F_Y(y_1, \ldots, y_n)$ can be expressed in terms of its marginal distributions and a copula function such that

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

To see this, first consider the so called probability-integral transformation which states that the random variables $U_i = F_{Y_i}(Y_i)$ are uniformly distributed (Embrechts, Mcneil, and Straumann 2002, p. 4). This is due to the fact that $F_{U_i}(u_i) = P(U_i \leq u_i) = P(F_{Y_i}(Y_i) \leq u_i) = P(Y_i \leq F_{Y_i}^{-1}(u_i)) = F_{Y_i}(F_{Y_i}^{-1}(u_i)) = u_i$. $F_{U_i}(u_i) = u_i$ is exactly the definition of the distribution function of a $Unif(0,1)$ distributed variable. Using this transformation we can write

$$\begin{aligned} F_Y(y_1, \ldots, y_N) &= P(Y_1 \leq y_1, \ldots, Y_N \leq y_N) \\ &= P(F_{Y_1}(Y_1) \leq F_{Y_1}(y_1), \ldots, F_{Y_N}(Y_N) \leq F_{Y_N}(y_N)) \\ &= P(U_1 \leq F_{Y_1}(y_1), \ldots, U_N \leq F_{Y_N}(y_N)) \\ &= C_U(F_{Y_1}(y_1), \ldots, F_{Y_N}(y_N)). \end{aligned} \tag{2.2}$$

Note that the transformation of the marginal distributions to a $Unif(0,1)$ distribution is somehow an arbitrary choice and other possibilities, e.g. a transformation to a Gaussian distribution, are also used in other areas of research (Mikosch 2006, p. 7).

If F_Y is a continuous function with marginal quantile functions $F_{Y_1}^{-1}, \ldots, F_{Y_N}^{-1}$ then the copula function $C_U(\mathbf{u})$ is uniquely determined by $C(\mathbf{u}) = F(F_{Y_1}^{-1}(u_1), \ldots, F_{Y_N}^{-1}(u_N))$. For other non-continuous functions this must not be the case and several valid copula functions can exist (Embrechts, Mcneil, and Straumann 2002, pp. 4-5).

The simplest form of a copula is the copula of a vector of independent variables for which we can write $F_Y(\boldsymbol{y}) = \prod_{i=1}^{N} F_{Y_i}(y_i)$. Using the derivation in (2.2), this results in the independence copula

$$C_U^{ind}(u_1, \ldots, u_N) = \prod_{i=1}^{N} u_i. \tag{2.3}$$

[2] We usually use the letter F to denote a cdf but the letter C is reserved for a cdf with uniformly distributed marginals.

Many other possible functional forms for a copula C_U exist in the literature. In this paper, we focus on a special class, the so called factor copulas which we present in section 2.2.

2.1.2. Copulas and dependency

Copulas are useful for the study of dependence between a set of random variables because the copula function is unaffected by monotonic transformations of the marginal variables. In contrast to the linear correlation coefficient, this property allows the definition of alternative scale-invariant dependence measures which only dependent on the copula function (Nelsen 1999, p. 125).

To see the scale-invariant property of the copula function, consider a random vector U with distribution function $C_U(\boldsymbol{u})$ and $C_U(F_1(y_1), \ldots, F_N(y_N)) = F_Y(y_1, \ldots, y_N)$. By applying some increasing functions $T_i(Y_i)$ to the random variables and using (2.2) we can write

$$\begin{aligned} F_{T(Y)}(\boldsymbol{y}) &= P(T_1(Y_1) \leq y_1, \ldots, T_N(Y_N) \leq y_N) \\ &= P(T_1(F_{Y_1}^{-1}(U_1)) \leq y_1, \ldots, T_N(F_{Y_N}^{-1}(U_N))) \\ &= C_U(F_{T_1(Y_1)}(y_1), \ldots, F_{T_N(Y_N)}(y_N)). \end{aligned} \tag{2.4}$$

Thus, although Y and $T(Y)$ have different joint distribution functions, they share the same copula function (Embrechts, Mcneil, and Straumann 2002, p. 6).

In the following we shortly present three common scale invariant measures of dependency between two variables Y_i and Y_j. In contrast to the linear correlation coefficient, these measures can be expressed solely as a function of the underlying bivariate copula.[3] Spearman's and Kendall's rank correlation both measure the degree of monotonic dependence between the two variables Y_i and Y_j with joint distribution function $F_Y(\boldsymbol{y})$. As the linear correlation coefficient they are symmetric and normalized in the interval $[0, 1]$ (Embrechts, Mcneil, and Straumann 2002, p. 15).

Kendall's rank correlation is based on the definition of concordance and discordance: Consider two pairs of independent and identically distributed (iid)

[3] For a proof of the copula representations see Embrechts, Mcneil, and Straumann (2002, pp. 16-18).

random vectors (Y_i^1, Y_j^1) and (Y_i^2, Y_j^2). Both pairs share the same joint distribution F_Y. The pair is concordant if $(Y_i^1 - Y_i^2)(Y_j^1 - Y_j^2) > 0$ and discordant if $(Y_i^1 - Y_i^2)(Y_j^1 - Y_j^2) < 0$. For the former case large (small) values of one pair occur with large (small) values of the other. For the latter case, large (small) values of one pair occur with small (large) values of the other (Nelsen 1999, pp. 125-126).

With this definition Kendall's rank correlation is given as the probability of concordance minus the probability of discordance:

$$\begin{aligned}\tau_{i,j} &= P((Y_i^1 - Y_i^2)(Y_j^1 - Y_j^2) < 0) - P((Y_i^1 - Y_i^2)(Y_j^1 - Y_j^2) > 0) \\ &= 2 * P((Y_i^1 - Y_i^2)(Y_j^1 - Y_j^2) > 0) - 1 \\ &= 4 \int\int_{[0,1]^2} C(u_i, u_j) dC(u_i, u_j) - 1.\end{aligned} \tag{2.5}$$

Spearman's rank correlation is given as the ordinary correlation coefficient between the probability-integral transforms $F_{Y_i}(Y_i)$ and $F_{Y_j}(Y_j)$:

$$\begin{aligned}\rho_{i,j}^S &= \rho(F_{Y_i}(Y_i), F_{Y_j}(Y_j)) = 12E(F_{Y_i}(Y_i), F_{Y_j}(Y_j)) - 3 \\ &= 12 \int\int_{[0,1]^2} u_i u_j dC(u_i, u_j) - 3.\end{aligned} \tag{2.6}$$

Finally, to capture dependencies in the joint lower or joint upper parts of the distribution, one defines the coefficients of upper and lower tail dependency:

$$\begin{aligned}\lambda_{i,j}^U &= \lim_{q \to 1} P(Y_i > F_{Y_i}^{-1}(q) | Y_j > F_{Y_j}^{-1}(q)) = \lim_{q \to 1} \frac{1 - 2q + C(q,q)}{1 - q}, \\ \lambda_{i,j}^L &= \lim_{q \to 0} P(Y_i \leq F_{Y_i}^{-1}(q) | Y_j \leq F_{Y_j}^{-1}(q)) = \lim_{q \to 0} \frac{C(q,q)}{q}.\end{aligned} \tag{2.7}$$

The coefficients measures the probability that extreme large (small) values occur in one variable, given extreme large (small) values in the other variable. In contrast to other common dependence measures, the coefficients of upper and lower tail dependency are defined on the interval $[0, 1]$. Since the limit only exists theoretically and not for observable data, one usually calculates upper and lower *quantile* dependency for some values of q close to 0 for lower and close

to 1 for upper quantile dependence (Joe 2015, pp. 62-63). The sample versions of this measures can be obtained by replacing the marginal distributions with the EDF and using the definition of the empirical copula (Nelsen 1999, p. 176):

$$\hat{C}_{i,j}(u,v) = \frac{1}{T+1}\sum_{t=1}^{T}\mathbf{1}(\hat{F}_i(x_{it}) \leq u, \hat{F}_j(x_{jt}) \leq v). \tag{2.8}$$

It counts the relative number of pairs in a sample $\{(x_{it}, x_{jt})\}_{t=1}^{T}$ whose rank is less than $u \cdot T$ and $v \cdot T$ respectively.

2.2. Factor copulas

Factor copulas are a special class of copula models for which the copula function $C_U(u_1, \ldots, u_N)$ is based on a latent factor structure as defined in Oh and Patton (2013) and Oh and Patton (2017). [4]

2.2.1. General set up

Consider a set of artificial variables $X_i, i = 1, \ldots, N$ which linearly depend on some latent factors $Z_k, k = 1, \ldots, K, K < N$ and some iid error e_i such that $X_i = \sum_{k=1}^{K} \beta_{ik} Z_k + \epsilon_i$. The linear coefficients β_{ik} are also called factor loadings. The latent variables Z_k and the error term ϵ_i follow some parametric distributions with parameter vectors $\boldsymbol{\gamma}_\epsilon$ and $\boldsymbol{\gamma}_{Z_k}$ and one can write: $\epsilon_i \overset{iid}{\sim} F_\epsilon(\boldsymbol{\gamma_\epsilon})$ and $Z_k \sim F_{Z_k}(\boldsymbol{\gamma_{Z_k}})$. While the variables X_i usually dependent on each other, the latent factors are independent from each other and from the error term.

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

The artificial variables X_i are only used to construct the factor copula function $C_U(u_1, \ldots, u_N; \boldsymbol{\theta})$. The parameters of the factor structure are chosen in such a way that the resulting copula function fits the copula of the random vector Y, such that $F_Y(y_1, \ldots, y_n) = C_U(F_{Y_1}(y_1), \ldots, F_{Y_N}(y_N); \boldsymbol{\theta})$. Once the factor copula

[4] Note, that Joe (2015, p. 128) developed another type of copula model under the same term *factor copula*. It is based on the assumption that the dependence structure can be defined through conditional distributions for which the conditioning set is based on some latent variables.

function is determined, the artificial variables and its marginal distributions $F_{X_i}(x)$ are of no interest.

The parameters of the factor model are collected in a parameter vector $\boldsymbol{\theta} = (\beta_{11}, \ldots, \beta_{i1}, \ldots \beta_{ik}, \boldsymbol{\gamma}'_{\boldsymbol{Z_1}}, \ldots, \boldsymbol{\gamma}'_{Z_K}, \boldsymbol{\gamma}'_{\epsilon})'$. It consists of all linear coefficients and the distributional parameters of the error term and the latent variables. 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.[5]

Using matrix notation, the model can be summarized in the following set of equations:

$$\begin{aligned} Y &= (Y_1, \ldots, Y_N)' \\ X &= (X_1, \ldots, X_N)' = \boldsymbol{\beta} \cdot Z + \boldsymbol{\epsilon} \\ F_Y(\boldsymbol{y}) &= C_U(F_{Y_1}(y_1), \ldots, F_{Y_N}(y_N); \boldsymbol{\theta}) \\ F_X(\boldsymbol{x}) &= C_U(F_{X_1}(x_1), \ldots, F_{X_N}(x_N); \boldsymbol{\theta}) \end{aligned} \tag{2.9}$$

To model the joint cdf $F_Y(\boldsymbol{y})$, a two-stage estimation process can be used: First, the marginal distributions $\hat{F}_{Y_i}$ are estimated parametrically or non-parametrically, e.g. by using some parametric model or the EDF. Second, the factor structure for the copula function is fitted to the data by finding the optimal $\hat{\boldsymbol{\theta}}$. In most cases, a closed form of the factor copula does not exist such that one cannot directly relate the copula parameters to the dependency measures as given by (2.5) - (2.7). Therefore, one can adapt simulation based estimation methods as described in section 2.4.

2.2.2. Model restrictions

An upper bound for the number of model parameters $P = |\boldsymbol{\theta}|$ to be estimated is given by the size of the factor matrix and the number of additional free distributional parameters such that $P \leq (N \cdot K + |\boldsymbol{\gamma_{Z_1}}| + \cdots + |\boldsymbol{\gamma_{Z_K}}| + |\boldsymbol{\gamma_{\epsilon}}|)$. To reduce the number of parameters, Oh and Patton (2017, pp. 148, 150) present two restrictions on the matrix of factor loadings $\boldsymbol{\beta}$: the restrictive *equidependence* and the less restrictive *block-equidependence* model.

[5] Oh and Patton (2017, p. 143) 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.

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 set up implies equal pairwise dependencies for all observable variables.

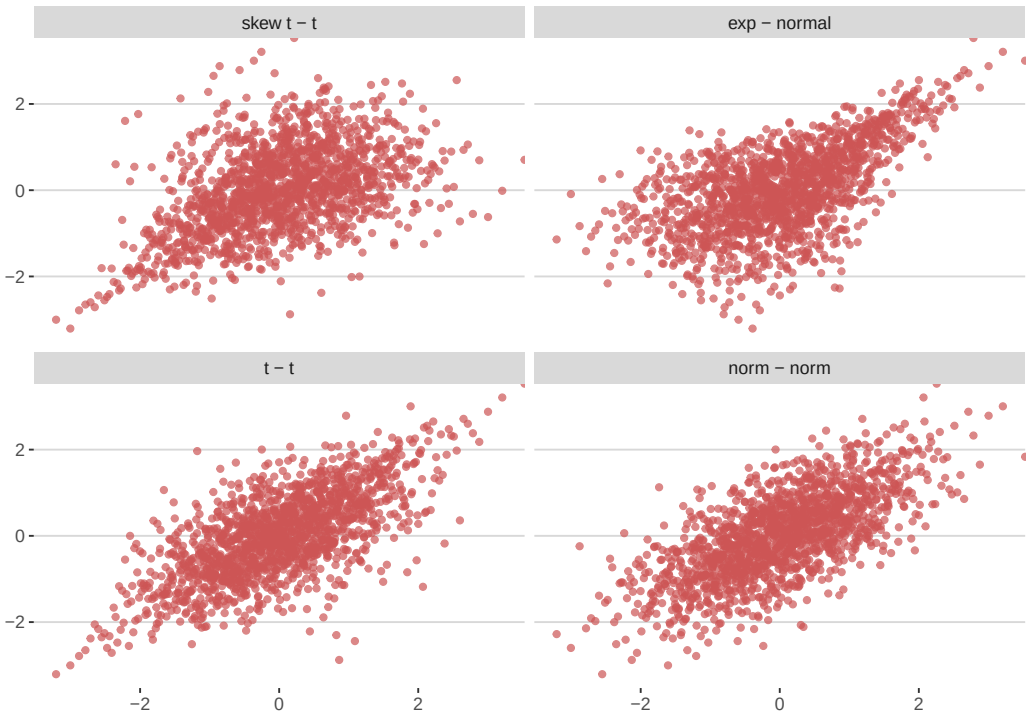


Figure 2.1: Simulations from different equidependence factor copula models with $N = 2$, $\beta = 1.5$, $Y_i \sim N(0, 1)$ and different distributions for the latent variable and the error term.

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 but the symmetry and tail dependency differs through the choice of the distributions and their parameters. The upper left panel shows realizations from a factor copula model with a skew-t distributed latent factor and a t-distributed error term.

The skew-t distribution is a generalization of the t-distribution and allows the specification of asymmetry via a parameter $\lambda \in (-1, 1)$. The normal and t-distribution can be seen as special cases of the skew-t distribution. For $\lambda = 0$

the t-distribution and for $df \to \infty$ the normal distribution is obtained (Hansen 1994, pp. 709, 710).

In this example, the degrees of freedom are set to $df = 4$ and the skewness parameter to $\lambda = -0.8$. This produces strong asymmetric tail dependencies. The factor copula in the bottom left panel has similar distributions. But here symmetric tail dependencies are produced since $\lambda = 0$. This results in an ordinary t-distribution for the latent variable. The bottom right panel shows realizations from a factor model for which both the error term and the latent variable are normal distributed. This yields a multivariate normal distribution for the artificial variables X and the corresponding copula is the Gaussian copula with no tail dependency or asymmetry. The last panel shows the combination of an exponential latent variable with a normal distributed error term. This produces strong upper tail dependencies.

The block-equidependence model, is less restrictive than the equidependence model. It is suitable for higher dimensional problems and for variables which can be naturally partitioned into different groups.[6] The model assumes a common latent factor for all groups and a group specific factor for each group. Thus, each artificial variable X_i is only affected by two factors. For the matrix of factor loadings, 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 structure implies equal pairwise intra-group dependencies while the pairwise inter-group dependencies can vary between the groups.

Formally, consider a partition of $\boldsymbol{Y} = (Y_1, \ldots, Y_N)'$ into M groups. A single variable can then be written as Y_i^j, where $i = 1, \ldots, N$, $j = 1, \ldots, M$. The value k_j is the number of variables in group j and it holds that $\sum_{j=1}^{M} k_j = N$. Then the factor copula model can be summarized as:

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

$$
\begin{aligned}
\boldsymbol{X} &= (X_1^1, \ldots, X_{k_1}^1, X_{k_1+1}^2, \ldots, X_{k_1+k_2}^2, \ldots, \ldots, X_N^M)' = \boldsymbol{\beta Z} + \boldsymbol{\epsilon} \\
\boldsymbol{Z} &= (Z_0, Z_1, \ldots, Z_M)' \\
X_i^j &= \beta_j Z_0 + \beta_{M+j} Z_j + \epsilon_i \\
\boldsymbol{\beta} &= \begin{pmatrix}
\beta^1 & \beta^{M+1} & 0 & \cdots & 0 \\
\beta^1 & \beta^{M+1} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\beta^1 & \beta^{M+1} & 0 & \cdots & 0 \\
\beta^2 & 0 & \beta^{M+2} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\beta^M & 0 & 0 & \cdots & \beta^{M+M} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\beta^M & 0 & 0 & \cdots & \beta^{M+M}
\end{pmatrix}.
\end{aligned}
\tag{2.10}
$$

The matrix $\boldsymbol{\beta}$ is of size $N \times (M+1)$ but with only $2M$ actual factor loadings.

2.3. Copula models for multivariate time series

Up to now, it was tacitly assumed, that observations sampled from the random variables Y are independent from each other, implying a static copula. But one can extent the copula model to univariate or multivariate time series processes for which each observation depends on past realizations. For the former case, the copula is used to estimate the joint distribution of a one dimensional time series $(Y_t, Y_{t+1}, \ldots, Y_{t+n})'$. For the latter, the interest lies in the conditional joint distribution of the time dependent random vector $Y_t = (Y_{1t}, \ldots, Y_{Nt})'$. The conditional cdf can be written as $F_{Y_t|\mathcal{F}_{t-1}}(\boldsymbol{y})$, where the σ-algebra $\mathcal{F}_{t-1}$ contains information from past realizations of Y and other exogenous variables (Patton 2012, pp. 4-6).

Sklar's theorem given by (2.1) shows the connection between a copula function, a multivariate distributions and its marginal cdfs. It can be extended to the multivariate time series case in which the conditional distribution is split into a conditional copula $C_{U_t|\mathcal{F}_{t-1}}(\boldsymbol{u})$ and conditional marginal distributions $F_{Y_{it}|\mathcal{F}_{t-1}}(y)$:

$$F_{Y_t|\mathcal{F}_{t-1}}(y_1, \ldots, y_N) = C_{U_t|\mathcal{F}_{t-1}}(F_{Y_{1t}|\mathcal{F}_{t-1}}(y_1), \ldots, F_{Y_{Nt}|\mathcal{F}_{t-1}}(y_N)). \tag{2.11}$$

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 the following sections, another approach is chosen. We built on a semi-parametric copula-based multivariate dynamic model as described in Chen and Fan (2006, pp. 126-129). First, the vectors with the conditional means and variances of $Y_t|\mathcal{F}_{t-1}$ are estimated parametrically and used to control for possible time dependencies. Second, the standardized residuals are calculated and their joint distribution is then modeled using non-parametric estimates of the marginal distributions and a parametric copula.

Hence, this model describes three levels of estimation: First, the dynamic parameters of the marginal models for the conditional time varying means and variances. Second, the estimation of the marginal distribution of the standardized residuals. Here, this is done non-parametrically with the EDF. Third, the parameters of the copula function for the standardized residuals.

If we denote the parametric conditional mean of a single variable as $\mu_{it} = E(Y_{it}|\mathcal{F}_{t-1}; \boldsymbol{\phi})$ and the parametric conditional standard deviation as $\sigma_{it} = \sqrt{V(Y_{it}|\mathcal{F}_{t-1}; \boldsymbol{\phi})}$ we can write the multivariate time series process as

$$Y_t = \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \eta_t, \tag{2.12}$$

with $\boldsymbol{\sigma}_t = \text{diag}(\sigma_{1t}, \ldots, \sigma_{Nt})$ and $\boldsymbol{\phi}$ the parameter vectors of the univariate time series models.

The standardized random vector of residuals $\eta_t = (\eta_{1t}, \ldots, \eta_{Nt})' = \boldsymbol{\sigma}_t^{-1}(Y_t - \boldsymbol{\mu}_t)$ is independent of past realizations since it is assumed that all dynamics are captured by the conditional mean and standard deviation. Using a parametric copula, the joint cdf of the residuals can be written as:

$$F_\eta(u_1, \ldots, u_t) = C_U(F_{\eta_1}(u_1), \ldots, F_{\eta_N}(u_N); \boldsymbol{\theta}). \tag{2.13}$$

The first level parameters $\boldsymbol{\phi}$ of the conditional means and variances are estimated with standard parametric models for univariate time series, such as models

from the GARCH family (Teräsvirta 2009, p. 17). Given a sample of observed residuals $\{\boldsymbol{\eta}_t\}_{t=1}^T$, its marginal distributions are estimated with the EDF:

$$\hat{F}_{\eta_i}(u) = \frac{1}{T+1} \sum_{t=1}^{T} \mathbf{1}(\eta_{it} \leq u). \tag{2.14}$$

It is assumed that the parametric copula for the residuals is estimated with a factor copula model and the SMM estimation technique which is presented in the following chapter.

2.4. Simulated method of moments estimation for factor copulas

Based on the time series model as shown in the previous section 2.3 and the factor copula model as defined in section 2.2, an estimation procedure known as Simulated Method of Moments (SMM) is presented.

In contrast to the factor copula model, in models for which the copula function is known in closed form, standard Maximum Likelihood (ML) estimation can be performed. Given the time series model as described by (2.12) multistage ML can be applied for estimating the first level conditional means and variances, the marginal distribution of the residuals and the copula. Chen and Fan (2006, p. 127) showed that, if the EDF is used for the estimation of the marginal distribution of the residuals, the first level estimates do not affect the asymptotic distribution of the semi-parametric ML estimator of the copula parameters.

Since the factor copula implied by (2.9) is not given in closed form approximating the Likelihood function is difficult and computationally demanding (Oh and Patton 2013, p. 694). On the other hand, it is easy and computationally cheap to simulate many values from the factor model. Therefore, an alternative approach is used. It is based on the Generalized Methods of Moments (GMM), for which its moment conditions are calculated with simulated values. As shown by Chen and Fan (2006) for the semi-parametric ML estimator, the conditional distribution of the SMM estimator is also not affected by the first level dynamic estimates of the time series model in (2.12) (Oh and Patton 2013, p. 692). This result means, that even under misspecified first level models, a valid copula can still be estimated.

2.4.1. SMM and the Generalized method of moments

SMM and GMM estimation is rooted in the Method of Moments. For many statistical parametric models, theoretical population moment conditions as functions of the model parameters and the random variables can be derived. For estimation, the expectation in these conditions can then be replaced by the respective sample average (Hall 2005, pp. 6-7).

Formally, for SMM the population moment conditions take the form $g(\boldsymbol{\theta}) = E(m(\eta_t)) - m_0(\boldsymbol{\theta})$ with $g(\boldsymbol{\theta}_0) = 0$. We define $m(\eta_t)$ to be a vector of moment functions applied to the random vector of residuals, $\boldsymbol{\theta}_0$ the vector of true parameters and $m_0(\boldsymbol{\theta})$ the theoretical moments as functions of the model parameters. Thus, given the true vector of parameters the moments match their theoretical counterparts. E.g. given a normal variable $X \sim N(\mu, \sigma)$ the condition can simply be written as $E(X) - \mu = 0$.

The moment conditions can be estimated by replacing the expectation with the sample average. If exactly as many moment conditions as model parameters exist, the system can be solved exactly. If more moment conditions than parameter are available no unique solution exist and one tries to minimizes the distance of the sample estimator $g_T(\boldsymbol{\theta})$ to 0. Given a sample of observed residuals $\{\boldsymbol{\theta}_t\}_{t=1}^T$ the distance is measured by the weighted quadratic form:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{GMM} &= \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) = g_T(\boldsymbol{\theta})' W_T g_T(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} \left(\frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{\eta}_t) - m_0(\boldsymbol{\theta})\right)' \hat{W} \left(\frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{\eta}_t) - m_0(\boldsymbol{\theta})\right).\end{aligned} \tag{2.15}$$

Under suitable assumptions and by applying a law of large numbers and the central limit theorem it can be shown that the estimator convergence in probability to the true parameters and that it is asymptotically normal distributed.

The weighting matrix is a positive semi-definite symmetric matrix which must converge in probability to a positive definite matrix of constants. By doing so it is ensured, that a positive weight is applied to all moment conditions. The variance of the GMM estimator depends on the choice of the weighting matrix. Therefore, W_T should be chosen in such a way, that it minimizes the asymptotic variance of the GMM estimator (Hall 2005, pp. 14, 43).

For the GMM, stating the vector of true moments $m_0(\boldsymbol{\theta})$ is often straightforward. In cases, where there exist no explicit mapping from the model parameters to the true moments but the DGP is known and simulations are easy to obtain one can use the SMM (Hall 2005, pp. 342-347). For this, the theoretical moments $m_0(\boldsymbol{\theta})$ are replaced by moments calculated with simulated data $\{\boldsymbol{X}\}_{s=1}^{S}$ from the model. The simulated moments are then compared to the observable sample moments. The SMM estimator is therefore similar to the GMM version but the moment condition changes to

$$g_T = \frac{1}{T}\sum_{t=1}^{T}(m(\boldsymbol{\eta}_t)) - \frac{1}{S}\sum_{s=1}^{S}(m(\boldsymbol{X}_s;\boldsymbol{\theta})). \tag{2.16}$$

For most cases the SMM estimator is consistent and asymptotically normal distributed. However, since the estimation error from the simulations cannot be neglected, the variance of the SMM is usually larger. The inefficiency vanishes if $\frac{T}{S} \to \infty$ (Hall 2005, pp. 344-345).

2.4.2. SMM for factor copulas

The presented estimation method and its asymptotic results are valid for well defined conditions based on the moments of the variables, e.g. the mean or the variance. For factor copulas, these moments cannot be formulated. The scale-invariant measures of dependency as defined in (2.5) - (2.7) are functions of the copula. For the estimation approach presented by Oh and Patton (2013), they are used to replace the moment conditions of the ordinary GMM approach. Therefore, the assumptions and the asymptotic results of the GMM estimator are not valid and a law of large numbers doesn't apply. However, as the authors proved, asymptotic results for the distribution of the copula parameters $\hat{\boldsymbol{\theta}}$ can still be derived.

For factor copulas, the "moment" conditions for the SMM estimation consist of vector functions of pairwise dependency measures separately calculated with the observed matrix of residuals $\boldsymbol{\theta}$ and S simulations $\boldsymbol{X}$ from the copula model as given by (2.9). The moment condition g_t in the GMM objective function given by (2.15) and the SMM adaption (2.16) is replaced by

$$g_T(\boldsymbol{\theta}) = m_T(\boldsymbol{\eta}) - m_S(\boldsymbol{X};\boldsymbol{\theta}), \tag{2.17}$$

where the vector function $m(\cdot)$ generates a vector of pairwise dependency statistics. As measure of dependency, the empirical counterparts of Spearman's Rho as given in 2.6 and lower and upper quantile dependency as given in 2.7 for $q \in (0.05, 0.1, 0.9, 0.95)$ are chosen. Other invariant measures such as Kendall's Tau can also be included.

Formally, we can write $m(\cdot) = \{\boldsymbol{\delta}_{i,j}\}_{i=1,j=i}^{N-1,N}$, with $\boldsymbol{\delta}_{i,j} = (\hat{\rho}_{i,j}^{S}, \hat{\lambda}_{i,j}^{0.05}, \hat{\lambda}_{i,j}^{0.1}, \hat{\lambda}_{i,j}^{0.90}, \hat{\lambda}_{i,j}^{0.95})'$. For an unrestricted factor model there exist $5 \cdot (0.5 \cdot N \cdot (N-1))$ dependency statistics - 5 for each unique pair of variables. For the simpler equidependence model, which assumes the same factor loading for all variables, one can average over all pairwise combinations such that we can write:

$$m(\cdot) = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \boldsymbol{\delta}_{i,j}. \tag{2.18}$$

This results in a single vector of dependencies.

For the bloc-equidependence model the final number of dependencies is $5 \cdot M$ because the model structure allows one to average over all intra- and inter-group dependencies. Using the same notation as in (2.10), we can write for each group $m = 1, \ldots, M$

$$\bar{\boldsymbol{\delta}}_m = \frac{1}{M}\Big(\underbrace{\sum_{r=1, r\neq m}^{M} \frac{1}{k_m k_r} \sum_{i=1}^{k_m} \sum_{j=1}^{k_r} \boldsymbol{\delta}_{im,jr}}_{\text{inter-group dependencies}} + \underbrace{\frac{2}{k_m(k_m-1)} \sum_{i=k_m}^{k_m-1} \sum_{j=i+1}^{k_m} \boldsymbol{\delta}_{im,jm}}_{\text{intra-group dependencies}}\Big), \tag{2.19}$$

where $\boldsymbol{\delta}_{im,jr}$ is the vector of dependencies for the ith variable in group s and the jth variable in group r. Finally, for the bloc-equidependence model, the moment function can be written as: $m(\cdot) = (\bar{\boldsymbol{\delta}}_1, \ldots, \bar{\boldsymbol{\delta}}_M)'$.

Oh and Patton (2013, p. 691ff) showed that under some regularity 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. Notably, it is assumed that the population version of the moment condition $g(\boldsymbol{\theta})$ is differentiable at the true parameter $\boldsymbol{\theta}_0$ while this is not true for the sample version $g_T(\boldsymbol{\theta})$.

The convergence of the estimator can be summarized as:

$$\hat{\boldsymbol{\theta}}_{SMM} \sim N(\boldsymbol{\theta}_0, (\frac{1}{T} + \frac{1}{S})\boldsymbol{\Omega}) \text{ for } T, S \to \infty, \tag{2.20}$$

with covariance matrix $\boldsymbol{\Omega} = (G'WG)^{-1}G'W'\Sigma WG(G'WG)^{-1}$, where $G = \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_0}$ is the Jacobian matrix of the first order derivatives of the moment condition $g(\boldsymbol{\theta})$ and Σ is the asymptotic variance of the vector of dependencies $m_T(\boldsymbol{\eta})$ calculated with the observed sample of residuals.

The weight matrix W can be chosen to be the identity matrix which results in some efficiency losses. On the other hand,the efficient weight matrix which minimizes the variance of the estimator is given by $W = \Sigma^{-1}$ and the overall variance reduces to $\Omega = (G'WG)^{-1}$. In contrast to ordinary GMM estimation, the efficient weight matrix does not depend on the parameters of the factor copula model and can be estimated consistently before estimating the parameters $\boldsymbol{\theta}$.

The asymptotic variance Σ of the sample moments can be estimated with an iid bootstrap procedure:

Algorithm 1: Bootstrap procedure for estimating the asymptotic variance $\Sigma_{T,B}$

Input: $\{\boldsymbol{\eta}_t\}_{t=1}^T$
Output: $\Sigma_{T,B}$
$\hat{m}_T \leftarrow m(\boldsymbol{\eta})$;
for $b \leftarrow 1$ **to** B **do**
  $\boldsymbol{\eta}_t^{(b)} \leftarrow$ sample T values with replacement from $\{\hat{\boldsymbol{\eta}}_t\}_{t=1}^T$;
  $\hat{m}_T^{(b)} \leftarrow m(\boldsymbol{\eta}^{(b)})$;
end
$\Sigma_{T,B} \leftarrow \frac{T}{B}\sum_{b=1}^B(\hat{m}_T^{(b)} - \hat{m}_T)(\hat{m}_T^{(b)} - \hat{m}_T)'$;
return $\Sigma_{T,B}$;

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

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

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. A structural break test for factor copulas

The factor copula as presented in the last sections is used in a dynamic model as given by (2.12). Here, the dependence structure of the residuals is assumed to be static since all dynamics are captured by the conditional mean and variance of the marginal distributions. But given the findings of time varying correlations in real datasets the assumption of a static copula seems unrealistic (Manner and Reznikova 2012, p. 655).

Different approaches exist to allow for a time varying dependence structure (Manner and Reznikova 2012, pp. 658 - 668): For instance, the parameters of the copula can be modeled as functions of the lagged variables, auto-regressive terms or as an independent stochastic process. But one can also test for a structural break in the copula parameters at some point in time t. While these approaches assume that the functional form of the copula doesn't change other methods explicitly allow multiple alternating copula functions.

The dynamic model as presented in 2.3 allows for a wide variety of parameterization and copula functions. Here, we focus on the factor copula model and the SMM estimation procedure as presented in the previous sections. In the following section a structural break test based on Manner, Stark, and Wied (2017) for a change of the parameters of the factor copula as given by (2.12) is presented. The underlying factor copula function doesn't change, while the copula's parameters are allowed to vary over time.

The test is based on recursive estimations of the copula model at each time point $t \in \{\lfloor \epsilon T \rfloor, \ldots, T\}$, with $0 < \epsilon < 1$. We write the recursive parameter as $\boldsymbol{\theta}_t$. For its estimation, all information from the first period up to point t is used. The strictly positive trimming parameter ϵ has to be chosen such that the model parameters don't fluctuate too much due to the larger estimation error of smaller sample sizes. In the following applications we usually choose ϵ such that the first recursive period starts at $t = 300$.

The null hypothesis states that the parameters of the factor copula model are stationary while the alternative assumes at least one significant break point in time:

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_T \qquad H_1 : \boldsymbol{\theta}_t \neq \boldsymbol{\theta}_{t+1} \text{ for some } t = \{1, \ldots, T\}. \tag{2.21}$$

For each period, the recursive estimator $\hat{\boldsymbol{\theta}}_t$ is compared to the the estimated parameters of the full model $\hat{\boldsymbol{\theta}}_T$. If the recursive estimator is significantly far away from the full model estimates, H_0 is rejected. Formally, the recursive statistics are calculated as the scaled squared distances between the full model and the recursive estimator. The final test statistic is then given as the maximum distance:

$$P = \max_{\epsilon T \leq t \leq T} (\frac{t}{T})^2 T(\hat{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_T)'(\hat{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_T). \tag{2.22}$$

Manner, Stark, and Wied (2017, p. 10) also propose a non-parametric alternative test statistic which is solely based on the moment functions calculated with the residuals up to point t as described in section 2.4.2. It does not involve the recursive estimation of a copula model:

$$M = \max_{\epsilon T \leq t \leq T} (\frac{t}{T})^2 T(m_t(\boldsymbol{\eta}) - m_T(\boldsymbol{\eta}))'(m_t(\boldsymbol{\eta}) - m_T(\boldsymbol{\eta})), \tag{2.23}$$

where $m_t(\boldsymbol{\eta})$ is the recursive vector of dependencies calculated with the residuals up to time t.

For the break point detection, the test statistic is compared to a critical value (CV). The CV is based on the $1-\alpha$ quantile of the distribution of M or P given that H_0 is true. It can be simulated with a bootstrap procedure as described in algorithm (2). If the test statistic is larger than the CV, H_0 is rejected and the corresponding time point of the significant test statistic is chosen as the breakpoint.

Manner, Stark, and Wied (2017, p. 13) show, that the test statistics converges in distribution to

$$\begin{aligned} P &\xrightarrow{d} \max_{\epsilon T \leq t \leq T} (\boldsymbol{A}^*(\frac{t}{T}) - \frac{t}{T}\boldsymbol{A}^*(1))'(\boldsymbol{A}^*(\frac{t}{T}) - \frac{t}{T}\boldsymbol{A}^*(1)), \\ M &\xrightarrow{d} \max_{\epsilon T \leq t \leq T} (\boldsymbol{A}(\frac{t}{T}) - \frac{t}{T}\boldsymbol{A}(1))'(\boldsymbol{A}(\frac{t}{T}) - \frac{t}{T}\boldsymbol{A}(1)), \end{aligned} \tag{2.24}$$

with $\boldsymbol{A}^*(\frac{t}{T}) = ((G'WG)^{-1}G'W(A(\frac{t}{T}) - \frac{t}{T}\sqrt{(\frac{T}{S})}A(1))$ and $T, S \to \infty$. As for the SMM estimator, G is the matrix of first order derivatives of the population moment condition $g(\boldsymbol{\theta})$, W is the weighting matrix of the SMM estimator and

$A(\cdot)$ is some Gaussian process (Manner, Stark, and Wied 2017, p. 39). The necessary regulatory assumptions to derive this result are similar in spirit to the assumptions used to derive the asymptotic distribution of the parameters of the factor copula model in section 2.4.2.

Given a sample of residuals, the distribution of the test statistics under the null hypothesis can be estimated using the following bootstrap procedure:

Algorithm 2: Bootstrap procedure for estimating the distribution of the test statistics under H_0

Input: $\{\boldsymbol{\eta}_t\}_{t=1}^T$
Output: $\{\boldsymbol{K}_b\}_{b=1}^B$
$\hat{m}_T \leftarrow m_T(\boldsymbol{\eta})$;
if *test based on* P **then**
  $L \leftarrow (G_T' W_T G_T)^{-1} G_T'$;
else
  $L \leftarrow 1$;
end
for $b \leftarrow 1$ **to** B **do**
  $\boldsymbol{\eta}^{(b)} \leftarrow$ sample T values with replacement from $\{\hat{\boldsymbol{\eta}}_t\}_{t=1}^T$;
  for $t \leftarrow \lfloor \epsilon T \rfloor$ **to** T **do**
    $\hat{m}_t^{(b)} \leftarrow m(\boldsymbol{\eta}_t^{(b)})$;
    $A_t^{(b)*} \leftarrow L \frac{t}{T} \sqrt{T} (\hat{m}_t^{(b)} - \hat{m}_T)$;
    $K_t^{(b)} \leftarrow (A_t^{(b)*} - \frac{t}{T} A_T^{(b)*})' (A_t^{(b)*} - \frac{t}{T} A_T^{(b)*})$;
  end
  $K^{(b)} \leftarrow \max(K_{\lfloor \epsilon T \rfloor}^{(b)}, \ldots, K_T^{(b)})$;
end
return $\{\boldsymbol{K}_b\}_{b=1}^B$;

The procedures calculates B bootstrap versions of the maximum scaled squared distance between the recursive moments $\hat{m}_t$ and the full sample estimate $\hat{m}_T$. For the copula based test $(G'WG)^{-1}G'W$ is estimated only once using the full sample estimates and the estimation procedures as described in section 2.4.2 and in algorithm (1).

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

In this chapter we focus on the implementation of the previous methods and procedures. Using the programming language *R* (R Core Team 2017) an open-source package was built, such that the methods can be easily tested, installed and distributed .[7] First, the functions for the configuration and simulation of factor copulas are shortly presented. Second, an overview of the numerical optimization strategy is given. Explained code examples of the main functions can be found in the appendix section A.2. Finally, the validity of the package and the discussed procedures is tested in two simulation studies.

The package consists of a set of high level functions which can be used to construct, simulate and fit various factor copula models. The specification of the factor copula model is handled by the functions `config_factor`, `config_error` and `config_beta`. The two functions `fc_create` and `fc_fit` can be used to either simulate values from a factor copula or to fit a model to a dataset. For conducting the break test as described in section 2.5, the library offers the functions `fc_critval`, `fc_mstat` and `fc_pstat`. The former simulates critical values for either the moments or copula based test. The latter two calculate the recursive test statistics. A matrix of recursive parameter estimates $\hat{\boldsymbol{\theta}}_t$ must be provided for the calculation of the copula based test statistic. It can be obtained by recursively applying `fc_fit` to the data.

Obtaining the recursive estimates is computationally costly. To speed up the estimation, the HPC cluster of the University of Cologne was utilized for all empirical applications(Achter, Borowski, Nieroda, et al. 2018).

3.1. Copula specification and simulation

With the functions `config_factor`, `config_error` and `config_beta`, the user can define the distribution of the latent variables Z and the error term ϵ as well as the matrix of factor loadings $\boldsymbol{\beta}$. For the specification of the distributions, the function name of any available random number generator (e.g. `rnorm`, `rt`, `rst`) can be used.

[7] See also the appendix notes on data access in section A.1.

Additional arguments such as the distributional parameters must be declared in a named list. The parameters can either be pre-defined or passed as non-evaluated expressions similar to the internal `formula` interface of the R programming language. To distinguish free model parameters from fixed distributional parameters, an additional character vector with the name of the model parameters has to be passed to `config_factor` and `config_error`.

The random number generating functions must have an additional argument `n` which defines the number of observations to be simulated. It is not necessary to set this argument explicitly since the number of simulations is controlled by the value S in other functions of the package.

For the matrix of factor loadings the user can either manually construct a character matrix of parameters or one can use the function `config_beta`. Given a vector `k` and the number of latent variables K this functions constructs a suitable character matrix of zeros and parameter names. The vector k is of length N and defines the group for each observable variable. Therefore, an equidependence model can be specified with $k_N = (1, 1, \dots, 1)$, an unrestricted model with $k_N = (1, \dots, N)$ and a bloc-equidependence model with $k_N = (1, 1, \dots, 2, 2, \dots, M, M, \dots)$, where M is the number of groups.

Given the copula specifications, the function `fc_create` returns itself a random number generating function. To simulate values from it, the user has to specify a vector `theta` of parameters, the number of simulations S and an optional random seed. Fixing the seed at some integer value can be useful to reproduce the same simulation results over several runs of a program. The vector `theta` must be a *named* vector for which the names correspond to the model parameters specified during the configuration of the model.

The simulation of new values is a time consuming part. To avoid unnecessary calls to the underlying random number generators, the function remembers the state of the latent variables Z and the error term ϵ over repeated function calls and only updates the values if necessary. Thus if neither the seed nor the distributional parameters in `theta` change, the function uses the same random values from a previous call. This reduces the computation time at the cost of a higher memory consumption.

3.2. Optimization strategy

While the function `fc_create` can be used to simulate values from a factor copula model given a vector of parameters, the function `fc_fit` estimates $\boldsymbol{\theta}$ via SMM as described in section 2.4. Given a copula specification and the observed residuals, it finds the optimal $\hat{\boldsymbol{\theta}}$ which minimizes the weighted squared distance between the dependency vectors calculated with simulated data from the factor model and with observable data as presented in equation (2.16).

Optionally, the function also estimates standard errors using the bootstrap algorithm given by (1). For simplicity, the weight matrix W is set to the identity matrix. Previous studies showed no significant improvement when the theoretically efficient weight matrix $W = \hat{\boldsymbol{\Sigma}}_{T,B}^{-1}$ is used (Oh and Patton 2013, p. 694).

As a back end, the method builds on top the `NLopt` optimization library which implements various algorithms for global, local or derivative-free optimization (Johnson 2018; Ypma 2014). The specific choice of the optimization algorithms and the stopping criteria can be altered by the user via the arguments `control.first.stage` and `control.second.stage`. The authors of the library recommend a two step optimization procedure for global optimization: First, a global optimizer approximates the parameter region in which the global optimum lies. Second, a local derivative free optimizer is then applied using the approximated solution from the first stage.

By default, `fc_fit` uses the *Multi-Level Single-Linkage* algorithm in the first step. The algorithm creates a sequence of optimal distributed starting values which are then passed to a local optimizer (Kucherenko and Sytsko 2005). For the local optimizer and the second stage, the *Subplex* algorithm is used, which is based on the popular Nelder-Mead Simplex procedure (Rowan 1990).

During the optimization process, many values from the factor copula model are simulated. To avoid numerical instabilities, the random seed is kept fixed, such that always the same random values are drawn (Gouriéroux and Monfort 1996, p. 29). As described in the previous section, the memory functionality of `fc_create` can save some computational costs, since redraws from the distributions are avoided if the distributional parameter don't change. This can improve the overall performance of the optimization process. This is especially true, if only the factor loadings $\boldsymbol{\beta}$ are optimized while the distributional

parameters are kept fixed. In this case, the random number generators are only called once at the beginning of the optimization routine.

3.3. Simulation study

To illustrate the discussed methods and the validity of the package, two simulation studies are performed: First, an equidependence factor copula model with varying dimensions and sample sizes is estimated repeatedly to show the consistency property of the SMM procedure. Second, both the moments and copula based structural break test are applied once to simulated data from a bloc-equidependence model as described in section 2.2.2.

3.3.1. Convergence for the equidependence model

The DGP of the first study is based on a simple time-invariant equidependence model with standard-normal marginal distributions, one skew-t distributed latent variable and a t-distributed error term. The copula model produces strong asymmetric tail-dependencies. This is achieved by fixing the degrees of freedom at $df = 4$ and the skewness at $\lambda = -0.8$, similar to the upper left graph of figure **??**. The single factor loading is set to $\beta = 1.5$. The only free parameter to be estimated is the factor loading. The distributional parameters are kept fixed at their true values. Thus we can write $\theta = \beta$. The model equations can be written as:

$$\begin{aligned} Y_i &\sim N(0,1) \ \forall \ i \in 1, \dots, N \\ Z &\sim \text{skew-t}(df = 4, \lambda = -0.8) \\ \epsilon &\sim \text{t}(df = 4) \\ X_i &= \beta Z + \epsilon \end{aligned} \tag{3.1}$$

We repeat the simulation and estimation of β over a grid of different values for the number of variables N and the sample size T. Each simulation consists of $C = 1000$ Monte-Carlo replications. The number of simulations in the SMM is set to $S = 25000$. Finally, we get a vector of estimates $\hat{\beta}_{t,n,c}$ with $t \in \{100, 1000, 10000\}, n \in \{2, 3, 10\}, c \in \{1, \dots, C\}$.

Figure 3.1 shows the results for the first study. For each combination of N and T the kernel density estimator over all Monte-Carlo simulations of is plotted.

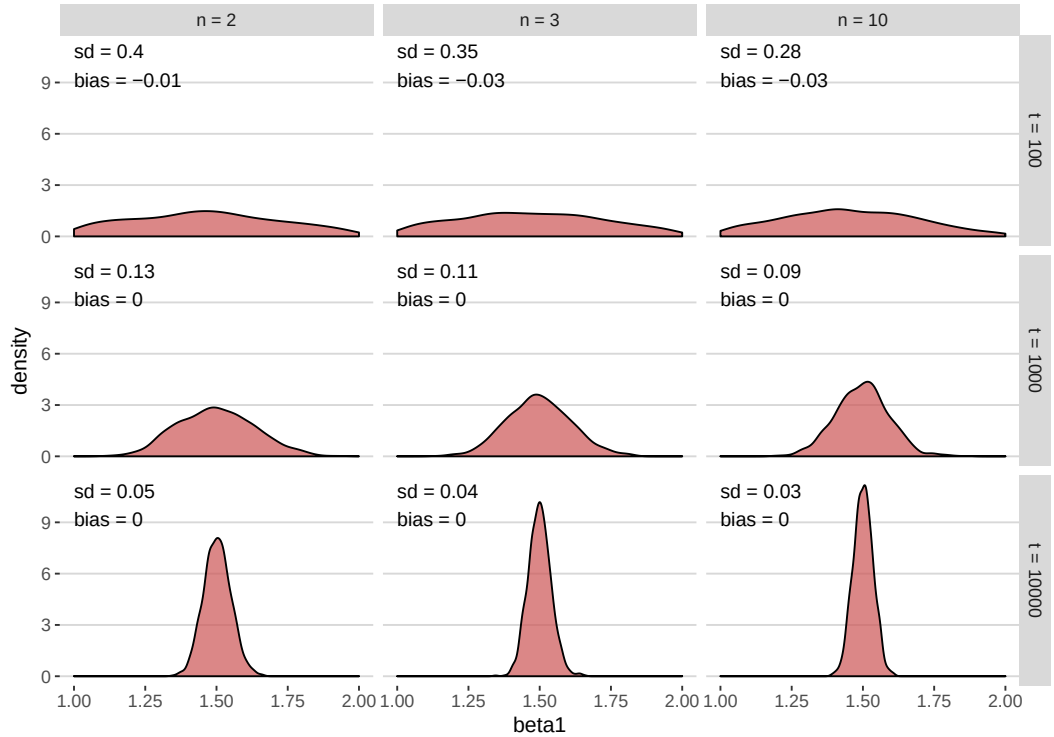


Figure 3.1: Approximated Monte-Carlo density estimators for $\hat{\beta}$ of an equidependence skew t - t factor copula model with $\beta = 1.5$, $S = 25000$, standard normal distributed marginals and different values for N and T. Each simulation is based on 1000 Monte-Carlo replications.

The bias $\hat{b}_{t,n} = \frac{1}{C}\sum_{c=1}^{C}\hat{\beta}_{t,n,c} - \beta$ and the standard deviation is printed in the top left corner.

The density mass centers around the true value of $\beta = 1.5$. For all simulations the bias is zero or close to it. As the sample size increases, the deviation gets smaller and the SMM estimator converges to the true parameter. But small sample sizes clearly give unreliable results.

Notably, more variables improve the quality of the estimator as more information is available to estimate the single parameter. This is due to the fact that the number of latent variables and parameters is constant and does not increase with the number of dimensions. As stated in (2.18) for the equidependence model one averages over all pairwise dependence vectors. Therefore, the estimation error of the mean decreases if N and the number of pairwise combinations $0.5\dot{N}(N-1)$ increases.

3.3.2. Structural break in a bloc-equidependence model

For the second study, a more sophisticated model is presented 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) the DGP is based on a *bloc-equidependence* model as described in section 2.2.2.

The model consists of $N = 21$ standard normal distributed variables which are partitioned in $M = 3$ groups of equal size. Each variable is affected by a common and a group specific factor. The common latent factor is skew-t distributed with strong asymmetric tail-dependency. The three group specific latent factors and the error term are t-distributed. Due to the bloc-equidependence structure, the number of factor loadings reduces from $0.5 * N * (N - 1) = 210$ to just $2M = 6$. Again, all distributional parameters are kept fixed at their true values and only the factor loadings are estimated such that the recursive estimator can be written as $\boldsymbol{\theta}_t = (\beta_{1,t}, \ldots, \beta_{6,t})'$.

We chose $T = 1500$, $S = 25 \times T$ and a breakpoint in the parameters of the copula at $t = 1000$. Before the break, $\boldsymbol{\theta}_{pre} = (0, 1, 1, 0, 1, 1)'$ and after the break $\boldsymbol{\theta}_{post} = (1.5, 1, 1, 1.5, 1, 1)'$. Thus, only the intra- and interdependence for the first group increases from 0 to 1.5 while the remaining factor loadings stay constant.

The test statistics and critical values are based on three different recursive calculations over a range from $t = 300$ to $T = 1500$: First, all parameters are estimated via recursive SMM. Second, only a subset of $\boldsymbol{\theta}_t$ is estimated recursively while fixing the common and group specific factor for the first group at their full sample estimates $\hat{\beta}_{1,T}$ and $\hat{\beta}_{4,T}$. Hence, for this calculation $\boldsymbol{\theta}_t = (\beta_{2,t}, \beta_{3,t}, \beta_{5,t}, \beta_{6,t})$. Finally, the non-parametric test statistics based on the moment functions are calculated.

We expect, that a breakpoint is detected around $t = 1000$ only for the first and third recursive calculations. For the second calculation no breakpoint should be detected since the factor loadings of the second and third group are not affected by the simulated structural break.

Figure 3.2 shows the test statistics of the three different break tests for a single recursive run of the simulation for $t = 300, \ldots, 1500$. The horizontal solid line indicates the critical value. Each critical value is calculated using $B = 2000$

bootstrap samples as described in algorithm (2). The vertical line indicates the theoretical breakpoint at $t = 1000$.

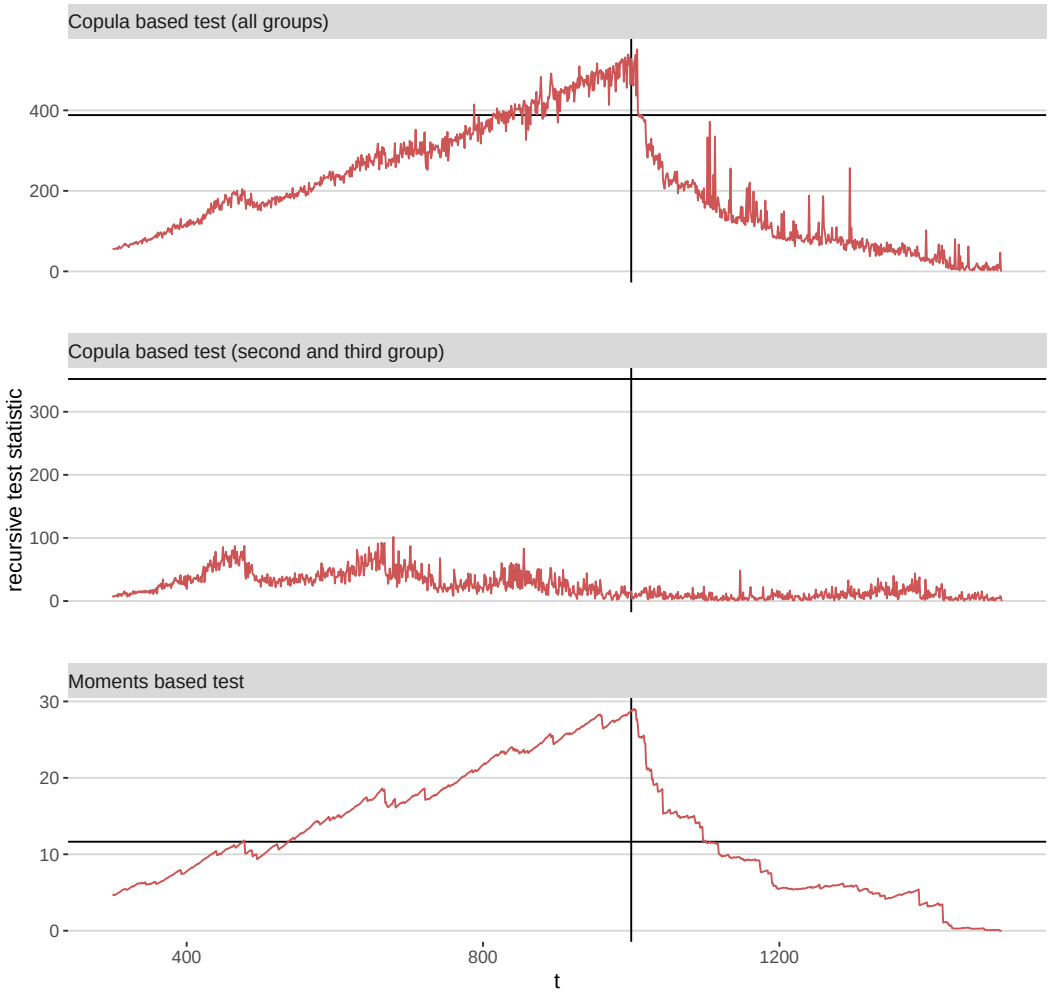


Figure 3.2: Parametric and non-parametric structural break test for a bloc-equidependence model with $N = 21$ and $M = 3$ groups of equal size. The theoretical breakpoint is at $t = 1000$ and is modeled as a change of the intra- and interdependence of the first group from 0 to 1.5. The first panel shows the copula based test based on all groups. The second panel the copula based test for only the second, and third group which are not affected by the structural break. The lower panel shows the moment based test.

As expected, both the full copula and the moments based test statistics in the top and lower panel fluctuate strongly. For both tests the maximum distance occurs close at the true breakpoint and lies above the CV. The CV for the moments based test is relatively far away from the maximum while this is not true for the full copula based test. This could indicate that the moments based

test is less restrictive. This observation can also be made in the empirical example in the following chapter. The copula based test statistics for the second and third group are clearly below their critical value. Hence, for this groups the null hypothesis of no parameter change cannot be rejected.

Often, numerical problems arise for large parameter vectors or parameter vectors which include distributional parameters, such as the degrees of freedom or the skewness. Therefore, the test statistics can be substantially distorted with unusual extreme outliers above the CV. We advise to always perform a manual graphical analysis of the final test statistics. In addition, clear outliers could be detected by performing a smoothing method such as running medians or smoothing splines (see also section 5).

Finally, table 3.1 shows the results for a copula model estimated on the full sample, before and after the theoretical breakpoint. For all models we fixed the distributional parameters at their true values. The standard errors $(\frac{1}{T} + \frac{1}{S})\mathbf{\Omega}$ as given by (2.20) were estimated with $B = 2000$ bootstrap replications and are reported in parenthesis.

coefficient	pre-break ($t \leq 1000$)	post-break ($t > 1000$)	full sample
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 sample. Standard errors in parenthesis (estimated with $B = 2000$ bootstrap samples).

Comparing the estimates and the value of the objective function at the optimal value, one can notice that the model after the breakpoint and for the full sample are less precise. For the former case, this can be due to the small sample size. For the latter case, this can be due to the misspecified copula, which assumes

no change in parameters. The estimates for the pre- and post-break sample are close to their true values. The standard errors for the post-break period are larger due to the smaller sample size. However the coefficients are still significant.

4. Modelling dependencies in topic salience over time

In this chapter, we apply the previously discussed methods to a real aggregated dataset collected from the social media platform Facebook. Typically, copula models are applied to model the dependency structures of financial data. For example, the structural break test was used to detect changing dependencies between daily returns of stock market assets during the financial crisis (Manner, Stark, and Wied 2017, p. 18). The aim of this analysis is to explore other ways of applying the discussed methods in areas outside from applications in finance and risk management.

The analysis in this paper is broadly related to studies of issue or topic salience in politics and the agenda-setting theory. Issue salience generally refers to the importance of certain topics for individuals or political actors (Wlezien 2005, p. 557). Here, by "importance", we refer to the prominence or publicity of an issue. Traditionally, agenda setting is a term which stands for the potency of the media to put policy issues on the public agenda. Therefore, this theory can explain positive dependencies between measures of press media coverage and survey measures of topic salience in a society. The growth of social media channels, blogs and online news sites in the internet has lead to a more diverse mechanism of agenda-setting but it also allows for more precise measurements of salience and coverage over time (W., Lauren, S., et al. 2014, pp. 193-195).

Specifically, the case of the German refugee crisis in late 2015 and its perception by the major German parties on the online platform Facebook is analyzed. In 2014 a steady rise of incoming refugees to Germany and Europe could be observed. The majority of them fled from the war in Syria. In 2015 the number of refugees increased rapidly with around 200000 refugees entering Germany in November 2015 (compare to figure 4.1 and bpb (2018)).

Using text data from public social media posts of politicians and parties we

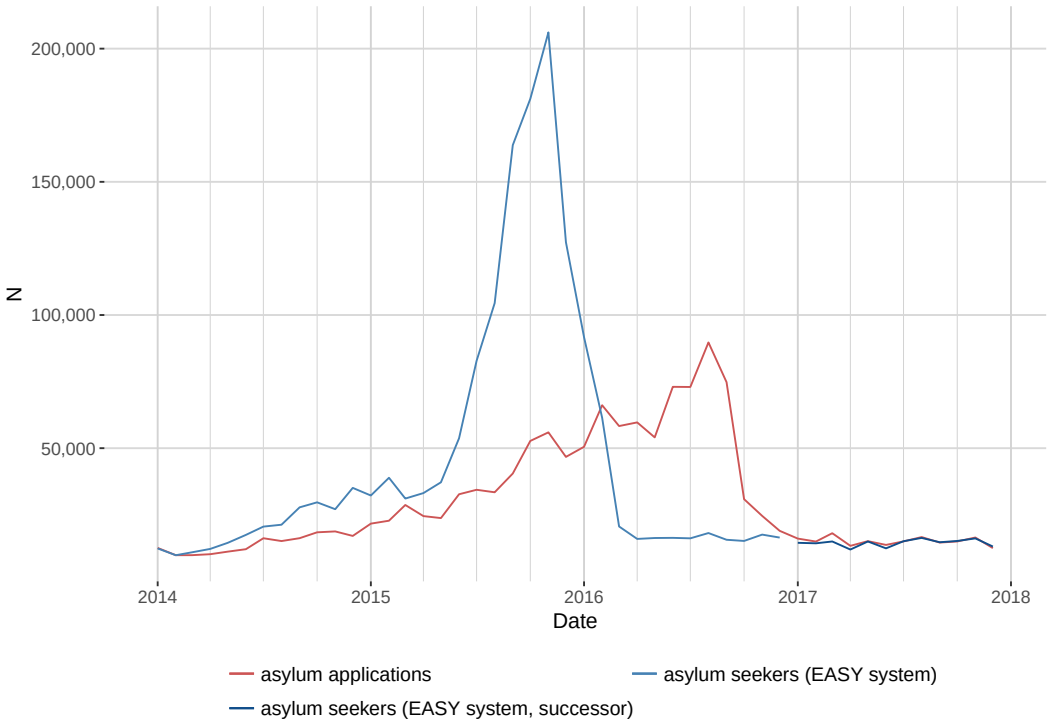


Figure 4.1: Monthly asylum seekers and applications from 2014 - 2017. Upon arrival in Germany an asylum seeker is first registered by the authorities (EASY system). Afterwards, a formal application process can be initialized. The actual number of asylum seekers is smaller, since the old EASY system contains around 13% - 18% duplicate entries due to its decentralized structure. After January 2017 in reaction to the crisis, a new registration system was initialized. Source: bpb (2018)

derive a measure of issue salience for refugee and asylum related topics. The statistic is calculated separately for each political party and day and hence it measures the importance of the topic within the political discourse of each party. We then analyze whether and how the dependencies between the measures changed due to the refugee crisis.

In the following, we use the dynamic model as described by (2.12) together with the factor copula function as given by (2.9). It is the same set up as in the previous chapters. First, we give an overview over the raw dataset. Second, we give a description of the feature generation process and present the first stage models of the conditional mean and variance. Third, estimation results for various factor copula models applied to the residual data are presented. Finally, we report the results for both the non-parametric and parametric structural break test.

4.1. The *btw17* social media dataset

The raw data 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 account for each of the politicians was manually researched. Only candidates from the six factions in the *Bundestag* (CDU/CSU, SPD, Die Linke, Bündnis 90/ Die Grünen, AfD, FDP) are part of the study. Around 84% of all 2516 candidates have an account on Facebook (Stier, Bleier, Bonart, et al. 2018, p. 16).

Due to API (application programming interface) and privacy restrictions, only information from public pages could be accessed such that around 52% of the social media accounts could be considered for the data collection. In addition to the candidates pages, 113 official pages from the political parties, both on the federal and regional level, were included.

The data collection took place on several days between 2017-11-21 and 2018-02-06. The web-scraping software is built on top of the *restfb* Java client library and makes calls to Facebook's official Graph API (Allen and Bartels 2018). The posts are stored in a document orientated database on cloud-servers located in Germany. Besides the actual content of the post, it includes a time-stamp, the user-id of the author and the number of likes, shares and comments it has received upon collecting it from the API.

For this analysis, the data is restricted on cleaned textual posts only.[8]. This results in almost 664 thousand posts tagged with the party membership of their authors. The first two panels of figure 4.2 show the monthly number of active accounts and the monthly number of posts for each party. An account was defined active if it published at least one post for a month considered.

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 a drop after the election. This trend has two reasons: First, we only collected accounts from politicians which were candidates in the btw17. Therefore, other politicians which were only active before the btw17 are not considered here. Second, we suspect that many politicians opened an account

[8] A post can also contain a photo, an album or an event with no additional text. The post's text was cleaned by removing links and stop-words, transforming umlauts and converting it to lowercase letters.

just for the election campaign. After the election many candidates closed their accounts due to a failure in the election or because campaigning time was over.

A similar pattern can also be observed for the monthly number of posts. Until the election year roughly 10000 posts from political parties were published per month. In the month of the election it was around 4 times more.

The bottom panel of figure 4.2 shows the absolute number of monthly posts related to refugee and asylum issues. As in W., Lauren, S., et al. (2014, pp. 192-198) we define a key identifying term for an issue and simply use it to search and filter the data.

Specifically, we define a post to be "refugee related" if it matches the regular expression `flucht|fluecht`. For example, a match occurs if a post contains the words *flüchtlingskrise* (refugee crisis), *fluchtursachen* (causes of flight), *flüchten* (to flee) or *flüchtlingsheime* (refugee hostels). A sharp rise of refugee related posts can be observed in autumn and winter of 2015 and 2016. Notably, this pattern is very similar to the overall number of refugees who entered Germany in 2015 with the peak occurring during in an interval of only a few weeks (compare to figure 4.1).

Table 4.1 summarizes some overall aggregated statistics for each party. It shows the overall number of posts, accounts, summed up likes and shares for posts and the within party share of refugee related posts over the observation period from 2014-01-01 to 2017-12-31. Although 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 and engagement in terms of likes and shares. The social democratic party *SPD* has over twice more posts than the *AfD* but roughly generates only half of the likes and only a fifth of the shares. Looking at the share of "refugee related" posts one can see that both the far left and far right parties talk more about this topic than the average.

4.2. Data processing and descriptive analysis

Separately for each party we count the daily number of "refugee related" posts and divide it by the overall daily number of posts from this party. Thus, for each day and party, we get a relative within-party frequency of refugee related posts. We use this statistic to measure the issue salience for each party. For example, a value close to 1 indicates that all discussions are related to refugee

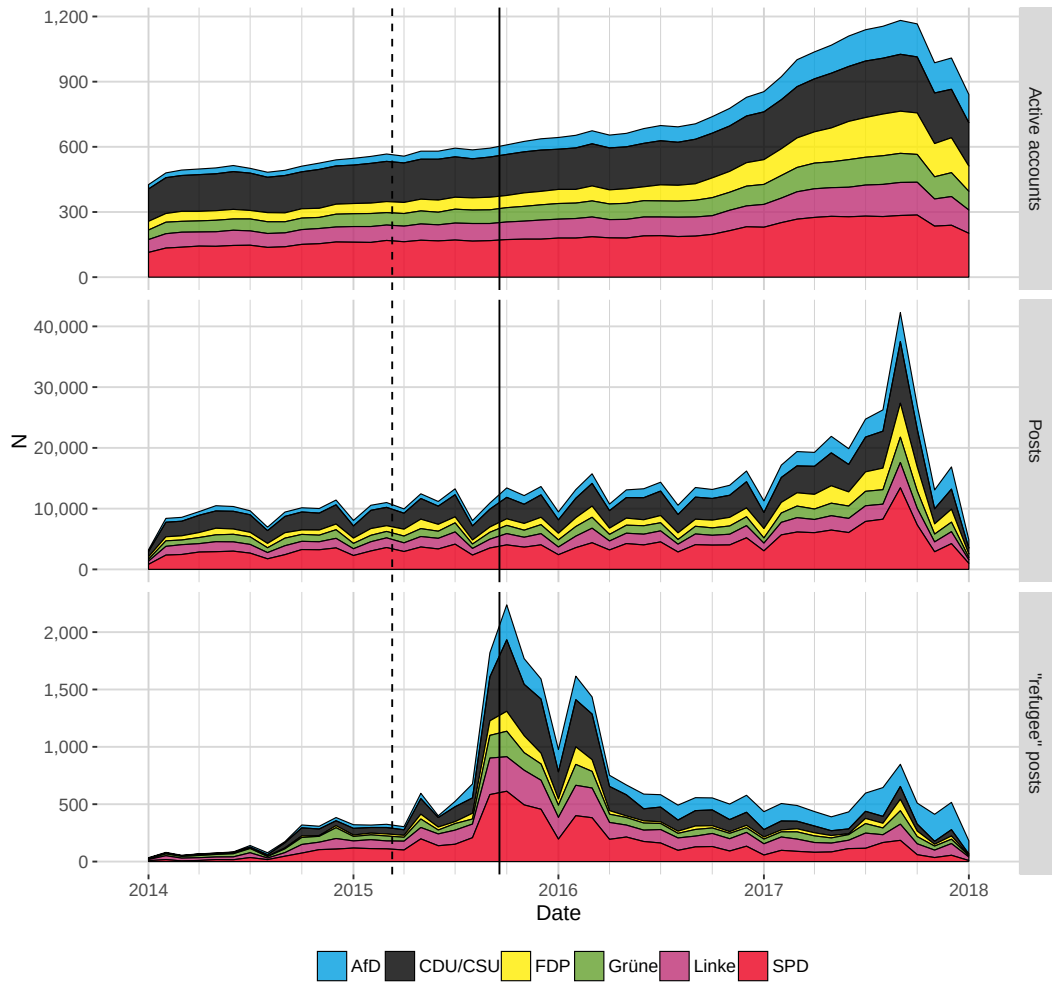


Figure 4.2: Number of active accounts and posts per party and month. The bottom panel shows the number of posts matching the regular expression `flucht|fluecht`. The vertical solid and dashed lines indicate the breakpoint detected by the non-parametric moments and the parametric copula based test.

and asylum issues while a value close to 0 indicates that this topic is of no importance in the political discourse.

First, the univariate models for each of the six salience measures are estimated. The factor copula dependency analysis is then performed on the residual information. The EDF is used to model the marginal distributions of the residuals.

Standard ARIMA-GARCH models are applied to the univariate time series. In this type of models the conditional mean $\mu_t = E(Y_t|\mathcal{F}_{t-1})$ of some possibly integrated variable is assumed to be a linear function of its own lags and lags

party	posts	accounts	likes (in Million)	shares (in Million)	share of posts related to "refugees" (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

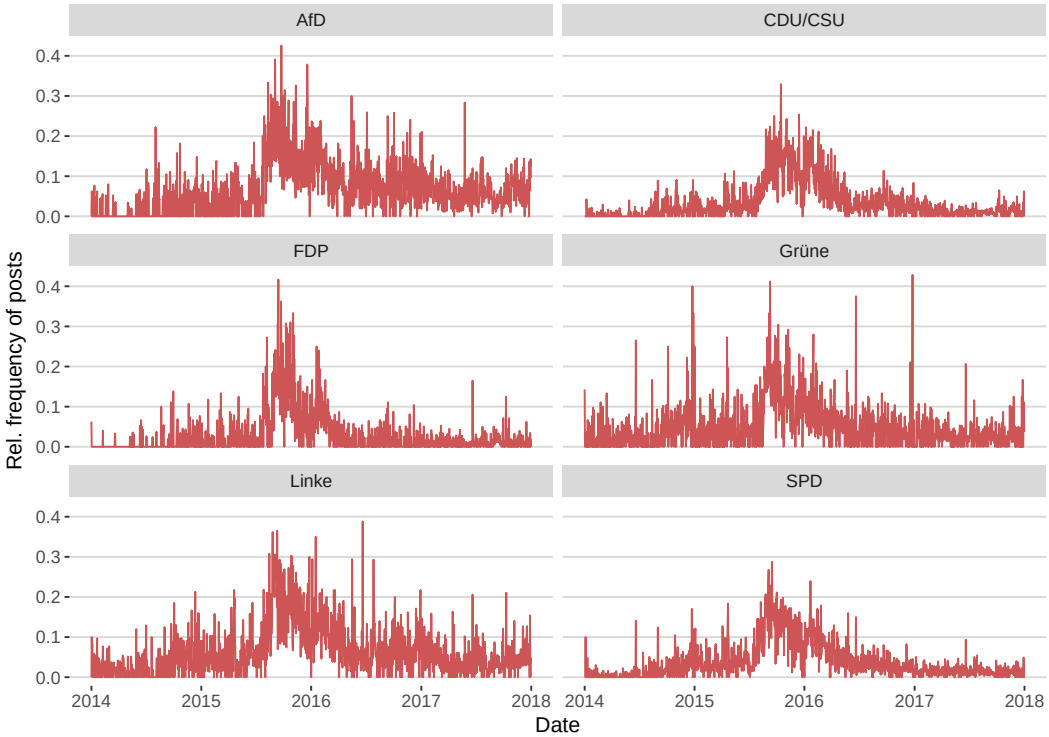
Table 4.1: Overall number of posts, active accounts, likes and shares over the observation period from "2014-01-01" - "2017-12-31". The last two columns show the fraction of posts matching the regular expression `flucht|fluecht`.

of the error term. To also account for volatility clustering, the conditional variance $\sigma_t^2 = V(Y_t|\mathcal{F}_{t-1})$ is assumed to be a linear function of its own lags and squared lags of the error term (Teräsvirta 2009, pp. 17-19).

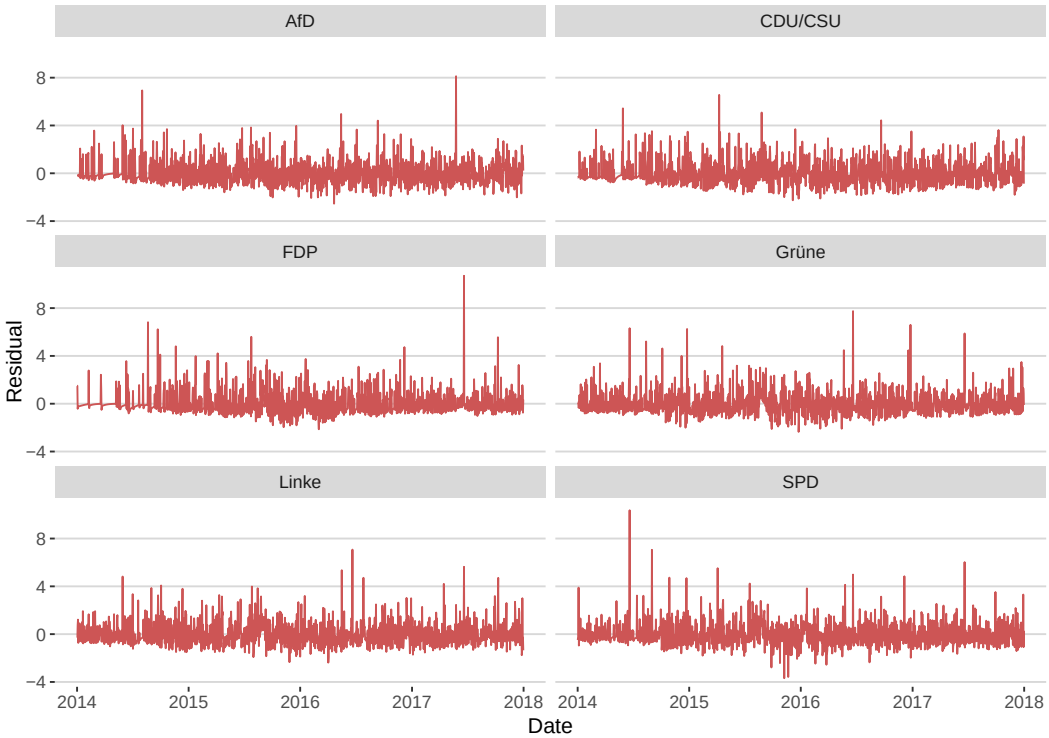
For the conditional variance we assume a GARCH(1, 1) process. For the model of the conditional mean we estimate different ARIMA models. The orders of the auto-regressive and moving average terms of the mean model are determined by testing various candidates over a small grid of order values. The final model is selected by comparing the Bayesian information criteria (BIC). Table 4.2 summarizes the final parameter values for the six time series model applied to the measures of topic salience.

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

Table 4.2: Parameters for the ARIMA-GARCH(1,1) model used to model the marginal distributions of the btw17 dataset.



(a) Relative daily within-party frequency of refugee related posts.



(b) Residuals of ARIMA-GARCH models applied to the time series.

Figure 4.3: Univariate time series and estimated residuals for the btw17 dataset. The upper panel shows the daily within-party frequency of refugee related topics. The residuals are estimated by applying ARIMA-GARCH models to the time series.

Figure 4.3 shows the time series of topic salience and the residuals for each of the six parties after applying the ARIMA-GARCH(1, 1) models. Inline with the bottom panel of figure 4.2, refugee related topics start to become important in mid 2015 with peaks of up to 40% importance in late 2015. Especially for the left and right wing parties, "Linke" and "AfD", the topic stays important even after the crisis.

Finally, in table 4.3 several pairwise sample dependencies for the residuals are presented. The same dependency statistics as for the SMM procedures as described in section 2.4.2 are used. Looking at the rank correlation one finds only very weak to weak positive dependencies of up to 0.20 for the two largest parties CDU/CSU and SPD. The quantile dependencies are all larger 0 but relatively small. Distinct patterns of asymmetric tail dependencies are not visible.

	Rank -	Quantile-dependence			
Pairs	correlation	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

Table 4.3: Pairwise sample dependencies for the daily topic salience of each party applied to the standardized residuals.

4.3. Results

To get a first idea of the characteristics of the data various factor copula models are fitted to the complete sample. To determine the number of latent factors K, we use an approach by Oh and Patton (2017, p. 148) and analyze the ordered eigenvalues of the residual rank-correlation matrix. One eigenvalue is above the threshold. Therefore, the analysis is restricted to one-factor copula models (see also figure A.2). We separately fitted equidependence and unrestricted models for various distributions of the factor and error term. Table 4.4 summarizes the estimation results for the factor copula models.

	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 4.4: Estimation results for different one-factor copula specifications applied to the residuals of the btw17 dataset.

For the equidependence models, the single factor loading is similar but relatively small for all three models. The degrees of freedom for both the t and skew-t model are large. This indicates, that no tail dependencies are present. The best fit in terms of the Q value of the objective function is given by the skew-t model. The skewness parameter is negative which could indicate stronger dependencies in the lower parts of the distribution.

The nonrestrictive models all have a larger Q value. The factor loadings are relatively stable between the models, ranging from 0.23 for the AfD party to around 0.65 for the two largest parties SPD and CDU/CSU. The average of the factor loadings for each unrestricted model is similar to the corresponding

estimate of the equidependence model. The degrees of freedom are still large but smaller compared to the equidependence model. The skew-t model is again characterized by negative asymmetries.

For the recursive estimation of the parameters, we proceed with the equidependence model which has the lowest Q value of all estimated models. To avoid numerical instabilities, we fix the distributional parameters df and λ at their full sample estimates. Therefore, we only test for a break in the single factor loading. Possible breaks in the tail dependency or symmetry are left out.

Four break tests are performed: A restrictive and non-restrictive non-parametric test based on the moment functions and two parametric tests based on the recursive estimates of a skewt-t and a norm-norm factor copula's loadings. Table 4.5 summarizes the results.

	Moments based test		Equidep. copula based test	
	unrestricted	restricted	skewt-t	norm-norm
test statistic	6.05	116.95	1.51	1.85
95%-CV	1.98 (0.00)	61.90 (0.00)	1.67 (0.069)	1.60 (0.028)
breakpoint	2015-09-18	2015-09-18	2015-03-13	2015-03-11

Table 4.5: Results for the parametric and non-parametric break point detection tests applied to the residuals of the btw17 dataset (p-value in parenthesis).

The two non-parametric test clearly detect a breakpoint in the dependence structure for 2015-09-18.[9] Around this date, the fraction of posts which include refugee related topics was the highest for the sample. The test statistics are around two to three times larger than the critical value. For the critical value the type I error was set to $\alpha = 0.05$. The test based on the skewt-t copula parameter cannot reject the null hypothesis. The p-value of around 0.07 is sightly larger than the chosen alpha value of 0.05. The highest value of the test statistic occurred at 2015-03-13 a few month before the date detected by the non-parametric tests. Conducting the break test based on a one factor equidependence norm-norm factor copula yields a p-value of 0.028 for a breakpoint around 2015-03-11.

[9]The dates of the breakpoints are also marked as vertical lines in figure 4.2.

As already seen in the simulation study in section 3.3.2, the non-parametric test seems to be less sensitive. The different results can possibly also be explained by the fact, that the copula based tests only tested for a break in the single factor loading. To better assess the results, we further estimated the dependence structure before and after the break using a skew-t equidependence factor copula (see table 4.6). We choose the breakpoint detected by the moment based tests since this has the lowest p-value. For convenience, the full sample estimates as stated in table 4.4 are replicated here. For all parameters, standard errors were estimated with the bootstrap procedure.

We excluded the estimation of the degrees of freedom and used the full sample estimates. Estimating this distributional parameter often resulted in very large standard errors and unstable numerical results. In this setting, the parameter estimates also changed if the *inverse* degrees of freedom over the range of $[0, 0.5)$ instead of the regular degrees of freedom over the range $(2, \infty]$ were estimated.

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 4.6: Estimation results for the btw17 dataset. A bloc-equidependence factor copula model is estimated before and after the breakpoint at $t = 626$ (2015-09-18). Standard errors in parenthesis (estimated with $B = 2000$ bootstrap samples). Degrees of freedom were fixed at the full model estimate of $df = 96$.

For all three models, the skewness parameters λ are negative. However, comparing the size of the estimates to their standard errors shows that the skewness parameters are not significantly different from 0. The estimates for the factor loading is significant and does not fluctuate between the pre and post-break period. Despite the "breakpoint" detected by the moment based tests, the pre and post-break copula models seem to be very similar.

Given the large degrees of freedom and the insignificant skewness parameter, it seems that the dependence structure can be modeled as a one factor

equidependence copula for which the latent variable is normal distributed. This special case of a one-factor copula model can also be described with a Gaussian copula. Estimating the pre- and post-break Gaussian one-factor model for the earlier breakpoint detected by the parametric break test at 2015-03-11 yields an increase in the factor loading from $\hat{\beta}_{pre} = 0.35$ to $\hat{\beta}_{post} = 0.54$. This indicates an increase in positive dependencies for the time of the refugee crisis and afterwards.

5. Discussion and summary

In this chapter, the presented methods are shortly reviewed, the results are critically assessed and finally we give some outlook on future improvements and research.

The results of the empirical chapters 3.3 and 4 are all based on a general statistical model for analyzing the dependence structure of multivariate time series. This models allow for a multistage estimation process: First the time varying conditional means and variances of the marginal distributions are estimated using some parametric model for univariate time series. Second, the standardized residuals are obtained and their multivariate cdf is modeled using the non-parametric EDF for the marginal distributions and a parametric copula model for the dependence structure.

In this paper, the focus was put on the special class of factor copula functions. This type of copula function is based on a linear factor structure for which the dependencies are described by a lower dimensional set of latent variable. The parameters of a factor copula model involve the matrix of factor loadings and the distributional parameters of the latent factors and the error term. The choice of the distributions and the number of latent factors are hyper-parameters of the model which have to be determined prior to the estimation.

Since the parameter vector of a factor copula model can grow very large, two model restrictions were discussed: The equidependence model which assumes equal pairwise dependencies for all variables and the bloc-equidependence model. For the latter, it is assumed that the variables can be naturally partitioned into groups and that the pairwise intra-group and inter-group dependencies are equal.

The latent factor model results in complex dependence structures and in most cases a closed form mapping from the model parameters to measures of dependency doesn't exist. In addition, obtaining the likelihood function of a factor copula model is difficult and computationally expensive. However, it is computationally cheap to simulate random values from the factor model. This leads to the Simulated Method of Moments estimation procedure:This method is rooted in GMM estimation but the population moment function is replaced by a sample equivalent calculated with simulations from the model. The classical GMM and SMM are based on functions involving moments of the variables. But the SMM for factor copulas is based on scale-free measures of dependencies as functions of the variable's EDFs. Nevertheless, asymptotic results can still be derived and it can be shown that the parameter estimates are asymptotically normal distributed.

For many applications it is of interest whether the dependence structure is time invariant. Therefore, a structural break test for a change in the parameters of the factor copula model was presented. It requires the recursive estimation of the factor copula parameters and compares them to the full sample estimates. If the distance is significantly large, a structural break is detected. The parametric test allows for the detection of breaks in subsets of the parameter vector but it is computationally costly due to its recursive model estimates. As an alternative, the test can also be formulated non-parametrically involving only finite sample versions of dependency measures applied to the residual data.

To enable other researcher to assess and enhance the discussed methods a software library for the programming language R is provided. To confirm the validity of the methods and their implementations, two simulation studies were performed. The first study demonstrated the convergence properties of the copula parameters for an equidependence model. Increasing the sample size or the number of observations reduced the variance of the estimator. The second study simulated a structural break in the copula for a bloc-equidependence model with 21 variables in 3 different groups. Both the non-parametric and the parametric test are able to detect the breakpoint with high accuracy.

The simulation studies in this paper show that the factor copula model is theoretically suitable for modelling high dimensional dependence structures. However, in contrast to classical GMM estimation, the objective function of the copula's SMM estimator is not continuous and differentiable. Therefore, one can only rely on derivative free optimization algorithms. This reduces

the performance and stability of the estimation process. The computational demand is especially high for the parametric break test which requires the recursive estimation of a copula model for almost all points in time.

To test the methods on real world data, a case study with textual data from the social media site Facebook is performed. The data consists of public posts of German politicians and political parties from 2014 to 2017. For each major German political party we derive dynamic measures of topic salience for refugee and asylum related issues and ask whether the refugee crisis in late 2015 influenced the dependence structure between the measures of topic salience.

Both the non-parametric and the parametric break test based on a norm-norm one factor equidependence copula detected a breakpoint in the parameters of the factor copula for March and September 2015 respectively. November 2015 marks the peak of the refugee crisis with around 200000 refugees entering Germany. For each break date a factor copula is estimated before and after the break. For the later breakpoint, detected by the non-parametric test, no change in the copula parameters can be found. This finding somehow contradicts the result of the test. However for the early parametric breakpoint, an increase in the single factor loading by around 48% can be observed.

This finding indicates that in 2014 and early 2015, just before the peak of the refugee crisis, the dependencies were weaker compared to the time of the crisis and afterwards. Before the crisis the within-party importance of refugee and asylum related issues was more independent from the political discourse of the other parties. However, the external shock of the crisis led to a stronger convergence of the issue salience.

The real-data application has several drawbacks: Interpreting the actual size of the factor loadings is difficult. That hinders the assessment of actual real changes due to the breakpoint and allows only for relatively fuzzy conclusions. In addition, we forwent the implementation and conduction of a so called J test for over-identifying restrictions for the equidependence model (Oh and Patton 2013, p. 693).

The undesirable properties of the objective function require the usage of global and local derivative free optimization algorithms. During the empirical analysis we regularly encountered unstable numerical results. For the break test, this can affect the recursive test statistics and extreme outliers above the critical value are possible. Numerical instabilities arise especially if some distributional

parameters such as the skewness or the degrees of freedom are included in the estimation process. It seems that it requires much more data than typically available to estimate this parameters consistently.

Therefore, much care has to be taken in the optimization process and the tuning parameters of the optimization algorithms. Future research should concentrate on improving the stability of the SMM method and the performance of estimating many recursive models for the structural break test. For example, a promising approach is undertaken by Frazier and Zhu (2017). Their method tries to overcome the problem of discontinuity by proposing an adapted simulation algorithm which allows the application of automatic differentiation techniques and hence allows the use of efficient derivative-based optimization procedures.

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 simulation studies and 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`.

Almost all estimation procedures (for instance the bootstrap algorithms and the recursive model estimation) are embarrassingly parallel. Therefore, the HPC cluster of the University of Cologne was utilized to massively speed up the estimation process (Achter, Borowski, Nieroda, et al. 2018).

To simplify the workflow and the communication with the cluster, we wrote the R-package `cheopsr`. It allows the execution of parallel jobs and the collection of results 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 access rights to the HPC cluster are obligatory.

The methods for the simulation and estimation of factor copula models and the break test are available via a separate R-package `factorcopula`. The package can be easily downloaded and installed from Github. Further notes can be found under: https://github.com/bonartm/factorcopula.

A.2. Code examples

In the following explained example, the *R* package *factorcopula* is used to define a bloc-equidependence factor copula with three latent variables. Data from the model is simulated for some specific vector $\boldsymbol{\theta}$. Finally, the parameter vector is estimated with the SMM.

The first factor is skewed-t distributed, the second and third are standard normal distributed. The error term is t distributed with 4 degrees of freedom.

The distributional parameters `dfInv` and `lambda` of the skewed-t distribution are free parameters of the model.

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

A model with $K = 3$, $N = 6$, $k_1 = k_2 = 3$ and $M = 2$ is constructed. Together with the specification of the latent variables this results in a bloc-equidependence model with a skewed-t distribution for the common factor and a standard normal distribution for each of the two group specific factors.

```
k <- c(1, 1, 1, 2, 2, 2)
beta <- config_beta(k, 3)
beta
```

```
##      first
## [1,] "beta1" "beta3" "0"
## [2,] "beta1" "beta3" "0"
## [3,] "beta1" "beta3" "0"
## [4,] "beta2" "0"     "beta4"
## [5,] "beta2" "0"     "beta4"
## [6,] "beta2" "0"     "beta4"
```

Next, $S = 1000$ random values from the copula model are simulated and the marginal distributions are transformed to standard normal ones. The factor loadings for the common factor are $\beta_1 = \beta_2 = 1$ and the group specific loadings $\beta_3 = \beta_4 = 1.5$. The distributional parameters of the common latent factor are set to $dfInv = 0.25$ and $\lambda = -0.8$. This produces asymmetric tail dependencies.

```
copfun <- fc_create(Z, eps, beta)
theta <- c(beta1 = 1, beta2 = 1, beta3 = 1.5, beta4 = 1.5,
           dfInv = 0.25, lambda = -0.5)
Y <- qnorm(copfun(theta, S = 1000))
```

Finally, we define some lower and upper bounds for the numerical estimates and approximate the parameters of the model via SMM. Standard errors are estimated with 1000 bootstrap samples. To get more precise parameter

estimates, the stopping criteria for the first or second stage optimization algorithms can be increased.

```
lower <- c(beta1 = 0, beta2 = 0, beta3 = 0, beta4 = 0,
           dfInv = 0.01, lambda = -0.9)
upper <- c(beta1 = 5, beta2 = 5, beta3 = 5, beta4 = 5,
           dfInv = 0.49, lambda = 0.9)
model <- fc_fit(Y, Z, eps, beta, lower, upper, k = k,
                S = 25000, se = TRUE, B = 1000)
round(model$theta.second.stage, 2)
# beta1  beta2  beta3  beta4  dfInv lambda
#  1.26   0.96   1.40   1.64   0.30  -0.55
round(model$se, 2)
# beta1  beta2  beta3  beta4  dfInv lambda
#  0.52  0.44   0.44   0.44   0.23   0.86
```

A.3. Additional figures

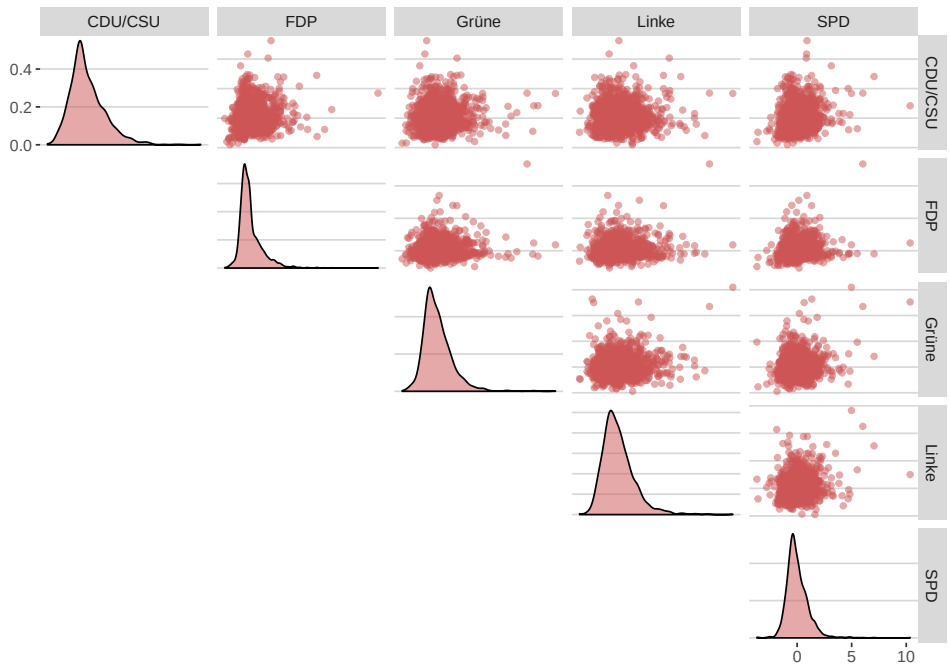


Figure A.1: Pairwise scatter-plot of the estimated residuals for the btw17 dataset

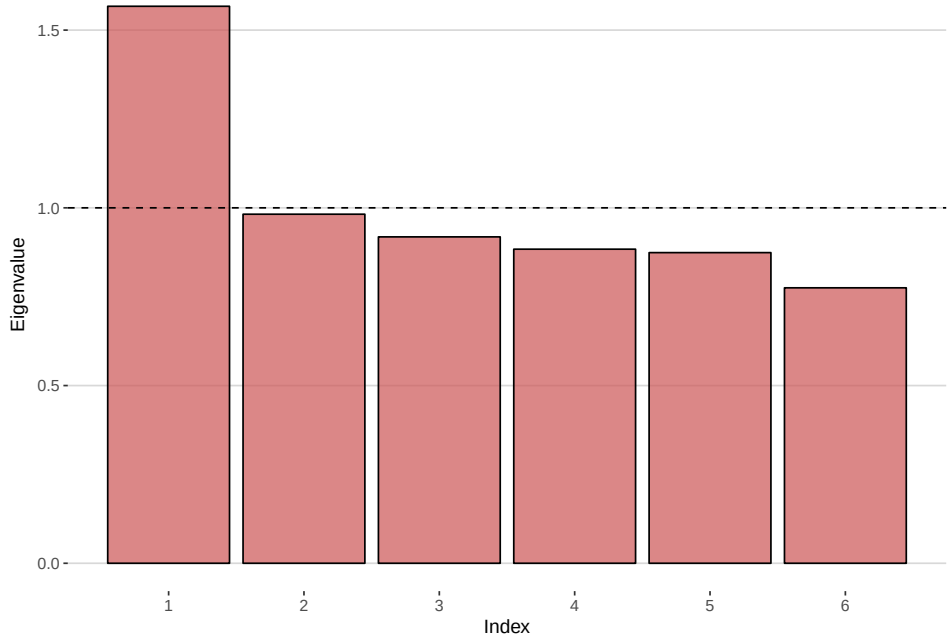


Figure A.2: Scree-plot of ranked eigenvalues based on the pairwise rank-correlation matrix

B. References

Achter, Viktor, Stefan Borowski, Lech Nieroda, et al. (2018). *CHEOPS Cologne High Efficient Operating Platform for Science*. https://rrzk.uni-koeln.de/sites/rrzk/HPC_Projekte/CHEOPS_Brief_Instructions.pdf. [Online techreport; accessed 12-May-2018].

Allen, Mark and Norbert Bartels (2018). *RestFB - a pure Java Facebook Graph API client*. http://restfb.com. [Online documentation; accessed 12-May-2018].

bpb, Bundeszentrale für politische Bildung (2018). *Zahlen zu Asyl in Deutschland*. https://www.bpb.de/gesellschaft/migration/flucht/218788/zahlen-zu-asyl-in-deutschland. [Online source; accessed 09-June-2018].

Chen, Xiaohong and Yanqin 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.

Embrechts, Paul (2009). "Copulas: A Personal View". In: *Journal of Risk and Insurance* 76.3, pp. 639–650.

Embrechts, Paul, Alexander Mcneil, and Daniel Straumann (2002). "Correlation and dependence in risk management: Properties and pitfalls". In: *RISK Management: Value at Risk and Beyond*. Cambridge University Press, pp. 176–223.

Fermanian, Jean-Davi (2017). "Recent Developments in Copula Models". In: *Econometrics* 5.3, pp. 1–3.

Frazier, David T. and Dan Zhu (2017). *Derivative-Based Optimization with a Non-Smooth Simulated Criterion*. https://arxiv.org/abs/1708.02365. [Online working paper; accessed 12-May-2018].

Gouriéroux, Christian and Alain Monfort (1996). *Simulation-based Econometric Methods*. CORE lectures. Oxford University Press.

Hall, Alastair R. (2005). *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press.

Hansen, Bruce E. (1994). "Autoregressive Conditional Density Estimation". In: *International Economic Review* 35.3, pp. 705–730.

Joe, Harry (2015). *Dependence Modeling with Copulas*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability 134. Taylor & Francis.

Johnson, Steven G. (2018). *The NLopt nonlinear-optimization package*. https://nlopt.readthedocs.io/en/latest/. [Online documentation; accessed 12-May-2018].

Kucherenko, Sergei and Yury Sytsko (2005). "Application of Deterministic Low-Discrepancy Sequences in Global Optimization". In: *Computational Optimization and Applications* 30.3, pp. 297–318.

Manner, Hans and Olga Reznikova (2012). "A Survey on Time-Varying Copulas: Specification, Simulations, and Application". In: *Econometric Reviews* 31.6, pp. 654–687.

Manner, Hans, Florian Stark, and Dominik Wied (2017). *Testing for Structural Breaks in Factor Copula Models*. https://www.wiwi.uni-due.de/fileadmin/fileupload/VWL-WIPO/WiWi-Kolloquium/Manner_Stark_Wied__2017_.pdf. [Online working paper; accessed 12-May-2018].

Mikosch, Thomas (2006). "Copulas: Tales and facts". In: *Extremes* 9.1, pp. 3–20.

Nelsen, Roger B. (1999). *An Introduction to Copulas*. Lecture notes in statistics 139. Berlin, Heidelberg: Springer.

Oh, Dong Hwan and Andrew 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, Andrew J. (2009). "Copula–Based Models for Financial Time Series". In: *Handbook of Financial Time Series*. Ed. by Thomas Mikosch, Jens-Peter Kreiß, Richard A. Davis, et al. Berlin, Heidelberg: Springer, pp. 767–785.

– (2012). "A Review of Copula Models for Economic Time Series". In: *J. Multivar. Anal.* 110, pp. 4–18.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/. [Online documentation; accessed 12-May-2018]. Vienna, Austria: R Foundation for Statistical Computing.

Rowan, Thomas Harvey (1990). "Functional Stability Analysis of Numerical Algorithms". PhD thesis. Austin, TX, USA.

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

Sklar, Abe (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, Sebastian, Arnim Bleier, Malte Bonart, et al. (2018). "Systematically Monitoring Social Media: the case of the German federal election 2017". In: *GESIS Papers* 2018/04, p. 25.

Teräsvirta, Timo (2009). "An Introduction to Univariate GARCH Models". In: *Handbook of Financial Time Series*. Ed. by Thomas Mikosch, Jens-Peter Kreiß, Richard A. Davis, et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 17–42.

W., Russell Neuman, Guggenheim Lauren, Mo Jang S., et al. (2014). "The Dynamics of Public Attention: Agenda Setting Theory Meets Big Data". In: *Journal of Communication* 64.2, pp. 193–214.

Wlezien, Christopher (2005). "On the salience of political issues: The problem with 'most important problem'". In: *Electoral Studies* 24.4, pp. 555–579.

Ypma, Jelmer (2014). *Introduction to nloptr: an R interface to NLopt*. https://cran.r-project.org/web/packages/nloptr/vignettes/nloptr.pdf. [Online documentation; accessed 12-May-2018].

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 10. Juni 2018

(Malte Bonart)