High Dimensional Financial Engineering: Dependence Modeling and Sequential Surveillance

DISSERTATION

zur Erlangung des akademischen Grades

doctor rerum politicarum
(Doktor der Wirtschaftswissenschaft)
im Fach Statistik

eingereicht an der
Wirtschaftswissenschaftlichen Fakultät
Humboldt-Universität zu Berlin

von
Herrn M.Sc. M.Sc. Yafei Xu

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Dr. Sabine Kunst

Dekan der Wirtschaftswissenschaftlichen Fakultät:
Prof. Dr. Christian D. Schade

Gutachter:
1. Prof. Dr. Ostap Okhrin
2. Prof. Dr. Bernd Droge

eingereicht am: 23. Oktober 2017
Tag der mündlichen Prüfung:

Abstract

This dissertation focuses on the high dimensional financial engineering, especially in dependence modeling and sequential surveillance.

In aspect of dependence modeling, an introduction of high dimensional copula concentrating on state-of-the-art research in copula is presented. Factor copula, hierarchical Archimedean copula and vine copula are explicated, including their statistical inference. An empirical study in risk management by employing the introduced copulas is given.

A more complex application in financial engineering using high dimensional copula is concentrated on the pricing of the portfolio-like credit derivative, i.e. credit default swap index (CDX) tranches. In this part, the convex combination of copulas is proposed in CDX tranche pricing with components stemming from elliptical copula family (Gaussian and Student-t), Archimedean copula family (Frank, Gumbel, Clayton and Joe) and hierarchical Archimedean copula family used in some publications. By comparison of two diverse credit derivatives, one can find that the convex combination of copulas has cutting edge in pricing CDX and CDO with components from Archimedean copula family with asymmetric tail dependence structures, e.g. Clayton, Gumbel and Joe copulas.

In financial surveillance part, the chapter focuses on the monitoring of high dimensional portfolios (in 5, 29 and 90 dimensions) by development of a nonparametric multivariate statistical process control chart, i.e. energy test based control chart (ETCC). The main features in ETCC are in three aspects. Firstly, the ETCC is nonparametric control chart which means it requires no pre-knowledge on the processes compared to many traditional parametric control chart, e.g. CUSUM and EWMA. Secondly, ETCC monitors multivariate processes, which are more often in the real life. Since multivariate processes bring more characteristics than the unique process, hence ETCC has strong potential in many application areas. Thirdly and the most important virtue of the ETCC is that it monitors the mean and covariance jointly, not separately, compared to many other control charts. Since its powerful detection capacity in covariance change, hence it has good performance for financial sequences during crisis period, where volatility is the main trigger of shift.

In order to support the further research and practice of nonparametric multivariate statistical process control chart devised in this dissertation, an R package "`EnergyOnlineCPM`" is developed. At moment, this package has been accepted and published in the Comprehensive R Archive Network (CRAN), which is the first package that can online monitor the shift in mean and covariance jointly.

Keywords: Credit Default Swap Index Tranche; Copula, Convex Combination of Copulas; Nonparametric Multivariate Statistical Process Control; R Package

Acknowledgements

First and foremost, I would like to express my deepest appreciation to the first supervisor Prof. Dr. Ostap Okhrin, who introduced me to the world of dependence modeling and financial surveillance. Under his assiduous, patient and comprehensive supervision, we harvested three papers and an R CRAN-package during the doctoral study. Hence I am thankful of his continuous support and help to my doctoral study in paper publication, fund and placement applications. His instruction, guidance and detailed paper correction helped me throughout the whole process of my writing of the dissertation. Without his encouragement and consistent advisory this thesis would have not been possible. I see him as one of my best friends in my life forever.

I would also like to thank Prof. Dr. Bernd Droge for his important comments and criticism. He is an extremely distinguished professor at Humboldt University of Berlin, who has industriously invested his time in teaching and supervising his students in fields of statistics and econometrics. Many of his courses, e.g. econometric methods, microeconometrics, panel data analysis and time series analysis, are always highly popular among students. Almost all of my econometrics training is obtained in his classroom.

Further, I would like to thank my family members, my father Mr. Gui Xi Xu and mother Mrs. Wen Hua Zhang, whose weekly phone call from Lianyungang to Berlin always brings me spiritual support and always makes me feel the deep love from the family. Without their selfless engagement and constant support, the accomplishment of the doctoral study would have been much more difficult. Dear father and mother, I love you!

Many other mentors during my doctoral study include Prof. Dr. Fabrizio Durante, Prof. Dr. Wolfgang Schmid, Prof. Dr. Wolfgang Karl Härdle, Dr. Iryna Okhrin and Dr. F. Marta L. Di Lascio. I appreciate their kind help and wish them all the best. Meanwhile, I am highly thankful of the Chinese Scholarship Council (CSC) which awarded me the Chinese Government Scholarship sponsoring my doctoral study in Germany.

Yafei Xu,

14. Oct. 2017 in Berlin.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The dissertation includes four projects which I worked on during my PhD study.

Chapter 2 reviews the latest proceeding of research in high dimensional copulas. At the beginning the bivariate copulas are given as a fundamental followed with the multivariate copulas which are the concentration of the paper. In multivariate copula sections, the hierarchical Archimedean copula, the factor copula and vine copula are introduced. In the following section the estimation methods for multivariate copulas including parametric and nonparametric routines, are presented. Also the introduction of the goodness of fit tests in copula context is given. An empirical study of multivariate copulas in risk management is performed thereafter.

Chapter 3 focuses on the pricing of portfolio-like credit derivative, credit default swap index (CDX) tranches. Among the pricing models for CDX tranches, the model equipped with the Gaussian copula has become the market benchmark for a long time. Albeit thereafter some other copulae were employed to improve the Gaussian model, yet a lot of them have suffered from shortcomings, especially in destitution of heterogeneous sectoral dependence, asymmetric dependence and fat tail dependence. For increasing the pricing accuracy and also keeping the model parsimonious, we propose in this paper an approach of convex combination of copulae (cc-copula) in pricing CDX tranches. Copulae from elliptical and Archimedean families were chosen as the components to construct the cc-copula models. In order to support the effectiveness of the cc-copula models, two distinct empirical studies were conducted to reproduce the spreads of the CDX tranches of two different contracts covering crisis and non-crisis periods. The results evince that the cc-copula based pricing models have dominant performance compared with the benchmark models.

Chapter 4 presents a nonparametric control chart based on the change point model, for multivariate statistical process control (MSPC). The main constituent of the chart is the energy test that focuses on the discrepancy between empirical characteristic functions of two random vectors. This new multivariate control chart highlights in three aspects. Firstly, it is nonparametric, requiring no pre-knowledge of the data generating processes. Secondly, this control chart monitors the whole distribution, and not only specific characteristics like mean or covariance. Thirdly, it is designed for online detection (Phase II), which is central for real time surveillance of stream data. Simulation study discusses in-control and out-of-control measures in context of mean shift and covariance shift. In the real application, three financial data sets (in 5, 29, 90 dimensions) were employed to analyze the performance of the control chart for financial surveillance. The results from both simulation and empirical studies, compared with benchmarks, strongly advocate the proposed control chart.

Chapter 5 introduces an R package ”`EnergyOnlineCPM`” for the energy-test-based control chart (ETCC) presented in Chapter 4. This package integrates energy test into change point model to realize the sequential surveillance of many processes simultaneously. In CRAN this is the first R package for nonparametric multivariate statistical process control, which monitors the distributional changes. The main virtue of this package is that it monitors mean shift or covariance shift not separately, but jointly. In the first section, a review of the main R packages for change point analysis (Phase I) and change point model (Phase II) is conducted. Thereafter, the installation and an example of mean shift detection using `EnergyOnlineCPM` are presented.

Chapter 2

Copula in High Dimensions

This chapter is based on the paper ”Copula in High Dimensions: An Introduction” by O. Okhrin, A. Ristig and Y.F. Xu (2017) published in *Applied Quantitative Finance.*

2.1 Introduction

Researches of dependence modeling were burgeoning during the last decade. The traditional approaches that concentrate on the elliptical distributions such as Gaussian models are giving way to copula-based models. Albeit these Gaussian models sometimes own the convenience in model construction and computation, yet an abundant amount of empirical evidences do not support the underlying assumptions. De facto, shortcomings in the elliptical and especially Gaussian family are mainly in lack of asymmetrical and tail dependence which have been deeply discussed in numerous papers. Furthermore and of great importance, margins of elliptical distributions belong to the same elliptical family.

The seminal result of Sklar (1959) provides a partial solution to these problems. It allows to separate the marginal distributions from the dependency structure between the random variables. Since the theory on modeling and estimation of univariate distributions is well established compared to the multivariate case, the initial problem reduces to modeling the dependency by copulas. In particular, this approach dramatically widens the class of candidate distributions and allows a simple construction of distributions with less parameters than imposed by elliptical models.

In the beginning of the copula hype, researches were mainly focused on the bivariate dependence but as time passes problems raised by the financial, technological, biological industries dictated the rules of further developments, namely moves to higher dimensions. Nonetheless, it has been realized as clearly stated in Mai and Scherer (2013), that "the step from one-dimensional modeling is clearly large. But, unfortunately, the step from two to three (or even more) dimensions is not a bit smaller.".

Numerous steps are done in order to contribute to research on high-dimensional modeling approaches and these main branches have been established: pair copula construction, see Joe (1996a), Bedford and Cooke (2001), Bedford and Cooke (2002) and Kurowicka and Cooke (2006), hierarchical Archimedean copula, see Savu and Trede (2010a), Hofert (2011) and Okhrin et al. (2013a), and factor copula, see Krupskii and Joe (2013) and Oh and Patton (2015).

This chapter attempts at discussing such non-standard multivariate copula models and the subsequent sections are organized as follows. We introduce bivariate copulae and review modern multivariate copula families. Then, corresponding estimation methods and goodness of fit tests are presented. Last but not least, we study a risk management topic empirically.

2.2 Bivariate Copula

Modeling the dependence between only two random variables using copulae is the subject of this section. There are several equivalent definitions of the copula function. We define it as a bivariate distribution function and the simplest one is as follows:

Definition 2.1. The copula $C(u, v)$ is a bivariate distribution with margins being $U[0, 1]$.

Term copula was mentioned for the first time in the seminal result of Sklar (1959). The separation of the bivariate distribution function into the copula function and margins is formally stated in the subsequent theorem. One possible proof is presented in Nelsen (2006, Section 2.3), for others we refer to Durante et al. (2012), Durante et al. (2013) and Durante and Sempi (2005, Chapter 2)

Theorem 2.2. *Let F be a bivariate distribution function with margins F_1 and F_2, then there exists a copula C such that*

$$F(x_1, x_2) = C\{F_1(x_1), F_2(x_2)\}, \quad x_1, x_2 \in \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}. \tag{2.1}$$

If F_1 and F_2 are continuous than C is unique. Otherwise C is uniquely determined on $F_1(\overline{\mathbb{R}}) \times F_2(\overline{\mathbb{R}})$.

Conversely, if C is a copula and F_1 and F_2 are univariate distribution functions, then function F in (2.1) is a bivariate distribution function with margins F_1 and F_2.

As indicated above, the theorem allows decomposing any continuous bivariate distribution into its marginal distributions and the dependency structure. Since by definition, the latter is the copula function with uniform margins, it follows that the copula density can be determined in the usual way

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2}, \quad u_1, u_2 \in [0, 1]. \tag{2.2}$$

Being armed with the Theorem 2.2 and (2.2), the density function $f(\cdot)$ of the bivariate distribution F can be rewritten in terms of copula

$$f(x_1, x_2) = c\{F_1(x_1), F_2(x_2)\} f_1(x_1) f_2(x_2), \quad x_1, x_2 \in \overline{\mathbb{R}}.$$

A very important property of copulae is given in Nelsen (2006, Theorem 2.4.3) stating that copulae are invariant under strictly monotone transformations of margins. Seen from this angle, copulae capture only those features of the dependency which are invariant under increasing transformations.

2.2.1 Copula Families

Naturally, there is an infinite number of different copula functions satisfying the properties of Definition 2.1 and the number of them being deeply studied is expanding. In this section, we discuss three copula classes namely simple, elliptical and Archimedean copulae.

Simplest Copulae

To form basic intuition for copula functions, we first study some extreme special cases, like stochastically independent, perfect positive or negative dependent random variables. According to Theorem 2.2, the copula of two stochastically independent random variables X_1 and X_2 is given by the product (independence) copula defined as

$$\Pi(u_1, u_2) = u_1 u_2, \quad u_1, u_2 \in [0, 1].$$

The contour diagrams of the bivariate density function with product copula and either Gaussian or t_3-distributed margins are given in Figure 2.1. Two additional extremes are the lower and upper Fréchet-Hoeffding bounds. They represent the perfect negative and positive dependence of two random variables respectively

$$W(u_1,u_2) = \max(0, u_1+u_2-1) \quad \text{and} \quad M(u_1,u_2) = \min(u_1,u_2), \quad u_1,u_2 \in [0,1].$$

If $C = W$ and $(X_1, X_2) \sim C(F_1, F_2)$ then X_2 is a decreasing function of X_1. Similarly, if $C = M$, then X_2 is an increasing function of X_1. In general, we can argue that an arbitrary copula which represents some dependency structure lies between these two bounds, i.e.

$$W(u_1,u_2) \leq C(u_1,u_2) \leq M(u_1,u_2), \quad u_1,u_2 \in [0,1].$$

The bounds serve as benchmarks for the evaluation of the dependency magnitude. There are numerous techniques for building new copulae by mixing at least two of the presented simplest copula. For example, copula families B11 and B12, see Joe (1997), arise as a combination of the upper Fréchet-Hoeffding bound and the product copula

$$\begin{aligned} C_{B11}(u_1,u_2,\theta) &= \theta M(u_1,u_2) + (1-\theta)\Pi(u_1,u_2) = \theta \min\{u_1,u_2\} + (1-\theta)u_1u_2, \\ C_{B12}(u_1,u_2,\theta) &= M(u_1,u_2)^{\theta}\Pi(u_1,u_2)^{1-\theta} = (\min\{u_1,u_2\})^{\theta}(u_1u_2)^{1-\theta}, \quad u_1,u_2,\theta \in [0,1]. \end{aligned}$$

Family B11 builds on the fact that every convex combination of copulas is a copula as well. Family B12 is also known as Spearman or Cuadras-Augé copula, which is a weighted geometric mean of the upper Fréchet-Hoeffding bound and the product copula. Further generalization is done by using power mean over the upper Fréchet-Hoeffding bound and the product copula

$$\begin{aligned} C_p(u_1,u_2,\theta_1,\theta_2) &= \{\theta_1 M^{\theta_2}(u_1,u_2) + (1-\theta_1)\Pi^{\theta_2}(u_1,u_2)\}^{1/\theta_2} \\ &= \{\theta_1 \min(u_1,u_2)^{\theta_2} + (1-\theta_1)(u_1u_2)^{\theta_2}\}^{1/\theta_2}, \end{aligned}$$

with $\theta_1 \in [0,1]$, $\theta_2 \in \mathbb{R}$. Last but not least, a convex combination of the Fréchet-Hoeffding lower bound, upper bound and product copula forms the Fréchet copula

$$C_F(u_1,u_2,\theta_1,\theta_2) = \theta_1 W(u_1,u_2) + (1-\theta_1-\theta_2)\Pi(u_1,u_2) + \theta_2 M(u_1,u_2),$$

subject to $0 \leq \theta_1 + \theta_2 \leq 1$. Note that any bivariate copula can be approximated by the Fréchet family and a bound of the resulting approximation error can be estimated. Nelsen (2006, Chapter 3) provides further methods for constructing multivariate copulas and discusses convex combination in more detail.

Elliptical Family

Due to the popularity of the Gaussian and t-distribution in several applications, elliptical copulae play an important role as well. The construction of this type of copulae is directly based on Sklar's Theorem showing how new bivariate distributions can be constructed. The copula-based modeling approach substantially widens the family of elliptical distributions by keeping the same elliptical copula function and varying the marginal distributions or vice versa.

To determine the copula function of a given bivariate distribution, we employ the transformation

$$C(u_1,u_2) = F\{F_1^{-1}(u_1), F_2^{-1}(u_2)\}, \quad u_1,u_2 \in [0,1], \tag{2.3}$$

where F_i^{-1}, $i=1,2$, are (generalized) inverses of the marginal distribution functions. Based on (2.3), arbitrary elliptical distributions can be derived. The problem, however, is that such copulae depend on the inverse distribution functions of the marginals which are rarely available in an explicit form.

For instance, from Formula 2.3 follows that the Gaussian copula and its density are given by

$$\begin{aligned} C_N(u_1,u_2,\delta) &= \Phi_\delta(\Phi^{-1}(u_1), \Phi^{-1}(u_2)), \\ c_N(u_1,u_2,\delta) &= (1-\delta^2)^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}(1-\delta^2)^{-1}(u_1^2+u_2^2-2\delta u_1 u_2)\Big\} \\ &\quad \times \exp\Big\{\frac{1}{2}(u_1^2+u_2^2)\Big\}, \quad \text{for all } u_1,u_2 \in [0,1],\ \delta \in [-1,1], \end{aligned}$$

where Φ is the distribution function of $\mathrm{N}(0,1)$, Φ^{-1} is the functional inverse of Φ and Φ_δ denotes the bivariate standard normal distribution function with correlation coefficient δ. In the bivariate case, the t-copula and its density are given by

$$\begin{aligned} C_t(u_1,u_2,\nu,\delta) &= \int_{-\infty}^{t_\nu^{-1}(u_1)} \int_{-\infty}^{t_\nu^{-1}(u_2)} \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi\nu\sqrt{(1-\delta^2)}} \\ &\quad \times \left\{1+\frac{x_1^2-2\delta x_1 x_2+x_2^2}{(1-\delta^2)\nu}\right\}^{-\frac{\nu}{2}-1} dx_1 dx_2, \\ c_t(u_1,u_2,\nu,\delta) &= \frac{f_{\nu,\delta}\{t_\nu^{-1}(u_1), t_\nu^{-1}(u_2)\}}{f_\nu\{t^{-1}(u_1)\} f_\nu\{t^{-1}(u_2)\}}, \quad u_1,u_2,\delta \in [0,1], \end{aligned}$$

where δ denotes the correlation coefficient, ν is the number of degrees of freedom. $f_{\nu,\delta}$ and f_ν are joint and marginal t-distributions respectively, while t_ν^{-1} denotes the quantile function of the t_ν distribution. In-depth analysis of the t-copula is done in Rachev et al. (2008) and Luo and Shevchenko (2010). Long-tailed distributed margins lead to more mass and variability in the tail areas of the corresponding bivariate distribution.

However, the contour-curves of the t-copula are symmetric, which reflects the ellipticity of the underlying copula. This property is theoretically supported by Nelsen (2006, Theorem 2.3.7), stating that a bivariate copula is elliptical and thus, has reflection symmetry, if and only if

$$C(u_1, u_2, \theta) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2, \theta), \quad u_1, u_2 \in [0, 1].$$

The next class of copulae and their generalizations provide an important flexible and rich family of alternatives to elliptical copulae.

Archimedean Family

In contrast to elliptical copulae, Archimedean copulae are not constructed via (2.3), but are related to Laplace transforms of bivariate distribution functions. The function $C : [0, 1]^2 \to [0, 1]$ defined as

$$C(u_1, u_2) = \phi\{\phi^{-1}(u_1) + \phi^{-1}(u_2)\}, \quad u_1, u_2 \in [0, 1]$$

is a 2-dimensional Archimedean copula, where $\phi \in \mathcal{L} = \{\phi : [0; \infty) \to [0, 1] \,|\, \phi(0) = 1,\ \phi(\infty) = 0;\ (-1)^j \phi^{(j)} \geq 0;\ j = 1, \ldots, \infty\}$ is referred to as the generator of the copula. The generator usually depends on some parameters, however, mostly generators with a single parameter θ are considered. Nelsen (2006) and Joe (2014) provide a thoroughly classified list of popular generators for Archimedean copulae and discuss their properties.

The useful applications in finance, see Patton (2012), appearing to be the Gumbel copula with the generator function $\phi(x, \theta) = \exp\{-x^{1/\theta}\}$, $1 \leq \theta < \infty$, $x \in [0, 1]$, leading to the copula function

$$C(u_1, u_2, \theta) = \exp\left\{-\left[(-\log u_1)^\theta + (-\log u_2)^\theta\right]^{1/\theta}\right\}, \quad u_1, u_2 \in [0, 1].$$

Genest and Rivest (1989) showed that a bivariate distribution based on the Gumbel copula with extreme valued marginal distributions is the only bivariate extreme value distribution belonging to the Archimedean family. Moreover, all distributions based on Archimedean copulae belong to its domain of attraction under common regularity conditions. In contrary to elliptical copulae, the Gumbel copula leads to asymmetric contour diagrams in Figure 2.1. It exhibits a stronger linkage between positive values, however, more variability and more mass in the negative tail area. Opposite is observed for the Clayton copula with the generator $\phi(x, \theta) = (\theta x + 1)^{-\frac{1}{\theta}}$ with $-1 < \theta < \infty$, $\theta \neq 0$,

$x \in [0,1]$, and copula function

$$C(u_1, u_2, \theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}}, \quad u_1, u_2 \in [0,1].$$

Also, the Frank generator $\phi(x, \theta) = \theta^{-1} \log\{1 - (1 - e^{-\theta})e^{-x}\}$ with $0 \leq \theta < \infty$, $x \in [0,1]$, enjoys increased popularity and induces the copula function

$$C(u_1, u_2, \theta) = -\theta^{-1} \log \left\{ \frac{1 - e^{-\theta} - (1 - e^{-\theta u_1})(1 - e^{-\theta u_2})}{1 - e^{-\theta}} \right\}, \quad u_1, u_2 \in [0,1].$$

The respective Frank copula is the only elliptical Archimedean copula.

2.2.2 Bivariate Copula and Dependence Measures

Since copulae define the dependence structure between random variables, there is a relationship between copulae and different dependency measures. The classical measures for continuous random variables are Kendall's τ and Spearman's ρ. Similarly as copula functions, these measures are invariant under strictly increasing transformations. They are equal to 1 or -1 under perfect positive or negative dependence respectively. In contrast to τ and ρ, the Pearson correlation coefficient measures the linear dependence and, therefore, is not suitable for measuring non-linear relationships. Next, we discuss the relationship between τ, ρ and the underlying copula function.

Definition 2.3. Let F be a continuous bivariate cumulative distribution function with the copula C. Moreover, let $(X_1, X_2) \sim F$ and $(X_1', X_2') \sim F$ be independent random pairs. Then Kendall's τ is given by

$$\begin{aligned} \tau_2 &= \mathrm{P}\{(X_1 - X_1')(X_2 - X_2') > 0\} - \mathrm{P}\{(X_1 - X_1')(X_2 - X_2') < 0\} \\ &= 2\,\mathrm{P}\{(X_1 - X_1')(X_2 - X_2') > 0\} - 1 = 4 \int\limits_{[0,1]^2} C(u_1, u_2)\, dC(u_1, u_2) - 1. \end{aligned}$$

Kendall's τ represents the difference between the probability of two random concordant pairs and the probability of two random discordant pairs. For most copula functions with a single parameter θ there is a one-to-one relationship between θ and the Kendall's τ_2. For example, it holds that

$$\tau_2(\text{Gaussian and } t) = \frac{2}{\pi} \arcsin \delta, \quad \tau_2(\text{Archimedean}) = 4 \int_0^1 \frac{\phi^{-1}(t)}{(\phi^{-1})'}\, dt + 1,$$

$$\tau_2(\Pi) = 0, \qquad \tau_2(W) = 1, \qquad \tau_2(M) = -1.$$

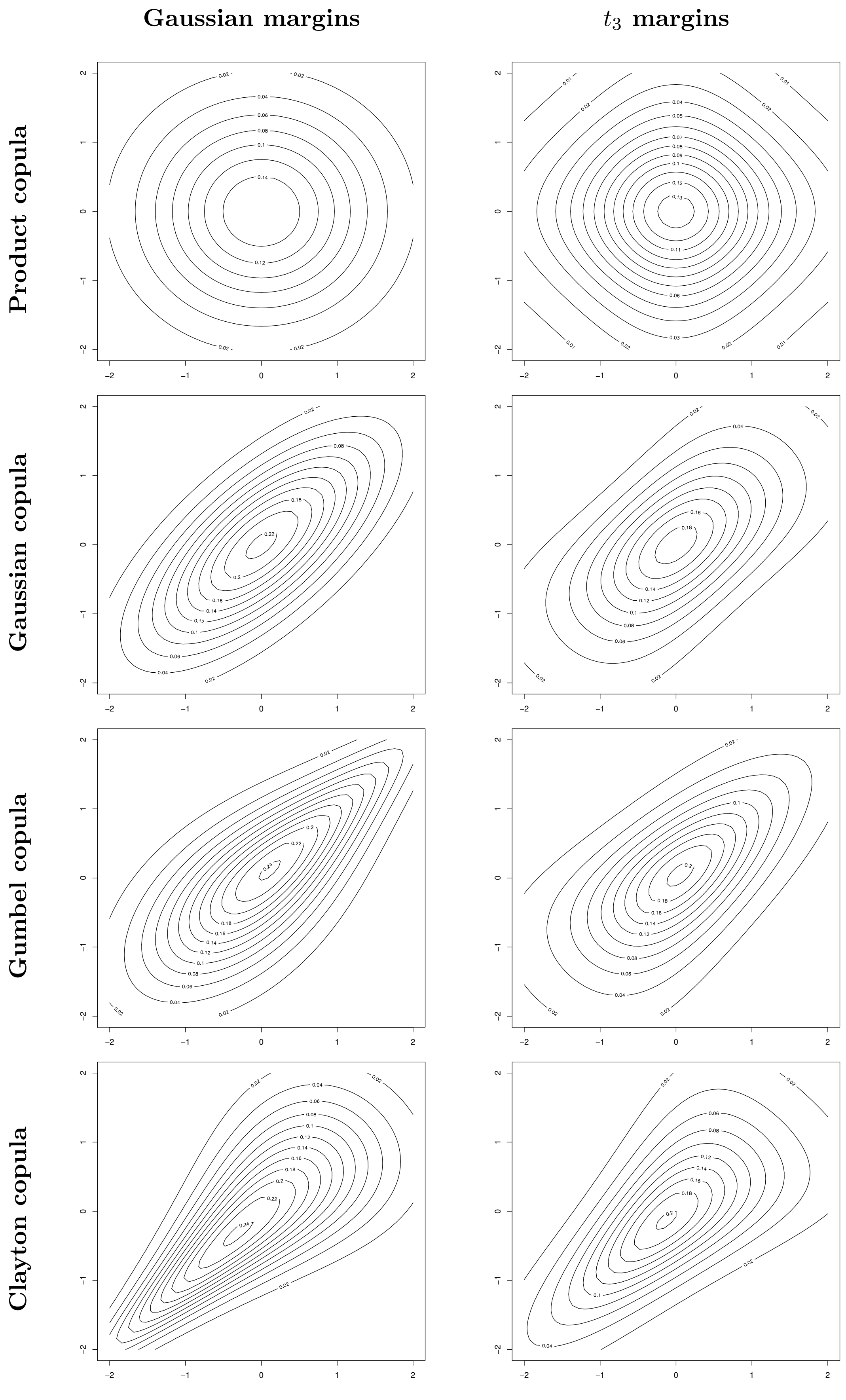


FIGURE 2.1: Contour diagrams for product, Gaussian, Gumbel and Clayton copulae with Gaussian (left column) and t_3 distributed (right column) margins.

For instance, this implies that an unknown copula parameter θ of the Gaussian, t and an arbitrary Archimedean copulae can be estimated using a type of method of moments procedure with a single moment condition. This requires, however, an estimator of τ_2, c.f. Kendall (1970). Naturally, it is computed by

$$\tau_{2n} = \frac{4}{n(n-1)} P_n - 1,$$

where n stands for the sample size and P_n denotes the number of concordant pairs, e.g. such pairs (X_1, X_2) and (X_1', X_2') that $(X_1 - X_1')(X_2 - X_2') > 0$. Next we provide the definition and similar results for the Spearman's ρ.

Definition 2.4. Let F be a continuous bivariate distribution function with the copula C and the univariate margins F_1 and F_2 respectively. Assume that $(X_1, X_2) \sim F$. Then the Spearman's ρ is given by

$$\rho_2 = 12 \int\limits_{\overline{\mathbb{R}}^2} F_1(x_1) F_2(x_2)\, dF(x_1, x_2) = 12 \int\limits_{[0,1]^2} u_1 u_2\, dC(u_1, u_2) - 3.$$

Similarly as for Kendall's τ, the relationship between Spearman's ρ and specific copulae is given through

$$\begin{aligned} \rho_2(\text{Gaussian and } t) &= \frac{6}{\pi} \arcsin \frac{\delta}{2}, \\ \rho_2(\Pi) &= 0, \qquad \rho_2(W) = 1, \qquad \rho_2(M) = -1. \end{aligned}$$

Unfortunately, there is no explicit representation of Spearman's ρ_2 for Archimedean in terms of generator functions as by Kendall's τ. The estimator of ρ is easily computed using

$$\rho_{2n} = \frac{12}{n(n+1)(n-1)} \sum_{i=1}^{n} R_i S_i - 3\frac{n+1}{n-1},$$

where R_i and S_i denote the ranks of two samples. The exact regions determined by Kendall's τ and Spearman's ρ has been recently given by Schreyer et al. (2017).

2.3 Multivariate Copula: Primer and State-of-the-Art

As mentioned in the introduction, step from bivariate copulas to multivariate is large. Nevertheless, many works have been written properly different high-dimensional copulas. This section introduces simple multivariate models and most prominent families like hierarchical Archimedean copula (HAC), pair-copula construction and factor copula.

A d-dimensional copula is also the distribution function on $[0,1]^d$ having all marginal distributions uniform on $[0,1]$. In Sklar's Theorem, the importance of copulas in the area of multivariate distributions is re-stated in an exquisite way.

Theorem 2.5. *Let F be a multivariate distribution function with margins $F_1, \ldots, F_d$, then there exists the copula C such that*

$$F(x_1, \ldots, x_d) = C\{F_1(x_1), \ldots, F_k(x_d)\}, \quad x_1, \ldots, x_d \in \overline{\mathbb{R}}.$$

If F_i are continuous for $i = 1, \ldots, d$ then C is unique. Otherwise C is uniquely determined on $F_1(\overline{\mathbb{R}}) \times \cdots \times F_d(\overline{\mathbb{R}})$.

Conversely, if C is a copula and $F_1, \ldots, F_d$ are univariate distribution functions, then function F defined above is a multivariate distribution function with margins $F_1, \ldots, F_d$.

As in the bivariate case, the representation in Sklar's Theorem can be used for constructing new multivariate distributions by changing either the copula function of marginal distributions. For an arbitrary continuous multivariate distribution we can determine its copula from the transformation

$$C(u_1, \ldots, u_d) = F\{F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\}, \quad u_1, \ldots, u_d \in [0,1], \tag{2.4}$$

where F_i^{-1} are inverse marginal distribution functions. Copula density and density of the multivariate distribution with respect to copula are

$$\begin{aligned} c(u_1, \ldots, u_d) &= \frac{\partial^k C(u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d}, \quad u_1, \ldots, u_d \in [0,1], \\ f(x_1, \ldots, x_d) &= c\{F_1(x_1), \ldots, F_d(x_d)\} \prod_{i=1}^{d} f_i(x_i), \quad x_1, \ldots, x_d \in \overline{\mathbb{R}}. \end{aligned}$$

For the multivariate case as well as for the bivariate case copula functions are invariant under monotone transformations.

2.3.1 Extensions of Simple and Elliptical Bivariate Copulae

The independence copula and the upper and lower Fréchet-Hoeffding bounds can be straightforwardly generalized to the multivariate case. The independence copula is defined by the product $\Pi(u_1, \ldots, u_d) = \prod_{i=1}^d u_i$ and the bounds are given by

$$\begin{aligned} W(u_1, \ldots, u_d) &= \max\big(0, \sum_{i=1}^{d} u_i + 1 - d\big), \\ M(u_1, \ldots, u_d) &= \min(u_1, \ldots, u_d), \quad u_1, \ldots, u_d \in [0,1]. \end{aligned}$$

An arbitrary copula $C(u_1, \dots, u_d)$ lies between the Fréchet-Hoeffdings bounds

$$W(u_1, \dots, u_d) \leq C(u_1, \dots, u_d) \leq M(u_1, \dots, u_d),$$

where the Fréchet-Hoeffding lower bound is not a copula function for $d > 2$ though. The generalization of elliptical copulas to $d > 2$ is straightforward as well. For example, the Gaussian case yields

$$\begin{aligned} C_N(u_1, \dots, u_d, \Sigma) &= \Phi_\Sigma\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\}, \\ c_N(u_1, \dots, u_d, \Sigma) &= |\Sigma|^{-1/2} \\ &\exp\Big[-\frac{1}{2}\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\}^\top (\Sigma^{-1} - I)\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_k)\}\Big] \end{aligned}$$

for all $u_1, \dots, u_d \in [0, 1]$, where Φ_Σ is a d-dimensional Gaussian distribution with zero mean and correlation matrix Σ. Individual dispersion is imposed via the marginal distributions. Note that in the multivariate case the implementation of elliptical copulas can be involved due to technical difficulties with multivariate cdf's.

2.3.2 Hierarchical Archimedean Copula

A simple multivariate generalization of the Archimedean copulas is defined as

$$C(u_1, \dots, u_d) = \phi\{\phi^{-1}(u_1) + \dots + \phi^{-1}(u_d)\}, \quad u_1, \dots, u_d \in [0, 1], \tag{2.5}$$

where $\phi \in \mathcal{L}$. This definition provides a simple, but rather limited technique for the construction of multivariate copulas, since a possibly complicated multivariate dependence structure is determined by a single copula parameter. Furthermore, multivariate Archimedean copulas imply that the variables are exchangeable. This means, that the distribution of $(u_1, \dots, u_d)$ is the same as of $(u_{j_1}, \dots, u_{j_d})$ for all $j_\ell \neq j_v$. This is certainly not an acceptable assumption in practical applications.

A more flexible method is provided by hierarchical Archimedean copulas (HAC) sometimes also called nested Archimedean copula which replace a uniform margin of a simple Archimedean copula by an additional Archimedean copula. The iterative substitution of margins by copulas widens the spectrum of attainable dependence structures. For example, the copula function for fully nested HAC is given by

$$\begin{aligned} C(u_1, \dots, u_d) &= \phi_{d-1}\big\{\phi_{d-1}^{-1} \circ \phi_{d-2}\big(\dots [\phi_2^{-1} \circ \phi_1\{\phi_1^{-1}(u_1) + \phi_1^{-1}(u_2)\} \\ &\quad + \phi_2^{-1}(u_3)] + \dots + \phi_{d-2}^{-1}(u_{d-1})\big) + \phi_{d-1}^{-1}(u_d)\big\} \\ &= \phi_{d-1}[\phi_{d-1}^{-1} \circ C(\{\phi_1, \dots, \phi_{d-2}\})(u_1, \dots, u_{d-1}) + \phi_{d-1}^{-1}(u_d)] \end{aligned} \tag{2.6}$$

for $\phi_{d-i}^{-1} \circ \phi_{d-j} \in \mathcal{L}^*$, $i < j$, where

$$\mathcal{L}^* = \{\omega : [0;\infty) \to [0,\infty) \,|\, \omega(0) = 0,\, \omega(\infty) = \infty;\, (-1)^{j-1}\omega^{(j)} \geq 0;\, j = 1,\dots,\infty\},$$

As indicated above, contrarily to the usual Archimedean copula (2.5), HAC define the dependency structure in a recursive way. At the lowest level of the so called HAC-tree, the dependency between the two variables is modeled by a copula function with the generator ϕ_1, i.e. $z_1 = C(u_1, u_2) = \phi_1\{\phi_1^{-1}(u_1) + \phi_1^{-1}(u_2)\}$. At the second level, an another copula function is used to model the dependency between z_1 and u_3, etc. The generators ϕ_i can come from the same family and differ only through the parameter or, to introduce more flexibility, come from different generator families, c.f. Hofert (2011). As an alternative to the fully nested model, so-called partially nested copulas combine arbitrarily many copula functions at each copula level. For example the following 4-dimensional copula, where the first and the last two variables are joined by individual copulas with generators ϕ_{12} and ϕ_{34}. Further, the resulted copulas are combined by a copula with the generator ϕ.

$$C(u_1, u_2, u_3, u_4) = \phi\big(\phi^{-1}[\phi_{12}\{\phi_{12}^{-1}(u_1) + \phi_{12}^{-1}(u_2)\}] + \phi^{-1}[\phi_{34}\{\phi_{34}^{-1}(u_3) + \phi_{34}^{-1}(u_4)\}]\big).$$

The estimation of HAC is a challenging task, since both the copula structure and parameters of the generator functions have to be estimated. The variety of possible structures does not permit the enumeration of all possible structures and selecting that structure-parameter combination with the largest log-likelihood value.

Okhrin et al. (2013a) first propose methods for determining the optimal structure of HAC with (non-)parametrically estimated margins and provide asymptotic theory for the estimated parameters. The basic idea of the estimation procedure uses the fact that HAC are recursively defined and that dependencies decreases from the lowest to the highest hierarchical level for common parametric families. To sketch the procedure suppose margins are known: Parameters related to strongly dependent random variables are estimated first and the variables grouped at the bottom of the HAC-tree. The determined HAC-tree is spanned by at least a two random variables and the tree itself determines a univariate random variable. After removing all random variables spanning the tree from the set of variables and adding the univariate random variable determined by the tree, the parameter of the subsequent level is determined by the selecting that pair of variables with the strongest dependency again. An additional level is added to the tree referring to the pair of variables with the strongest dependence and the set of variables is modified as explained above. The sketched steps are iteratively repeated until the HAC-tree is spanned by all random variables. This method is implemented in the `HAC` package for `R`, see Okhrin and Ristig (2014a).

Segers and Uyttendaele (2014) introduce an algorithm for non-parametric structure determination by firstly decomposing the HAC's tree structure into four variants of trivariate structures. Then, the whole tree structure is subsequently determined based on testing the distance between trivariate copulas and Kendall's distribution function. Górecki et al. (2016) generalize the approach of Okhrin et al. (2013a) and propose an algorithm for simultaneous estimation of the structure and parameters based on the inversion of Kendall's τ_2, i.e. based on the link between Kendall's τ_2 and Archimedean generators.

Properties and simulation procedures are comprehensively studied in Joe (1997), Whelan (2004), Savu and Trede (2010a), Hofert (2011), Okhrin et al. (2013b), Rezapour (2015) and Górecki et al. (2016). Note that HAC became a standard tool for pricing credit derivatives in academia such as collateralized debt obligations, see Hering et al. (2010), Hofert (2011) and Choroś-Tomczyk et al. (2013).

Brechmann (2014) proposed hierarchical Kendall copula, which does not suffer from parameter restriction, but are slightly more complicated in estimation. Similar approach to avoid parameter restrictions and family limitations are proposed by using Lévy subordinated HAC, see Hering et al. (2010) and the corresponding application see Zhu et al. (2016).

2.3.3 Factor Copula

In classical factor analysis, a function links the observed and latent variables under the assumption that the latent variables explain the observed variables, e.g., see Johnson and Wichern (2013) and Härdle and Simar (2015). For example, a random variable X_i, $i = 1, \ldots, d$, is generated by an additive factor model, if

$$X_i = \sum_{j=1}^{m} \alpha_{ij} W_j + \varepsilon_i, \tag{2.7}$$

where W_j, $j = 1, \ldots, m$, are latent common factors and ε_i, $i = 1, \ldots, d$, are mutually independent idiosyncratic disturbances. The basic idea of factor models and their natural interpretation can be exported to the copula world in order to induce dependencies between independent idiosyncratic disturbances via common factors. Factor copula models, however, can be split in two complementary groups both having strength and weaknesses. On the one hand, there are (implicit) factor copula models inducing dependencies among random variables via a functional which links latent factors and idiosyncratic disturbances. Such models are a straightforward extension of factor models from multivariate analysis. On the other hand, factor copulas and dependencies also arise from integrating the product of conditionally independent distributions –given a latent factor– with respect to this factor. This approach benefits from the fact, that the copula collapses to the product copula in case of known factors.

Oh and Patton (2015) concentrate on (implicit) factor copulas for $X = (X_1, \dots, X_d)^\top$ arising from a functional relation between the factor(s) and mutual independent idiosyncratic errors. In this sense, the dependence component of the joint distribution of X is implied from the factors' distribution, the distribution of the idiosyncratic disturbances and the link function. In particular, X follows a multivariate distribution specified via a copula, i.e. $X \sim F(x_1, \dots, x_d) = C\{F_1(x_1), \dots, F_d(x_d)\}$. For instance, the additive single factor copula model is represented as

$$\begin{aligned} &X_i = W + \varepsilon_i, \;\; i = 1, \dots, d, \\ &W \sim F_W(\theta_W), \;\; \varepsilon_i \overset{i.i.d.}{\sim} F_\varepsilon(\theta_\varepsilon), \;\; W \perp \varepsilon_i, \quad \text{for all} \quad i = 1, \dots, d,, \end{aligned} \tag{2.8}$$

where W is the single common factor following the distribution of $F_W(\theta_W)$ and $\varepsilon_1, \dots, \varepsilon_d$ are mutually independent shock with distribution function $F_\varepsilon(\theta_\varepsilon)$. This model is extended to the non-linear factor copula based on the following representation,

$$\begin{aligned} &Z_i = h(W, \varepsilon_i), \;\; i = 1, 2, \dots, d, \\ &W \sim F_W(\theta_W), \;\; \varepsilon_i \overset{i.i.d.}{\sim} F_\varepsilon(\theta_\varepsilon), \;\; W \perp \varepsilon_i, \quad \text{for all} \quad i = 1, \dots, d, \end{aligned} \tag{2.9}$$

where h is a non necessarily linear link function. Thus, the dependence structure can be built in a more flexible way compared to the linear additive version. Model (2.8) implies a joint Gaussian random vector $X = (X_1, \dots, X_d)^\top$, if the common factor and the idiosyncratic factor are both Gaussian. Therefore, a joint density function is available as well.

Nonetheless, a nice analytical expression of the joint density function for a factor copula with non-Gaussian margins and non-Gaussian factor is rarely available which makes parameter estimation demanding. Oh and Patton (2013) propose an estimation method for copula models without analytical form of the density function. This relies on a simulated method of moments approach building on the simplicity to draw random samples from a factor model. The proposed estimator for $(\theta_W^\top, \theta_\varepsilon^\top)^\top$ is found numerically by minimizing the distance between scale free empirical dependence measures between X_k and X_ℓ, such as $\tau_{2n}^{k\ell}$, $k = 1, \dots, d; \ell = k+1, \dots, d$, and those obtained from a drawn sample. Oh and Patton (2013) prove under weak regularity conditions that the simulated method of moment estimator is consistent and asymptotically normal. However, as argued by Genest et al. (1995), method of moment estimators of copula parameters can be highly inefficient.

Another form of factor copulae relies on the assumption that the observed variables $U_1, \ldots, U_d$ are conditionally independent given latent factors $V_1, \ldots, V_m$. Note that all random variables U_i, $i = 1, \ldots, d$, and V_j, $j = 1, \ldots, m$, are assumed to uniformly distributed. Then, the conditional distribution of U_i given m factors $V_1, \ldots, V_m$ is given by $C_{U_i|V_1,\ldots,V_m}$. By using $C_{U_i|V_1,\ldots,V_m}$, the dependence structure of the observed variables $U_1, \ldots, U_d$ can be specified by the following copula function, such that

$$C(u_1, \ldots, u_d) = \int_{[0,1]^m} \prod_{i=1}^{d} C_{U_i|V_1,\ldots,V_m}(u_i|v_1, \ldots, v_m) dv_1 \cdots dv_m \quad \text{with} \quad u_i \in (0,1), \tag{2.10}$$

where the factors are out integrated. For the special case $m = 1$, the copula function (2.10) can be simplified to the form

$$C(u_1, \ldots, u_d) = \int_{[0,1]} \prod_{i=1}^{d} C_{U_i|V_1}(u_i|v_1) dv_1 \quad \text{with} \quad u_i \in (0,1). \tag{2.11}$$

Let C_{U_i,V_1} and c_{U_i,V_1} be the joint cdf and density of the pairs of random variables (U_i, V_1), $i = 1, \ldots, d$. Moreover, let the conditional distribution of U_i given V_1 be denoted by $C_{U_i|V_1}(u_i|v_1) = \partial C_{U_i,V_1}(u_i, v)/\partial v|_{v=v_1}$. Then, the copula density of $C(u_1, \ldots, u_d)$ can be represented by

$$c(u_1, \ldots, u_d) = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \cdots \partial u_d} = \int_{[0,1]} \prod_{i=1}^{d} c_{U_i,V_1}(u_i, v_1) dv_1 \quad \text{with} \quad u_i \in (0,1), \tag{2.12}$$

where $c_{U_i,V_1}(u_i, v_1) = \partial C(u_i|v_1)/\partial u_i$. Seen from this angle, the dependencies between d observed variables is determined by d bivariate copulas $C_{U_i,V_1}(u_i, v)$. Based on a parametric copula density $c(\cdot;\theta)$, Krupskii and Joe (2013) separate the parameter estimation into two steps. In the first step, the margins are estimated parametrically or non-parametrically. In the second step, the maximum likelihood (ML) method is employed to estimate the parameter θ.

Numerous literature about the factor copula's theory and applications can be referred to. Andersen et al. (2003), Hull and White (2004) and Laurent and Gregory (2005) have contributed works on generalization of one factor copula models. A comprehensive review of the factor copula theory is given in Joe (2014). Some applications by using factor copula models can be referred to Li (2000) for credit derivative pricing, Krupskii and Joe (2013) for fitting stock returns and Oh and Patton (2015) for measuring systemic risk.

2.3.4 Vine Copula

Vine copula or pair-copula constructions are originally proposed in Joe (1996a) and developed in depth by Bedford and Cooke (2001), Bedford and Cooke (2002), Kurowicka and Cooke (2006) and Aas et al. (2009). The catchy name is due to similarities of the graphical representation of vine copulae and botanical vines. The fundamental idea of the vine copula is to construct a d-dimensional copula by decomposing the dependence structure into $d(d-1)/2$ bivariate copulas.

Let S be the index subset of $D = \{1, \dots, d\}$ referring to the index set of conditioning variables and T be the index set of conditioned variables with $T \cup S = D$. Let $\sharp M$ denote the cardinality of set M. The cdf of variables with index in S is denoted by F_S, so that $F(x) = F_D(x)$. The conditional cdf of variables with index in T conditional on S is denoted $F_{T|S}$. A similar notation is used for the corresponding copulas. To derive a vine copula for a given $x = (x_1, \dots, x_d)^\top$ in the spirit of Joe (2014), we start from a d-dimensional distribution function, i.e.

$$F(x) = \int_{(-\infty, x_S]} F_{T|S}(x_T|y_S) dF_S(y_S), \tag{2.13}$$

and replace the conditional distribution $F_{T|S}(x_T|x_S)$ by the corresponding $\sharp T$-dimensional copula $F_{T|S}(x_T|x_S) = C_{T;S}\{F_{j|S}(x_j|x_S) : j \in T\}$. The copula $C_{T;S}\{F_{j|S}(x_j|x_S) : j \in T\}$ is implied by Sklar's Theorem with margins $F_{j|S}(x_j|x_S)$, $j \in T$. It is not a conditional distribution although with conditional distribution as margins. This yields a copula-based representation of the joint d-dimensional distribution function from (2.13), which is given by

$$F(x) = \int_{(-\infty, x_S]} C_{T;S}\{F_{j|S}(x_j|y_S) : j \in T\} dF_S(y_S). \tag{2.14}$$

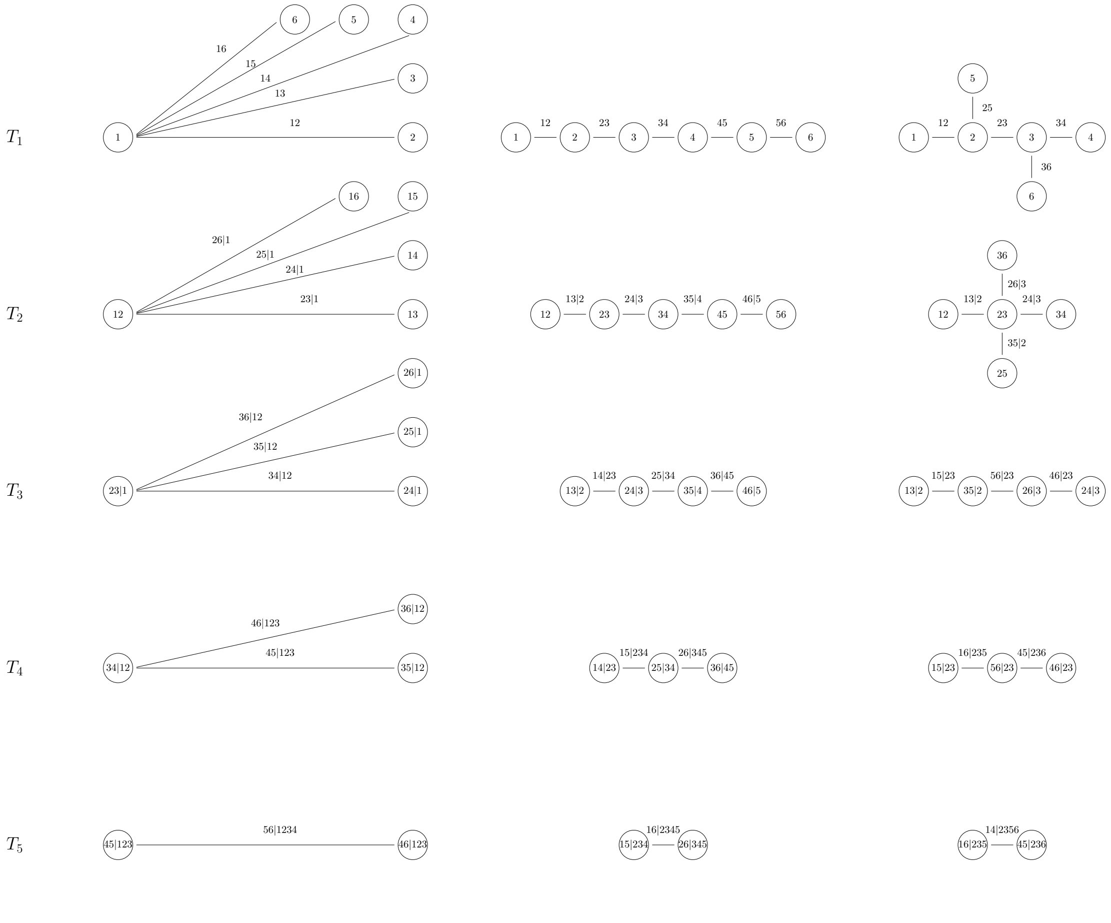
Note that the support of the integral in (2.13) and (2.14) is a cube $(-\infty, x_S] \in \mathbb{R}^{\sharp S}$. Converting all univariate margins to uniform distributed random variables allows rewriting $F(x)$ as a d-dimensional copula

$$C(u) = \int_{[0, u_S]} C_{T;S}\{G_{j|S}(u_j|v_S) : j \in T\} dC_S(v_S), \tag{2.15}$$

where $G_{j|S}(u_j|v_S)$ is a conditional distribution from copula $C_{S\cup\{j\}}$. If $T = \{i_1, i_2\}$, then

$$C_{S\cup\{i_1,i_2\}}(u_{S\cup\{i_1,i_2\}}) = \int_{[0, u_S]} C_{i_1,i_2;S}\{G_{i_1|S}(u_{i_1}|v_S), G_{i_2|S}(u_{i_2}|v_S)\} dC_S(v_S). \tag{2.16}$$

Since the essential idea of vine copula is based on building a joint dependence structure by $d(d-1)/2$ bivariate copulae, (4.11) is an important building block in the construction of vines referring to a $(\sharp S + 2)$-dimensional copula built from a bivariate copula $C_{i_1,i_2;S}$.



C-Vine D-Vine R-Vine

Figure 2.2: Vine tree structures of C-vine, D-vine and R-vine.

In case of continuous random variables, the d-dimensional distribution function from (2.13) admits a density function $f(x_1, \ldots, x_d)$, which can be decomposed and represented by bivariate copula densities in an analogue manner. Examples of density decompositions for the 6-dimensional case related to so called C-vine (canonical vine), D-vine (drawable vine) and R-vine (regular vine) copulas are given as follows. The C-vine structure is illustrated in the left column of Figure 2.2 and its density decomposition is

$$
\begin{aligned}
c\{F_1(x_1),\ldots,F_6(x_6)\} = {} & c_{12}\{F_1(x_1),F_2(x_2)\}\cdot c_{13}\{F_1(x_1),F_3(x_3)\} \\
& \cdot c_{14}\{F_1(x_1),F_4(x_4)\}\cdot c_{15}\{F_1(x_1),F_5(x_5)\}\cdot c_{16}\{F_1(x_1),F_6(x_6)\} \\
& \cdot c_{23;1}\{F(x_2|x_1),F(x_3|x_1)\}\cdot c_{24;1}\{F(x_2|x_1),F(x_4|x_1)\} \\
& \cdot c_{25;1}\{F(x_2|x_1),F(x_5|x_1)\}\cdot c_{26;1}\{F(x_2|x_1),F(x_6|x_1)\} \\
& \cdot c_{34;12}\{F(x_3|x_{12}),F(x_4|x_{12})\}\cdot c_{35;12}\{F(x_3|x_{12}),F(x_5|x_{12})\} \\
& \cdot c_{36;12}\{F(x_3|x_{12}),F(x_6|x_{12})\}\cdot c_{45;123}\{F(x_4|x_{123}),F(x_5|x_{123})\} \\
& \cdot c_{46;123}\{F(x_4|x_{123}),F(x_6|x_{123})\}\cdot c_{56;1234}\{F(x_5|x_{1234}),F(x_6|x_{1234})\}.
\end{aligned}
\tag{2.17}
$$

The density of the D-vine structure –given in the centred column of Figure 2.2– is

$$
\begin{aligned}
c\{F_1(x_1),\ldots,F_6(x_6)\} = {} & c_{12}\{F_1(x_1),F_2(x_2)\}\cdot c_{23}\{F_2(x_2),F_3(x_3)\} \\
& \cdot c_{34}\{F_3(x_3),F_4(x_4)\}\cdot c_{45}\{F_4(x_4),F_5(x_5)\}\cdot c_{56}\{F_5(x_5),F_6(x_6)\} \\
& \cdot c_{13;2}\{F(x_1|x_3),F(x_2|x_3)\}\cdot c_{24;3}\{F(x_2|x_3),F(x_4|x_3)\} \\
& \cdot c_{35;4}\{F(x_3|x_4),F(x_5|x_4)\}\cdot c_{46;5}\{F(x_4|x_5),F(x_6|x_5)\} \\
& \cdot c_{14;23}\{F(x_2|x_{23}),F(x_4|x_{23})\}\cdot c_{25;34}\{F(x_2|x_{34}),F(x_5|x_{34})\} \\
& \cdot c_{36;45}\{F(x_3|x_{45}),F(x_6|x_{45})\}\cdot c_{15;234}\{F(x_1|x_{234}),F(x_5|x_{234})\} \\
& \cdot c_{26;345}\{F(x_2|x_{345}),F(x_6|x_{345})\}\cdot c_{16;2345}\{F(x_1|x_{2345}),F(x_6|x_{2345})\}.
\end{aligned}
\tag{2.18}
$$

The density of the R-vine structure illustrated in the right column of Figure 2.2 is

$$
\begin{aligned}
c\{F_1(x_1),\ldots,F_6(x_6)\} = {} & c_{12}\{F_1(x_1),F_2(x_2)\}\cdot c_{23}\{F_2(x_2),F_3(x_3)\} \\
& \cdot c_{34}\{F_3(x_3),F_4(x_4)\}\cdot c_{25}\{F_2(x_2),F_5(x_5)\}\cdot c_{36}\{F_3(x_3),F_6(x_6)\} \\
& \cdot c_{13;2}\{F(x_1|x_2),F(x_3|x_2)\}\cdot c_{24;3}\{F(x_2|x_3),F(x_4|x_3)\} \\
& \cdot c_{26;3}\{F(x_2|x_3),F(x_6|x_3)\}\cdot c_{35;2}\{F(x_3|x_2),F(x_5|x_2)\} \\
& \cdot c_{15;23}\{F(x_1|x_{23}),F(x_5|x_{23})\}\cdot c_{56;23}\{F(x_5|x_{23}),F(x_6|x_{23})\} \\
& \cdot c_{46;23}\{F(x_4|x_{23}),F(x_6|x_{23})\}\cdot c_{16;235}\{F(x_1|x_{235}),F(x_6|x_{235})\} \\
& \cdot c_{45;236}\{F(x_4|x_{236}),F(x_5|x_{236})\}\cdot c_{14;2356}\{F(x_1|x_{2356}),F(x_4|x_{2356})\}.
\end{aligned}
\tag{2.19}
$$

In particular, the C-vine and D-vine have an intuitive graphical representation which can be immediately related to the decomposition of the copula density function into the product of bivariate copula densities. For example, the product of bivariate copula densities from the first two lines of the right hand side of Equation 2.17 refers to a C-vine represented in the upper left graphic of Figure 2.2. The formula and the corresponding graphic illustrate that the first variable X_1 is pairwise coupled with the second, third ... and sixth random variable. The subsequent two lines (3-4) of Equation 2.17 are related to the second graphic of the left column of Figure 2.2. Conditional on X_1, random variable X_2 is pairwise coupled with X_3, X_4, X_5 and X_6. Connecting the remaining graphics with formulas is left to the reader. While the "formula-graphic" matching follows a similar scheme in case of the D-vine, the R-vine belongs to a more general vine copula class and contains the C-vine and D-vine as special cases. A rigorous definition of an R-vine copula can be found in Joe (2014).

In fact, vines can be estimated by either full or stage-wise ML such as the inference function for margins (IFM) method discussed below in Section 2.4. Nonetheless, the inference approach derived in Haff (2013) namely the stepwise semi-parametric estimator deserves to be mentioned in more detail. Here, the marginal distributions are non-parametrically estimated by the empirical distribution function such as for factor copulae or HAC. In order to obtain a consistent and asymptotically Gaussian distributed estimator of a parametric vine copula, a so called simplifying assumption is required. The latter permits replacing "conditional" bivariate copula densities with unconditional densities. Then, it can be straightforwardly shown, that the log-likelihood can be maximized in a stage-wise manner. This is due to the decomposition of the density into the product of bivariate copula densities, so that the log-likelihood function is a sum of logarithmized copula densities. Coming back to the C-vine example from Figure 2.2. At the first stage, all parameters of bivariate copulas represented in the upper left graphic of Figure 2.2 are estimated, i.e. the parameters of the copulae for $(X_1, X_2), \dots, (X_1, X_6)$. Keeping the corresponding parameters fixed at estimated values, the four parameters of copulae referring to the pairs from the second graphic of the left column of Figure 2.2 are estimated. Holding these parameters fixed at estimated values again, all vine parameters of the remaining bivariate densities can be estimated iteratively. Literature on pair-copula construction is spreading steadily, and most recent information about it can be found on vine copula homepage `http://www.statistics.ma.tum.de/en/research/vine-copula-models/`.

2.4 Estimation Methods

The estimation of a copula-based multivariate distribution involves both the estimation of the copula parameters $\boldsymbol{\theta}$ and the estimation of the margins F_j, $j = 1, \dots, d$. The properties and goodness of the estimator of $\boldsymbol{\theta}$ heavily depend on the estimators of F_j, $j = 1, \dots, d$. We distinguish between a parametric and a non-parametric specification of the margins. If we are interested only in the dependency structure, the estimator of $\boldsymbol{\theta}$ should be independent of any parametric models for the margins. However, Joe (1997) argues that complete distribution models and, therefore, parametric models for margins are actually more appropriate for applications.

In the bivariate case, a standard method of estimating the univariate parameter θ is based on Kendall's τ_2 statistic by Genest and Rivest (1993). The estimator of τ_2 complemented by the method of moments allows to estimate the parameters. However, as shown in Genest et al. (1995), the ML method leads to substantially more efficient estimators. For non-parametrically estimated margins, Genest et al. (1995) show the consistency and asymptotic normality of ML estimators and derive the moments of the asymptotic distribution. The ML procedure can be performed simultaneously for the parameters of the margins and of the copula function. Alternatively, a two-stage procedure can be applied, where the parameters of margins are estimated at the first stage and the copula parameters at the second stage, see Joe (1997) and Joe (2005). Chen and Fan (2006) and Chen et al. (2006) analyze the case of non-parametrically estimated margins. Fermanian and Scaillet (2003) and Chen and Huang (2007) consider a fully non-parametric estimation of the copula. Next we provide details on both approaches. Note that estimation procedures for HAC, conditional-independence-based factor copulas and vines are in fact generalizations of the subsequent approaches taking specific needs of the copula into account, e.g., parameter restrictions.

2.4.1 Parametric Margins

Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \dots, \boldsymbol{\alpha}_d^\top)^\top$ denote the vector of parameters of marginal distributions and $\boldsymbol{\theta}$ parameters of the copula. The classical full ML estimator $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\theta}^\top)^\top$ solves the system of equations

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\eta}, \mathbf{X})}{\partial \boldsymbol{\eta}} &= \mathbf{0}, \\
\text{where} \quad \mathcal{L}(\boldsymbol{\eta}, \mathbf{X}) &= \sum_{i=1}^{n} \log \left\{ c(F_1(x_{1i}, \boldsymbol{\alpha}_1), \dots, F_d(x_{di}, \boldsymbol{\alpha}_d), \boldsymbol{\theta}) \prod_{j=1}^{d} f_j(x_{ji}, \boldsymbol{\alpha}_j) \right\} \\
&= \sum_{i=1}^{n} \Big\{ \log c(F_1(x_{1i}, \boldsymbol{\alpha}_1), \dots, F_d(x_{di}, \boldsymbol{\alpha}_k), \boldsymbol{\theta}) + \sum_{j=1}^{d} \log f_j(x_{ji}, \boldsymbol{\alpha}_j) \Big\}.
\end{aligned}$$

Following the standard theory on ML estimation, the estimator $\hat{\boldsymbol{\eta}}$ is efficient and asymptotically normal. However, it is often computationally demanding to solve the system simultaneously. Alternatively the multistage optimization proposed in Joe (1997, Chapter 10), also known as inference functions for margins, can be applied: Firstly, the parameters of the margins are separately estimated under the assumption that the copula is the product copula. Secondly, the parameters of the copula are estimated replacing the parameters of margins by estimates from the first step and treating them as known quantities. The above optimization problem is then replaced by

$$\left(\frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\alpha}_1^\top}, \dots, \frac{\partial \mathcal{L}_d}{\partial \boldsymbol{\alpha}_d^\top}, \frac{\partial \mathcal{L}_{d+1}}{\partial \boldsymbol{\theta}^\top}\right)^\top = \mathbf{0}, \tag{2.20}$$

$$\begin{aligned}
\text{where} \qquad & \mathcal{L}_j = \sum_{i=1}^{n} l_j(\mathbf{X}_i), \quad \text{for} \quad j = 1, \dots, d+1, \\
& l_j(\mathbf{X}_i) = \log f_j(x_{ji}, \boldsymbol{\alpha}_j), \quad \text{for} \quad j = 1, \dots, d, i = 1, \dots, n, \\
\text{and} \qquad & l_{d+1}(\mathbf{X}_i) = \log c\big\{F_1(x_{1i}, \boldsymbol{\alpha}_1), \dots, F_d(x_{di}, \boldsymbol{\alpha}_d), \boldsymbol{\theta}\big\}, \quad \text{for} \quad i = 1, \dots, n.
\end{aligned}$$

The first d components in (2.20) correspond to the usual ML estimation of the parameters of the marginal distributions. The last component reflects the estimation of the copula parameters. Detailed discussion on this method can be found in Joe (1997, Chapter 10). Note, that this procedure does not lead to efficient estimators, however, as argued by Joe (1997) the loss in the efficiency is modest and mainly depends on the strength of dependencies. This method is a special case of the generalized method of moments with an identity weighting matrix, see Cherubini et al. (2004, Section 4.5). The advantage of the two-stage procedure lies in the dramatic reduction of the numerical complexity.

2.4.2 Non-parametric Margins

In this section, we consider a non-parametric estimation of the marginal distributions also referred to as canonical ML. The asymptotic properties of the multistage estimator for $\boldsymbol{\theta}$ do not depend explicitly on the type of the non-parametric estimator, but on its convergence properties. Here, we use the rectangular kernel (histogram) resulting in the estimator

$$\widehat{F}_j(x) = (n+1)^{-1} \sum_{i=1}^{n} \mathbf{1}(x_{ji} \leq x), \quad j = 1, \dots, d.$$

The factor $n/(n+1)$ is used to restrict fitted values to the open unit interval. This is necessary as several copula densities are not bounded at zero and/or one. Let $\widehat{F}_1, \dots, \widehat{F}_d$ denote the non-parametric estimators of $F_1, \dots, F_d$. The canonical ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ solves the system $\partial \mathcal{L} / \partial \boldsymbol{\theta}^\top = \mathbf{0}$ by maximizing the pseudo log-likelihood with estimated margins $\widehat{F}_1, \dots, \widehat{F}_d$, i.e.

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{n} l(\mathbf{X}_i) \quad \text{for} \quad j = 1, \dots, p, \\
l(\mathbf{X}_i) &= \log c\big\{\widehat{F}_1(x_{1i}), \dots, \widehat{F}_d(x_{di}), \boldsymbol{\theta}\big\}, \quad \text{for} \quad i = 1, \dots, n.
\end{aligned}$$

As in the parametric case, the semi-parametric estimator $\hat{\boldsymbol{\theta}}$ is asymptotically normal under suitable regularity conditions. This method was first used in Oakes (1994) and then investigated by Genest et al. (1995) and Shih and Louis (1995). Additional properties of the estimator, such as the covariance matrix, are stated in these papers.

2.5 Goodness-of-Fit Tests for Copulae

Having a dataset and an estimated copula at hand, it arises the natural question whether the selected copula describes the data properly. For this purpose, a series of different goodness-of-fit tests has been developed in the last decade. Under the H_0-hypothesis one assumes that the true copula belongs to some parametric family $H_0 : C \in C_0$.

The most natural test approach is to measure the deviation of the parametric copula from the empirical one given through

$$C_n(u_1, \ldots, u_d) = n^{-1} \sum_{i=1}^{n} \prod_{j=1}^{d} I\{\widehat{F}_j(x_{ij}) \leq u_j\}.$$

Gaensler and Stute (1987) and Radulovic and Wegkamp (2004) show that C_n is a consistent estimation of the true underlying copula. Several tests are based on the empirical copula process, which is defined as follows

$$\mathbb{C}_n(u_1, \ldots, u_d) = \sqrt{n}\{C_n(u_1, \ldots, u_d) - C_{\hat{\theta}}(u_1, \ldots, u_d)\}.$$

Fermanian (2005) and Genest and Rèmillard (2008) propose to compute different measures to quantify the deviation of the assumed parametric copula from the empirical copula, one of those is Cramér-von Mises distance

$$S_n^E = \int_{[0,1]^d} \mathbb{C}_n(u_1, \ldots, u_d)^2 dC_n(u_1, \ldots, u_d)$$

or the weighted Cramér-von Mises distance, with tuning parameters $m \geq 0$ and $\zeta_m \geq 0$ given as

$$R_n^E = \int_{[0,1]^d} \left\{ \frac{\mathbb{C}_n(u_1, \ldots, u_d)}{[C_{\hat{\theta}}(u_1, \ldots, u_d)\{1 - C_{\hat{\theta}}(u_1, \ldots, u_d)\} + \zeta_m]^m} \right\}^2 dC_n(u_1, \ldots, u_d).$$

The usual Kolmogorov-Smirnov distance as for classical univariate tests is also applicable here

$$T_n^E = \sup_{\{u_1, \ldots, u_d\} \in [0,1]^d} |\mathbb{C}_n(u_1, \ldots, u_d)|.$$

The other group of tests developed and investigated by Genest and Rivest (1993), Wang and Wells (2000), Genest et al. (2006) are based on the probability integral transform and in particular on so called Kendall's transform. Having

$$(X_1, \dots, X_d) \sim F(x_1, \dots, x_d) = C_\theta\{F_1(x_1), \dots, F_d(x_d)\},$$

one concludes similar to $F_i(X_i) \sim U(0,1)$ that the copula-based random variable is

$$C_\theta\{F_1(X_1), \dots, F_d(X_d)\} \sim K_\theta(v)$$

where $K_\theta(v)$ is the univariate Kendall's distribution (not necessarily uniform), see Barbe et al. (1996), Jouini and Clemen (1996). Empirically, the distribution function K can be estimated as

$$K_n(v) = n^{-1} \sum_{i=1}^{n} I\left[C_n\{\widehat{F}_1(x_{i1}), \dots, \widehat{F}_d(x_{id})\} \leq v\right], \quad v \in [0,1].$$

Further usual test statistics for the univariate distributions like Cramér-von Mises or Kolmogorov-Smirnov, see Genest et al. (2006), can be applied

$$S_n^{(K)} = \int_0^1 \mathbb{K}_n(v)^2 dK_{\hat{\theta}}(v), \qquad T_n^{(K)} = \sup_{v \in [0,1]} |\mathbb{K}_n(v)|,$$

where $\mathbb{K}_n = \sqrt{n}(K_n - K_{\hat{\theta}})$ is the Kendall's process. Here is, however, a little challenge in using this tests: as in testing for Kendall's distribution one tests in null hypothesis has $H_0^{''} : K \in \mathcal{K}_0 = \{K_\theta : \theta \in \Theta\}$, and as $H_0 \subset H_0^{''}$, the non-rejection of $H_0^{''}$ does not imply non rejection of H_0. For the bivariate Archimedean copulas $H_0^{''}$ and H_0 are equivalent.

Another series of goodness-of-fit tests, is constructed via the other important integral transform, that dates back to Rosenblatt (1952). Based on the conditional distribution of U_i by

$$\begin{aligned} C_d(u_i|u_1, \dots, u_{i-1}) &= \mathrm{P}\{U_i \leq u_i | U_1 = u_1 \dots U_{i-1} = u_{i-1}\} \\ &= \frac{\partial^{i-1} C(u_1, \dots, u_i, 1, \dots, 1)/\partial u_1 \dots \partial u_{i-1}}{\partial^{i-1} C(u_1, \dots, u_{i-1}, 1, \dots, 1)/\partial u_1 \dots \partial u_{i-1}}, \end{aligned}$$

the Rosenblatt transform is defined as follows.

Definition 2.6. Rosenblatt's probability integral transform of a copula C is the mapping $\mathfrak{R} : (0,1)^d \to (0,1)^d$, $\mathfrak{R}(u_1, \dots, u_d) = (e_1, \dots, e_d)$ with $e_1 = u_1$ and $e_i = C_d(u_i|u_1, \dots, u_{i-1})$, $\forall i = 2, \dots, d$.

Under this definition, the null hypothesis $H_0 : C \in \mathcal{C}_0$ can be rewritten as H_{0R} :

$(e_1, \ldots, e_d)^\top \sim \Pi$. The first test based on the Rosenblatt transform exploits information, that under H_0 transformed observations should be exactly uniform distributed and independent, which is not the case, as those variables as not mutually independent and only approximately uniform. Nevertheless, two tests use Anderson-Darling test statistics, see Breymann et al. (2003), and are constructed as

$$T_n = -n - \sum_{i=1}^{n} \frac{2i-1}{n} [\log G_{(i)} + \log\{1 - G_{(n+1-i)}\}]$$

where G_i might be constructed in two ways. In the first possibility

$$G_{i,Gamma} = \Gamma_d \left\{ \sum_{j=1}^{d} (-\log e_{ij}) \right\},$$

where $\Gamma_d(\cdot)$ is the Gamma distribution with shape d and scale 1. The second way takes

$$G_{i,\chi^2} = \chi^2_d \left[\sum_{j=1}^{d} \{\Phi^{-1}(e_{ij})\}^2 \right],$$

where χ^2_d refers to the Chi-squared distribution with d degrees of freedom and Φ is standard normal distribution. Another possibility compares the variables not via the Anderson-Darling test statistics, but by purely deviations between estimated density functions, as in Patton et al. (2004), where the test statistics is constructed by

$$C_n^{Ch} = \frac{n\sqrt{h}\hat{J}_n - c_n}{\sigma}$$

with c_n and σ are normalization factors and $\hat{J}_n = \int_0^1 \{\frac{1}{n}\sum_{i=1}^{n} K_h(w, G_{i,\chi^2}) - 1\}^2 dw$.

As discussed by Dobrić and Schmid (2007), the problem with those tests is that they have almost no power and even do not capture the type 1 error. Much better power have tests, that work directly on the copulas of the Rosenblatt transformed data, see Genest et al. (2009). The idea is to compute Cramer-von Mises statistics of the following form

$$S_n = n \int_{[0,1]^d} \{D_n(u) - \Pi(u)\}^2 du$$

$$S_n^{(C)} = n \int_{[0,1]^d} \{D_n(u) - \Pi(u)\}^2 dD_n(u)$$

where the empirical distribution function

$$D_n(u) = D_n(u_1, \ldots, u_d) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} I(e_{ij} \leq u_j)$$

should be “close” to product copula Π under H_0.

Different from previous test are those based on the kernel density estimators, and just to mention one, let us consider test developed by Scaillet (2007), where the test statistics is given through

$$J_n = \int_{[0,1]^d} \{\hat{c}(u) - K_H * c(u;\hat{\theta})\} w(u) du,$$

with “$*$” being a convolution operator and $w(u)$ a weight function. The kernel function $K_H(y) = K(H^{-1}y)/\det(H)$ where K is the bivariate quadratic kernel with the bandwidth $H = 2.6073 n^{-1/6} \widehat{\Sigma}^{1/2}$ and $\widehat{\Sigma}$ being a sample covariance matrix. The copula density is estimated non-parametrically as

$$\hat{c}(u) = n^{-1} \sum_{i=1}^{n} K_H[u - \{\widehat{F}_1(X_{i1}), \ldots, \widehat{F}_d(X_{id})\}^\top],$$

where $\widehat{F}_j$ refers to an estimated marginal distribution, $j = 1, \ldots, d$. The most recent goodness of fit test for copulas have been proposed recently by Zhang et al. (2016), where one compares the two-step pseudo maximum likelihood:

$$\hat{\theta} =_{\theta \in \Theta} \sum_{i=1}^{n} \mathcal{L}\{\widehat{F}_1(X_{i1}), \ldots, \widehat{F}_d(X_{id}); \theta\}.$$

with the delete-one-block pseudo maximum likelihood $\hat{\theta}_{-b}$, $1 \leq b \leq B$:

$$\hat{\theta}_{-b} =_{\theta \in \Theta} \sum_{b' \neq b}^{B} \sum_{i=1}^{m} \mathcal{L}\{\widehat{F}_1(X_{i1}), \ldots, \widehat{F}_d(X_{id}); \theta\}, \quad b = 1, \ldots, B.$$

Further, ”in-sample” and ”out-of-sample” pseudo-likelihoods are compared with the following test statistic:

$$T_n(m) = \sum_{b=1}^{B} \sum_{i=1}^{m} \left[\mathcal{L}\{\widehat{F}_1(X_{i1}), \ldots, \widehat{F}_d(X_{id}); \hat{\theta}\} - \mathcal{L}\{\widehat{F}_1(X_{i1}), \ldots, \widehat{F}_d(X_{id}); \hat{\theta}_{-b}\} \right].$$

This leads to some challenges, like computation of $[\frac{n}{m}]$ dependence parameters, but Zhang et al. (2016) proposed an asymptotically equivalent test statistics based on variability and sensitivity matrices. As most of the above mentioned tests, have complicated asymptotic distributions, p-values of the tests can be performed via the parametric bootstrap sketched in the subsequent procedure:

Step 1 Generate bootstrap sample $\left\{\epsilon_i^{(k)}, i = 1, \ldots, n\right\}$ from copula $C(u;\hat{\theta})$ under H_0 with $\hat{\theta}$ and estimated marginal distribution $\widehat{F}$ obtained from original data;

Step 2 Based on $\left\{\epsilon_i^{(k)}, i = 1, \ldots, n\right\}$ from Step 1, estimate θ of the copula under H_0, and compute test statistics under consideration, say R_n^k;

Step 3 Repeat Steps $(1 - 2)$ N-times and obtain N statistics $R_n^k, k = 1, \ldots, N$;

Step 4 Compute an empirical p-value as $p_e = N^{-1} \sum_{k=1}^{N} I\left(|R_n^k| \geq |R_n|\right)$ with R_n being the test statistic estimated from original data.

2.6 Empirical Study

Value-at-Risk (VaR) is an important measure in risk management. The traditional models for VaR estimation assume that the assets returns in a portfolio are jointly normally distributed. However, numerous empirical studies show that Gaussian based models are not sufficient to describe data characteristics, especially when extreme events happen such as financial crisis. The weak points of the Gaussian based models include the lack of asymmetry and tail dependence. Therefore copula methods come into the focus.

Twelve different copulas are used in this study to construct dependence structures. The employed families include the Gaussian copula, t-copula, Archimedean copulas (Clayton, Gumbel, Joe), HAC (Gumbel, Clayton, Frank), C- and D-vine structures and two factor copulas linked individually by a bivariate Gumbel and Clayton copula.

The data set utilized in this study includes five time series of stock close prices containing ADI (Analog Devices, Inc.), AVB (Avalonbay Communities Inc.), EQR (Equity Residential), LLY (Eli Lilly and Company) and TXN (Texas Instruments Inc.), from Yahoo finance. Here, ADI and TXN belong to high-tech industry, AVB and EQR to real estate industry and LLY to pharmacy industry. The time window spans from 20070113 to 20160116.

Let $w = (w_1, \ldots, w_d)^\top \in \mathbb{R}^d$ denote the long position vector of a d-dimensional portfolio, $S_t = (S_{1,t}, \ldots, S_{d,t})^\top$ stand for the vector of asset prices at time $t \in \{1, \ldots, T\}$ and $X_{i,t} = \log(S_{i,t}/S_{i,t-1})$ for the one period log-return of the i-th asset at time t. Then, $L_t = \sum_{i=1}^{d} w_i X_{i,t}$ denotes the portfolio return. The distribution function of the univariate random variable L_t is denoted by $F_{L_t}(x) = P(L_t \leq x)$ and the Value-at-Risk at level α for the portfolio is defined as the inverse of $F_{L_t}(x)$, namely $\mathrm{VaR}_t(\alpha) = F_{L_t}^{-1}(\alpha)$.

	AVB	EQR	TXN	ADI	AVB	EQR	TXN	ADI	AVB	EQR	TXN	ADI
EQR	0.867				0.686				0.866			
TXN	0.359	0.375			0.260	0.264			0.376	0.381		
ADI	0.384	0.399	0.752		0.277	0.285	0.583		0.398	0.410	0.770	
LLY	0.358	0.370	0.358	0.362	0.268	0.260	0.272	0.270	0.390	0.376	0.393	0.391

Table 2.1: Pairwise dependence measures including Pearson's correlation (left), Kendall's correlation (center) and Spearman's correlation (right).

Copula Performance in Risk Management

From the above formulations can be concluded that the idiosyncratic dependence of the log-return process $\{X_t\}_{t=1}^T$ is crucial for the appropriate estimation of the VaR. To remove temporal dependence from X_t, the single log-return processes are filtered through GARCH$(1,1)$ processes,

$$\begin{aligned} X_{i,t} &= \mu_{i,t} + \sigma_{i,t}\epsilon_{i,t}, & (2.21)\\ \sigma_{i,t}^2 &= a_i + \alpha_i(X_{i,t-1} - \mu_{i,t-1})^2 + \beta_i\sigma_{i,t-1}^2. & (2.22)\end{aligned}$$

The GARCH$(1,1)$-filtered log-returns are illustrated in Figure 2.3. Obviously, assets coming from the same sector have high correlation according to the GARCH residuals. For example, the AVB-EQR and TXN-ADI pairs have strong correlation coming from real estate industry and high technology industry respectively. The strong correlation is also observed in Table 2.1 presenting three dependence measures for pairs of AVB-EQR and TXN-ADI. LLY is from pharmacy industry and shows weak correlation with the other four companies according to the scatter-plots and the contours.

The performance of different copulas utilized for VaR estimation is evaluated via backtesting based on the exceeding ratio

$$\mathrm{ER}^\alpha = (T-w)^{-1}\sum_{t=w}^{T} \mathbf{1}\{l_t < \widehat{\mathrm{VaR}}_t(\alpha)\}, \tag{2.23}$$

where w is the sliding window size and l_t is the realization of L_t. For the twelve copulas, Table 2.2 presents the ERs which optimal if it equals α. The Gaussian copula performs best for $\alpha = 0.05$, the HAC-Clayton copula has reached the most appropriate ER for $\alpha \in \{0.01, 0.005\}$ and the Clayton copula for $\alpha = 0.001$. The Factor-Gumbel copula provides the worst ER values for all values of α. Vines perform neither outstanding good nor bad. It deserves to be mentioned that copulas exhibiting upper-tail dependence show higher ER values, for instance, Joe copula, HAC-Gumbel copula and Factor-Gumbel

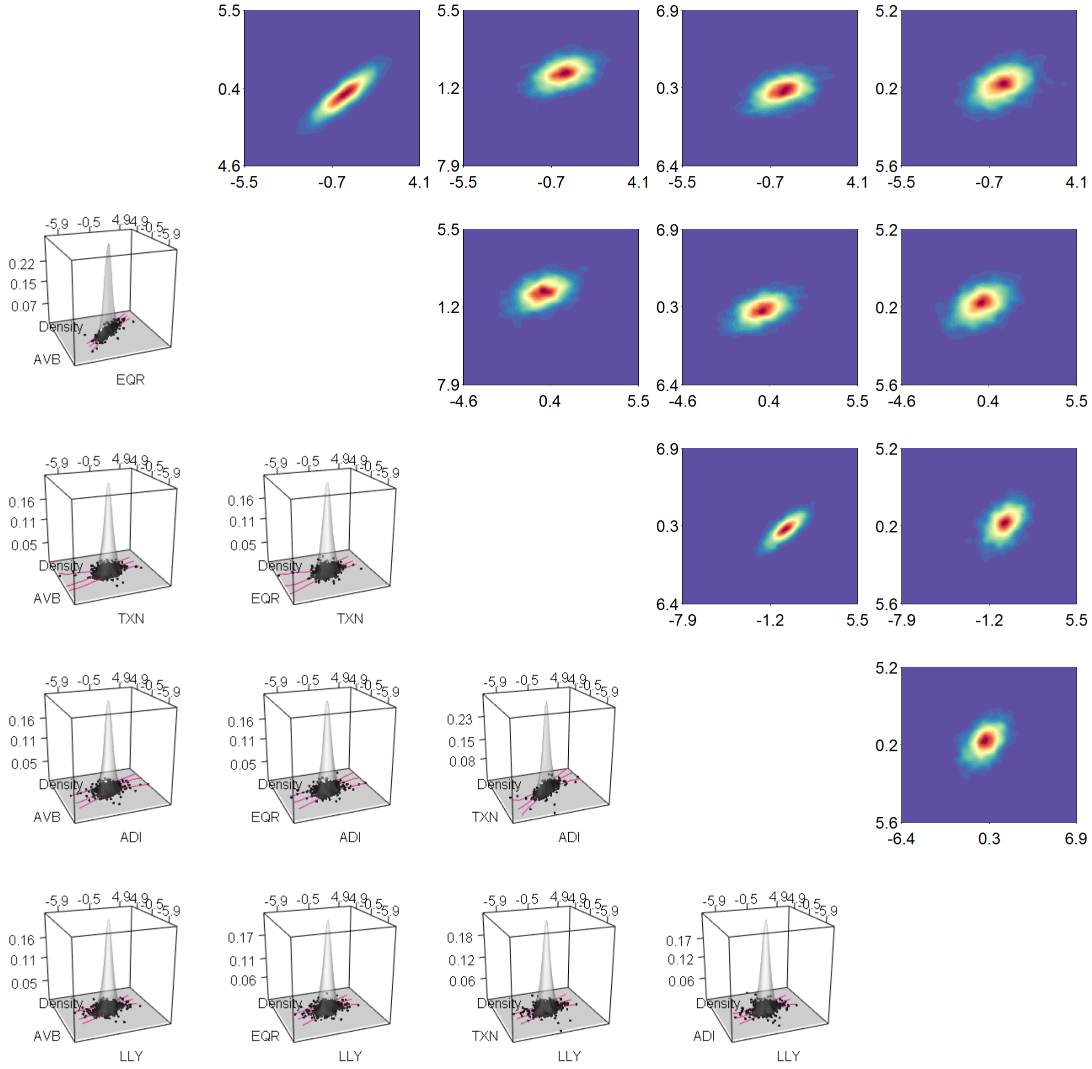


FIGURE 2.3: The lower triangular plots give 2-dimensional kernel density estimations containing scatter plots of pairwise GARCH$(1,1)$-filtered log-returns with quantile regressions under 0.05, 0.5, 0.95 quantiles. The upper triangular plots give pairwise contours of five variables.

copula. Even though some copulas are based on more parameters and thus, offer more flexibility, the increase of parameters does not essentially improve the ER.

2.7 Conclusion

This work discusses bivariate copula and focuses on three high dimensional copula models including the hierarchical Archimedean copula, the factor copula and the vine copula. The three models are developed in-depth with their advantages in modeling high dimensional data for diverse research fields. For the sake of comparison, an empirical study

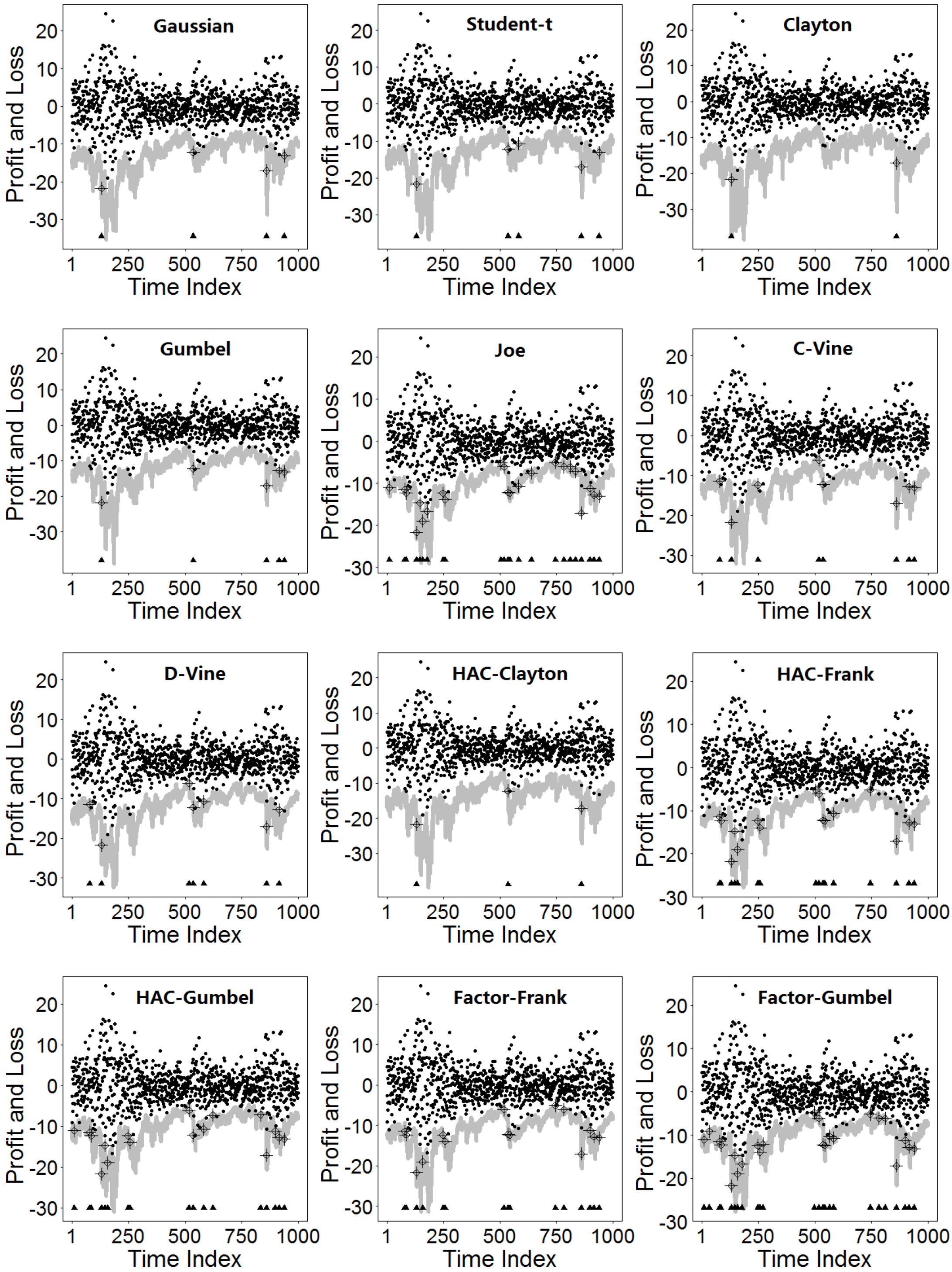


FIGURE 2.4: VaRs for $\alpha = 0.001$ are constructed based on 1000 back-testing points with copulas of Gaussian, t, Clayton, Gumbel, Joe, C-Vine, D-Vine, HAC-Clayton, HAC-Frank, HAC-Gumbel, Factor-Frank, Factor-Gumbel, illustrated by row.

Copula	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.005$	$\alpha = 0.001$
Gaussian	**0.050**	0.018	0.009	0.004
t	0.048	0.014	0.011	0.005
Clayton	0.047	0.017	0.011	**0.002**
Gumbel	0.048	0.025	0.013	0.005
Joe	0.065	0.032	0.030	0.023
C-Vine	0.045	0.019	0.015	0.008
D-Vine	0.044	0.018	0.012	0.007
HAC-Clayton	0.044	**0.013**	**0.008**	0.003
HAC-Frank	0.055	0.033	0.026	0.016
HAC-Gumbel	0.070	0.036	0.028	0.017
Factor-Frank	0.046	0.026	0.017	0.015
Factor-Gumbel	0.086	0.042	0.032	0.024

TABLE 2.2: Exceeding ratios based on $\alpha \in \{0.05,\ 0.01,\ 0.005,\ 0.001\}$.

from risk management is presented. In this study, the estimation of Value-at-Risk is performed under 12 different copula models including the discussed state-of-art copulas as well as some classical benchmarks such as some of the elliptical and Archimedean family. Considered in toto, the hierarchical Archimedean copula with Clayton generator performs better than the alternatives in terms of the exceeding ratios measure.

Chapter 3

Pricing CDX Tranches with Convex Combination of Copulae

This chapter is based on the paper ”A Comparison Study of Pricing Credit Default Swap Index Tranches with Convex Combination of Copulae” by O. Okhrin and Y.F. Xu (2017) published in *The North American Journal of Economics and Finance.*

3.1 Introduction

In recent years, the financial innovation has been accelerated significantly with introduction of many new types of financial vehicles. In credit derivative market new vehicles, for instance, credit default swap index (CDX) has attracted more and more attention. The opportunity and challenge for investors are coexistent in this product. From one perspective, CDX provides credit investors possibility to diversify their credit portfolio's risk in contrary to a single CDS contract. It has a multi-name protection for the credit portfolios by employing a slicing technique termed as *tranche* under a large pool of debtors. From another perspective, the complex mechanism of pricing CDX contract brings investors challenges in the accurate pricing of the product, where one of the core questions is in modeling of the dependence of random default times.

In studies of the CDX pricing, the cynosure is in the dependence modeling of random default times. Since CDX has analogous pricing philosophy to CDO (collateralized debt obligation), therefore literature for CDX pricing can be referred to those for CDO pricing. Firstly proposed in Li (1999) and Li (2000), the Gaussian factor copula model in CDO pricing focuses on modeling the multi-name default times with a high dimensional exchangeable Gaussian copula combined with a transformation of the single-name survival function. Although being simple in dependence modeling, there are a lot of drawbacks in the Gaussian copula, thoroughly discussed in the literature over the last decade. These drawbacks include the destitution of the heterogeneity of dependence between sectors and the asymmetric tail-dependence. This makes the exchangeable Gaussian copula based pricing not accurate.

In order to overcome drawbacks listed above, various new methods have been proposed. These models specified the defaults dependence structure by choosing new copulae possessing partly or whole features such as the heterogeneity of dependence in different sectors and the asymmetrical tail-dependence. In choosing new copulae, literature is abundant, such as the Student-t copula model, see Demarta and McNeil (2005), Schloegl and O'Kane (2005), the double-t copula model in Hull and White (2004), the Clayton copula model in Schönbucher and Schubert (2000); Lindskog and McNeil (2001); Schönbucher (2002), the hierarchical Archimedean copula model in Hofert and Scherer (2011); Hofert (2010); Choroś-Tomczyk et al. (2013), just to name a few.

This paper focuses on the CDX pricing approach based on the convex combination of copulae (cc-copula). Within this project we intend to convexly combine different copulae in order to acquire advantageous properties from component copulae. In the cc-copula models different copula families were convexly combined together so that the merits of different copulae can be utilized together for default dependence modeling. Two empirical studies were conducted in this work. The first empirical study used the data set of the CDX NA IG (Credit Swap Index North America Investment Grade) Series 19 tranches managed by Markit. The CDX NA IG Series 19 containing 125 names dispersed in 5 diverse sectors, was issued on 20120920 and will end on 20171220. The second empirical study employed the data set of the Markit iTraxx Europe Index Series 8 tranches managed by Markit. Similar to the first data set, the Markit iTraxx Europe Index Series 8 containing 125 names dispersed in 5 diverse sectors, covering the period of 20071023-20080701. The main purpose of this paper is to employ the cc-copula models in reproduction of the spreads of CDX tranches to achieve higher accuracy of CDX tranche pricing. We calibrated the parameters in the cc-copula models with numerical optimization, whose objective function is root-mean-square error (RMSE) based on the theoretical spreads and the real market spreads.

This paper is structured as follows. Section 2 introduces the fundamental of copula. Section 3 discusses the CDX structure and the pricing mechanism. Section 4 includes two important empirical studies, where the computation of tranche spread, the parameter calibration and the performance comparison of models are introduced. Section 5 concludes.

3.2 Copula Models

3.2.1 Basics of Copula

Copula is a function which joints marginal distributions into a multivariate distribution and is in essence a multivariate cumulative distribution function with all marginals being uniformly distributed. To construct a multivariate cumulative distribution function is equivalent to separately choose the copula function and the corresponding margins, according to the Sklar's Theorem.

Theorem 3.1. *Sklar's Theorem, c.f. Sklar (1959)*
Every multivariate cumulative distribution function $H(x_1, \dots, x_d) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d)$ of a random vector $(X_1, X_2, \dots, X_d)$ can be expressed in terms of its marginals $F_i(x) = \mathbb{P}(X_i \leq x)$ and a copula C, such that

$$H(x_1, \dots, x_d) \quad = \quad C\left\{F_1(x_1), \dots, F_d(x_d)\right\}. \tag{3.1}$$

If $F_i(\cdot)$ are continuous, then C is unique.

Reader interested in the copula theory is referred to Nelsen (2006) and Joe (2014) and in copula application of finance to Cherubini et al. (2004).

Two elliptical copulae used in this work are Gaussian copula and Student-t copula. The first one is given by,

$$C_{gs}(u_1, \dots, u_d; G) = \Phi_d\left\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d); G\right\}, \ u_k \in [0,1], \ k = 1, \dots, d, \tag{3.2}$$

where G is a $(d \times d)$ correlation matrix, Φ_d a d-dimensional standard Gaussian CDF and Φ a one dimensional standard Gaussian CDF. Gaussian copula is symmetric with zero tail dependence.

Let $\nu \in (1, +\infty)$ be the degree of freedom and $R = (1 - \frac{2}{\nu})\mathrm{Var}(X)$ the $(d \times d)$ correlation matrix, $X = (X_1, \dots, X_d)^\top \in \mathbb{R}^d$. The Student-$t$ copula can be represented as follows,

$$\begin{aligned} &C_t(u_1, \dots, u_d; \nu, \mu, R) \\ &\quad = \int_{-\infty}^{t^{-1}(u_1)} \cdots \int_{-\infty}^{t^{-1}(u_d)} \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^d|R|}} \left\{1 + \frac{(x-\mu)^T R^{-1}(x-\mu)}{\nu}\right\}^{-\frac{\nu+d}{2}} \mathrm{d}x, \\ &\quad = T_d\left\{t^{-1}(u_1; \nu), \dots, t^{-1}(u_d; \nu); \nu, \mu, R\right\}, \end{aligned} \tag{3.3}$$

where T_d is a d-dimensional Student-t CDF and t^{-1} is an inverse of a one dimensional Student-t distribution function. Student-t copula has a non-zero tail dependence.

Archimedean Copula	Representation $C(u_1, \dots, u_d; \theta)$	Generator Function $\varphi^{-1}(t; \theta)$	Parameter θ
Frank	$-\frac{1}{\theta} \log \left[1 + \frac{\prod_{k=1}^{d}\{\exp(-\theta u_k)-1\}}{\{\exp(-\theta)-1\}^{d-1}}\right]$	$-\log \left\{ \frac{\exp(-\theta t)-1}{\exp(\theta)-1} \right\}$	$(-\infty, +\infty)\backslash\{0\}$
Clayton	$\left(\sum_{k=1}^{d} u_k^{-\theta} - d + 1\right)^{-\theta^{-1}}$	$\frac{1}{\theta}(t^{-\theta} - 1)$	$[-1/(d-1), \infty)\backslash\{0\}$
Gumbel	$\exp\left\{-\sum_{k=1}^{d}(-\log u_k)^{\theta}\right\}^{\theta^{-1}}$	$\{-\log(t)\}^{\theta}$	$[1, +\infty)$
Joe	$1 - \left\{\sum_{k=1}^{d}(1-uk)^{\theta} - \prod_{k=1}^{d}(1-u_k)^{\theta}\right\}^{\frac{1}{\theta}}$	$-\log\left\{1 - (1-t)^{\theta}\right\}$	$[1, +\infty)$

Table 3.1: Structures of common Archimedean copulae.

Another important family is the Archimedean copula family, which can be constructed as

$$C_A(u_1, \dots, u_d; \theta) = \begin{cases} \varphi\left\{\varphi^{-1}(u_1;\theta) + \cdots + \varphi^{-1}(u_d;\theta);\theta\right\} & \text{if } \sum_{k=1}^{d} \varphi(u_k;\theta) \le \varphi(0;\theta), \\ 0 & \text{else,} \end{cases} \tag{3.4}$$

where the decreasing function $\varphi\colon [0, +\infty] \to [0, 1]$ is the generator function with $\varphi(0) = 1$ and $\varphi(+\infty) = 0$. Here four most well-known Archimedean copulae were considered, i.e. Frank, Clayton, Gumbel and Joe. Table 3.1 lists the representations, generator functions and parameter spaces of these four common Archimedean copulae.

Frank copula is the only elliptically contoured Archimedean copula owning no tail dependence. Clayton copula has lower tail dependence but no upper tail dependence and this is important for modeling losses. Gumbel copula is the only extreme value copula, and often used in modeling gains. Joe copula has upper tail dependence.

As mentioned above, a simple multivariate Archimedean copula has two weak points. Firstly, it typically uses a single parameter of the generator function $\varphi(\cdot)$ to specify the dependence structure. Secondly, Archimedean copula implies that the distribution of $(U_1, \dots, U_d)^{\top}$ is the same as that of $(U_{i_1}, \dots, U_{i_d})^{\top}$ for all $i_l \neq i_h$, $l, h \in \{1, \dots, d\}$, which is not common in the practice. A much more flexible model is the hierarchical Archimedean copula (HAC), $C(u_1, \dots, u_d; \theta, s)$, where s stands for the HAC's structure, and θ is the set of copula parameters. Details of HAC can be referred to Savu and Trede (2010b), Okhrin et al. (2013a) and Okhrin and Ristig (2014b). A special case of HAC, the d-dimensional fully nested HAC, is shown as follows,

$$\begin{aligned} C_{fnHAC}(u_1, \dots, u_d) &= C[C[\dots C\{C(u_1, u_2; \varphi_1), u_3; \varphi_2\}, \cdots, u_{d-1}; \varphi_{d-2}], u_d; \varphi_{d-1}] \\ &= \varphi_{d-1}[\varphi_{d-1}^{-1}[\varphi_{d-2}[\dots[\varphi_2^{-1}[\varphi_1\{\varphi_1^{-1}(u_1) + \varphi_1^{-1}(u_2)\}] + \varphi_2^{-1}(u_3)] \\ &\quad + \cdots + \varphi_{d-2}^{-1}(u_{d-1})]] + \varphi_{d-1}^{-1}(u_d)]. \end{aligned} \tag{3.5}$$

3.2.2 Convex Combination of Copulae

It is known that a convex combination of distribution functions is again a distribution function, same holds for copulae, see Joe (1996b), thus let

$$C(u_1,\ldots,u_d;\theta_1,\ldots,\theta_I)=\sum_{i=1}^{I}\lambda_i C_i(u_1,\ldots,u_d;\theta_i),\ \sum_{i=1}^{I}\lambda_i=1,\ u_k\in[0,1],\ k=1,\ldots,d, \tag{3.6}$$

where λ_i is the weight parameter of the i-th component copula and I stands for the number of the component copulae in the cc-copula. $C_i(u_1,\ldots,u_d;\theta_i)$ is the i-th component copula with the parameter θ_i. And $C(u_1,\ldots,u_d;\theta_1,\ldots,\theta_I)$ can be thought as a complicated but flexible joint distribution composing known copula functions of $C_i(u_1,\ldots,u_d;\theta_i)$, $i=1,\ldots,I$, hence the convex combined copula $C(u_1,\ldots,u_d;\theta_1,\ldots,\theta_I)$ will inherit features from its component copulae, $C_i(u_1,\ldots,u_d;\theta_i)$, which is practical and reasonable in finance for capturing different joint behaviors such as the heterogeneity of dependence and the asymmetrical tail-dependence.

Example 3.1. *A cc-copula with Clayton and Joe component copulae is given through*

$$C(u_1,u_2;\theta_1,\theta_2)=\lambda C_{Clayton}(u_1,u_2;\theta_1)+(1-\lambda)C_{Joe}(u_1,u_2;\theta_2). \tag{3.7}$$

Example 3.1 gives a cc-copula with Clayton copula and Joe copula as components. Figure 3.1 illustrates this example. In this copula there are three parameters, i.e. θ_1,θ_2 used for the dependence structure in Clayton copula and Joe copula separately. The third parameter, λ, is used for the convex combination of the two components, which can control the attributes inheriting from the both component copulae. For instance in Figure 3.1, both copula structure parameters, θ_1,θ_2, were given as known constants, $\theta_1=\theta_2=0.7$. And the ten weight parameter were set that $\lambda\in\{0.1,0.2,\ldots,0.9,1.0\}$. It is clear that when λ is small, say 0.1, then the Joe copula owns a large weight in the cc-copula. This implies that the upper triangular panels contain figures with more observations accumulated in the upper tail area. This means that this cc-copula is an upper tail dependence characterized copula. Analogously, when λ is large then, say $\lambda=0.9$, then the Clayton copula will own larger weight, hence the cc-copula will have the lower tail dependence structure, which can be advocated by the contour plot in the first upper triangular panels.

Therefore, competing against the classical elliptical copula (zero-tail dependence, see Gaussian copula) and the common Archimdean copula (only upper-tail dependence or lower-tail dependence, see Gumbel copula, Joe copula, Clayton copula), the cc-copula

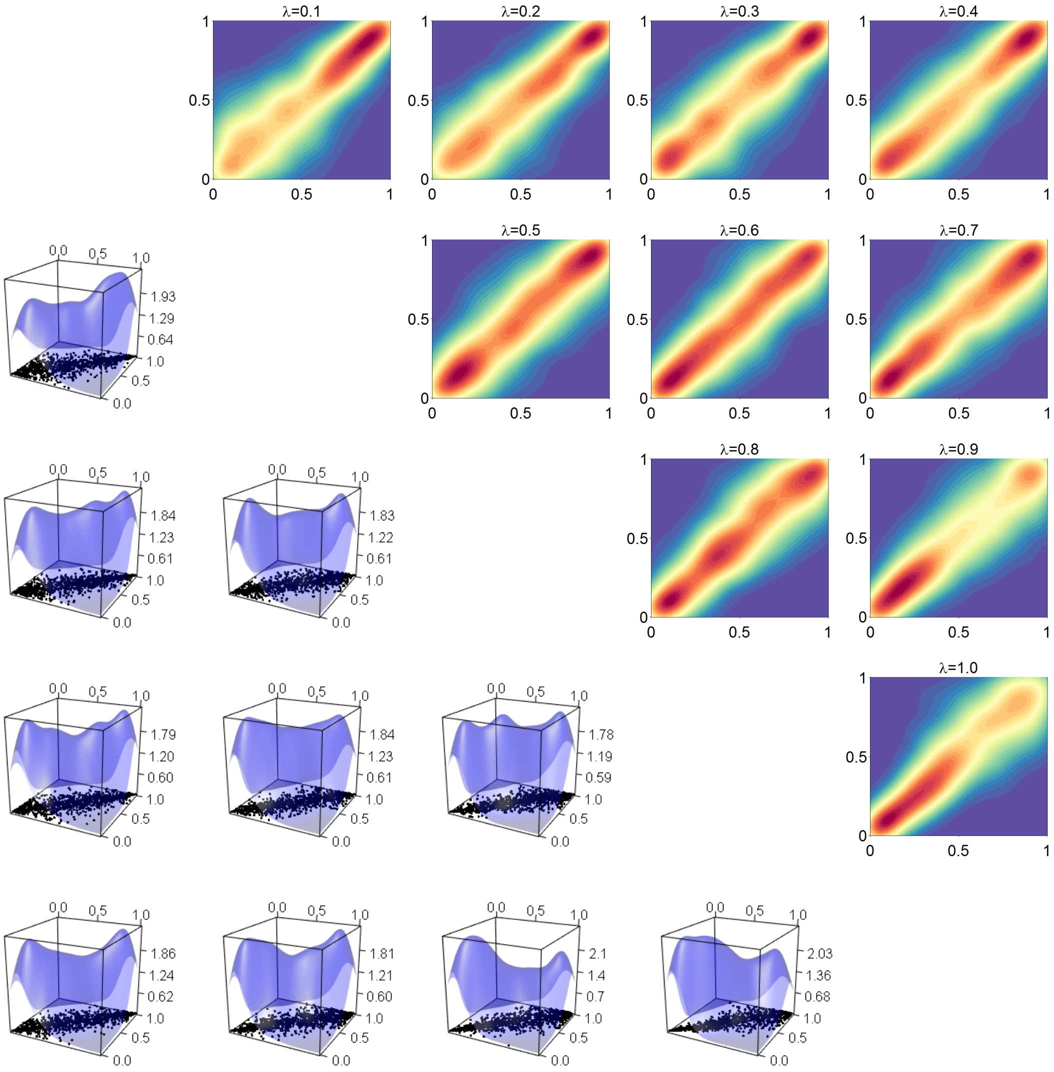


FIGURE 3.1: The lower triangular graphs illustrate two dimensional kernel density estimations containing scatter plots of (U_1, U_2). The scatter points were obtained from 1000 simulations of the cc-copula of Clayton-Joe with Kendall's $\tau = 0.7$ for the both component copulae and $\lambda \in \{0.1, 0.2, \ldots, 0.9, 1\}$, i.e. $C(u_1, u_2; \theta_1, \theta_2) = \lambda C_{Clayton}(u_1, u_2; \theta_1) + (1 - \lambda) C_{Joe}(u_1, u_2; \theta_2)$. The upper triangular panels introduce the corresponding contours of the scatter points under 10 λs.

model with its adaptivity and flexibility in inheriting of assets from different component copulae, has a promising application future.

3.3 Credit Default Swap Index

The CDX is a structured credit derivative which can be used to protect against default of the multi-name credit. The portfolio's default risk is divided into slices using the tranche technique, which slices the risk into different hierarchies with a ranking. The CDX issuer is the protection buyer which pays a fixed premium periodically and receives payment for the contingent loss of the credit portfolio. The CDX investor is the protection seller who receives the premium payments from the CDX issuer and takes responsibility to cover the issuer's contingent loss of the credit portfolio.

The tranche technique uses attachment points and detachment points to define hierarchies of the product, which gives the loss percentages of the credit portfolio. The sliced hierarchy is also termed as the tranche. In CDX NA IG product, four attachment points are $l_q = (0, 0.03, 0.07, 0.15)^\top$, thus the corresponding detachment points are $u_q = (0.03, 0.07, 0.15, 1)^\top$. When contingent loss happens between an attachment point and a detachment point of a hierarchy then the notional will be decreased and the periodic payments for the portfolio protection buyer will be reduced either. When contingent loss increases over the detachment point of a hierarchy, then the protection seller pays no premium any more and the protection buyer covers the corresponding losses.

3.3.1 CDX Pricing

Firstly, let a credit portfolio containing d reference entities with overall N notional principal being equally distributed on entities, i.e. every entity shares $1/d$ of the overall investment. In the meanwhile let the maturity of the CDS index tranches be T, i.e. the length of the contract duration, and premiums are payed at points $t_j,\ j = 1, \ldots, J$ and it was set $t_0 = 0$. In the practice, credit events can occur at any point of the interval $[0, t_J],\ t_J = T$. For simplicity we assumed that the default occurred in the midpoint of the two premium payment dates, i.e. $(t_j + t_{j+1})/2$, see Choroś-Tomczyk et al. (2013). Then let the random variable $\tau_k, k = 1, \ldots, d$, be the default time of the k-th entity standing for the survival length and r be the constant recovery rate.

The portfolio loss process L_{t_j} is given through

$$L_{t_j} = \frac{1}{d} \sum_{k=1}^{d} (1 - r) \mathbf{1}_{\{\tau_k \leq t_j\}},\ j = 0, \ldots, J, \tag{3.8}$$

where the indicator function $\mathbf{1}_{\{\cdot\}}$ stands for the default indication of the k-th entity. Let $q = 1, \ldots, Q$ be the index of the q-th tranche and L_{q,t_j} the tranche loss of the q-th tranche at t_j. As the tranche loss is a function of the portfolio loss process, the q-th tranche loss is given as follows,

$$L_{q,t_j} = \min\{\max\{L_{t_j} - l_q, 0\}, u_q - l_q\},\ j = 1, \ldots, J,\ q = 1, \ldots, Q. \tag{3.9}$$

In the run of a CDX tranche, if credit events of underlying entities occur then the premium to be paid in the next period needs to be adjusted according to the outstanding notional P_{q,t_j}

$$P_{q,t_j} = u_q - l_q - L_{q,t_j}. \tag{3.10}$$

Under the non-arbitrage assumption the expectation of the accumulative payments generated by the protection buyer and seller should be equal. In the CDX pricing study two terminologies for these two expectations were used, the default leg DL_q which represents the expectation of the aggregated compensation payments from the protection seller side, and the premium leg PL_q which stands for the expectation of the aggregated premium payments from the protection buyer side. The default leg DL_q is thus formalized as follows,

$$DL_q \quad = \quad \mathbb{E}\left\{\sum_{j=1}^{J} \beta_{t_j} N(L_{q,t_j} - L_{q,t_{j-1}})\right\},\ q = 1, \ldots, Q, \tag{3.11}$$

where β_{t_j} is the discount function dependent on the survival length at each payment point.

As in the market practice the protection buyer of a tranche needs to pay upfront payment for every tranche based on the quotation convention of the CDX NA IG Series 19 and iTraxx Europe Index Series 8, therefore the premium legs for diverse tranches equal to

$$PL_q \quad = \quad \mathbb{E}\left\{(u_q - l_q)NS_q^{CDX} - \sum_{j=1}^{J} B_q\beta_{t_j}(t_j - t_{j-1})N(P_{q,t_j} + P_{q,t_{j-1}})/2\right\}, \tag{3.12}$$

where in (3.12) S_q^{CDX} is the upfront payment rate.

According to the non-arbitrage assumption, the default leg (3.11) should equals the premium leg (3.12), leading to

$$PL_q \quad = \quad DL_q, \tag{3.13}$$

then plugging (3.12) and (3.11) into (3.13) one obtains

$$\mathbb{E}\left\{(u_q - l_q)NS_q^{CDX} - \sum_{j=1}^{J} B_q\beta_{t_j}(t_j - t_{j-1})N(P_{q,t_j} + P_{q,t_{j-1}})/2\right\}$$
$$= \mathbb{E}\left\{\sum_{j=1}^{J} \beta_{t_j} N(L_{q,t_j} - L_{q,t_{j-1}})\right\}.$$

Hence the q-th CDS index tranche upfront payment rate S_q^{CDX} can be extracted as follows,

$$(u_q - l_q)S_q^{CDX}N + \mathbb{E}\left\{\sum_{j=1}^{J} 0.5B_q\beta_{t_j}(t_j - t_{j-1})N(P_{q,t_j} + P_{q,t_{j-1}})\right\} \quad = \quad DL_q,\tag{3.14}$$

therefore the S_q^{CDX} is obtained from (3.14) as follows,

$$S_q^{CDX} \quad = \quad \mathbb{E}\left[\frac{\sum_{j=1}^{J}\beta_{t_j}\{(L_{q,t_j} - L_{q,t_{j-1}}) - 0.5B_q(t_j - t_{j-1})(P_{q,t_j} + P_{q,t_{j-1}})\}}{u_q - l_q}\right]\tag{3.15}$$

3.3.2 Modeling of Joint Defaults

As mentioned at the beginning, $\tau_k, k = 1, \dots, d$ is the random variable of survival length (or termed as the default time) of the k-th entity in the reference pool, then let F_k be denoted as the CDF of τ_k and $S_k(t)$ as a survival function. The marginal defaults are assumed to follow homogeneous Poisson process with intensity h, therefore survival times till default has a distribution function of the form

$$F_k(t) = 1 - \exp\left(-ht\right).\tag{3.16}$$

Next the copula function is employed for modeling the joint behavior of default times, $(\tau_1, \dots, \tau_d)^\top$.

As in (3.16), $\exp\left(-h\tau_k\right)$ is uniformly distributed over $[0, 1]$, thus let $U_k = \exp(-h\tau_k),\ k = 1, \dots, d$. The joint CDF of $(U_1, \dots, U_d)^\top$ is represented as

$$\mathbb{P}(U_1 \leq u_1, \dots, U_d \leq u_d) \quad = \quad C(u_1, \dots, u_d).$$

Samples of $(U_1, \dots, U_d)^\top$ are obtained from the copula function $C(u_1, \dots, u_d)$, and using the fact that $U_k = \exp(-h\tau_k),\ k = 1, \dots, d$ one can obtain

$$(\tau_1, \dots, \tau_d)^\top \quad = \quad \left(\frac{-\log U_1}{h}, \dots, \frac{-\log U_d}{h}\right)^\top.\tag{3.17}$$

By using (3.11), (3.12) and (3.13) the expectation of $\mathbb{E}(L_{q,t_j}),\ q = 1, \dots, Q$ and $j = 1, \dots, J$, is estimated through

$$\hat{\mathbb{E}}(L_{q,t_j}) \quad = \quad \frac{1}{M}\sum_{m=1}^{M}\left(\min\left[\max\left\{\frac{1}{d}\sum_{k=1}^{d}(1-r)\mathbf{1}_{\{\mathfrak{z}_k^m \leq t_j\}} - l_q, 0\right\}, u_q - l_q\right]\right),\tag{3.18}$$

where $(\mathfrak{z}_1^m, \dots, \mathfrak{z}_d^m)^\top$ is the m-th Monte Carlo sample of the default times $(\tau_1, \dots, \tau_d)^\top$. Therefore at last the empirical representations for spreads of CDS index tranches (upfront rate version) is obtained with the following formula.

$$\widehat{S}_q^{CDX} = \widehat{\mathbb{E}}\left[\frac{\sum_{j=1}^{J}\beta_{t_j}\{(L_{q,t_j} - L_{q,t_{j-1}}) - 0.5B_q(t_j - t_{j-1})(P_{q,t_j} + P_{q,t_{j-1}})\}}{u_q - l_q}\right]. \tag{3.19}$$

3.4 Two Empirical Studies

3.4.1 Data Set of Empirical Study 1

In the first empirical study, the data of CDX NA IG index was employed. The CDX NA IG index based tranche has four different maturity structures (3, 5, 7 and 10 years) and its underlying entity pool contains overall $d = 125$ CDS contracts. In this paper the maturity with 5 years of the CDX NA IG Series 19 was used, which was issued on 20120920 and ends on 20171220. And the pricing for all $Q = 4$ CDS index tranches was computed with 10 randomly chosen evaluation date points (20140601, 20140703, 20140815, 20140923, 20141011, 20141117, 20141201, 20150107, 20150210, 20150315). In the pricing it was assumed that the risk-free rate as 0.0014 (consistent with the mean of LIBOR of the ten dates) and recovery rate as 0.40 being consistent with its usage in Markit company which administrates the CDX NA IG index, see Markit$^{\text{TM}}$ (2008). The illustration of the spreads of the four tranches and the corresponding CDS is given in Figure 3.9 and the data set is given in Table 3.2.

3.4.2 Data Set of Empirical Study 2

In the second empirical study the Markit iTraxx Europe Index Series 8 from the Bloomberg Terminal was employed. The iTraxx Europe index based tranche has four different maturity structures, 3, 5, 7 and 10 years and its underlying pool contains overall $d = 125$ CDS contracts. Every six months the underlying pool is updated for eliminating the already default entities. In this paper the maturity with 5 years of the iTraxx Europe Series 8 was chosen, which was issued on 20070920 and ended on 20120920, whose running period covers the financial crisis which was thought that CDOs (collateralized debt obligations) were important triggers. And the pricing was conducted for all $Q = 5$ CDS index tranches with 12 randomly chosen evaluation dates on 20071023, 20071102, 20071109, 20071206, 20080111, 20080204, 20080222, 20080318, 20080404, 20080407, 20080530, 20080701. The historical data of $Q = 5$ CDS index tranches on these 12 pricing dates is given in Figure 3.3 (a) and 3.3 (b). In the pricing it was assumed the

Date	0-3%	3-7%	7-15%	15-100%	CDS
2014/06/01	4.250	2.000	0.036	0.014	39
2014/070/3	3.750	1.375	0.048	0.015	37
2014/08/15	4.094	1.719	0.050	0.014	38
2014/09/23	3.750	1.375	0.056	0.012	37
2014/10/11	5.775	1.810	0.050	0.012	41
2014/11/17	4.188	0.985	0.057	0.015	35
2014/12/01	3.183	0.747	0.060	0.016	32
2015/01/07	7.065	0.875	0.055	0.013	39
2015/02/10	7.559	0.563	0.055	0.014	37
2015/03/15	6.874	0.073	0.064	0.015	34

Table 3.2: Spreads of four tranches of the CDX NA IG Series 19 and the corresponding CDS spreads.

risk-free rate as 0.03 and recovery rate as 0.40 which is consistent with it was used in Markit company which administrates the Markit iTraxx Europe Index.

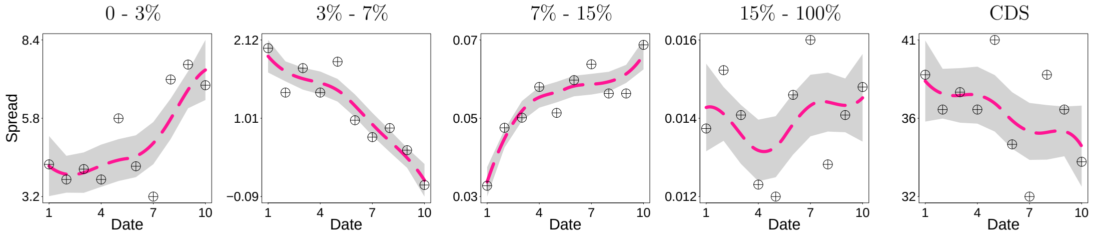
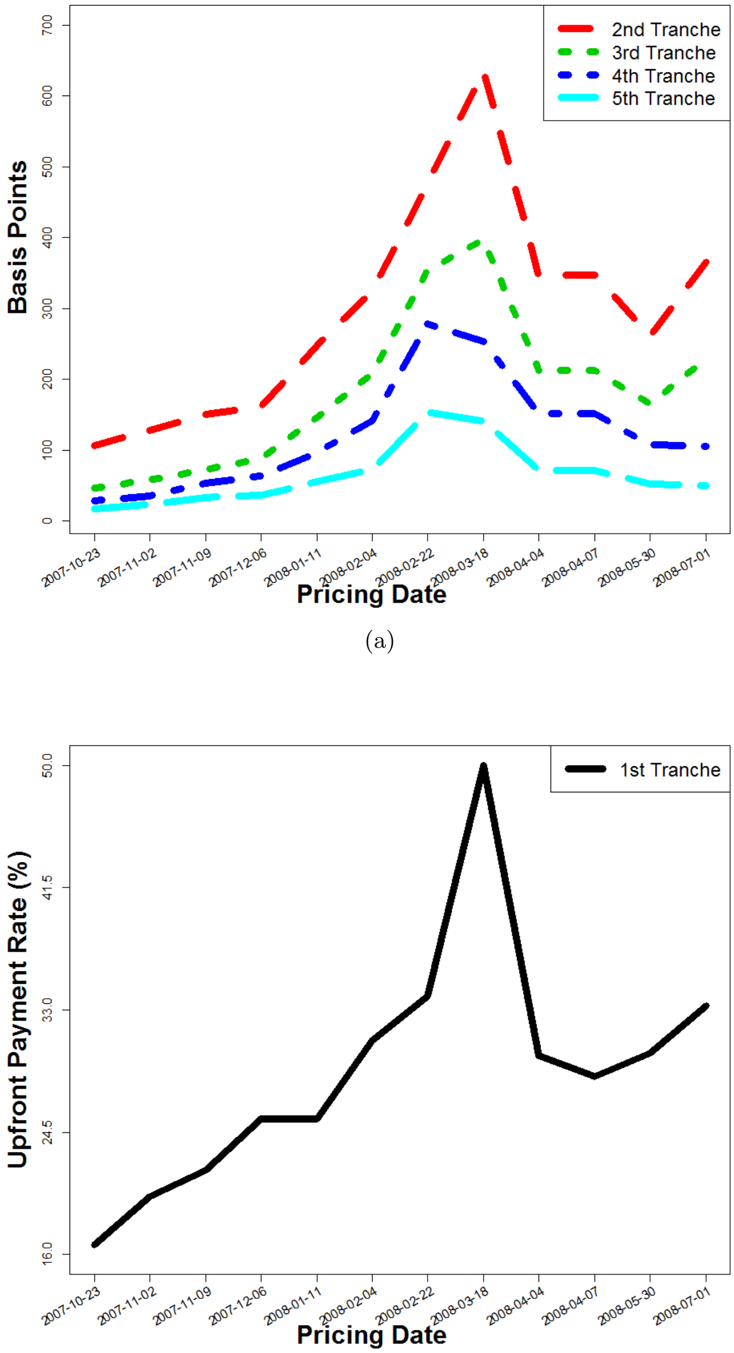


Figure 3.2: Spreads of four tranches of the CDX NA IG Series 19 and the corresponding CDS spreads are illustrated with scatter points, at ten dates 20140601, 20140703, 20140815, 20140923, 20141011, 20141117, 20141201, 20150107, 20150210, 20150315. The dashed line gives a local polynomial regression with its confidence boundaries constraining the gray shading area.

3.4.3 Employed Models

Overall 43 copula models used in the study are described below. In the following the notations are set as ga: Gaussian; t: Student-t; fr: Frank; cl: Clayton; gu: Gumbel; jo: Joe; gai, $i = 1, \dots, 6$: Gaussian with the correlation matrix R_{gai}, $i = 1, \dots, 6$;



(a)

(b)

FIGURE 3.3: Tranche spreads at 12 pricing dates of Markit iTraxx Europe Index Series 8. (*a*) Tranche spreads for four tranches ($q = 2, 3, 4, 5$) of Markit iTraxx Europe Index Series 8 from 20071023 to 20080701 by the Bloomberg Terminal. (*b*) Tranche spreads for equity tranche of Markit iTraxx Europe Index Series 8 from 20071023 to 20080701 by the Bloomberg Terminal.

$tj,\ j = 1, \ldots, 6$: Student-t with the same correlation matrix structure as R_{gai}; ng: HAC with the Gumbel generator function.

From the elliptical family of copulae an exchangeable Gaussian copula and an exchangeable Student-t copula were chosen in Model 1 and 2.

Model 1. Gaussian copula,

$$C(u_1, \ldots, u_d; \theta) \quad = \quad C_{ga}(u_1, \ldots, u_d; R_{ga}), \tag{3.20}$$

where R_{ga} is the correlation matrix with equal correlation in off-diagonal elements.

Model 2. Student-t copula,

$$C(u_1, \ldots, u_d; \theta) \quad = \quad C_t(u_1, \ldots, u_d; R_t, \nu). \tag{3.21}$$

where R_t is the correlation matrix with equal correlation in off-diagonal elements.

For Gaussian copulae with diverse dependence structures are given in Model 3 to Model 8.

Model 3. Gaussian copula with sectoral dependence illustrated in Figure 3.4 (a),

$$C(u_1, \ldots, u_d; \theta) \quad = \quad C_{ga1}(u_1, \ldots, u_d; R_{ga1}). \tag{3.22}$$

Here two parameters were used, ρ_2 for controlling the dependence within a sector and ρ_1 to specify the dependence between sectors. The correlation matrix of Model 3 is given in Figure 3.4 (a).

Model 4. Gaussian copula with sectoral dependence as in Figure 3.4 (b),

$$C(u_1, \ldots, u_d, u_{d+1}; \theta) \quad = \quad C_{ga2}(u_1, \ldots, u_d, u_{d+1}; R_{ga2}). \tag{3.23}$$

It was set that the random recovery U_{d+1} shown in (3.23) is uniformly distributed. The parameter ρ_1 is the unique parameter for the dependence structure as given in Figure 3.4 (b).

Model 5. Gaussian copula with sectoral dependence in Figure 3.4 (c),

$$C(u_1, \ldots, u_d, u_{d+1}; \theta) \quad = \quad C_{ga3}(u_1, \ldots, u_d, u_{d+1}; R_{ga3}). \tag{3.24}$$

This model is a generalization of Model 4 that let the parameter ρ_2 specify the dependence within and between sectors. U_{d+1} is a random recovery as in latter model.

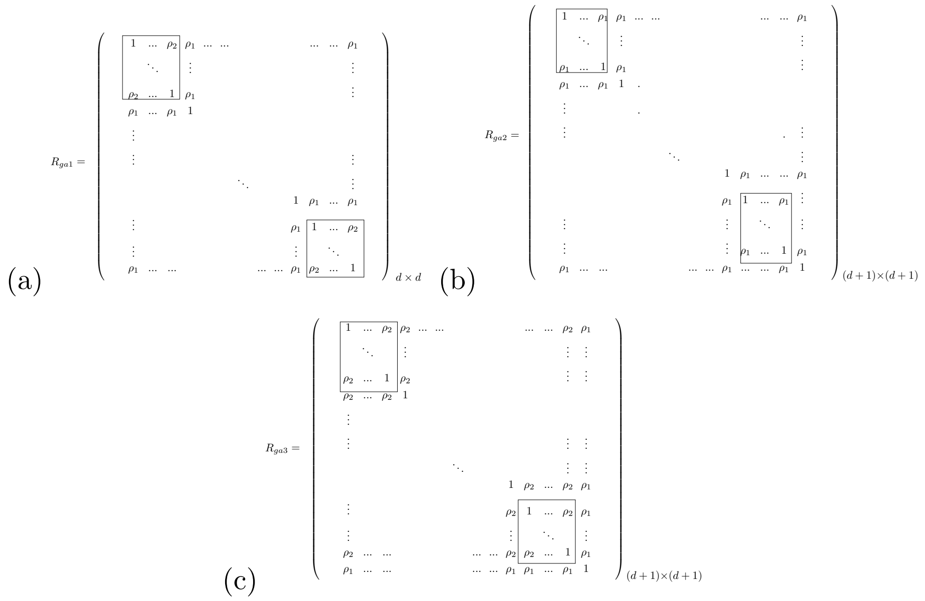


FIGURE 3.4: The structure of the correlation matrix (a) R_{ga1} was utilized in Model 3 and Model 9. And the structure of the correlation matrix (b) R_{ga2} was utilized in Model 4 and Model 10. The structure of the correlation matrix (c) R_{ga3} was utilized in Model 5 and Model 11.

Parameter ρ_1 controls the dependence structure between U_{d+1} and $(U_1, \dots, U_d)^\top$. The corresponding correlation matrix is illustrated in Figure 3.4 (c).

Model 6. Gaussian copula with sectoral dependence as in Figure 3.5 (a),

$$C(u_1, \dots, u_d, u_{d+1}, \dots, u_{d+5}; \theta) \quad = \quad C_{ga4}(u_1, \dots, u_d, u_{d+1}, \dots, u_{d+5}; R_{ga4}). \quad (3.25)$$

As diverse sectors may have heterogeneous recovery rates, therefore Model 6 let $(U_{d+1}, \dots, U_{d+5})$ be six different uniformly distributed random recovery rates for each vector separately. Figure 3.5 (a) presents the correlation matrix for Model 6 where the parameter ρ_2 is responsible for within sector dependence and the parameter ρ_1 for between sectors dependence.

Model 7. Gaussian copula with sectoral dependence as in Figure 3.5 (b),

$$C(u_1, \dots, u_d, u_{d+1}, \dots, u_{d+6}; \theta) \quad = \quad C_{ga5}(u_1, \dots, u_d, u_{d+1}, \dots, u_{d+5}; R_{ga5}). \quad (3.26)$$

Model 7 still keeps the six heterogeneous recovery rates setting but it was modified that the parameter ρ_3 was used to specify the dependence structure within sectors and the parameter ρ_2 to control the dependence between $U_s,\ s = d+1, \dots, d+5$ and 5 different sectors. At last the parameter ρ_1 was used to specify the dependence between blocks as described in Figure 3.5 (b).

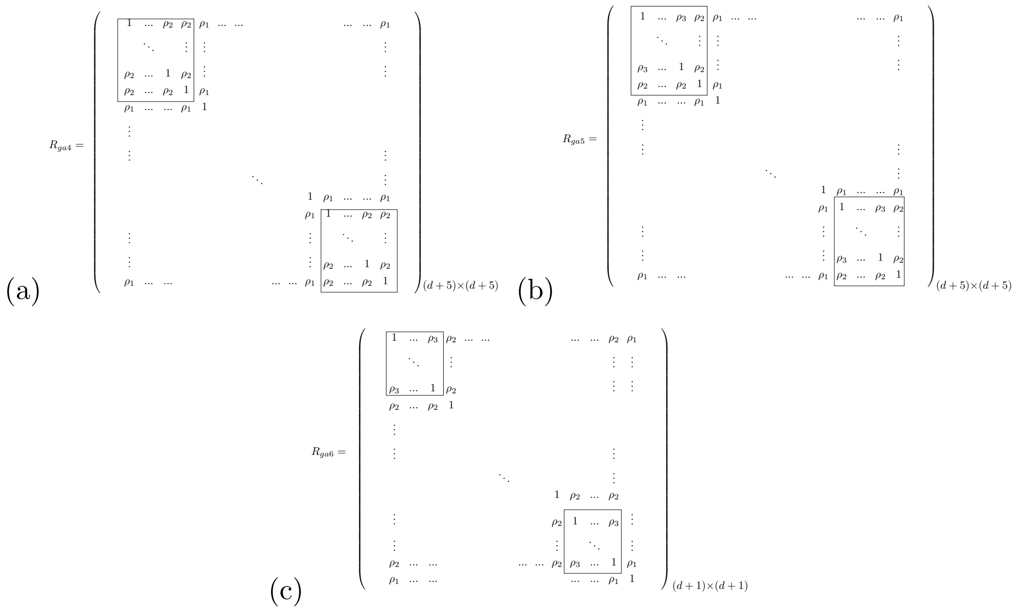


FIGURE 3.5: the structure of the correlation matrix (a) R_{ga4} was utilized in Model 6 and Model 12. The structure of the correlation matrix (b) R_{ga5} was utilized in Model 7 and Model 13. And the structure of the correlation matrix (c) R_{ga6} was utilized in Model 8 and Model 14.

Model 8. Gaussian copula with sectoral dependence as in Figure 3.5 (c),

$$C(u_1,\ldots,u_d,u_{d+1};\theta) \quad = \quad C_{ga6}(u_1,\ldots,u_d,u_{d+1};R_{ga6}). \tag{3.27}$$

This model still uses 3 parameters to specify the dependence structure of $(U_1,\ldots, U_d, U_{d+1})^\top$. For the within-sector dependence, the parameter ρ_3 and the parameter ρ_2 were used to control the between-sector dependence. At last the parameter ρ_1 was used for the dependence between U_{d+1}, which stands for the single random recovery rate, and $(U_1,\ldots,U_d)^\top$.

As the Gaussian copula has zero tail-dependence, therefore another member of elliptical copula with the fat tail-dependence feature, the Student-t copula, was considered. Models 9-14 are the Student-t copulae, denoted by C_{t1}, C_{t2}, C_{t3}, C_{t4}, C_{t5}, C_{t6}, with the same correlation matrix structures shown in Figures 3.4 and 3.5. Ten different degrees of freedom were obtained by calibration for the Student-t copula of Model 2. Then these ten calibrated parameters were plugged into Models 9-14 as the known parameters. The cc-copula models with the Student-t copula as component copula used a fixed parameter of degree of freedom equal to 3.

After models were constructed by elliptical family of copula, in the following the Archimedean copula based models are given. As introduced before the Archimedean copula members share different tail-dependence structures. Model 15 to Model 18 are four diverse

Archimedean copula models represented as follows,

$$C(u_1,\dots,u_d;\theta_a) \quad = \quad C_a(u_1,\dots,u_d;\theta_a), \tag{3.28}$$

where $a = cl, jo, gu, fr$, standing separately for Clayton, Joe, Gumbel and Frank copula.

Model 19 to Model 22 are HAC copulae used in the empirical study.

Model 19. Gumbel HAC,

$$C(u_1,\dots,u_d,u_{d+1};\theta) \quad = \quad C^1_{ng2}\{C^2_{ng2}(u_1,\dots,u_d;\rho_{\mathcal{K}2}),u_{d+1};\rho_{\mathcal{K}1}\}, \tag{3.29}$$

where C^1_{ng2} is the root copula and C^2_{ng2} is the child copula. Model 19 is a Gumbel HAC copula with one parameter $\rho_{\mathcal{K}1}$ for dependence between sectors and random recovery rate U_{d+1}. And $\rho_{\mathcal{K}2}$ is used for dependence of d entities.

Model 20. Gumbel HAC,

$$\begin{aligned} C(u_1,\dots,u_d;\theta) \quad = \quad & C^1_{ng3}\,\{ \\ & C^2_{ng3}\,(u_1,\dots,u_{s_1};\rho_{\mathcal{K}2})\,, \\ & C^2_{ng3}\,(u_{s_1+1},\dots,u_{s_1+s_2};\rho_{\mathcal{K}2})\,,\dots, \\ & C^2_{ng3}\,(u_{s_1+\dots+s_5+1},\dots,u_d;\rho_{\mathcal{K}2})\,;\rho_{\mathcal{K}1}\}, \end{aligned} \tag{3.30}$$

where $s_i,\ i = 1,\dots,5$ is the number of entities in the i-th sector, C^1_{ng3} means the root copula in the HAC with a Gumbel generator function and C^2_{ng3} means the child copula in this model. Model 20 is a HAC without random recovery using a root copula and 5 child copulae. The parameter $\rho_{\mathcal{K}2}$ is for dependence within a sector and $\rho_{\mathcal{K}1}$ for dependence between sectors.

Model 21. Gumbel HAC,

$$\begin{aligned} C(u_1,\dots,u_d,u_{d+1},\dots,u_{d+5};\theta) \quad = \quad & C^1_{ng4}\,\{ \\ & C^2_{ng4}\,(u_1,\dots,u_{s_1},u_{d+1};\rho_{\mathcal{K}2})\,, \\ & C^2_{ng4}\,(u_{s_1+1},\dots,u_{s_1+s_2},u_{d+2};\rho_{\mathcal{K}2})\,,\dots, \\ & C^2_{ng4}\,(u_{s_1+\dots+s_5+1},\dots,u_d,u_{d+5};\rho_{\mathcal{K}2})\,;\rho_{\mathcal{K}1}\}. \end{aligned}$$

This model has five random recoveries, i.e. for each sector a single random recovery following uniform distribution.

Model 22. Gumbel HAC,

$$\begin{aligned} C(u_1,\ldots,u_d,u_{d+1},\ldots,u_{d+5};\theta) \;=\; & C^1_{ng5}\,[\\ & C^2_{ng5}\left\{u_{d+1}, C^3_{ng5}\left(u_1,\ldots,u_{s_1};\rho_{\mathcal{K}3}\right);\rho_{\mathcal{K}2}\right\}, \\ & C^2_{ng5}\left\{u_{d+2}, C^3_{ng5}\left(u_{s_1+1},\ldots,u_{s_1+s_2};\rho_{\mathcal{K}3}\right);\rho_{\mathcal{K}2}\right\},\ldots, \\ & C^2_{ng5}\left\{u_{d+5}, C^3_{ng5}\left(u_{s_1+\cdots+s_5+1},\ldots,u_d;\rho_{\mathcal{K}3}\right);\rho_{\mathcal{K}2}\right\};\rho_{\mathcal{K}1}]. \end{aligned}$$

Model 22 is a HAC model with a Gumbel generator function using five random recoveries, $(U_{d+1},\ldots,U_{d+5})^\top$, and three dependence parameters. $\rho_{\mathcal{K}3}$ was utilized for within-sector dependence, i.e. all five sectors share the same dependence parameter in every sector. $\rho_{\mathcal{K}2}$ was employed for dependence between the i-th random recovery and the i-th sector, where $i = 1,\ldots,5$. The parameter $\rho_{\mathcal{K}1}$ controls the dependence between the second layer child copulae.

Next the cc-copula models from Model 23 to Model 43 are given. In a cc-copula, six copulae were employed as the component copulae containing the exchangeable Gaussian copula, the Student-t copula with degree of freedom equal to 3, the Frank copula, the Clayton copula, the Gumbel copula and the Joe copula. It was set λ, $\lambda \in [0,1]$ as the weight for the component copulae, then a general formula for cc-copula models with two components to be used can be given as follows,

$$\begin{aligned} & C_{comp1-comp2}(u_1,\ldots,u_d;\theta) \\ = \; & \lambda C_{comp1}(u_1,\ldots,u_d;\theta_1) + (1-\lambda)C_{comp2}(u_1,\ldots,u_d;\theta_2), \end{aligned} \tag{3.31}$$

where the $comp1$, $comp2 \in \{ga, t, fr, cl, gu, jo\}$ and parameters θ_1 and θ_2 belong correspondingly to the component copula 1 and 2. An example of a cc-copula is given as follows,

Model 23. cc-copula with two Gaussian components,

$$C_{ga-ga}(u_1,\ldots,u_d;\theta) \;=\; \lambda C_{ga}(u_1,\ldots,u_d;\theta_1) + (1-\lambda)C_{ga}(u_1,\ldots,u_d;\theta_2). \tag{3.32}$$

According to the convention in (3.31), C_{ga-ga} in Model 23 means that this model is constructed by two Gaussian (ga) copulae. All the 43 copula models used in this paper are listed in the Table 3.3.

3.4.4 Parameter Calibration

HAC, Archimedean copulae, elliptical copulae and cc-copula have been introduced, which can be applied in CDS index tranche pricing by using the copula to construct

Model	Notation	Model	Notation	Model	Notation	Model	Notation
1	C_{ga}	12	C_{t4}	23	C_{ga-ga}	34	C_{fr-fr}
2	C_t	13	C_{t5}	24	C_{ga-t}	35	C_{fr-cl}
3	C_{ga1}	14	C_{t6}	25	C_{ga-fr}	36	C_{fr-gu}
4	C_{ga2}	15	C_{fr}	26	C_{ga-cl}	37	C_{fr-jo}
5	C_{ga3}	16	C_{cl}	27	C_{ga-gu}	38	C_{cl-cl}
6	C_{ga4}	17	C_{gu}	28	C_{ga-jo}	39	C_{cl-gu}
7	C_{ga5}	18	C_{jo}	29	C_{t-t}	40	C_{cl-jo}
8	C_{ga6}	19	C_{ng2}	30	C_{t-fr}	41	C_{gu-gu}
9	C_{t1}	20	C_{ng3}	31	C_{t-cl}	42	C_{gu-jo}
10	C_{t2}	21	C_{ng4}	32	C_{t-gu}	43	C_{jo-jo}
11	C_{t3}	22	C_{ng5}	33	C_{t-jo}		

Table 3.3: Abbreviations: ga: Gaussian, t: Student-t, fr: Frank, cl: Clayton, gu: Gumbel, jo: Joe, gai, $i = 1, \dots, 6$: Gaussian with the correlation matrix R_{gai}, $i = 1, \dots, 6$, tj, $j = 1, \dots, 6$: Student-t with the same correlation matrix structure as R_{gai}, $i = 1, \dots, 6$, ng: HAC with the Gumbel generator function.

the dependence structure of default times $(\tau_1, \dots, \tau_d)^\top$. In this work it was assumed the hazard function as a constant scalar h and this quantity is implied from the market spreads of the CDX contract. For a detailed method of implication of h it is referred to Hofert and Scherer (2011).

The exact computation of tranche prices can be performed by the following algorithm.

Algorithm:

1. Choose a copula model C listed in the Table 3.3.
2. Sample by $M = 10^4$ runs of Monte Carlo simulation according to $(U_1, \dots, U_d)^\top \sim C$.
3. Obtain samples of $(u_{m,1}, \dots, u_{m,d})^\top$, $m = 1, \dots, M$.
4. Compute (3.11) to (3.12) using samples obtained from the step 3.

For models embedded with one random recovery such as (3.23), (3.24), (3.27), (3.29), and with five random recoveries such as (3.25), (3.26), (3.31), (3.31) one needs to obtain samples respectively according to $(U_1, \dots, U_d, U_{d+1})^\top \sim C$ and $(U_1, \dots, U_d, U_{d+1}, \dots, U_{d+5})^\top \sim C$ in step (2) of algorithm.

After $(U_1, \dots, U_d)^\top \sim C$ was sampled from copulae, then (3.17) was used to obtain samples of default times $(\tau_1, \dots, \tau_d)^\top$ which can be utilized to compute the portfolio

Rank	Model	MRMSE	Rank	Model	MRMSE	Rank	Model	MRMSE	Rank	Model	MRMSE
1	C_{cl-jo}	0.0980	12	C_{ga-ga}	0.1585	23	C_{jo}	0.4693	34	C_{t4}	1.0222
2	C_{fr-gu}	0.1037	13	C_{cl-cl}	0.1717	24	C_{ng3}	0.4798	35	C_{ga2}	1.0309
3	C_{cl-gu}	0.1062	14	C_{fr-cl}	0.1803	25	C_{ng2}	0.4967	36	C_{t1}	1.0851
4	C_{t-cl}	0.1142	15	C_{t-gu}	0.3433	26	C_{fr-fr}	0.7303	37	C_{ga4}	1.1020
5	C_{ga-jo}	0.1170	16	C_{gu-gu}	0.3574	27	C_{t}	0.8785	38	C_{ga1}	1.1060
6	C_{fr-jo}	0.1182	17	C_{t-jo}	0.3621	28	C_{ga6}	0.9387	39	C_{t4}	1.1140
7	C_{t-fr}	0.1228	18	C_{ng5}	0.3705	29	C_{t6}	0.9728	40	C_{fr}	1.2916
8	C_{ga-cl}	0.1344	19	C_{ng4}	0.3805	30	C_{ga5}	0.9772	41	C_{t-t}	1.4206
9	C_{ga-t}	0.1399	20	C_{gu-jo}	0.3852	31	C_{t5}	0.9854	42	C_{cl}	1.4770
10	C_{ga-gu}	0.1433	21	C_{gu}	0.4052	32	C_{t2}	1.0004	43	C_{ga}	2.9570
11	C_{ga-fr}	0.1536	22	C_{jo-jo}	0.4207	33	C_{ga3}	1.0194			

TABLE 3.4: Empirical study 1: The ranking of 43 copula based models under the mean RMSE (MRMSE). Abbreviations: ga: Gaussian, t: Student-t, fr: Frank, cl: Clayton, gu: Gumbel, jo: Joe, gai, $i = 1, \ldots, 6$: Gaussian with the correlation matrix R_{gai}, $i = 1, \ldots, 6$, tj, $j = 1, \ldots, 6$: Student-t with the same correlation matrix structure as R_{gai}, $i = 1, \ldots, 6$, ng: HAC with the Gumbel generator function.

loss in (3.8), q-th tranche loss in (3.9) and the outstanding notional in (3.10). At last by (3.11) and (3.12) the q-th default leg DL_q and the q-th premium leg PL_q for CDS index tranche pricing can be obtained. Here it uses the notation $\widehat{S}_q^{CDX}$, defined under (3.19), as the tranche spreads (upfront rate version) by Monte Carlo simulation under models listed in Table 3.3 and S_q^{Market} as the real market tranche spread (upfront rate version). And for the parameter calibration, the following measure was utilized, which is a root-mean-square error (RMSE) such that,

$$RMSE = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left(\widehat{S}_q^{CDX} - S_q^{Market}\right)^2}. \tag{3.33}$$

According to the minimization of RMSE in (3.33) the calibration was performed.

As it is given that RMSE is an argument representation, therefore it is needed to perform numerical optimization to calibrate parameters. For all these models the grid search with the multi-core parallel computation in the optimization was employed.

Following the Equation (3.33) through all computation date points, the mean RMSE (MRMSE) can be used for ranking the performance of all 43 models. The MRMSE rankings for two data sets are given in Table 3.4 and 3.5. It is worth mentioning that the two error ranking tables are special case of the Bayesian model comparison methods introduced in Bunnin et al. (2002) and Duembgen and Rogers (2014).

3.4.5 Results and Analysis for Empirical Study 1

In Table 3.4 the MRMSEs based on ten pricing points were calculated and a ranking based on the mean of the RMSEs is given. In Table 3.6 and 3.7 the parameter calibration of the 21 cc-copula models is given. According to the ranking of MRMSE and the parameter calibration, we interpret results as follows:

1. The cc-copula with Archimedean components obtained advantage in CDX pricing. In Table 3.4 it is found that according to the mean of RMSEs the top three best performed models are correspondingly C_{cl-jo}, C_{fr-gu} and C_{cl-gu}. And it is shown that the top 17 models are all cc-copula models. Especially it can be seen that the top three models are not only the cc-copula models but also their components are all from the Archimedean family and it is quite clear that if a model belongs to a member in the top five rank then there must be at least one component copula coming from a Gumbel copula or a Joe copula or a Clayton copula, the copulae with lower or upper tail-dependence. The comparison in RMSE measure of the best three models and the worst three models is shown in the first two rows of Figure 5. It is clear that the gap between the best and the worst in RMSE gauge is quite large.

2. Dearth of asymmetric tail dependence led to the failure for the elliptical family in the MRMSE ranking. In the ranking in Table 3.4 another result is that the group of elliptical copulae perform the worst, see last two rows of Figure 5. One can see that the worst ten models are almost all elliptical copulae. And under the same structures, the Gaussian copula models and the Student-t copula models are compared pair by pair, and it is found that in every structure introduced by Figure 3.4 and 3.5 the Student-t copula models performed similarly to the Gaussian copula models. The last column in Table 3.4 shows that the elliptical copulae are not appropriate for modeling the defaults dependence under the context of CDX NA IG Series 19 index tranche. It can be seen clearly that the Gaussian copulae and the Student-t copulae rank in quite low place.

3. The cc-copula models as a group outperformed the competing models. Hierarchical Archimedean copulae performed better than elliptical copulae, see Figure 5. The best HAC model is C_{ng5} ranked at the 18th place being better than the best single parameter Archimedean copula C_{gu} ranked in place of 21. And the best performed elliptical copula is C_{ga6} ranking at the 28th place. Elliptical family performed the worst. Single parameter Archimedean copula models showed bifurcating performance. The copulae with upper tail dependence structure (Gumbel and Joe copulae) show fair performance, while the lower tail-dependent model

(Clayton copula) and the zero tail-dependent model (Frank copula) belonged to the tail group.

4. The cc-copula models had adaptivity and flexibility to CDX pricing. The Figure 3.8 shows the relationship between the weight parameter and the date in upper triangular. We can observe that the cc-copula model can reduce the RMSE by adjusting the weight parameter in order to be more flexible in different environment. The elliptical copula, the Archimedean copula and the HAC cannot own such properties, which could be a reason of obtaining higher MRMSE in the ranking. Last but not the least, cc-copula models performed stably through ten pricing dates. According to the Figure 3.7 it can be observed that the cc-copula models' RMSEs vary more stable than the other models. The elliptical models vary stronger through ten pricing dates.

Therefore from the above analysis, some conclusions can be obtained. Firstly, the cc-copula model is superior against elliptical copula model, single parameter Archimedean copula model and HAC model, according to the mean RMSE ranking. Secondly, among the well performed cc-copula models the model employing a Gumbel or Joe or Clayton copula has better performance as the both components share the asymmetrical tail-dependence. At last it is concluded that the elliptical copula model are not appropriate for the CDS index tranche pricing as its elliptical distribution and symmetrical tail-dependence.

3.4.6 Results and Analysis for Empirical Study 2

This sub-section provides the empirical results of the iTraxx Europe Series 8 index tranche pricing. In Tables 3.8 and 3.9, the computation results according to the RMSE introduced in (3.33) are shown. While Tables 3.10, 3.11, 3.12, 3.13 and 3.14 present the calibrated parameters under the approach given in Section 4.4. Table 3.5 provides the mean of RMSE based on 12 pricing days and a ranking based on the mean of the relative difference measures is given.

1. As can be seen from the Table 3.5, according to the mean of RMSE introduced in Formula (3.33) the top three best performed models are correspondingly C_{gu-jo}, C_{gu-gu} and C_{fr-jo}, and the top 14 models are all cc-copula models. The top five models are not only the cc-copula models but also their components are all from the Archimedean family and the models for the top ten contain at least one component copula coming from a Gumbel copula or a Joe copula, which are both right tail-dependent. From the model list in Table 3.3 one sees that Models 3, 6, 7,

Rank	Notation	MRMSE	Rank	Notation	MRMSE	Rank	Notation	MRMSE
1	C_{gu-jo}	0.5254	16	C_{ng3}	0.7803	31	C_{ga5}	1.0152
2	C_{gu-gu}	0.5279	17	C_{gu}	0.7994	32	C_{fr-fr}	1.0358
3	C_{fr-jo}	0.5279	18	C_{ga-t}	0.8083	33	C_{t2}	2.0747
4	C_{cl-jo}	0.5401	19	C_{t-cl}	0.8236	34	C_{t4}	2.3944
5	C_{fr-gu}	0.5492	20	C_{ng5}	0.8271	35	C_{ga2}	2.5520
6	C_{ga-gu}	0.5524	21	C_{ng2}	0.8450	36	C_{fr}	2.5583
7	C_{cl-gu}	0.5629	22	C_{ga-ga}	0.8563	37	C_{ga4}	2.6659
8	C_{t-gu}	0.5652	23	C_{cl-cl}	0.8697	38	C_{t}	2.7114
9	C_{jo-jo}	0.5817	24	C_{ga-cl}	0.8707	39	C_{ga}	2.7130
10	C_{ga-jo}	0.5894	25	C_{t6}	0.9469	40	C_{ga1}	2.7444
11	C_{t-fr}	0.6157	26	C_{t5}	0.9490	41	C_{t1}	2.7583
12	C_{t-jo}	0.6184	27	C_{ng4}	0.9618	42	C_{t-t}	2.8851
13	C_{fr-cl}	0.6614	28	C_{ga6}	0.9724	43	C_{cl}	3.0089
14	C_{ga-fr}	0.6858	29	C_{t3}	0.9888			
15	C_{jo}	0.7476	30	C_{ga3}	0.9971			

Table 3.5: Empirical study 2: The ranking of 43 copula based models under the mean RMSE (MRMSE). Abbreviations: ga: Gaussian, t: Student-t, fr: Frank, cl: Clayton, gu: Gumbel, jo: Joe, gai, $i = 1, \ldots, 6$: Gaussian with the correlation matrix R_{gai}, $i = 1, \ldots, 6$, tj, $j = 1, \ldots, 6$: Student-t with the same correlation matrix structure as R_{gai}, $i = 1, \ldots, 6$, ng: HAC with the Gumbel generator function.

8, Model 20 to Model 22, Models 9, 12, 13, 14 were specified for the heterogeneous dependence between sectors and cc-copula did not use a special parameter to do the same thing but the empirical results show that the cc-copula do consider these heterogeneity of dependence between sectors.

2. Table 3.5 highlights that the elliptical copulae perform worst as among worst 10 are almost all elliptical. The Gaussian copula models and the Student-t copula models were compared pairwise with the same structure, and in every structure introduced by Figure 3.4, 3.5, the Student-t copula models outperform the Gaussian copula models.

3. Hierarchical Archimedean copulae performed better than elliptical copulae and the best hierarchical Archimedean copula model is C_{ng3} ranked at the 16th place being not much better than the best single parameter Archimedean copula C_{jo}. And the best performed elliptical copula is C_{t6} ranking at the 25th place. The last column in Table 3.5 shows that the elliptical copulae are not appropriate for modeling the defaults dependence under the iTraxx Europe index tranche context as it is indicated that the Frank copula, the Gaussian copula and the Student-t copula rank in low ranking place.

4. Another interesting result from Table 3.12 shows that the calibrated parameter λ, which is the weight of the first component copula in a cc-copula model, gives a much larger weight in a cc-copula composing an elliptical copula and a Gumbel copula or a Joe copula to the Gumbel or Joe copula, i.e. the calibration automatically choose Gumbel or Joe rather than an elliptical copula, which means Gumbel and Joe copulae are appropriate for modeling default times of entities of the iTraxx Europe index components. The main reason is that the joint default times have a right tail-dependence. And from the results of parameters in Table 3.12 it can be verified that the joint defaults are not left tail-dependent as the λ in model C_{cl-gu} and the model C_{c-j}, which are correspondingly the cc-copula of a Clayton copula and a Gumbel or Joe copula, in 12 pricing days were mostly lower than 0.5, which can be an evidence of non-left-dependence. Another evidence is that the model C_{cl} performed the worst under the MRMSE.

Therefore, from the above analysis some conclusions can be drawn. Firstly, the cc-copula model is superior against elliptical copulae, single parameter Archimedean copulae and four hierarchical Archimedean copulae employed in Table 3.3 according to the MRMSE. Secondly, among the well performed cc-copula models the model employing a Gumbel or Joe copula has better performance since the both share the right tail-dependence. Thirdly, the joint default times has a right tail-dependence not a left one and an elliptical one, therefore the Clayton copula and the Frank copula is not appropriate for modeling the joint defaults under the iTraxx Europe index tranche context. At last we conclude that the elliptical copulae are not appropriate for the CDS index tranche pricing as its elliptical distribution and symmetrical tail-dependence.

3.5 Conclusion

The goal of this paper is to construct defaults dependence structure mainly with cc-copulae for the CDX tranche pricing. In this work totally 43 diverse copula models were employed, containing 21 cc-copulae with two component copulae coming from two elliptical copulae and four Archimedean copulae. At last all computation results were given out based on the MRMSE measure. It is found that cc-copula models have dominant performance compared with other copula models. In Figure 3.9, it is clear that the cc-copula models (clustering in blue group) are robustly best performed in different market regimes, in crisis and non-crisis. Especially those cc-copulae which own at least one asymmetrical component copula coming from the Gumbel, Joe or Clayton copula, show top performance. It is a clear evidence that joint defaults are asymmetrically tail-dependent. According to the Figure 3.9, in the other three families, the elliptical family (clustering in chocolate color group) performs the worst, which means that those copulae without tail-dependence feature and asymmetrical distribution are not suitable for CDX tranche pricing. The rest two families, the Archimedean copula family and the HAC family (clustering in green group), perform similarly and place in the middle of the ranking.

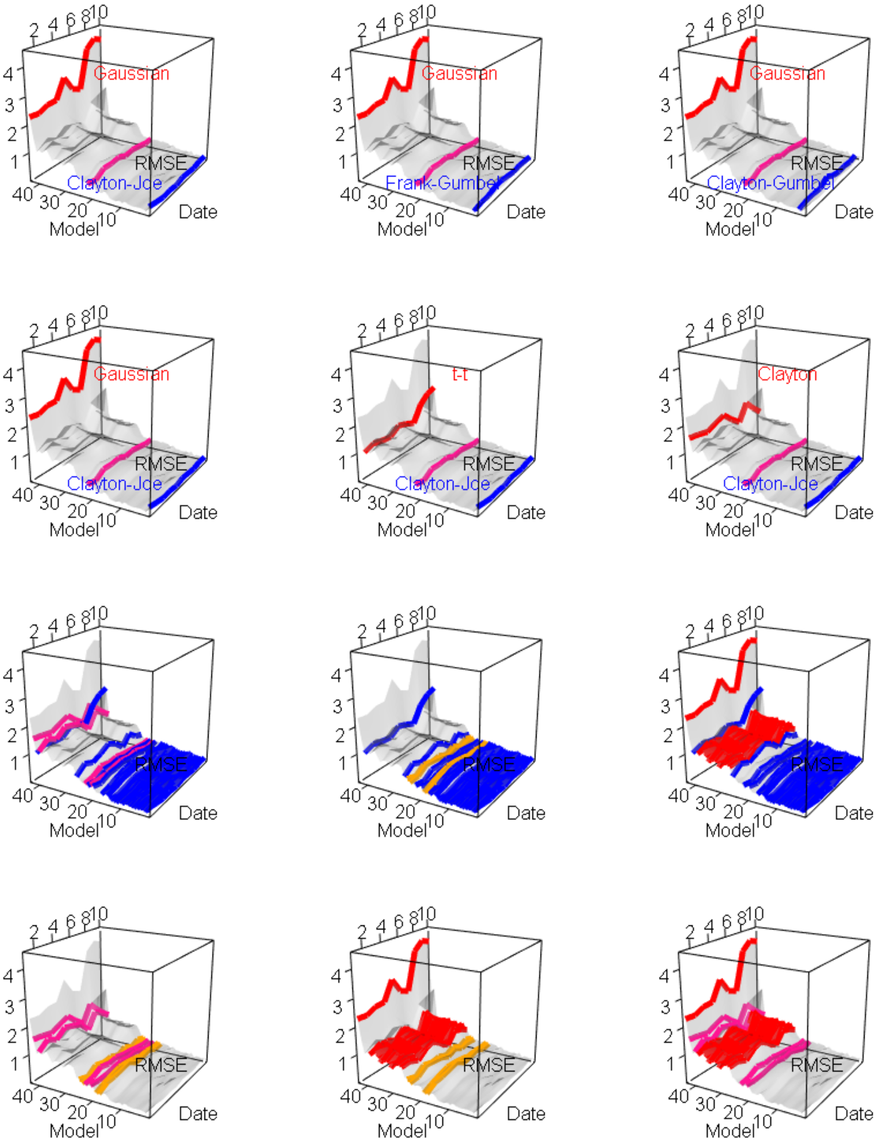


Figure 3.6: Comparison of RMSEs between the three best performed models (C_{cl-jo}, C_{fr-gu}, C_{cl-gu}) and the worst performed model (C_{ga}) at ten date points (see the first row). Comparison of RMSEs between the three worst performed models (C_{ga}, C_{t-t}, C_{cl}) and the best performed model (C_{cl-jo}) at ten date points (see the second row). The pink lines in the first and second row stand for the median performed model, C_{jo-jo}. Comparison of RMSEs of models between four different families (see the third and the fourth row). The red lines stand for the models from the elliptical family, the orange for the HAC family, the pink for the Archimedean family and the blue for the cc-copula family. The transparent gray terrain in every plot stands for the RMSE surface of all 43 models.

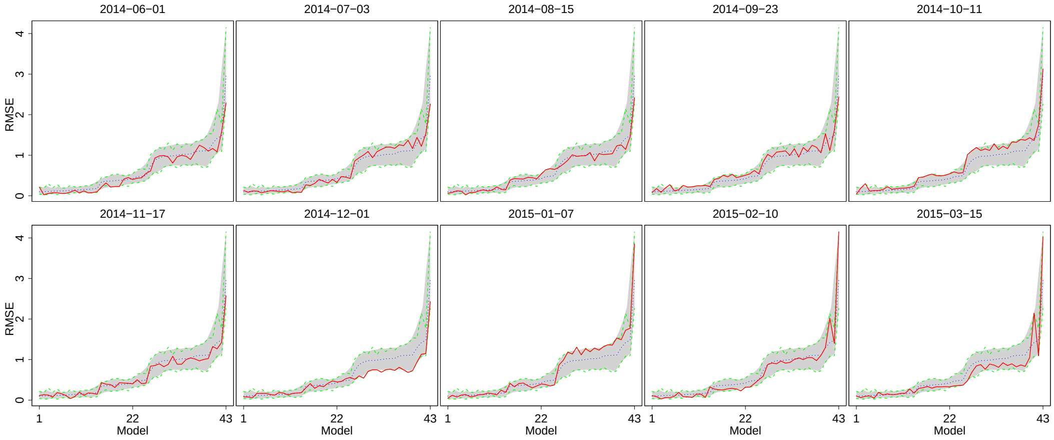


FIGURE 3.7: RMSEs' comparison of 43 models at ten pricing dates. The red line stands for the RMSE of the corresponding pricing date. The green line stands for the RMSE bounds of ten dates. The black dashed line shows the mean RMSE through ten pricing date points. The shading area is limited by 0.05 and 0.95 nonlinear local quantile regressions.

Model		20140601	20140703	20140815	20140923	20141011	20141117	20141201	20150107	20150210	20150315
C_{cl-cl}	θ_1	0.443	0.990	0.990	0.532	0.532	0.990	0.990	0.512	0.990	0.443
	θ_2	0.968	0.532	0.577	0.990	0.990	0.577	0.641	0.990	0.512	0.990
	λ	0.443	0.468	0.423	0.532	0.577	0.423	0.359	0.621	0.314	0.661
C_{cl-gu}	θ_1	0.532	0.532	0.641	0.577	0.463	0.577	0.597	0.488	0.512	0.557
	θ_2	0.488	0.853	0.946	0.921	0.537	0.710	0.794	0.879	0.621	0.819
	λ	0.379	0.532	0.641	0.577	0.448	0.577	0.597	0.597	0.666	0.774
C_{cl-jo}	θ_1	0.577	0.577	0.577	0.532	0.577	0.577	0.597	0.512	0.468	0.488
	θ_2	0.572	0.853	0.883	0.750	0.887	0.468	0.428	0.621	0.314	0.887
	λ	0.532	0.577	0.577	0.532	0.621	0.577	0.601	0.621	0.532	0.706
C_{fr-cl}	θ_1	0.314	0.359	0.379	0.403	0.359	0.379	0.448	0.270	0.206	0.294
	θ_2	0.968	0.990	0.990	0.990	0.990	0.990	0.990	0.990	0.946	0.990
	λ	0.488	0.532	0.552	0.577	0.577	0.597	0.666	0.597	0.508	0.730
C_{fr-fr}	θ_1	0.166	0.794	0.188	0.143	0.794	0.226	0.794	0.794	0.794	0.794
	θ_2	0.794	0.010	0.794	0.794	0.161	0.794	0.188	0.010	0.044	0.117
	λ	0.182	0.818	0.216	0.216	0.794	0.294	0.706	0.661	0.641	0.577
C_{fr-gu}	θ_1	0.314	0.359	0.423	0.403	0.314	0.379	0.423	0.314	0.294	0.359
	θ_2	0.468	0.710	0.968	0.750	0.572	0.572	0.750	0.901	0.537	0.857
	λ	0.359	0.512	0.641	0.577	0.488	0.552	0.641	0.641	0.601	0.750
C_{fr-jo}	θ_1	0.359	0.423	0.379	0.423	0.403	0.379	0.359	0.359	0.270	0.270
	θ_2	0.557	0.879	0.710	0.812	0.972	0.468	0.448	0.990	0.443	0.537
	λ	0.448	0.641	0.532	0.597	0.666	0.488	0.492	0.686	0.552	0.621
C_{gu-gu}	θ_1	0.270	0.246	0.294	0.314	0.250	0.181	0.339	0.206	0.161	0.113
	θ_2	0.294	0.294	0.270	0.206	0.290	0.250	0.206	0.250	0.319	0.339
	λ	0.577	0.250	0.443	0.601	0.750	0.216	0.121	0.921	0.863	0.774
C_{gu-jo}	θ_1	0.339	0.250	0.270	0.294	0.294	0.206	0.206	0.206	0.250	0.077
	θ_2	0.161	0.226	0.182	0.161	0.117	0.621	0.552	0.147	0.113	0.113
	λ	0.492	0.617	0.730	0.666	0.819	0.879	0.901	0.641	0.147	0.216
C_{ga-cl}	θ_1	0.226	0.946	0.990	0.946	0.314	0.990	0.990	0.990	0.990	0.990
	θ_2	0.968	0.488	0.557	0.532	0.990	0.621	0.621	0.532	0.468	0.557
	λ	0.463	0.512	0.443	0.512	0.641	0.334	0.334	0.359	0.359	0.226
C_{ga-fr}	θ_1	0.990	0.990	0.923	0.990	0.990	0.990	0.946	0.968	0.990	0.968
	θ_2	0.403	0.359	0.359	0.403	0.339	0.403	0.423	0.270	0.339	0.294
	λ	0.399	0.423	0.468	0.423	0.443	0.334	0.314	0.403	0.290	0.270

TABLE 3.6: Calibration of parameters of cc-copulae, i.e. θ_1, θ_2, λ. Abbreviations: *ga*: Gaussian, *t*: Student-*t*, *fr*: Frank, *cl*: Clayton, *gu*: Gumbel, *jo*: Joe.

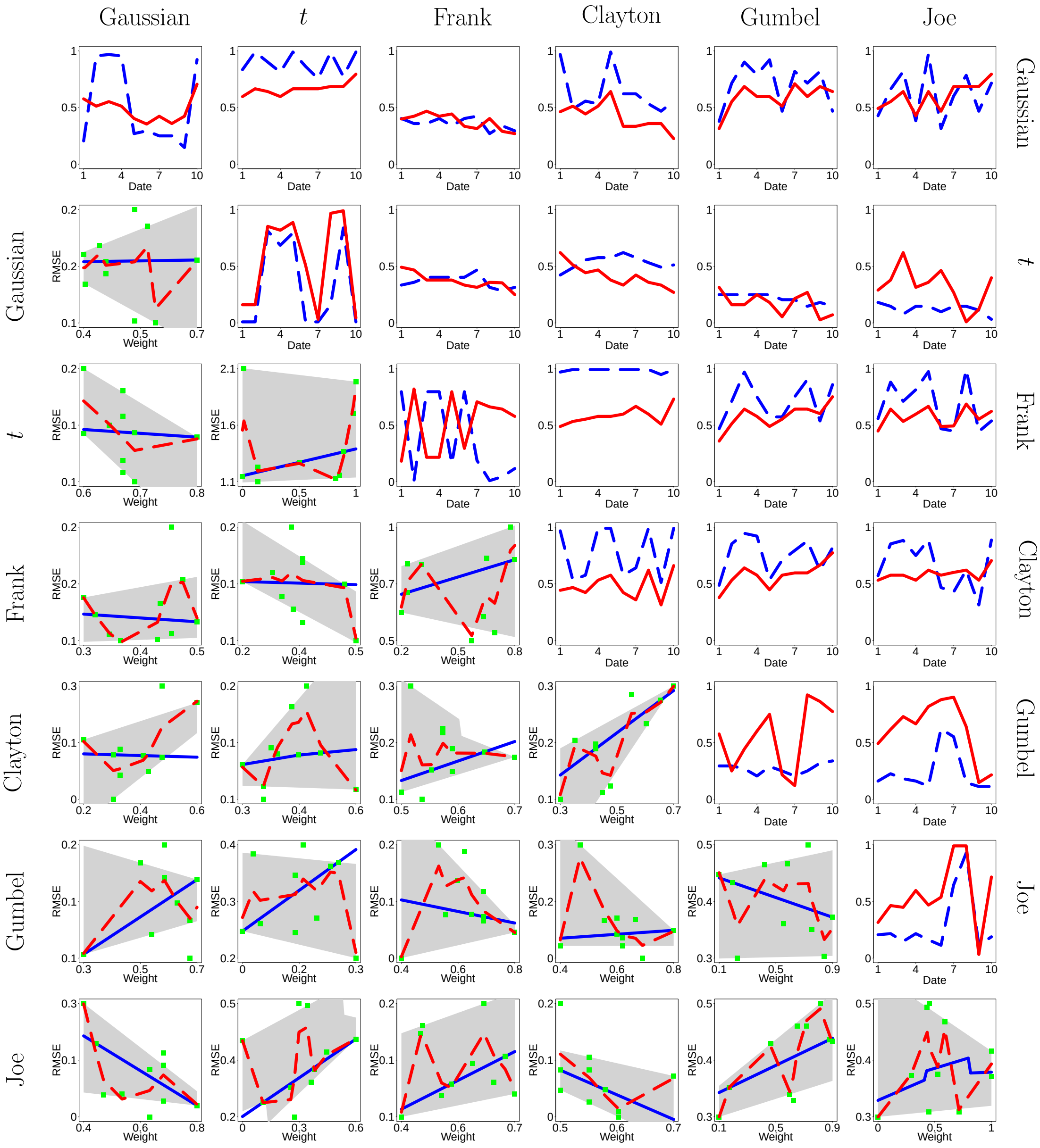


FIGURE 3.8: The lower triangular panels show the relationship between the weight (λ) and the RMSE for 21 cc-copula models. The blue solid line stands for the 0.5 quantile regression with 0.15 and 0.85 quantile regressions as the shading boundaries. The dashed red line is a local polynomial linear regression. The upper triangular graphs illustrate the two parameters' series. The red solid line stands for λ and the blue dashed for θ_1.

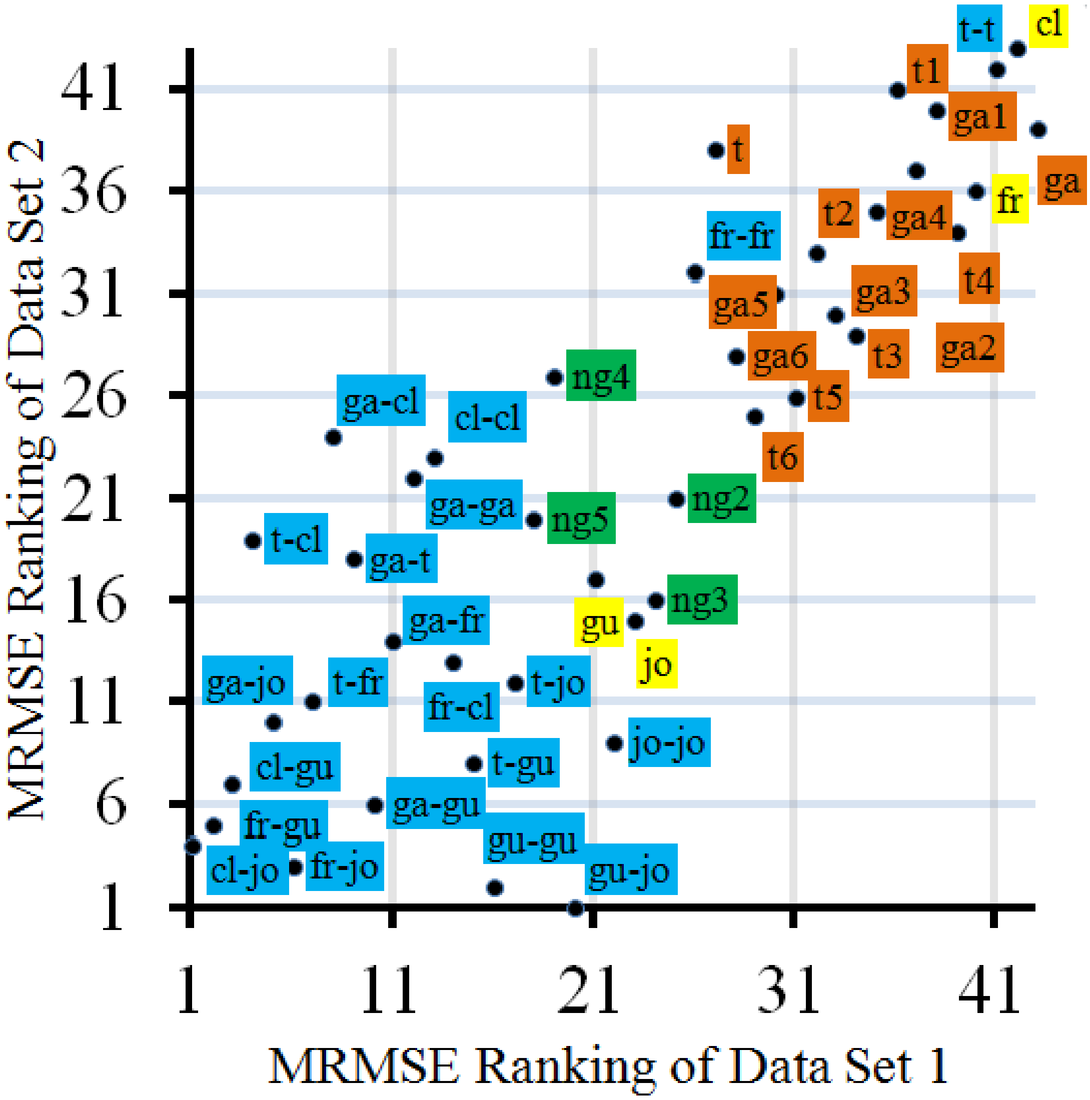


FIGURE 3.9: MRMSE ranking comparison based on the Table 3.4 and 3.5. Convex combination of copula models in blue, Archimedean copula models in yellow, hierarchical Archimedean copula models in green and elliptical copula models in chocolate.

Model		20140601	20140703	20140815	20140923	20141011	20141117	20141201	20150107	20150210	20150315
C_{ga-gu}	θ_1	0.294	0.270	0.359	0.314	0.270	0.250	0.339	0.206	0.206	0.206
	θ_2	0.379	0.715	0.901	0.790	0.921	0.468	0.819	0.715	0.819	0.468
	λ	0.314	0.552	0.686	0.597	0.597	0.512	0.710	0.597	0.686	0.641
C_{ga-ga}	θ_1	0.887	0.250	0.270	0.250	0.990	0.972	0.921	0.990	0.887	0.206
	θ_2	0.206	0.956	0.968	0.956	0.270	0.294	0.250	0.250	0.147	0.923
	λ	0.577	0.512	0.552	0.512	0.403	0.354	0.423	0.359	0.423	0.706
C_{ga-jo}	θ_1	0.290	0.270	0.314	0.226	0.314	0.250	0.314	0.250	0.250	0.250
	θ_2	0.428	0.661	0.818	0.383	0.972	0.314	0.601	0.784	0.468	0.715
	λ	0.492	0.552	0.641	0.428	0.641	0.468	0.686	0.686	0.686	0.794
C_{ga-t}	θ_1	0.314	0.339	0.314	0.314	0.339	0.294	0.294	0.250	0.206	0.250
	θ_2	0.834	0.990	0.901	0.812	0.990	0.863	0.754	0.990	0.774	0.990
	λ	0.597	0.666	0.641	0.597	0.666	0.666	0.666	0.686	0.686	0.794
C_{jo-jo}	θ_1	0.147	0.161	0.246	0.161	0.216	0.206	0.147	0.147	0.853	0.079
	θ_2	0.206	0.216	0.147	0.216	0.161	0.113	0.646	0.928	0.113	0.188
	λ	0.314	0.463	0.448	0.597	0.468	0.537	0.990	0.990	0.032	0.715
C_{t-cl}	θ_1	0.463	0.715	0.774	0.617	0.928	0.730	0.532	0.887	0.754	0.819
	θ_2	0.423	0.488	0.557	0.577	0.577	0.621	0.577	0.532	0.492	0.512
	λ	0.621	0.512	0.443	0.468	0.379	0.334	0.423	0.359	0.334	0.270
C_{t-fr}	θ_1	0.681	0.784	0.887	0.990	0.956	0.972	0.883	0.818	0.537	0.853
	θ_2	0.334	0.359	0.403	0.403	0.403	0.403	0.468	0.314	0.285	0.314
	λ	0.492	0.468	0.379	0.379	0.379	0.334	0.314	0.359	0.354	0.250
C_{t-gu}	θ_1	0.137	0.379	0.314	0.226	0.117	0.887	0.054	0.443	0.468	0.079
	θ_2	0.250	0.250	0.250	0.250	0.250	0.206	0.206	0.147	0.182	0.147
	λ	0.314	0.161	0.161	0.250	0.181	0.054	0.216	0.270	0.028	0.072
C_{t-jo}	θ_1	0.137	0.270	0.294	0.290	0.181	0.226	0.028	0.863	0.010	0.250
	θ_2	0.182	0.147	0.079	0.147	0.147	0.099	0.147	0.147	0.113	0.032
	λ	0.290	0.379	0.621	0.314	0.359	0.463	0.270	0.010	0.121	0.399
C_{t-t}	θ_1	0.863	0.794	0.010	0.010	0.010	0.077	0.839	0.010	0.010	0.077
	θ_2	0.010	0.010	0.812	0.686	0.794	0.010	0.010	0.161	0.839	0.010
	λ	0.161	0.161	0.853	0.818	0.887	0.512	0.032	0.968	0.990	0.044

TABLE 3.7: Calibration of parameters of cc-copulae, i.e. θ_1, θ_2, λ. Abbreviations: *ga*: Gaussian, *t*: Student-*t*, *fr*: Frank, *cl*: Clayton, *gu*: Gumbel, *jo*: Joe.

Model	*Notation*	*20071023*	*20071026*	*20071117*	*20071206*	*20080111*	*20080228*	*20080314*	*20080405*	*20080424*	*20080529*	*20080530*	*20080701*
1	C_{ga}	2.6364	3.3408	3.8685	3.5433	2.6748	2.0957	1.7642	2.2786	2.0728	2.0930	2.3526	3.8355
2	C_t	2.8039	3.4428	3.7744	3.2184	2.7395	2.0718	1.6934	2.2712	2.2760	2.0072	2.3348	3.9039
3	C_{ga1}	2.8023	3.4867	3.8514	3.5009	2.5876	2.1691	2.0252	2.3454	2.0440	2.0232	2.3041	3.7924
4	C_{ga2}	2.5391	3.0057	3.4617	3.0994	2.4720	2.2165	2.1415	1.8940	2.2355	2.2035	2.3655	2.9899
5	C_{ga3}	1.6070	1.5954	1.5554	1.5377	0.9710	0.7018	0.5904	0.4747	0.6204	0.6901	0.9500	0.6717
6	C_{ga4}	2.7136	3.0264	3.2811	3.1006	2.7897	2.4122	2.2159	2.2041	2.3757	2.2935	2.5053	3.0722
7	C_{ga5}	1.7111	1.6647	1.6644	1.6089	0.8870	0.7419	0.6270	0.4856	0.6969	0.5853	0.8386	0.6705
8	C_{ga6}	1.7436	1.6231	1.6359	1.3469	0.8617	0.6844	0.5711	0.4792	0.6200	0.6014	0.8632	0.6379
9	C_{fr}	2.7815	3.1873	3.4164	3.1603	2.6158	2.0224	1.6685	1.9767	2.1416	2.0128	2.3107	3.4054
10	C_{cl}	3.0302	3.6218	4.3013	3.7675	2.8532	2.5900	1.8167	2.4436	2.5464	2.3539	2.7680	4.0137
11	C_{gu}	0.4899	0.8384	1.2759	0.9499	0.6277	0.4709	0.4109	0.4611	0.6291	0.6490	0.7115	2.0788
12	C_{jo}	0.6361	0.8812	0.9367	0.5786	0.6130	0.4716	0.4027	0.4983	0.5679	0.5988	0.6559	2.1302
13	C_{ng2}	0.3557	1.0089	1.1733	0.9186	0.7440	0.3873	0.4637	0.4506	0.6081	0.4838	0.6432	2.1260
14	C_{ng3}	0.6882	0.9172	1.2679	0.6957	0.7178	0.7629	1.2664	0.6050	0.8637	0.9681	0.7432	2.0454
15	C_{ng4}	0.3669	0.8330	1.1313	0.8300	0.7391	0.7267	1.0132	0.4871	0.8623	0.8185	0.7356	1.5962
16	C_{ng5}	0.4491	0.8260	1.1922	1.0083	0.6302	0.6001	0.9353	0.5942	0.7699	0.7318	0.7073	1.4812
17	C_{ga-ga}	0.5985	0.8055	0.8255	0.8513	0.7176	0.7649	0.9012	0.6710	0.7797	0.6973	0.9598	1.7029
18	C_{ga-t}	0.6175	0.7216	0.7623	0.8850	0.7234	0.7694	0.8372	0.7113	0.7474	0.7025	0.9436	1.2783
19	C_{ga-fr}	0.4614	0.6310	0.6850	0.7659	0.5377	0.5895	0.6538	0.5974	0.5461	0.5366	0.7036	1.5210
20	C_{gs-c}	0.5993	0.8112	0.7903	0.8591	0.8073	0.8713	0.8248	0.6984	0.7978	0.7155	0.9920	1.6809
21	C_{ga-gu}	0.4382	0.6840	0.6186	0.4111	0.4359	0.4212	0.3754	0.3985	0.4668	0.4897	0.5163	1.3734
22	C_{ga-jo}	0.3765	0.6762	0.7687	0.7052	0.4753	0.3216	0.3722	0.3771	0.5079	0.3926	0.5573	1.5428

TABLE 3.8: RMSEs after the calibration for copula models from Model 1 to Model 22, $M = 10^4$.

Model	*Notation*	*20071023*	*20071026*	*20071117*	*20071206*	*20080111*	*20080228*	*20080314*	*20080405*	*20080424*	*20080529*	*20080530*	*20080701*
23	C_{t-t}	5.7815	5.4408	4.0507	3.8298	2.1894	1.6421	0.9903	1.3129	1.6365	1.6596	2.2572	3.8304
24	C_{t-fr}	0.4805	0.5305	0.6540	0.7454	0.5223	0.5419	0.5939	0.5001	0.4942	0.4732	0.6698	1.1828
25	C_{t-cl}	0.5628	0.6864	0.7068	0.8377	0.8175	0.8200	0.8395	0.7203	0.8075	0.7414	1.0853	1.2585
26	C_{t-gu}	0.5041	0.6917	0.4719	0.4974	0.3376	0.3654	0.3779	0.3513	0.4451	0.4389	0.5480	1.7529
27	C_{t-jo}	0.7308	0.7217	0.5580	0.4921	0.4814	0.4576	0.3520	0.4097	0.4736	0.5467	0.5537	1.6433
28	C_{fr-fr}	0.6814	1.0177	1.3154	1.0419	0.9411	0.8434	0.9101	0.8835	0.8518	0.8625	0.9406	2.1401
29	C_{fr-cl}	0.5358	0.5432	0.6560	0.7960	0.5371	0.5516	0.6010	0.5264	0.4966	0.5211	0.6910	1.4815
30	C_{fr-gu}	0.4267	0.4855	0.6245	0.6878	0.4149	0.3868	0.4408	0.3930	0.5136	0.4257	0.5517	1.2494
31	C_{fr-jo}	0.3334	0.4891	0.5957	0.7513	0.4177	0.3583	0.4035	0.4436	0.4690	0.4045	0.5565	1.2587
32	C_{cl-cl}	0.6335	0.8272	0.8235	0.8578	0.8260	0.8790	0.8636	0.7621	0.8206	0.7818	1.0465	1.3143
33	C_{cl-gu}	0.4043	0.9131	0.5349	0.6922	0.3732	0.3201	0.3904	0.3450	0.5384	0.4531	0.5109	1.2787
34	C_{cl-jo}	0.4754	0.5953	0.7214	0.5917	0.4153	0.3331	0.3943	0.2728	0.4667	0.4559	0.5170	1.3517
35	C_{gu-gu}	0.4416	0.5727	0.4310	0.3956	0.3541	0.3972	0.4118	0.3600	0.4756	0.4371	0.5074	1.5506
36	C_{gu-jo}	0.4465	0.5800	0.4961	0.4014	0.3477	0.3922	0.3487	0.3222	0.4145	0.4541	0.5496	1.5516
37	C_{jo-jo}	0.4964	0.6344	0.5243	0.5442	0.4708	0.4064	0.3616	0.4205	0.5202	0.5160	0.5088	1.5773
38	C_{t1}	2.7960	3.5066	3.6762	3.4915	2.6375	2.1405	1.8009	2.3879	2.3185	2.0525	2.3156	3.9759
39	C_{t2}	1.9565	2.5117	2.9483	2.6540	2.0149	1.7602	1.7140	1.4207	1.8202	1.6384	1.9116	2.5465
40	C_{t3}	1.6953	1.5694	1.5639	1.4425	0.9600	0.6797	0.5882	0.5019	0.6658	0.6501	0.8827	0.6659
41	C_{t4}	2.4961	2.7395	3.1500	2.8079	2.3363	2.0149	2.2010	1.9131	2.1548	1.9435	2.0546	2.9213
42	C_{t5}	1.6043	1.5006	1.5411	1.3852	0.8785	0.6759	0.5876	0.4881	0.6103	0.6078	0.8414	0.6672
43	C_{t6}	1.6503	1.5631	1.4718	1.3730	0.9253	0.6886	0.5708	0.4586	0.6172	0.5869	0.7740	0.6836

Table 3.9: RMSEs after the calibration for copula models from Model 23 to Model 43, $M = 10^4$.

Model	*Notation*	*Parameter*	*20071023*	*20071026*	*20071117*	*20071206*	*20080111*	*20080228*	*20080314*	*20080405*	*20080424*	*20080529*	*20080530*	*20080701*
1	C_{ga}	θ	0.2377	0.1288	0.1882	0.1981	0.3169	0.3466	0.8415	0.2872	0.3268	0.3367	0.2971	0.1684
2	C_t	θ	0.1053	0.0632	0.1474	0.1684	0.1895	0.2947	0.8211	0.2737	0.2737	0.2947	0.2737	0.1053
		df	18.0000	20.0000	19.0000	20.0000	14.0000	15.0000	5.0000	12.0000	20.0000	14.0000	18.0000	19.0000
3	C_{ga1}	ρ_1	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
		ρ_2	0.0833	0.0833	0.0648	0.0833	0.0648	0.0833	0.0648	0.0833	0.0833	0.0833	0.0833	0.0833
4.	C_{ga2}	ρ_1	0.1417	0.0709	0.1063	0.1063	0.1063	0.1417	0.9214	0.1772	0.1417	0.1417	0.1417	0.0709
5	C_{ga3}	ρ_1	0.0648	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0648	0.0833	0.0833	0.0833	0.0648
		ρ_2	-0.0463	-0.0463	-0.0463	-0.0463	-0.0648	-0.0648	-0.0648	-0.0648	-0.0648	-0.0648	-0.0648	-0.0463
6	C_{ga4}	ρ_1	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
		ρ_2	0.0648	0.0833	0.0648	0.0648	0.0648	0.0833	0.0648	0.0648	0.0833	0.0833	0.0648	0.0648
7	C_{ga5}	ρ_1	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833
		ρ_2	-0.0833	-0.0833	-0.0463	-0.0463	-0.0463	-0.0463	-0.0463	-0.0463	-0.0463	-0.0833	-0.0463	-0.0463
		ρ_3	0.0833	0.0833	0.0648	0.0648	0.0648	0.0648	0.0648	0.0648	0.0648	0.0833	0.0648	0.0648
8	C_{ga6}	ρ_1	0.0648	0.0833	0.0833	-0.0833	0.0833	0.0833	0.0833	0.0648	0.0833	0.0833	0.0833	0.0833
		ρ_2	-0.0648	-0.0833	-0.0833	0.0648	-0.0463	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0463	-0.0648
		ρ_3	0.0463	0.0463	0.0278	-0.0093	0.0648	-0.0463	-0.0648	-0.0463	-0.0463	-0.0463	0.0648	-0.0463
9	C_{fr}	θ	0.4753	0.4060	0.2476	0.4159	0.6039	0.5544	0.9207	0.8019	0.5049	0.5445	0.5544	0.1981
10	C_{cl}	θ	0.2575	0.1585	0.2080	0.2179	0.2872	0.3367	0.8118	0.3664	0.2971	0.2971	0.3070	0.0991
11	C_{gu}	θ	0.0991	0.1090	0.0991	0.1189	0.1981	0.2278	0.3763	0.2971	0.2377	0.2971	0.2080	0.0892
12	C_{jo}	θ	0.0694	0.0793	0.0892	0.0892	0.1486	0.1981	0.2971	0.2278	0.1882	0.1882	0.1288	0.0694

TABLE 3.10: Calibration of parameters of copulae, i.e. θ_1, θ_2, λ. Abbreviations: *ga*: Gaussian, *t*: Student-*t*, *fr*: Frank, *cl*: Clayton, *gu*: Gumbel, *jo*: Joe.

Model	*Notation*	*Parameter*	*20071023*	*20071026*	*20071117*	*20071206*	*20080111*	*20080228*	*20080314*	*20080405*	*20080424*	*20080529*	*20080530*	*20080701*
13	C_{ng2}	$\rho_{\mathcal{K}1}$	0.0536	0.0536	0.0680	0.0780	0.1066	0.1291	0.3193	0.1402	0.1338	0.1368	0.1268	0.0528
14	C_{ng3}	$\rho_{\mathcal{K}1}$	0.0100	0.0322	0.0544	0.1433	0.0100	0.1433	0.1816	0.1473	0.2056	0.2500	0.1816	0.0443
		$\rho_{\mathcal{K}2}$	0.1130	0.0786	0.1130	0.1473	0.1611	0.2500	0.3589	0.2944	0.2500	0.2944	0.2056	0.1130
15	C_{ng4}	$\rho_{\mathcal{K}1}$	0.0544	0.0443	0.0786	0.0786	0.1130	0.1473	0.2056	0.1473	0.1473	0.1473	0.1130	0.0786
		$\rho_{\mathcal{K}2}$	0.0544	0.0443	0.0786	0.0786	0.1130	0.1473	0.2056	0.1473	0.1473	0.1473	0.1130	0.0786
16	C_{ng5}	$\rho_{\mathcal{K}1}$	0.0322	0.0544	0.0544	0.0544	0.1130	0.0786	0.1611	0.1473	0.1130	0.0786	0.0786	0.0322
		$\rho_{\mathcal{K}2}$	0.0322	0.0544	0.0544	0.0544	0.1130	0.0786	0.1611	0.1473	0.1130	0.0786	0.0786	0.0322
		$\rho_{\mathcal{K}3}$	0.1130	0.0786	0.0786	0.1167	0.1130	0.2500	0.2944	0.2056	0.1816	0.2159	0.1473	0.1611
17	C_{ga-ga}	θ_1	0.8567	0.9557	0.0100	0.0100	0.1130	0.1130	0.9900	0.1130	0.1130	0.9678	0.9678	0.0544
		θ_2	0.0443	0.0100	0.9678	0.9011	0.9678	0.9456	0.1611	0.9900	0.9900	0.1473	0.1130	0.9900
		λ	0.2944	0.4478	0.5322	0.5322	0.5767	0.5767	0.5567	0.4233	0.5767	0.4233	0.3589	0.6456
18	C_{ga-t}	θ_1	0.0786	0.0786	0.0100	0.0100	0.1473	0.1130	0.1611	0.1130	0.1473	0.1816	0.1473	0.0786
		θ_2	0.7498	0.8567	0.8527	0.6611	0.9214	0.7056	0.9678	0.9900	0.9557	0.9722	0.8389	0.9678
		λ	0.7056	0.6411	0.5322	0.5322	0.5767	0.5322	0.3789	0.4233	0.5767	0.5767	0.6411	0.5522
19	C_{ga-fr}	θ_1	0.9678	0.9233	0.9900	0.8344	0.9678	0.9456	0.9678	0.9678	0.9900	0.9900	0.9900	0.9900
		θ_2	0.1130	0.0786	0.0544	0.0100	0.1130	0.1473	0.1473	0.1130	0.1816	0.1611	0.1473	0.0786
		λ	0.2700	0.3144	0.4233	0.5122	0.4678	0.4678	0.5322	0.4233	0.3789	0.4478	0.3344	0.3344
20	C_{ga-cl}	θ_1	0.7498	0.9900	0.9900	0.9214	0.1130	0.9678	0.2056	0.1130	0.9678	0.1167	0.8527	0.9900
		θ_2	0.0786	0.0722	0.0100	0.0100	0.9900	0.2500	0.9900	0.9900	0.2500	0.9678	0.1878	0.1130
		λ	0.3144	0.4033	0.4678	0.4678	0.5322	0.4678	0.3589	0.4233	0.4233	0.5522	0.4433	0.4678

TABLE 3.11: Calibration of parameters of copulae, i.e. θ_1, θ_2, λ. Abbreviations: *ga*: Gaussian, *t*: Student-*t*, *fr*: Frank, *cl*: Clayton, *gu*: Gumbel, *jo*: Joe.

Model	*Notation*	*Parameter*	*20071023*	*20071026*	*20071117*	*20071206*	*20080111*	*20080228*	*20080314*	*20080405*	*20080424*	*20080529*	*20080530*	*20080701*
21	C_{ga-gu}	θ_1	0.1367	0.0322	0.9678	0.9900	0.9900	0.8344	0.9678	0.9456	0.9900	0.9722	0.9011	0.0544
		θ_2	0.1130	0.3344	0.0544	0.0989	0.1473	0.2500	0.3144	0.2056	0.2500	0.2500	0.1816	0.8633
		λ	0.0100	0.5322	0.3344	0.2456	0.2700	0.0100	0.1816	0.2159	0.0786	0.0722	0.0722	0.5122
22	C_{ga-jo}	θ_1	0.0544	0.0100	0.0100	0.0100	0.3989	0.0767	0.8344	0.1367	0.0989	0.2500	0.4433	0.0786
		θ_2	0.1167	0.2159	0.7500	0.1367	0.1473	0.2056	0.2944	0.2056	0.2056	0.2056	0.1130	0.8633
		λ	0.3544	0.5367	0.5322	0.2456	0.0100	0.0443	0.0278	0.1130	0.0786	0.1167	0.1816	0.5767
23	C_{t-t}	θ_1	0.9900	0.8527	0.9900	0.9900	0.0100	0.9678	0.0322	0.0322	0.9900	0.9678	0.9900	0.0322
		θ_2	0.1816	0.9678	0.0278	0.9900	0.9900	0.0322	0.9900	0.9678	0.0544	0.0100	0.0100	0.9900
		λ	0.9214	0.0786	0.7498	0.4233	0.3589	0.5967	0.4033	0.4478	0.5078	0.5078	0.6456	0.3144
24	C_{t-fr}	θ_1	0.8633	0.8122	0.7841	0.8389	0.8833	0.7500	0.9011	0.9233	0.8567	0.9233	0.7944	0.9900
		θ_2	0.1473	0.1130	0.0322	0.0767	0.1473	0.1473	0.1611	0.1130	0.2056	0.2056	0.1473	0.0786
		λ	0.2500	0.3144	0.4678	0.3789	0.4233	0.4233	0.5122	0.4678	0.3589	0.4033	0.3589	0.3989
25	C_{t-cl}	θ_1	0.5522	0.8527	0.8870	0.6611	0.6411	0.9011	0.9900	0.9678	0.9011	0.8184	0.4278	0.9900
		θ_2	0.2502	0.0722	0.0100	0.0100	0.0322	0.2944	0.2700	0.2056	0.2500	0.2500	0.0767	0.1473
		λ	0.3144	0.4033	0.4678	0.4678	0.5767	0.4678	0.6211	0.5767	0.4678	0.4233	0.5322	0.4678
26	C_{t-gu}	θ_1	0.4478	0.4678	0.8789	0.9900	0.9722	0.7154	0.9678	0.9214	0.9011	0.6811	0.7544	0.9900
		θ_2	0.0786	0.0544	0.0443	0.0786	0.1473	0.2056	0.3589	0.2500	0.2500	0.2500	0.2056	0.1130
		λ	0.1211	0.1811	0.3833	0.2846	0.2256	0.1473	0.0443	0.0786	0.0786	0.1130	0.0544	0.2846
27	C_{t-jo}	θ_1	0.4033	0.4678	0.8184	0.7944	0.9722	0.4878	0.7154	0.1473	0.4033	0.2900	0.9214	0.9900
		θ_2	0.0322	0.0322	0.0322	0.0544	0.1473	0.1473	0.2500	0.2056	0.1816	0.1816	0.1473	0.0443
		λ	0.1811	0.2056	0.3344	0.2456	0.1167	0.1367	0.2056	0.0100	0.1611	0.1367	0.0786	0.3833

Table 3.12: Calibration of parameters of copulae, i.e. θ_1, θ_2, λ. Abbreviations: ga: Gaussian, t: Student-t, fr: Frank, cl: Clayton, gu: Gumbel, jo: Joe.

Model	*Notation*	*Parameter*	*20071023*	*20071026*	*20071117*	*20071206*	*20080111*	*20080228*	*20080314*	*20080405*	*20080424*	*20080529*	*20080530*	*20080701*
28	C_{fr-fr}	θ_1	0.7744	0.0100	0.0100	0.0100	0.7944	0.0443	0.7944	0.7944	0.0544	0.0322	0.0544	0.7744
		θ_2	0.0322	0.7944	0.7944	0.7944	0.0322	0.7944	0.0443	0.0443	0.7944	0.7744	0.7944	0.0544
		λ	0.2700	0.6856	0.6411	0.5967	0.4878	0.4433	0.6856	0.5122	0.4478	0.4033	0.5122	0.2256
29	C_{fr-cl}	θ_1	0.1473	0.0786	0.0767	0.0544	0.1473	0.1611	0.2056	0.1130	0.1816	0.1611	0.1473	0.0786
		θ_2	0.9900	0.9900	0.9900	0.9678	0.9900	0.9900	0.9900	0.9678	0.9900	0.9900	0.9678	0.9900
		λ	0.7300	0.6411	0.5767	0.5767	0.5767	0.5967	0.5322	0.5322	0.6211	0.5767	0.6211	0.5967
30	C_{fr-gu}	θ_1	0.0443	0.0767	0.0322	0.0322	0.1473	0.7056	0.7056	0.6611	0.1611	0.1816	0.1130	0.0786
		θ_2	0.2456	0.6456	0.6856	0.6211	0.7500	0.2500	0.3589	0.2944	0.8789	0.6656	0.2700	0.9900
		λ	0.5078	0.6211	0.4878	0.5122	0.5767	0.0100	0.0443	0.0100	0.5967	0.5567	0.1811	0.5522
31	C_{fr-jo}	θ_1	0.1211	0.0786	0.0322	0.0544	0.1130	0.2900	0.7544	0.5722	0.1816	0.1611	0.2502	0.0767
		θ_2	0.1130	0.7900	0.7056	0.7498	0.5722	0.2500	0.3589	0.2944	0.9900	0.9233	0.2056	0.9557
		λ	0.1130	0.6411	0.5322	0.5767	0.4878	0.0100	0.1211	0.0100	0.6211	0.5967	0.0100	0.6456
32	C_{cl-cl}	θ_1	0.1130	0.1611	0.9900	0.9678	0.9678	0.9557	0.9900	0.9900	0.9456	0.9900	0.9678	0.1473
		θ_2	0.9456	0.9900	0.0322	0.0100	0.2056	0.1816	0.2700	0.1611	0.2056	0.2256	0.2256	0.9900
		λ	0.6856	0.6411	0.4678	0.4678	0.4678	0.4433	0.5967	0.5767	0.4678	0.4233	0.3789	0.5322
33	C_{cl-gu}	θ_1	0.0786	0.6656	0.9900	0.0100	0.9900	0.6011	0.9900	0.9900	0.7900	0.9011	0.4433	0.1473
		θ_2	0.1130	0.0786	0.0544	0.5322	0.1473	0.2500	0.3589	0.2056	0.2500	0.2500	0.2056	0.9900
		λ	0.1167	0.0322	0.3789	0.4678	0.2700	0.0767	0.0443	0.1816	0.0443	0.1211	0.0544	0.5322
34	C_{cl-jo}	θ_1	0.1473	0.0322	0.0100	0.0544	0.9678	0.3144	0.8870	0.1367	0.3789	0.3389	0.4678	0.1473
		θ_2	0.1130	0.2056	0.4033	0.2700	0.1473	0.2056	0.2500	0.2500	0.2056	0.2056	0.1473	0.8527
		λ	0.2456	0.4678	0.5122	0.3789	0.0278	0.1167	0.2256	0.1816	0.0786	0.1473	0.1130	0.6211

TABLE 3.13: Calibration of parameters of copulae, i.e. θ_1, θ_2, λ. Abbreviations: ga: Gaussian, t: Student-t, fr: Frank, cl: Clayton, gu: Gumbel, jo: Joe.

Model	*Notation*	*Parameter*	*20071023*	*20071026*	*20071117*	*20071206*	*20080111*	*20080228*	*20080314*	*20080405*	*20080424*	*20080529*	*20080530*	*20080701*
35	C_{gu-gu}	θ_1	0.0322	0.8633	0.9278	0.1130	0.1473	0.8344	0.7841	0.8122	0.2056	0.5767	0.2056	0.0989
		θ_2	0.1130	0.1130	0.0989	0.9678	0.8189	0.2500	0.3144	0.2500	0.5522	0.2500	0.4433	0.9678
		λ	0.1611	0.0989	0.2456	0.7944	0.7300	0.0443	0.1816	0.0786	0.7498	0.1167	0.9456	0.6456
36	C_{gu-jo}	θ_1	0.0786	0.0544	0.8833	0.0786	0.1816	0.2944	0.3789	0.2500	0.2500	0.2500	0.6456	0.0443
		θ_2	0.1878	0.3789	0.0544	0.6611	0.6611	0.1611	0.2944	0.8527	0.6167	0.2700	0.1130	0.9233
		λ	0.8527	0.7544	0.2456	0.6611	0.8189	0.5122	0.5767	0.9214	0.9214	0.7500	0.2500	0.5078
37	C_{jo-jo}	θ_1	0.0443	0.0100	0.4678	0.0786	0.1130	0.2846	0.2500	0.1367	0.2159	0.1878	0.1367	0.9678
		θ_2	0.2100	0.1878	0.0322	0.1656	0.3544	0.1611	0.3144	0.3189	0.2056	0.2056	0.1473	0.0443
		λ	0.7544	0.5078	0.3789	0.7841	0.7154	0.3144	0.2700	0.4433	0.4278	0.0100	0.3389	0.4278
38	C_{t1}	ρ_1	0.1333	0.0444	0.1333	0.1333	0.2667	0.2667	0.8000	0.7111	0.2222	0.2667	0.2222	0.0444
		ρ_2	0.1778	0.3111	0.1333	0.1778	0.4000	0.4000	0.8444	0.8444	0.5333	0.4889	0.3556	0.1778
39	C_{t2}	ρ_1	0.0646	0.0364	0.0081	0.0606	0.0768	0.0929	0.0566	0.101	0.1616	0.1212	0.1333	0.0081
40	C_{t3}	ρ_1	-0.1333	-0.2667	-0.3111	-0.2222	-0.1333	-0.1778	-0.1333	-0.1333	-0.1333	-0.1333	-0.2222	-0.2667
		ρ_2	0.2667	0.3111	0.4444	0.4444	0.4889	0.4889	0.4444	0.4444	0.4444	0.4444	0.4889	0.4000
41	C_{t4}	ρ_1	0.0889	0.0889	0.0444	0.0889	0.0889	0.1778	0.1333	0.1333	0.1333	0.1333	0.1333	0.0444
		ρ_2	0.0889	0.0889	0.0444	0.0889	0.0889	0.1778	0.1333	0.1333	0.1778	0.1333	0.1333	0.0444
42	C_{t5}	ρ_1	0.1778	0.1778	0.2222	0.2222	0.2667	0.2667	0.2667	0.2667	0.2667	0.2222	0.2667	0.1778
		ρ_2	-0.2667	-0.3111	-0.3111	-0.3111	-0.2222	-0.1333	-0.1333	-0.1333	-0.2222	-0.1333	-0.2222	-0.3111
		ρ_3	0.4444	0.4889	0.7111	0.7111	0.6222	0.5333	0.7111	0.5778	0.6222	0.6667	0.6222	0.6222
43	C_{t6}	ρ_1	-0.1778	-0.2222	-0.3556	-0.3111	-0.1778	-0.1778	-0.0889	-0.1333	-0.0889	-0.1333	-0.1778	-0.2667
		ρ_2	0.2667	0.2222	0.5333	0.5333	0.3556	0.4889	0.4000	0.4000	0.3556	0.4444	0.4000	0.3111
		ρ_3	0.3111	0.4889	0.5778	0.5333	0.4000	0.6667	0.4444	0.4000	0.3556	0.4444	0.4000	0.5778

TABLE 3.14: Calibration of parameters of copulae, i.e. θ_1, θ_2, λ. Abbreviations: *ga*: Gaussian, *t*: Student-*t*, *fr*: Frank, *cl*: Clayton, *gu*: Gumbel, *jo*: Joe.

Chapter 4

Nonparametric Multivariate Control Chart for Financial Surveillance

This chapter is based on the paper ”A Nonparametric Multivariate Statistical Process Control Chart for Financial Surveillance” by O. Okhrin and Y.F. Xu (2017), submitted.

4.1 Introduction

Control chart plays a pivotal role in statistical process monitoring. Natural assumption of the sequence of $d-$dimensional vectors $X_1, \dots, X_t$, is identically independently distributed. The assumption of identical distribution is however, not always fulfilled, and different sub-sequences of vectors follow different distribution, e.g. portfolio of stock returns before and after crisis, characteristics of the product before and after re-calibration of the production machine, etc. In practice the number of these change points is often unknown, i.e. when exactly the crisis started, when the production machine was decalibrated, etc. Hence the problem, that the control chart tackles, is to identify these change points, what formally can be considered as separation of the series $X_1, \dots, X_t$ into diverse segments, where each adjacent pair of segments follows different distributions.

In the early stage, feature research on statistical process control chart can be referred to seminal papers by Shewhart (1931), Shewhart and Deming (1939), Page (1954), Roberts (1959). Since multivariate process becomes useful and common in practical quality engineering (Woodall and Montgomery (2014)) in recent decades, numerous papers have contributed to forward statistical process control (SPC) in multivariate context. A part of research is based on parametric assumptions, such as Crosier (1988) for multivariate CUSUM and Lowry et al. (1992) for multivariate EWMA and Zou and Tsung (2011) with underlying multivariate Gaussian distribution. Qiu and Hawkins (2001), Qiu and Hawkins (2003), Hawkins and Deng (2010) developed change point models with assumed pre-knowledge in in-control distribution. Another part of research focusing on online nonparametric multivariate change point models can be found in Zou et al. (2012), Holland and Hawkins (2014) and Zhou et al. (2015). A special accumulation of recent papers on nonparametric control chart can be referred to Chakraborti et al. (2015). An interested reader finds a comprehensive review of nonparametric control chart in Qiu (2017).

For a proper detection of the changes, different statistical tests with different advantages and disadvantages were used, e.g. Student-t test, Bartlett test and Generalized Likelihood Ratio test, see Hawkins et al. (2003), Hawkins and Zamba (2005a), and Hawkins and Zamba (2005b) respectively. Current research employs the energy test, which is nonparametric, simple in implementation and has good power. Székely and Rizzo (2004), Zech and Aslan (2003), Székely and Rizzo (2013) investigated the energy statistic and the related test and performed the power analysis for distributional equality. Further, Kim et al. (2009) shows the satisfactory performance of the test in the rolling window scheme with fixed window size in detection of change points in image data. Matteson and James (2014) and James and Matteson (2015) employ energy test combined with two different clustering schemes in change point retrospective analysis, i.e. the batch analysis (Phase I).

This paper proposes an *Energy Test Control Chart* (ETCC) that is the nonparametric control chart for online detection of multiple change points in multivariate time series. ETCC gathers three attractive features, which in most other tests are fulfilled separately. Firstly, it is nonparametric, what implies no need of pre-knowledge on the process comparing with traditional parametric control charts. Secondly, this control chart monitors *multivariate* time series which is pervasive in practice, e.g. in financial portfolio management. Last but not least, ETCC controls for more general changes in multivariate time series, i.e. simultaneous surveillance of mean and covariance. Proposed ETCC does *online* monitoring, which can be applied life in many areas using real-time data. To the best of our knowledge, this is the first nonparametric control chart which can simultaneously monitor mean and covariance changes in the multivariate distribution in online fashion.

Methodologically, the ETCC was integrated with the maximum energy divergence based permutation test. The later uses discrepancy between empirical characteristic functions of two random vectors with the empirical distribution of the test statistic being obtained using permutation samples. This differs from the commonly used rank test. Afterwards, the sequential detection of change points can be conducted under the algorithm introduced by change point model proposed by Hawkins et al. (2003) to perform online detection.

The simulation study investigates the ETCC in detecting mean and covariance shifts (in multivariate Gaussian, Student-t, Gamma and Laplace distributions). The performance of the ETCC was compared with the benchmark control charts including the spatial rank based EWMA (SREWMA) by Zou et al. (2012), the self-starting multivariate minimal spanning tree (SMMST) based control chart by Zhou et al. (2015) and the nonparametric multivariate change point (NPMVCP) model based control chart by Holland and Hawkins (2014). Results indicates a very good performance of the proposed chart.

In real-data application, the ETCC was employed in financial surveillance, i.e. monitoring high dimensional financial portfolios. Three data sets were used, separately in 5, 29, and 90 dimensions. The time windows of all three data sets covered the time span of 2007-2010 what contains the global financial crisis, with window width of more than 1000 observations. The result shows that the new control chart is capable to detect fast the abnormal distributional changes in the financial market. For the purpose of reproducible research and practice of nonparametric online MSPC, we contributed an `R` package '`EnergyOnlineCPM`' based on this paper, see Xu (2017).

The paper is structured as follows. Section 2 introduces the methodology of the energy test and the preliminary of change point model in two diverse phases (Phase I and II) providing information on benchmark models. Simulation study, application study and their corresponding results are presented in Section 3 and 4 respectively. Section 5 concludes. Some supplementary materials about the data meta information are provided in appendix.

4.2 Methodology

4.2.1 Energy Test

The main constituent of every control chart is the underlying test which is used to control characteristics of interest mean, variance, or the whole distribution. In the very general set-up having d-dimensional vectors $X \sim F_X$ and $Y \sim F_Y$, we aim at testing $\mathrm{H}_0 : \; F_X = F_Y$ versus $\mathrm{H}_0 : \; F_X \neq F_Y$. It is known that the corresponding characteristic functions ϕ_X and ϕ_Y of X and Y respectively are uniquely determined from distribution functions, see Serfling (2009). Usage of the divergence between ϕ_X and ϕ_Y to control for difference between distributions F_X and F_Y becomes an applicable routine. To test directly for the equivalence of F_X and F_Y fails under the curse of dimensionality and often requires the knowledge of F_X and F_Y. Székely and Rizzo (2005) used an integrated weighted distance between two characteristic functions, and showed that the larger the distance the higher the probability that the two random vectors are not identically distributed, i.e. $F_X \neq F_Y$.

Theorem 4.1. *Let $X \sim F_X$ and $Y \sim F_Y$ be two d-dimensional random vectors. X', Y' are independent copies of X and Y. The corresponding characteristic functions of the two random vectors are ϕ_X and ϕ_Y. If $0 < \alpha < 2$ with $\mathbb{E}||X||_2^\alpha < \infty$ and $\mathbb{E}||Y||_2^\alpha < \infty$ then*

$$\int_{\mathbb{R}^d} \frac{|\phi_X(p) - \phi_Y(p)|^2}{||p||_2^{d+\alpha}} dp = W(d,\alpha)\mathcal{E}^\alpha(X,Y), \tag{4.1}$$

where

$$\begin{aligned} W(d,\alpha) &= \frac{2\Pi^{\frac{d}{2}}\Gamma(1-\frac{\alpha}{2})}{\alpha 2^\alpha \Gamma(\frac{\alpha+d}{2})}, \textit{with } \Gamma(\cdot) \textit{ being the Gamma function,} \\ \mathcal{E}^\alpha(X,Y) &= 2\mathbb{E}||X-Y||_2^\alpha - \mathbb{E}||X-X'||_2^\alpha - \mathbb{E}||Y-Y'||_2^\alpha. \end{aligned} \tag{4.2}$$

Proof. See Lemma 1 in Appendix of Székely and Rizzo (2005). □

Theorem 4.2. *Under assumptions of Theorem 1, $\mathcal{E}^\alpha(X,Y) = 0$ iff X and Y are identically distributed.*

Proof. See Theorem 2 (ii) in Székely and Rizzo (2005). □

Therefore the metric $\mathcal{E}^\alpha(X,Y)$ can be used to measure the divergence between two distributions. Let the samples of random vectors X, Y be $S_X = \{x_1, \ldots, x_m\}$ and $S_Y = \{y_1, \ldots, y_n\}$ respectively. The empirical counterpart of (4.2) replaces expectations by the averages and leads to

$$\begin{aligned} \hat{\mathcal{E}}^\alpha(S_X, S_Y) &= \frac{mn}{m+n}\left(\frac{2}{mn}\sum_{i=1}^m\sum_{j=1}^n ||x_i - y_j||_2^\alpha \right. \\ &\left. -\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m ||x_i - x_j||_2^\alpha - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n ||y_i - y_j||_2^\alpha\right). \end{aligned} \tag{4.3}$$

From Theorem 2 one sees that the larger the value of $\hat{\mathcal{E}}^\alpha(S_X, S_Y)$ the higher is the likelihood that the components in S_X, S_Y are from different distributions. Hence $\hat{\mathcal{E}}^\alpha(S_X, S_Y)$

can be used as the test statistics. Since the theoretical distribution of the test statistics $\hat{\mathcal{E}}^{\alpha}$ is intractable, the permutation test is employed under the assumption of independent random vectors. In order to accomplish this, $P = (m+n)!$ bootstrap samples $(B, C) = (b_1, \dots, b_m, c_1, \dots, c_n) = (a_{(1)}, \dots, a_{(m+n)})$ can be generated by random shuffling of $\{a_1, \dots, a_{m+n}; a_i = x_i, i \in 1, \dots, m, a_i = y_{i-m}, i = m+1, \dots, m+n\} = \{x_1, \dots, x_m, y_1, \dots, y_n; x_i, y_i \in \mathbb{R}^d\}$. For every permutation sample (B, C) the energy test statistics $\hat{\mathcal{E}}^{\alpha}(B, C)$ is calculated, what leads to a P-vector of test statistics based on P different permutation samples. It allows to compute the empirical distribution of $\hat{\mathcal{E}}^{\alpha}(S_X, S_Y)$. The critical value can be then obtained by choosing a quantile following the given confidence level. For more details on the permutation test and its related empirical distribution please refer to Fisher (1937) and Pitman (1938).

4.2.2 Benchmark Tests

In this sub-section, test used in three recent nonparametric control charts are briefly reviewed, including the SREWMA by Zou et al. (2012), the SMMST by Zhou et al. (2015) and the NPMVCP by Holland and Hawkins (2014). Control charts based on these tests are considered as benchmark in later studies.

SREWMA by Zou et al. (2012)

Proposed by Zou et al. (2012) a nonparametric multivariate EWMA control chart is based on the spatial rank test to monitor the changes in the location parameter. It assumes that for a sequence of random vectors $X_{-g+1}, \dots, X_0, X_1, \dots, X_t \in \mathbb{R}^d$, where $X_{-g+1}, \dots, X_0$ are g vectors before the starting point X_1, the multivariate change point problem is represented as:

$$X_i \overset{\text{i.i.d.}}{\sim} \begin{cases} \mu_0 + \Omega\varepsilon_i & \text{if } i \le \tau, \\ \mu_1 + \Omega\varepsilon_i & \text{if } i > \tau, \end{cases} \tag{4.4}$$

where τ stands for the change index, Ω for a full-rank $d \times d$ transformation matrix with the inverse $M = \Omega^{-1}$. It is assumed that $\varepsilon_i \in \mathbb{R}^d$ are i.i.d. with $\text{Cov}(\varepsilon_i) = I_d$ and $\mathbb{E}(\varepsilon_i) = 0$. The test statistics and its asymptotic distribution are given as

$$Q_t^{R_E} = \frac{(2-\lambda)d}{\lambda\xi_t} V_t^{\top} V_t \to \chi_d^2, \ \lambda \to 0, \ \lambda t \to \infty, \tag{4.5}$$

with

$$\begin{aligned} V_t &= (1-\lambda)V_{t-1} + \lambda R_E(\hat{M}_{t-1}X_t),\ V_0 = 0, \\ \xi_t &= \hat{\mathbb{E}}\{R_F(MX_t)^\top R_F(MX_t)\}, \end{aligned}$$

where $\hat{M}_t = \hat{\Omega}_t^{-1}$, $R_E(X_t) = \frac{1}{g}\sum_{j=1}^{g} U(X_t - X_j)$ is the empirical version of the spatial rank for the $d-$vector X_t, and the theoretical counterpart is $R_F(X_t) = \mathbb{E}_{X_j}\{U(X_t - X_j)\}$ with U being the spatial sign function

$$U(X) = \begin{cases} (X^\top X)^{-1/2}X, & \text{for } X = 0, \\ 0 & \text{else}. \end{cases} \tag{4.6}$$

SMMST by Zhou et al. (2015)

The multivariate version Wald-Wolfowitz runs test by Friedman and Rafsky (1979) has been integrated into the (Hawkins et al. (2003)) control chart based on change point model as in Hawkins et al. (2003) by Zhou et al. (2015) to perform nonparametric multivariate location surveillance. The main idea of the multivariate Wald-Wolfowitz runs test is to use the minimal spanning tree (MST) approach to generalize the sorted list in univariate runs test to the multivariate context. That is, in the d-dimensional data set with $m+n$ observations $\{a_1, \ldots, a_{m+n}; a_i = x_i, i \in 1, \ldots, m, a_i = y_{i-m}, i = m+1, \ldots, m+n\} = \{x_1, \ldots, x_m, y_1, \ldots, y_n; x_i, y_i \in \mathbb{R}^d\}$ stemming from random vectors X and Y respectively, every observation is seen as a node and all the nodes can be connected by $(m+n)(m+n-1)/2$ edges. Friedman and Rafsky (1979) gives three steps to compute the test statistic:

1. Use the MST algorithm to construct the MST for all nodes in the data set, see Appendix in Friedman and Rafsky (1979).
2. Remove all edges, of which the two nodes are from different groups.
3. Compute the statistics $R = \#\{\text{disjoint sub-trees in the MST}\}$.

Further the change point problem in Zhou et al. (2015) is represented as

$$X_i \overset{\text{i.i.d.}}{\sim} \begin{cases} F_X & \text{if } i \leq \tau, \\ F_Y & \text{if } i > \tau. \end{cases} \tag{4.7}$$

The null hypothesis $\mathrm{H}_0 : F_X = F_Y$, will be rejected if R is smaller than a critical value, which follows normal distribution as

$$W = \frac{R - \mathbb{E}(R)}{\sqrt{\mathrm{Var}(R|C)}} \to N(0,1), \text{if } m, n \to \infty,$$

where C is determined by the node degrees and explicit formulas for expectation and variance can be found in Zhou et al. (2015).

NPMVCP by Holland and Hawkins (2014)

A nonparametric control chart using multivariate rank based test by Choi and Marden (1997) is proposed in Holland and Hawkins (2014). It gives the multivariate change point model (Hawkins et al. (2003)) to identify changes in a sequence, $X_1, \dots, X_t$, as follows

$$X_i \sim \begin{cases} F(\mu), & \text{if } i \le \tau, \\ F(\mu + \delta), & \text{if } i > \tau, \end{cases} \tag{4.8}$$

and $\mathrm{H}_0 : \delta = 0$, versus $\mathrm{H}_1 : \delta \neq 0$. The test statistics and its asymptotic distribution are given for $k \in \{1, \dots, t-1\}$ as:

$$\frac{tk}{t-k} \bar{r}_t^{(k)^\top} \tilde{\Sigma}_{k,t}^{-1} \bar{r}_t^{(k)} \to \chi_d^2, \text{ if } t \to \infty, \tag{4.9}$$

where $\tilde{\Sigma}_{k,t}$ is the pooled sample covariance matrix of the centered rank vector $\bar{r}_t^{(k)}$ computed using kernel function. At last Holland and Hawkins (2014) uses the test statistic

$$r_{k,t} \quad = \quad \bar{r}_t^{(k)^\top} \hat{\Sigma}_{k,t}^{-1} \bar{r}_t^{(k)},$$

where $\hat{\Sigma}_{k,t} = (\frac{t-k}{tk})\hat{\Sigma}_t$ is the unpooled estimator of covariance matrix of centered ranks. Simulation study by Holland and Hawkins (2014) shows that the power of using pooled or unpooled estimator of covariance matrix leads to similar performance. However for convenience of computation the unpooled covariance estimator $\hat{\Sigma}_{k,t}$ is employed.

4.2.3 Phase I Change Point Model

Let $\{x_1, ..., x_T\}$ denote a sample of observations with length of T. In Phase I detection and no new observation appears, detection is performed only based on sample $\{x_1, ..., x_T\}$ as historical data. Hence this type change point analysis is retrospective and static. Phase I analysis has many applications in bio-statistics and transportation statistics, see Székely and Rizzo (2005) and Matteson and James (2014).

For the sake of simplicity consider the case with only one change occurred at $\tau + 1$, then the change point detection problem can be represented in the following test hypotheses,

$$\begin{aligned} \mathrm{H}_0: &\quad X_i \sim F_0,\ 1 \leq i \leq T, \\ \mathrm{H}_1: &\quad X_i \sim \begin{cases} F_0, & 1 \leq i \leq \tau, \\ F_1, & \tau + 1 \leq i \leq T. \end{cases} \end{aligned}$$

A two-sample parametric or nonparametric test similar to those discussed in the previous section with test statistics $B_{i,T}$ is usually applied here. Before conducting the permutation test the significance level should be fixed. If $B_{i,T}$ is larger than a predefined critical value $h_{i,T}$, i.e. $B_{i,T} > h_{i,T}$, then the null hypothesis is rejected, meaning that the two sets of random vectors are not identically distributed. Then a detection point is admitted at i-th point. Since the change point location is unknown, hence the two-sample test will be performed at every point i, $1 \leq i < T$, i.e. conducting $T - 1$ dichotomizations. According to the change point model (Hawkins et al. (2003)), the test statistics is derived from $B_{i,T},\ i = 1, \ldots, T-1$, as the largest value, such that

$$B_T = \max_{1 \leq i < T} B_{i,T}.$$

The null hypothesis is rejected if $B_T > h_T$, where h_T is the critical value derived from the distribution of B_T. The Type I error α in this context means that the model signals a change point when actually there is actually no change occurs. The distribution of the test statistic B_T can be obtained either by its asymptotic distribution (if available) or by simulation methods e.g. permutation test scheme. At the end, the location of the change can be estimated by

$$\hat{\tau} = \arg \max_{1 \leq i < T} B_{i,T}.$$

4.2.4 Phase II Change Point Model

In contrary to the Phase I detection based on the fixed-sized sample $\{x_1, ..., x_T\}$, Phase II detection considers the dynamic sample $\{x_1, ..., x_t\}$ with an increasing size, i.e. the sample size t increases with time proceeding. For this reason Phase II detection is also termed as online detection and sequential detection, e.g. the stock price is updated with time, therefore the length of time series of returns is always increased. Hence the detection in Phase II concentrates on the dynamic stream data.

With the Phase I analysis in Section 4.2.3 at hand, Phase II can be extended from the Phase I to update the old sample size. That is whenever a new observation x_t arrives, a new sample $\{x_1, \ldots, x_T, x_{T+1}, \ldots, x_t\}$ is constructed and the new sample size is denoted here as t. For example, if the old sample is $\{x_1, \ldots, x_T\}$ with $t = T$ and the new arrival is x_{T+1}, then the new sample becomes $\{x_1, \ldots, x_T, x_{T+1}\}$ and t becomes $t = T + 1$. For every new arrival of observation the Phase I analysis will be performed based on the new sample $\{x_1, \ldots, x_T, x_{T+1}, \ldots, x_t\}$. For this sample, $t - 1$ two-sample tests will be performed and computed. Further $B_t = \max\{B_{1,t}, \ldots, B_{t-1,t}\}$. Hence the null hypothesis is rejected if $B_t > h_t$. The Type I error α can be thus represented with

$$\begin{aligned} \mathbb{P}(B_1 > h_1) &= \alpha, \ t = 1, \\ \mathbb{P}(B_t > h_t | B_{t-1} \leq h_{t-1}, \ldots, B_1 \leq h_1) &= \alpha, \ t > 1. \end{aligned} \tag{4.10}$$

In statistical process control, the in-control average run length (ARL_0), is the inverse of the Type I error, i.e. $ARL_0 = 1/\alpha$, which stands for the average step length of the detection until the first erroneous alarm signals.

4.3 Simulation Study

In the study of statistical process monitoring, the assessment of change-point detection methods uses mainly two measures, the in-control average run length (ARL$_0$) and the out-of-control average run length (ARL$_1$). ARL$_0$ assumes that the time series follows a distribution without changes in order to calculate the steps until the first erroneous signal flags, therefore the larger the ARL$_0$ the better the model. ARL$_1$ assumes that the process has a change point in a known point in order to compute the average run length until the model detects this pre-set change. Since there is a delay in detection, the detection method, therefore, is expected to have a small ARL$_1$ value.

The recent paper studying nonparametric multivariate control chart using the change point model (Hawkins et al. (2003)) is NPMVCP in Holland and Hawkins (2014). Therefore, we choose NPMVCP as the benchmark model for comparison in this paper, which is a mainstream nonparametric control chart for multivariate location shift detection. Since this paper used the code provided in `R` package `NPMVCP` in Holland and Hawkins (2014) without the usage of optimal quarantine technique, for the fair comparison, the quarantine was not considered for both models. Here the in-control length (ICL) is set as 32, and out-of-control length (OCL) is set separately as 100 and 200 consistent to the default set-up in `NPMVCP`. Choosing warm-up equal to 32, because firstly the defualt set-up in `NPMVCP` is fixed in 32 and secondly 32 is short for re-starting the control chart which is especially crucial in finanical surveillance.

As the test integrated in the ETCC is based on permutation samples, hence the choice of the number of simulation runs should be considered. As all the metrics were computed based on the i.i.d. samples, the mean of ARL_1s will converge under the law of large numbers. In order to choose an appropriate size of simulation, a simple simulation study was conducted. The DGP is a five dimensional standard Gaussian distribution, $\mathrm{N}_5(0, \mathcal{I})$, shifted to $\mathrm{N}_5(3, \mathcal{I})$, where $\mathcal{I}$ is the identity matrix and the warm-up is set to $\tau = 32$ identical to the setting in the package `NPMVCP`. As can be seen in Figure 4.1 the simulation runs larger than 50 led to the similar results and the mean of both control charts' ARL_1s arrived closely to the run of 50. Hence in this paper, the simulation size was chosen as 50 runs for sufficiency.

In the next three scenarios, we consider shifts in mean (whole vector and single constituent) and variance and compare the performance with NPMVCP. In the fourth scenario with mean shift we compare ETCC with SMMST and SREWMA.

a) In the mean shift scenario, the detection assessment sets the break of $\tau = 32$, i.e. ICL $= 32$ and OCL $\in \{100, 200\}$, and the distributions used in simulation are $\mathrm{N}(0, \mathcal{I})$, Student-$t_5(0, \mathcal{I})$ and Laplace$(0, \Sigma_L)$, $\Sigma_L = (a_{ij})$, $a_{ii} = 11$, $a_{ij} = 10$. The shifts are set as $\delta = 0, 0.25, 0.5, 0.75, 1, \ldots, 9$. Hence the in-control distributions, $\mathrm{N}(0, \mathcal{I})$, Student-$t_5(0, \mathcal{I})$ and Laplace$(0, \Sigma_L)$, will shift to $\mathrm{N}(\delta, \mathcal{I})$, Student-$t_5(\delta, \mathcal{I})$ and Laplace(δ, Σ_L) at the 32nd observation. To emphasize the performance of the ETCC under in-control situation, we give in Table 4.1 ARL_0s for ETCC and NPMVCP with corresponding empirical standard deviations computed over the simulation runs. As can be seen ARL_0s for ETCC are almost equal in OCL (100 and 200) where ARL_0s for NPMVCP are twice smaller with often more than twice bigger variance. The ARL performance (ARL_0 for $\delta = 0$ and ARL_1 else) is shown in Figure 4.2. For all three distributions the ETCC performs better in moderate to large shifts ($\delta \geq 2$) for three dimensional cases and in small to large shifts ($\delta \geq 0.75$) in ten dimensional cases, see Gaussian and t_5. With the increase of the dimension of data the performance of ETCC is steadily improving.

b) In scenario of the single component mean shift, the breaks are set at $\tau = 32$, i.e. ICL $=$ 32 and OCL $\in \{100, 200\}$, and the distributions used in simulation are $\mathrm{N}(0, \mathcal{I})$, Student-$t_5(0, \mathcal{I})$ and Laplace$(0, \Sigma_L)$, $\Sigma_L = (a_{ij})$, $a_{ii} = 11$, $a_{ij} = 10$. The shifts are set as the $\delta = 0, 0.25, 0.5, 0.75, 1, \ldots, 9$. The shifting method here is similar to the one in the first scenario, but only the last column shifts by δ while the other columns are kept unchanged $\mathrm{N}((0, \ldots, 0, \delta)^\top, \mathcal{I})$, Student-$t_5((0, \ldots, 0, \delta)^\top, \mathcal{I})$ and Laplace$((0, \ldots, 0, \delta)^\top, \Sigma_L)$. Single component mean shift scenario shows that NPMVCP performs well in small shifts and the ETCC performs well in moderate shift ($\delta \geq 2$), see Figure 4.3. However in all the categories, the ETCC outperforms the NPMVCP in ARL_0. The NPMVCP has only roughly 60 percent correct detection, which is far worse than the ETCC, where the disadvantageous performance of NPMVCP is consistent with the result in Holland and

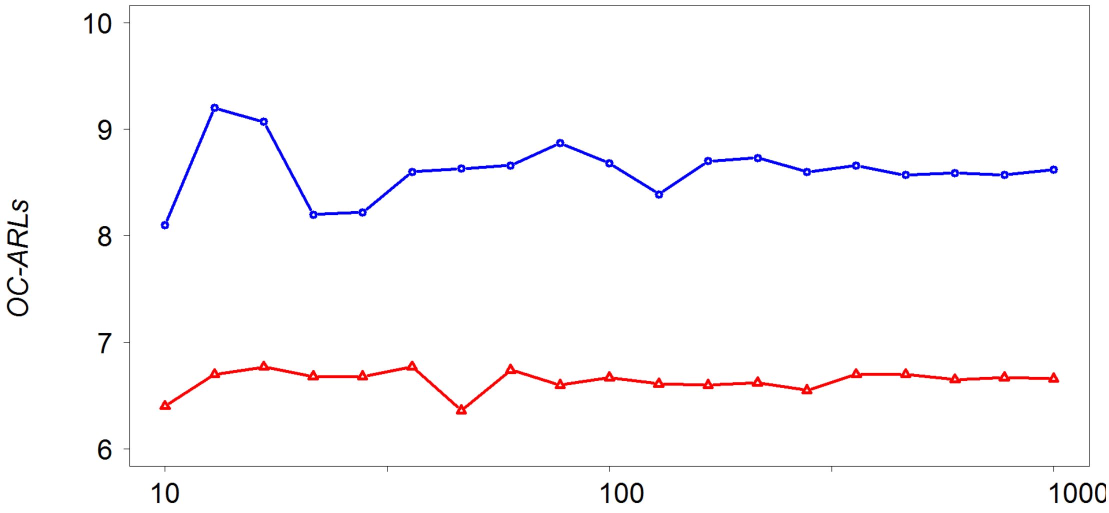


FIGURE 4.1: Comparison of ARL_1s under NPMVCP and ETCC through 10 to 1000 runs of simulation.

Hawkins (2014). According to the Table 4.1 and the above analysis, one can conclude that the ETCC in mean shift detection is capable and robust.

c) In covariance shift part, the DGPs are set as the $\text{N}(0, \mathcal{I})$ and Student-t_5 with $\sigma^2 = 0.25, 0.5, 0.75, 1, 2, \ldots, 11$. Here it means there is no change when $\sigma^2 = 1$. Hence the in-control distributions will shift to $\text{N}(0, \mathcal{I}\sigma^2)$ and Student-$t_5(0, \mathcal{I}\sigma^2)$. The ETCC outperforms the NPMVCP in most cases, see Figure 4.4 while NPMVCP has ability to detect the small covariance shift, e.g. in scale of $\sigma^2 = 2$. In larger covariance shifts or larger dimension data sets, the ETCC gave better results. The NPMVCP shows to be robust to the changes of dimensions or distributions, while the ETCC shows high sensitivity to the increase of dimension, and ARL_1 strongly decrease with dimension.

d) Additionally, as mentioned in Section 4.2.2, the ETCC is compared with another two nonparametric control charts, namely the SMMST and the SREWMA in scenario of 200 ARL_1-steps mean shift under Gaussian, t_5 and Gamma_5. Breaks τ are set as 40 and 90, and shifts $\delta = 1, 1.5, 2, 3, 4$. The results of SMMST and SREWMA are collected from the Table 2, 3, 4, 5 in Zhou et al. (2015). Using the same simulation setting in Zhou et al. (2015) we tested the performance of ETCC. Since our simulation based on independent samples hence for convenience we do not re-run the SMMST and SREWMA but just took the result from Zhou et al. (2015). In order to further support the robustness and capacity of the ETCC, Figure 4.5 provides another evidence. The ETCC performs generally better than the other benchmarks, especially in Gaussian and t_5 cases.

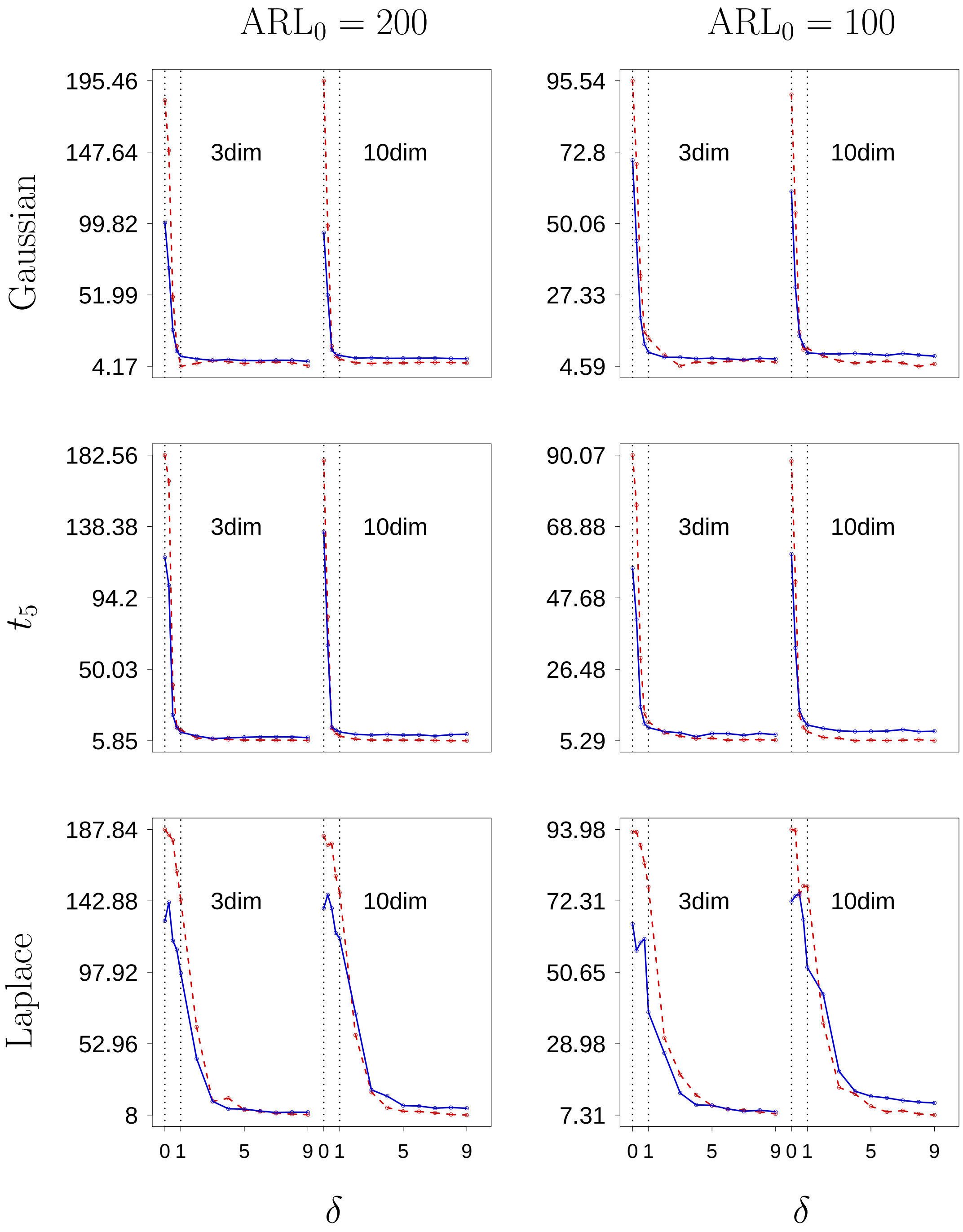


FIGURE 4.2: Simulation results for mean shift with DGPs of Gaussian, Student-t_5 and Laplace distributions. The blue line stands for NPMVCP and the red for ETCC.

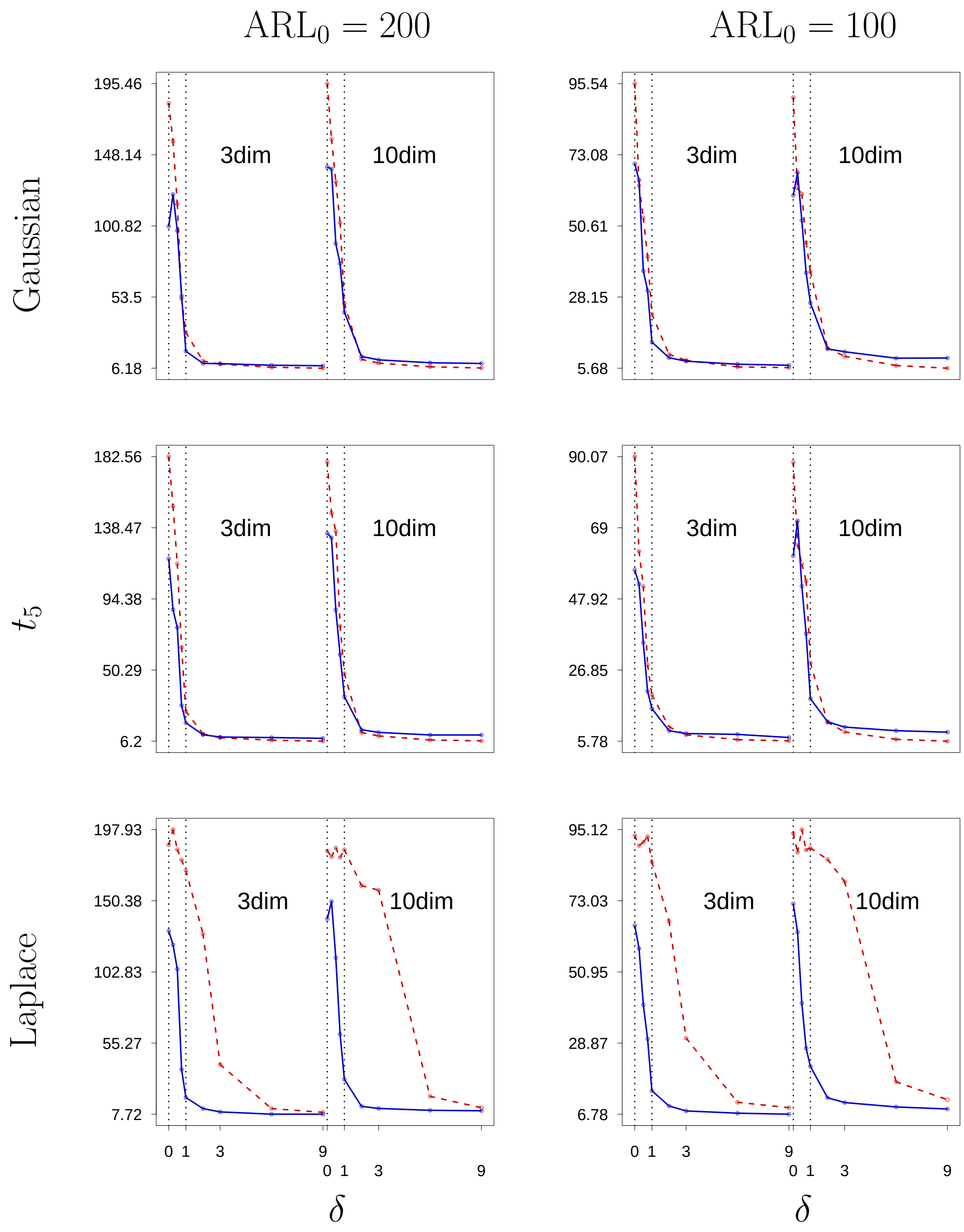


FIGURE 4.3: Single mean shift for multivariate Gaussian, Student-t_5 and Laplace with mean $\mu_k + \delta$, $\delta \in \{0, 0.25, 0.50, 0.75, 1, 2, 3, 6, 9\}$. The red line stands for the ETCC and the blue line for the Holland and Hawkins (2014).

		Gaussian		t		Laplace	
ARL_0	Dim.	ETCC	NPMVCP	ETCC	NPMVCP	ETCC	NPMVCP
200	3	182.36 (50.29)	124.82 (71.55)	182.56 (48.89)	118.66 (74.11)	187.84 (41.96)	135.62 (67.62)
	10	195.46 (19.71)	138.62 (70.29)	179.26 (52.06)	135.02 (69.45)	183.47 (45.77)	140.28 (67.84)
100	3	95.54 (16.91)	67.30 (40.25)	90.07 (27.13)	68.00 (34.34)	93.30 (20.45)	62.84 (35.35)
	10	91.13 (24.29)	58.24 (38.50)	88.38 (29.25)	74.12 (34.49)	93.98 (21.11)	69.34 (34.68)

Table 4.1: Comparison of ETCC against the NPMVCP model (Holland and Hawkins (2014)) in In-Control ARL for mean shift with 100 and 200 ARL_1-steps (with standard deviations in parentheses).

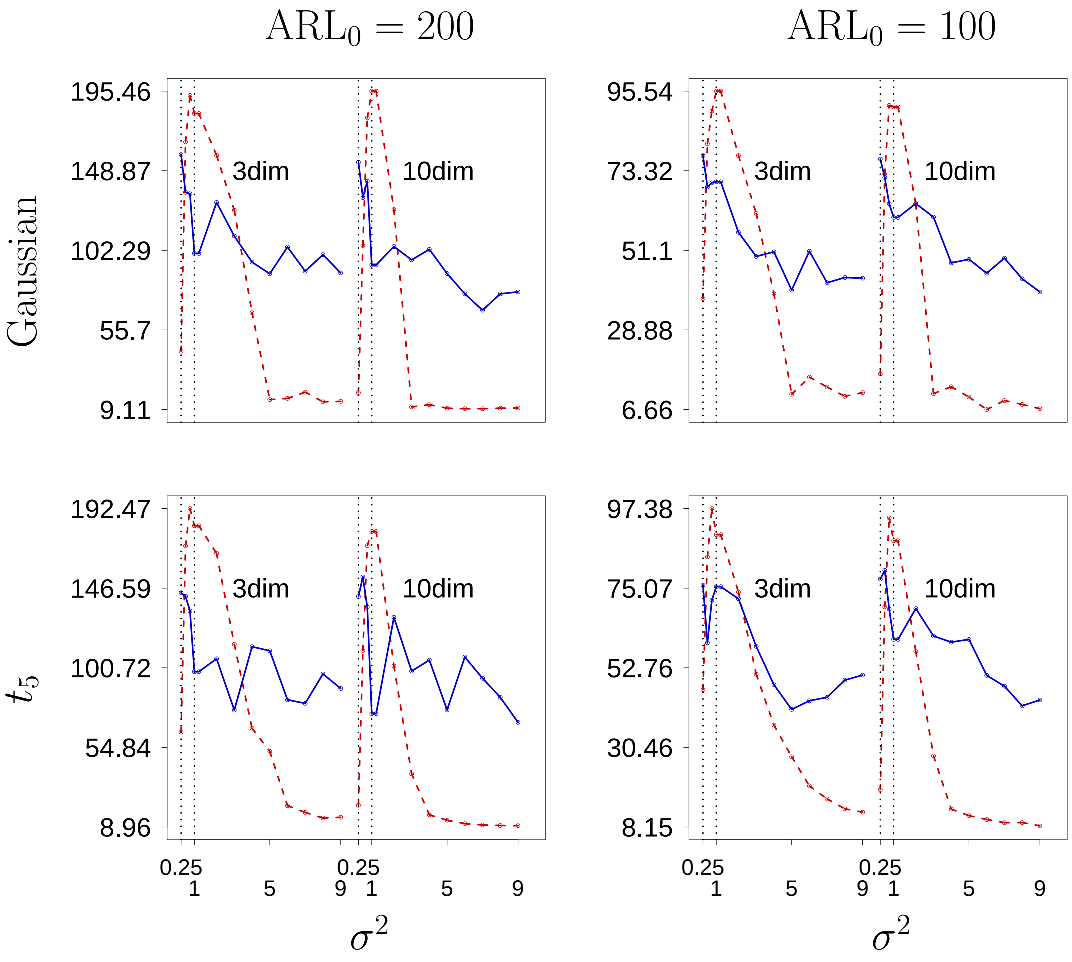


Figure 4.4: Simulation results for covariance shift with DGPs of Gaussian and Student-t_5. The blue line stands for NPMVCP and the red line for ETCC.

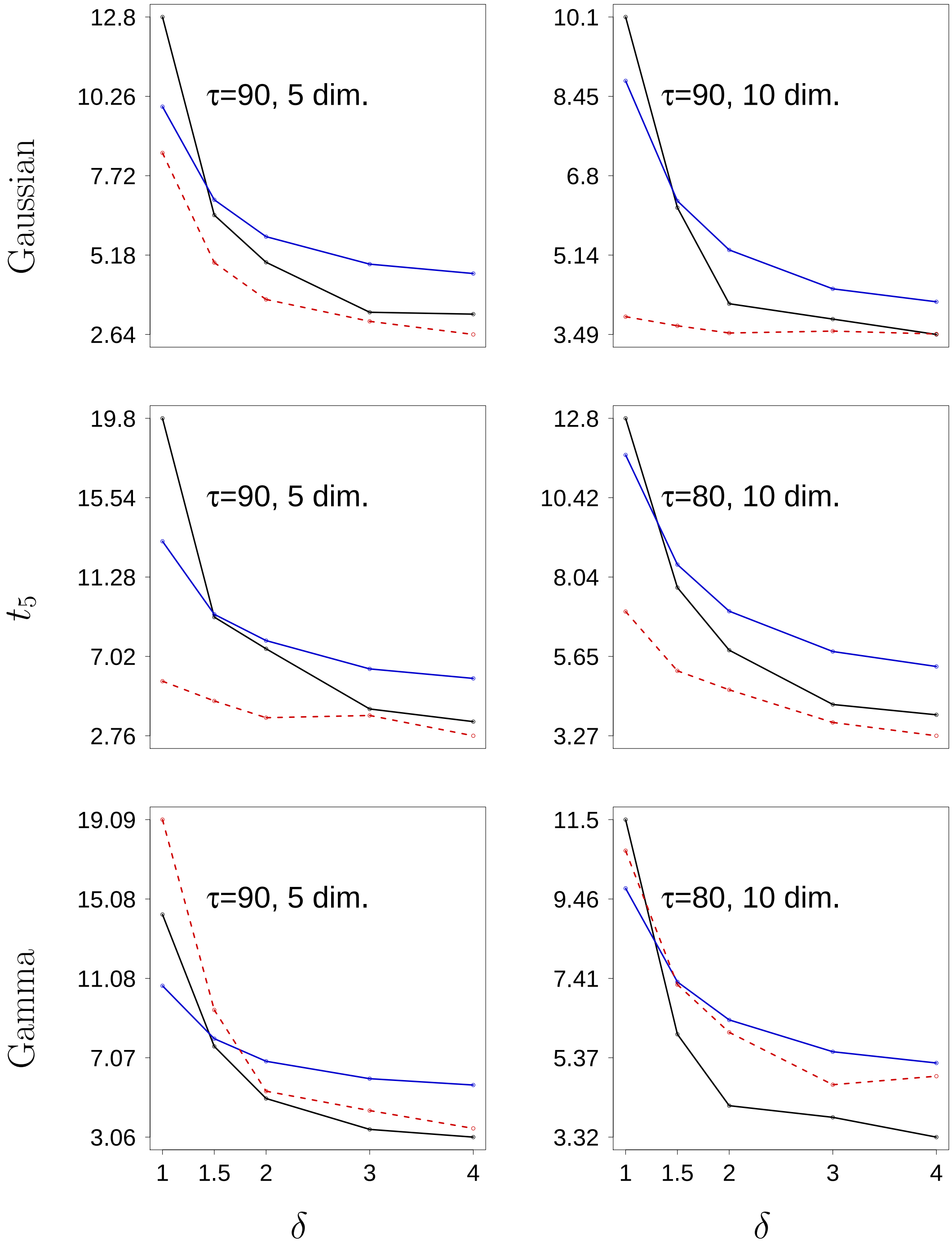


FIGURE 4.5: Comparison of simulation results of the ETCC (red) with SMMST (Zhou et al. (2015)) in black and SREWMA (Zou et al. (2012)) in blue appeared in Table 2, 3, 4 in Zhou et al. (2015).

4.4 Real Data Application in Financial Surveillance

4.4.1 Data Sets

In this section, three data sets were employed. The first data set contain close prices on the U.S. ETF (Exchange-Traded Fund) market of five tickers of DGT, EWD, GLD, IGV and IUSG, see Table 4.2 for names in Appendix. The data is obtained from the Wall Street Journal web site.

The second data set contains close prices from 29 components of DJIA (Dow Jones Industrial Average), see Table 4.3 in Appendix for the list. The third data set contains close prices from 90 components of S%P100, see Table 4.4 in Appendix. Both data sets were obtained from Yahoo Finance. The window length spans from 20070103-20101231, and contain 1007 observations for each data set. Therefore the global financial crisis occurred in 2008-2009 is covered by all three data sets, what is of our interest to check if the ETCC is capable to detect the market shift.

Visual representation of prices for both data sets is given in Figure 4.6, there is clearly visible sharp change prior to and during the crisis. Crucial assumption for the ETCC is the assumption of independent observations, that is often not met in practice, e.g. in Matteson and James (2014) and James and Matteson (2015). Therefore before using the ETCC, all three data sets need to be ajusted properly. In this work the VAR (Vector Auto Regression), see Sims (1980), was used to filter out the residuals from the raw data sets in the first step. We are aware of the fact that doing this some changes in the mean and variances are smoothed out and, therefore might be not detected. We hope, however, there was a more complex change in the joint distribution of the residuals that should be picked up by ETCC. Especially, according to the Figure 4.7, the volatility is the main reason of shift in multivariate residual sequences, therefore in this angle ETCC's strong detection capacity in covariance shift makes critical sense compared to NPMVCP, see Figure 4.4. VAR model generalizes the uni-variate AR model by allowing for more than one endogenous variable to capture the linear inter-dependency among multiple time series. A VAR of order z, denoted VAR(z), is given through

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_z y_{t-z} + e_t,$$

where c is a $d \times 1$ vector of constants, A_z is a time-invariant $d \times d$ matrix and e_t is a $d \times 1$ vector of error terms satisfying $\mathbb{E}(e_t) = 0$, $\mathbb{E}(e_t {e_t}^\top) = 0$ and $\mathbb{E}(e_t {e_{t-k}}^\top) = 0$.

After filtering out residuals by VAR model, multivariate portmanteau test (Hosking (1980)), was employed to test them for independence. The three data sets were fitted

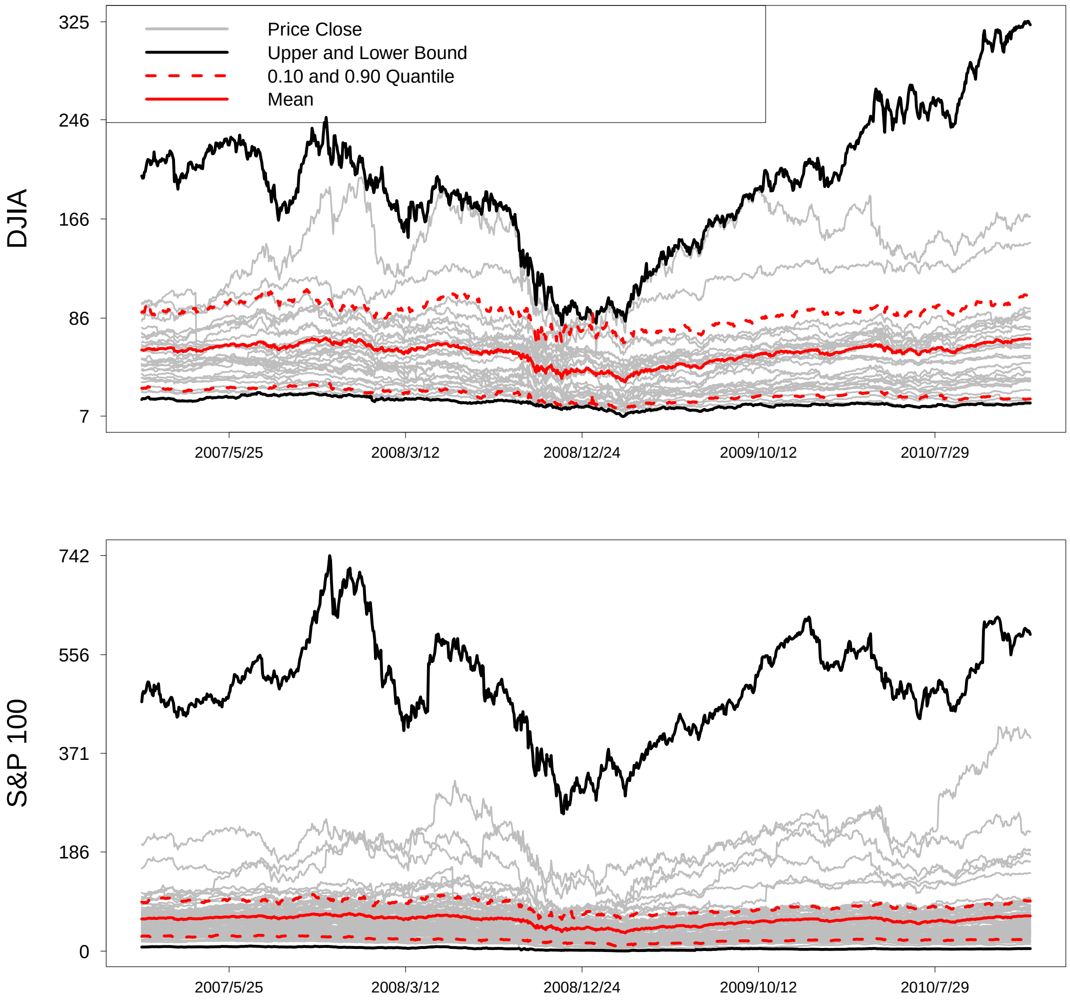


FIGURE 4.6: The upper panel presents the 29-dimensional DJIA data set. The lower panel illustrates the 90-dimensional SP100 data set.

by VAR models separately in VAR(10) for ETF data set, VAR(5) for DJIA and S&P100 data sets.

The ETCC is set in this application section with the warm-up of 32 observations and $\alpha = 0.005$ significant level, which are the same settings as in package `NPMVCP`.

4.4.2 Results and Analysis of Application

The real data applications can be seen as a complex scenario containing different changes (mean, covariance, etc.) of the whole distribution, therefore a control chart with the capacity to simultaneously detect different changes will have competitive edge. The findings in applications show similar to those in simulation study that the performance of the ETCC stands out.

First of all, the ETCC detected out the changes of the market regimes in five dimensional ETF data set. Figure 4.7 illustrate the detection points using the ETCC and the NPMVCP. The fact that the control chart by NPMVCP has more detection points than the proposed is supported by the results in simulation study. A possible reason is that the NPMVCP has more erroneous detection than the ETCC as depicted in the simulation study in the mean shift, see Table 4.1.

Secondly, the ETCC has strong detection power for covariance shift detection for high dimensional data, consistent with the results shown in simulation study. In Figure 4.7, it shows that NPMVCP has obvious large delay in detection of financial turmoil period (2008.09-2009.03). The first detection point for NPMVCP model on financial crisis is 20090205 with the location of the change point on 20081118, while the ETCC is on 20081007 detect the change point happened on 20081007. Therefore the ETCC can signal alarm for investors of the in-crisis, while the NPMVCP can not.

Thirdly, similar to the ETF data set, the ETCC detected out the change points of financial crisis in 29 and 90 dimensional data sets. In Figure 4.8 it shows the ETCC signaled detection points for in-crisis, separately on 20081005 for 29-dimensional DJIA data set (similar to result in James and Matteson (2015)) and on 20080917 for 90-dimensional S&P data set. Hence the ETCC can be used to serve as an alarm tool for the investors in financial market. In fact 20090309 is the lowest point in both DJIA and S&P100 indices, ETCC signaled on 20081005 for DJIA and 20080917 for S&P100, which means the ETCC gives the alarm of the market change in front of roughly five months. Then ETCC signaled the second change of the market regime on 20090901 (DJIA) and 20090916 (S&P100) separately, which means that the market is recovering upwards.

4.5 Conclusion

This paper proposes a nonparametric multivariate control chart to detect the multiple change points in high-dimensional stream data. It has four features. Firstly, it is a nonparametric control chart requiring no assumption on the process, compared with the classical parametric control chart. Secondly, it is oriented to Phase II change point detection which is central for real time surveillance of stream data and can be applied extensively, e.g. in industrial quality control, finance, medical science, geology et al. Thirdly, the control charts is designed for multivariate time series, which is more practical and informative for catching the essence of data as a whole than uni-variate time series.

Last but the most important feature of the ETCCs is that it monitors not only mean or only covariance, but monitors mean and covariance simultaneously, not separately.

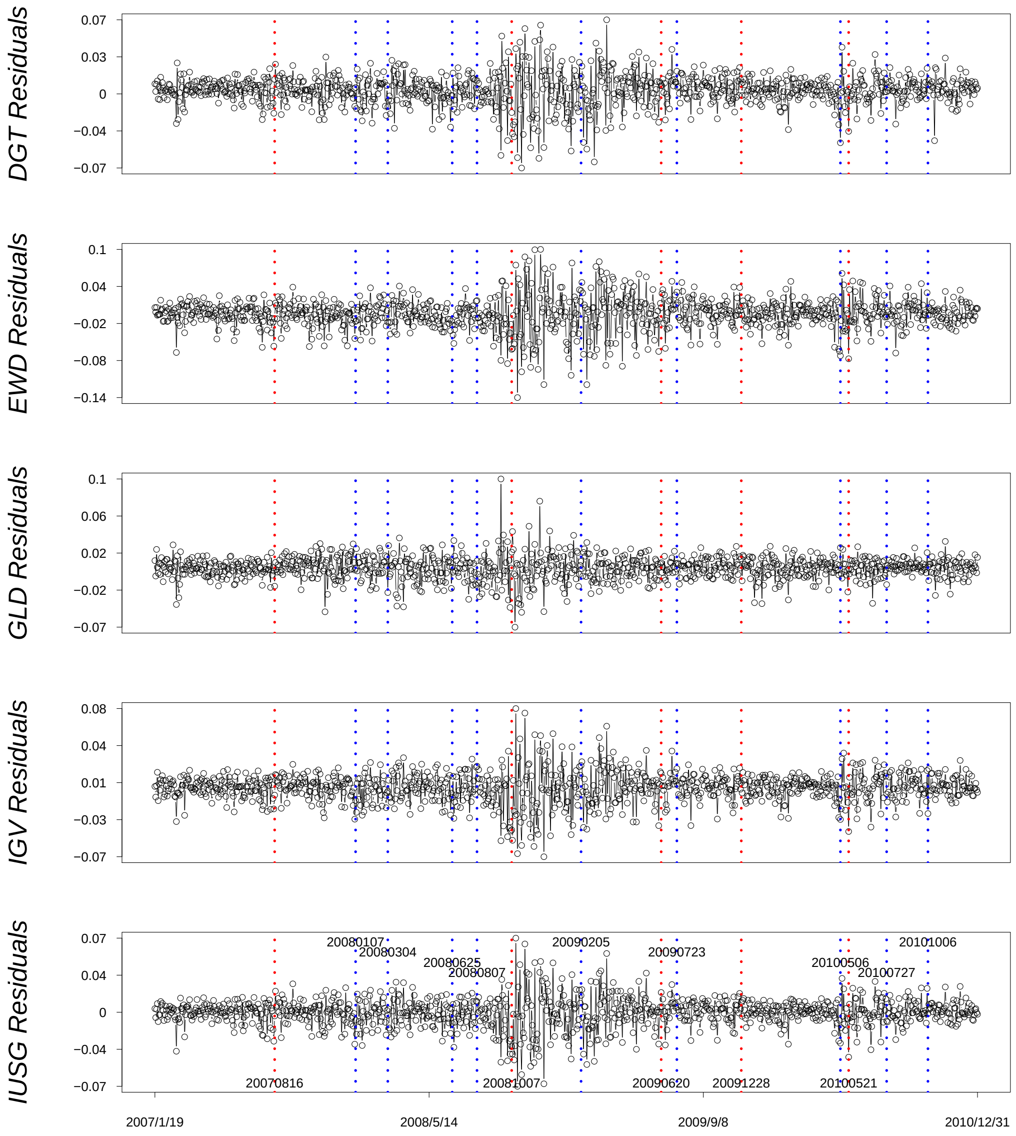


Figure 4.7: Change detection by the ETCC (red) and the NPMVCP (blue) control charts. DGT: SPDR Global Dow ETF, EWD: iShares MSCI Sweden Capped ETF, GLD: SPDR Gold Trust, IUSG: iShares Core SP U.S. Growth ETF, IGV: iShares North American Tech-Software ETF.

In simulation study the mean and covariance shifts were investigated and the control chart has shown outstanding performance compared to the benchmark models. In real data application, the ETCC was implemented for surveillance of three high-dimensional portfolios in 5, 29 and 90 dimensions separately. The ETCC shows the capacity to detect changes of the market regimes from quiescent period to volatile period, which provides reference to financial investors to take measures for the in-crisis investment. An `R` package '`EnergyOnlineCPM`' for Phase II nonparametric multivariate statistical process control is contributed in this paper.

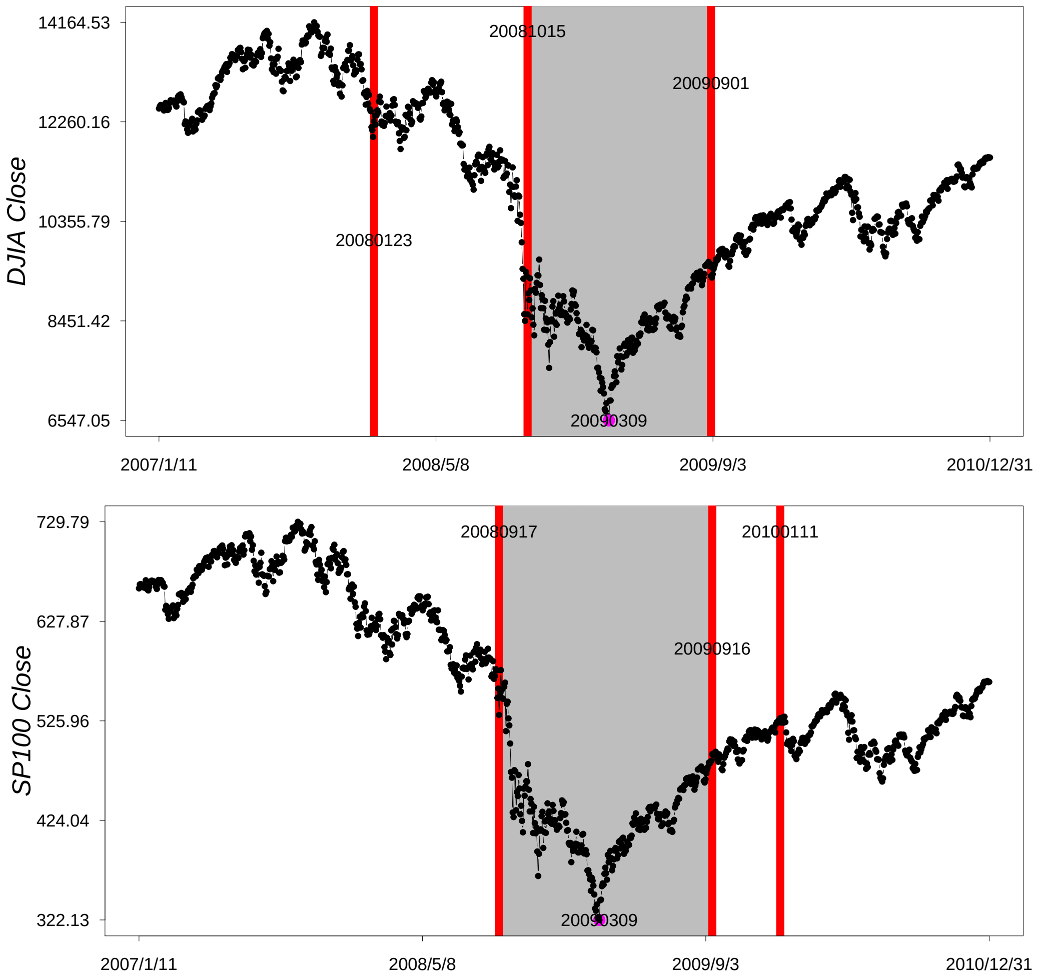


FIGURE 4.8: ETCC for detection points of DJIA (upper) and SP100 (lower) data sets. The red line stands for the detection point. The pink point stands for the lowest point in each index.

4.6 Information of Data Sets and Supplemental Tables

Symbol	Company
DGT	SPDR Global Dow ETF
EWD	iShares MSCI Sweden Capped ETF
GLD	SPDR Gold Trust
IGV	iShares Core S&P U.S. Growth ETF
IUSG	iShares North American Tech-Software ETF

TABLE 4.2: Related information of components of 5-dimensional data set of ETFs.

Company	Exchange	Symbol	Industry
Apple	NASDAQ	AAPL	Consumer electronics
American Express	NYSE	AXP	Consumer finance
Boeing	NYSE	BA	Aerospace anddefense
Caterpillar	NYSE	CAT	Construction andmining equipment
Cisco Systems	NASDAQ	CSCO	Computer networking
Chevron	NYSE	CVX	Oil & gas
DuPont	NYSE	DD	Chemical industry
Walt Disney	NYSE	DIS	Broadcasting andentertainment
General Electric	NYSE	GE	Conglomerate
Goldman Sachs	NYSE	GS	Banking,Financial services
The Home Depot	NYSE	HD	Home improvementretailer
IBM	NYSE	IBM	Computers andtechnology
Intel	NASDAQ	INTC	Semiconductors
Johnson & Johnson	NYSE	JNJ	Pharmaceuticals
JPMorgan Chase	NYSE	JPM	Banking
Coca-Cola	NYSE	KO	Beverages
McDonald's	NYSE	MCD	Fast food
3M	NYSE	MMM	Conglomerate
Merck	NYSE	MRK	Pharmaceuticals
Microsoft	NASDAQ	MSFT	Software
Nike	NYSE	NKE	Apparel
Pfizer	NYSE	PFE	Pharmaceuticals
Procter & Gamble	NYSE	PG	Consumer goods
Travelers	NYSE	TRV	Insurance
UnitedHealth Group	NYSE	UNH	Managed health care
United Technologies	NYSE	UTX	Conglomerate
Verizon	NYSE	VZ	Telecommunication
Walmart	NYSE	WMT	Retail
ExxonMobil	NYSE	XOM	Oil & gas

Table 4.3: Related information of components of 29-dimensional data set from DJIA.

Symbol	Company	Symbol	Company	Symbol	Company
ABT	Abbott Laboratories	EMR	Emerson Electric Co.	MS	Morgan Stanley
ACN	Accenture plc	EXC	Exelon	MSFT	Microsoft
AGN	Allergan plc	F	Ford Motor	NEE	NextEra Energy
AIG	American International Group Inc.	FDX	FedEx	NKE	Nike
ALL	Allstate Corp.	FOX	21st Century Fox	ORCL	Oracle Corporation
AMGN	Amgen Inc.	GD	General Dynamics	OXY	Occidental Petroleum Corp.
AMZN	Amazon.com	GE	General Electric Co.	PCLN	Priceline Group Inc/The
AXP	American Express Inc.	GILD	Gilead Sciences	PEP	Pepsico Inc.
BA	Boeing Co.	GOOG	Alphabet Inc	PFE	Pfizer Inc
BAC	Bank of America Corp	GS	Goldman Sachs	PG	Procter & Gamble Co
BIIB	Biogen Idec	HAL	Halliburton	QCOM	Qualcomm Inc.
BK	The Bank of New York Mellon	HD	Home Depot	RTN	Raytheon Company
BLK	BlackRock Inc	HON	Honeywell	SBUX	Starbucks Corporation
BMY	Bristol-Myers Squibb	IBM	International Business Machines	SLB	Schlumberger
C	Citigroup Inc	INTC	Intel Corporation	SO	Southern Company
CAT	Caterpillar Inc	JNJ	Johnson & Johnson Inc	SPG	Simon Property Group, Inc.
CELG	Celgene Corp	JPM	JP Morgan Chase & Co	T	AT&T Inc
CL	Colgate-Palmolive Co.	KO	The Coca-Cola Company	TGT	Target Corp.
CMCSA	Comcast Corporation	LLY	Eli Lilly and Company	TWX	Time Warner Inc.
COF	Capital One Financial Corp.	LMT	Lockheed-Martin	TXN	Texas Instruments
COP	ConocoPhillips	LOW	Lowe's	UNH	UnitedHealth Group Inc.
COST	Costco	MA	MasterCard Inc	UNP	Union Pacific Corp.
CSCO	Cisco Systems	MCD	McDonald's Corp	UPS	United Parcel Service Inc
CVS	CVS Health	MDLZ	Mondelez International	USB	US Bancorp
CVX	Chevron	MDT	Medtronic Inc.	UTX	United Technologies Corp
DD	DuPont	MET	Metlife Inc.	VZ	Verizon Communications Inc
DHR	Danaher	MMM	3M Company	WBA	Walgreens Boots Alliance
DIS	The Walt Disney Company	MO	Altria Group	WFC	Wells Fargo
DOW	Dow Chemical	MON	Monsanto	WMT	Wal-Mart
DUK	Duke Energy	MRK	Merck & Co.	XOM	Exxon Mobil Corp

Table 4.4: Related information of components of 90-dimensional data set from S&P100.

Chapter 5

`EnergyOnlineCPM`: An R Package for Nonparametric Control Chart

This chapter is based on the R package "`EnergyOnlineCPM`" by O. Okhrin and Y.F. Xu (2017) published in *The Comprehensive R Archive Network* (CRAN).

5.1 Introduction

For research of control chart, many R packages have been provided. In this section, an introduction of the proposed model based R package '`EnergyOnlineCPM`' is presented. A review of main control charts' R packages, installation of R package '`EnergyOnlineCPM`' and an example of usage are given in the following.

In nowadays many packages are devised for control chart. We review some main packages based on R programming language. Zeileis et al. (2005)'s '`strucchange`' is used for univariate change point analysis (Phase I) for mean monitoring. Erdman et al. (2007)'s '`bcp`' focused still on Phase I change point analysis for univariate data but used Bayesian method for mean surveillance. '`changepoint`' in Killick and Eckley (2011) is used for mean or/and variance monitoring based on (non)parametric model in Phase I. '`cpm`' in Ross et al. (2013) is used for Phase II analysis but only for univariate data set. '`spc`' in Knoth (2016) collects some parametric control chart models using for Phase II monitoring of mean or/and variance. '`NPMVCP`' in Holland (2013) is a package for multivariate data monitoring using a nonparametric model for surveillance of location changes. '`ecp`' in James and Matteson (2015) is used for uni/multivariate Phase I data using a nonparametric model to surveillance distribution changes.

Energy statistic (Székely and Rizzo (2004)) is attracting attention for empirical discrepancy of characteristic functions. At the moment there are two R packages for energy statistic, James and Matteson (2015) and Rizzo and Székely (2016). Rizzo and Székely (2016) is focused on the energy tests and James and Matteson (2015) concentrates on the Phase I change point model used for retrospective analysis. The package '`EnergyOnlineCPM`' is the first package which centers on the nonparametric Phase II change point model to online detect multiple change points for high dimensional time series based on the maximum energy test statistic using permutation samples.

5.2 Installation and Example

The installation of the package is convenient. The package is already published in CRAN (The Comprehensive R Archive Network) and it can be installed on the R terminal with following lines. Please note the package requires R version >= 3.3.2.

```
install.packages("EnergyOnlineCPM")
library(EnergyOnlineCPM)
```

Next we show an example of using '`EnergyOnlineCPM`' to detect a simulated data set with five dimensions. The data-driven-process is set as a process with three segments. The first segment has 20 readings following $\mathrm{N}(1_{5\times1}, \mathcal{I}_{5\times5})$. The second segment has 30 observations following $\mathrm{N}(2_{5\times1}, \mathcal{I}_{5\times5})$. The third segment follows $\mathrm{N}(1_{5\times1}, \mathcal{I}_{5\times5})$, the same with the first segment, but has 50 observations. Therefore the 20-th and 50-th points are two theoretical change points. The task for '`EnergyOnlineCPM`' is to detect these two points with least delayed steps. The script is given as follows.

```
library(MASS)
simNr = 300 # simulate 300 length time series

# simulate 300 length 5 dimensional standard Gaussian series
Sigma2 = matrix(c(1,0,0,0,0, 0,1,0,0,0,
                  0,0,1,0,0, 0,0,0,1,0, 0,0,0,0,1),5,5)
Mean2 = rep(1,5)
sim2 = (mvrnorm(n = simNr, Mean2, Sigma2))

# simulate 300 length 5 dimensional standard Gaussian series
Sigma3 = matrix(c(1,0,0,0,0, 0,1,0,0,0,
                  0,0,1,0,0, 0,0,0,1,0, 0,0,0,0,1),5,5)
```

```
Mean3 = rep(0,5)
sim3 = (mvrnorm(n = simNr, Mean3, Sigma3))

# construct a data set of length equal to 90.
# first 20 points are from standard Gaussian.
# second 30 points from a Gaussian with a mean shift with 2.
# last 40 points are from standard Gaussian.
data1 = rbind(sim2[1:20,], (sim3+2)[1:30,], sim2[1:50,])

# set warm-up number as 20,
# permutation 200 times, significant level 0.005
wNr    = 20
permNr = 200
alpha  = 1/200
maxEnergyCPMv(data1, wNr, permNr, alpha)
```

After running the codes above, a plot can be drawn, which shows the change points and detection points for the first univariate column in data set. The middle segment between blue lines shows a process with mean equal to 2, while the other two segments' means are all equal to 1. The two red lines give the detection time points. Installation, user manual, examples and more information can be referred to the user manual Xu (2017) and `https://sites.google.com/site/EnergyOnlineCPM/`.

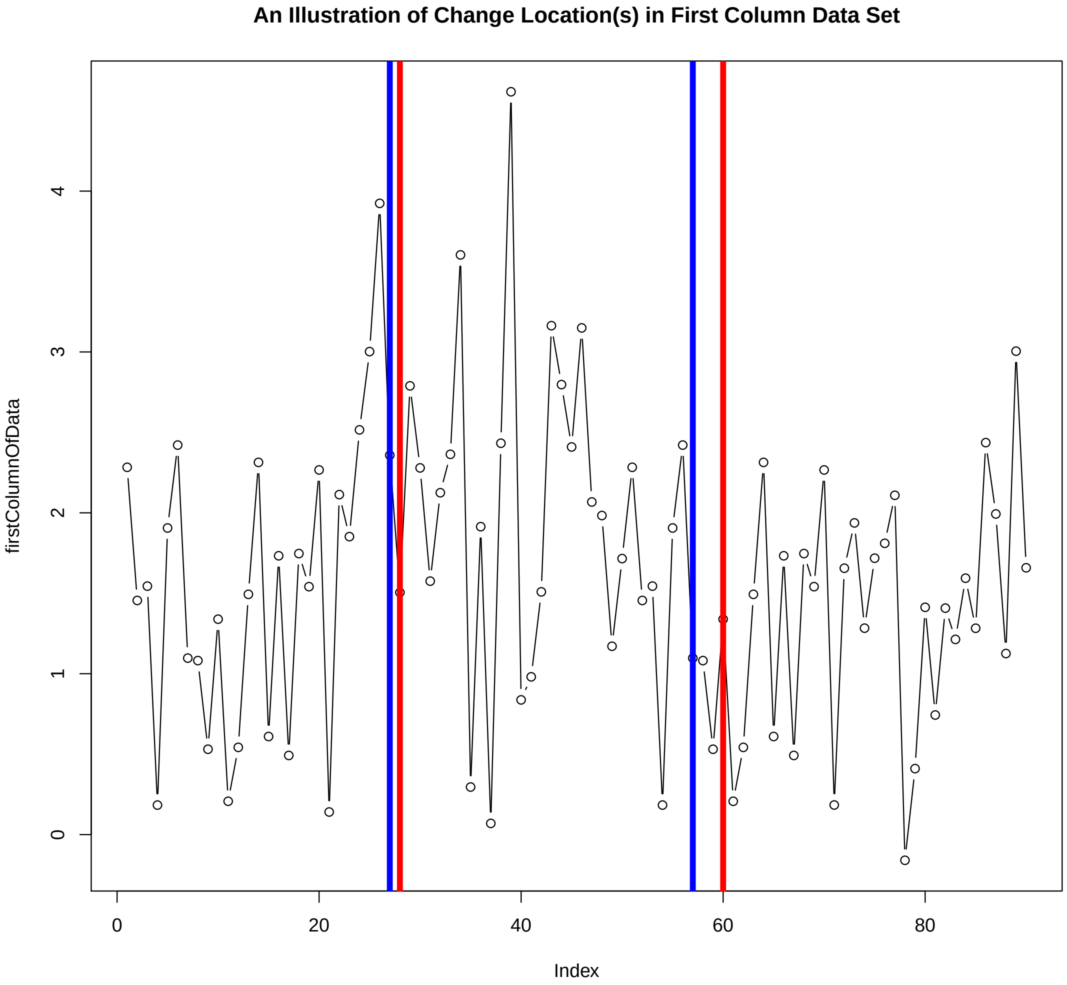


Figure 5.1: An example of change detection of a 5-dimensional data. The blue line stands for the estimated change point in DGP and the red for the detection point by the package.

Bibliography

Aas, K., Czado, C., Frigessi, A. and Bakken, H. (2009). Pair-copula constructions of multiple dependence, *Insurance: Mathematics and Economics* **44**(2): 182–198.

Andersen, L., Sidenius, J. and Basu, S. (2003). Credit derivatives: All your hedges in one basket, *Risk* **16**: 67–72.

Barbe, P., Genest, C., Ghoudi, K. and Rémillard, B. (1996). On Kendalls's process, *Journal of Multivariate Analysis* **58**: 197–229.

Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines, *Annals of Mathematical and Artificial Intelligence* **32**: 245–268.

Bedford, T. and Cooke, R. M. (2002). Vines – a new graphical model for dependent random variables, *Annals of Statistics* **30**(4): 1031–1068.

Brechmann, E. C. (2014). Hierarchical kendall copulas: Properties and inference, *Canadian Journal of Statistics* **42**(1): 78–108.

Breymann, W., Dias, A. and Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance, *Quantitative Finance* **1**: 1–14.

Bunnin, F. O., Guo, Y. and Ren, Y. (2002). Option pricing under model and parameter uncertainty using predictive densities, *Statistics and Computing* **12**(1): 37–44.

Chakraborti, S., Qiu, P. and Mukherjee, A. (2015). Editorial to the special issue: Nonparametric statistical process control charts, *Quality and Reliability Engineering International* **31**(1): 1–2.

Chen, S. X. and Huang, T. (2007). Nonparametric estimation of copula functions for dependence modeling, *The Canadian Journal of Statistics* **35**(2): 265–282.

Chen, X. and Fan, Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspesification, *Journal of Econometrics* **135**(1–2): 125–154.

Chen, X., Fan, Y. and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models, *Journal of the American Statistical Association* **101**(475): 1228–1240.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula Methods in Finance*, John Wiley & Sons, New York.

Choi, K. and Marden, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance, *Journal of the American Statistical Association* **92**(440): 1581–1590.

Choroś-Tomczyk, B., Härdle, W. K. and Okhrin, O. (2013). Valuation of collateralized debt obligations with hierarchical Archimedean copulae, *Journal of Empirical Finance* **24**(C): 42–62.

Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes, *Technometrics* **30**(3): 291–303.

Demarta, S. and McNeil, A. J. (2005). The t copula and related copulas, *International Statistical Review* **73**: 111–129.

Dobrić, J. and Schmid, F. (2007). A goodness of fit test for copulas based on Rosenblatt's transformation, *Computational Statistics & Data Analysis* **51**(9): 4633–4642.

Duembgen, M. and Rogers, L. (2014). Estimate nothing, *Quantitative Finance* **14**(12): 2065–2072.

Durante, F., Fernández-Sánchez, J. and Sempi, C. (2012). A topological proof of Sklar's theorem, *Applied Mathematical Letters* **26**: 945–948.

Durante, F., Fernández-Sánchez, J. and Sempi, C. (2013). Sklar's theorem obtained via regularization techniques, *Nonlinear Analysis: Theory, Methods & Applications* **75**(2): 769–774.

Durante, F. and Sempi, C. (2005). *Principles of Copula Theory*, Chapman and Hall/CRC.

Erdman, C., Emerson, J. W. et al. (2007). bcp: an R package for performing a Bayesian analysis of change point problems, *Journal of Statistical Software* **23**(3): 1–13.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas, *Journal of Multivariate Analysis* **95**(1): 119–152.

Fermanian, J.-D. and Scaillet, O. (2003). Nonparametric estimation of copulas for time series, *Journal of Risk* **5**: 25–54.

Fisher, R. A. (1937). *The design of experiments*, Oliver And Boyd; Edinburgh; London.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests, *The Annals of Statistics* pp. 697–717.

Gaensler, P. and Stute, W. (1987). *Seminar on empirical processes*, Springer Basel AG, Boca Raton.

Genest, C., Ghoudi, K. and Rivest, L.-P. (1995). A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika* **82**(3): 543–552.

Genest, C., Quessy, J.-F. and Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation, *Scandinavian Journal of Statistics* **33**: 337–366.

Genest, C. and Rèmillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models, *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **6**(44): 1096–1127.

Genest, C., Rémillard, B. and Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study, *Insurance: Mathematics and Economics* **44**: 199–213.

Genest, C. and Rivest, L.-P. (1989). A characterization of Gumbel family of extreme value distributions, *Statistics & Probability Letters* **8**(3): 207–211.

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas, *Journal of the American Statistical Association* **88**(3): 1034–1043.

Górecki, J., Hofert, M. and Holeňa, M. (2016). An approach to structure determination and estimation of hierarchical Archimedean copulas and its application to bayesian classification, *Journal of Intelligent Information Systems* **46**(1): 21–59.

Haff, I. H. (2013). Parameter estimation for pair-copula constructions, *Bernoulli* **19**(2): 462–491.

Härdle, W. K. and Simar, L. (2015). *Applied Multivariate Statistical Analysis*, Springer, Berlin.

Hawkins, D. M. and Deng, Q. (2010). A nonparametric change-point control chart, *Journal of Quality Technology* pp. 165–173.

Hawkins, D. M., Qiu, P. and Kang, C. W. (2003). The changepoint model for statistical process control, *Journal of Quality Technology* **35**(4): 355.

Hawkins, D. M. and Zamba, K. (2005a). A change-point model for a shift in variance, *Journal of Quality Technology* **37**(1): 21.

Hawkins, D. M. and Zamba, K. (2005b). Statistical process control for shifts in mean or variance using a changepoint formulation, *Technometrics* **47**(2): 164–173.

Hering, C., Hofert, M., Mai, J.-F. and Scherer, M. (2010). Constructing hierarchical Archimedean copulas with Lévy subordinators, *Journal of Multivariate Analysis* **101**(6): 1428–1433.

Hofert, J. M. (2010). Sampling nested Archimedean copulas with applications to CDO pricing, *Dissertation of Ulm University* .

Hofert, M. (2011). Efficiently sampling nested Archimedean copulas, *Computational Statistics & Data Analysis* **55**(1): 57–70.

Hofert, M. and Scherer, M. (2011). CDO pricing with nested Archimedean copulas, *Quantitative Finance* **11**(5): 775–87.

Holland, M. D. (2013). NPMVCP: Nonparametric multivariate change point model, *Reference manual* .
URL: *ftp://cran.r-project.org/pub/R/web/packages/NPMVCP/index.html*

Holland, M. and Hawkins, D. (2014). A control chart based on a nonparametric multivariate change-point model, *Journal of Quality Technology* **46**: 1975–1987.

Hosking, J. R. (1980). The multivariate portmanteau statistic, *Journal of the American Statistical Association* **75**(371): 602–608.

Hull, J. and White, A. (2004). Valuation of a CDO and an n-th to default CDS without Monte Carlo simulation, *Journal of Derivatives* **12**(2): 8–23.

James, N. and Matteson, D. (2015). ecp: An R package for nonparametric multiple change point analysis of multivariate data, *Journal of Statistical Software* **62**(1): 1–25.

Joe, H. (1996a). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters, *in* L. Rüschendorf, B. Schweizer and M. Taylor (eds), *Distribution with fixed marginals and related topics*, IMS Lecture Notes – Monograph Series, Institute of Mathematical Statistics.

Joe, H. (1996b). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters, *Lecture Notes-Monograph Series* pp. 120–141.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models, *J. Multivariate Anal.* **94**(2): 401–419.

Joe, H. (2014). *Dependence Modeling with Copulas*, Chapman and Hall/CRC, Boca Raton.

Johnson, R. A. and Wichern, D. W. (2013). *Applied Multivariate Statistical Analysis*, 6 edn, Pearson, Harlow.

Jouini, M. and Clemen, R. (1996). Copula models for aggregating expert opinions, *Operation Research* **3**(44): 444–457.

Kendall, M. (1970). *Rank Correlation Methods*, Griffin, London.

Killick, R. and Eckley, I. (2011). Changepoint analysis with the changepoint package in R, *The R User Conference, useR! 2011 August 16-18 2011 University of Warwick, Coventry, UK*, p. 51.

Kim, A. Y., Marzban, C., Percival, D. B. and Stuetzle, W. (2009). Using labeled data to evaluate change detectors in a multivariate streaming environment, *Signal Processing* **89**: 2529–2536.

Knoth, S. (2016). spc: Statistical process control - collection of some useful functions, *Reference manual* .
URL: *https://cran.r-project.org/web/packages/spc/index.html*

Krupskii, P. and Joe, H. (2013). Factor copula models for multivariate data, *Journal of Multivariate Analysis* **120**: 85–101.

Kurowicka, M. and Cooke, R. M. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*, John Wiley & Sons, New York.

Laurent, J.-P. and Gregory, J. (2005). Basket default swaps, CDO's and factor copulas, *Journal of Risk* **7**(4): 103–122.

Li, D. X. (1999). The valuation of basket credit derivatives, *CreditMetrics Monitor* **4**: 34–50.

Li, D. X. (2000). On default correlation: A copula function approach, *Journal of Fixed Income* **9**: 43–54.

Lindskog, F. and McNeil, A. (2001). Common poisson shock models: Applications to insurance and credit risk modelling, *ASTIN BULLETIN* **33**: 209–238.

Lowry, C. A., Woodall, W. H., Champ, C. W. and Rigdon, S. E. (1992). A multivariate exponentially weighted moving average control chart, *Technometrics* **34**(1): 46–53.

Luo, X. and Shevchenko, P. V. (2010). The t-copula with multiple parameters of degrees of freedom: Bivariate characteristics and application to risk management, *Quantitative Finance* **10**: 1039–1054.

Mai, J.-F. and Scherer, M. (2013). What makes dependence modeling challenging? Pitfalls and ways to circumvent them, *Statistics & Risk Modeling* **30**(4): 287–306.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data, *Journal of the American Statistical Association* **109**(505): 334–345.

Nelsen, R. (2006). *An Introduction to Copulas*, Springer- New York.

Oakes, D. (1994). Multivariate survival distributions, *Journal of Nonparametric Statistics* **3**(3-4): 343–354.

Oh, D. H. and Patton, A. (2015). Modelling dependence in high dimensions with factor copulas, *Finance and Economics Discussion Series 2015-051. Washington: Board of Governors of the Federal Reserve System* .

Oh, D. H. and Patton, A. J. (2013). Simulated method of moments estimation for copula-based multivariate models, *Journal of the American Statistical Association* **108**(502): 689–700.

Okhrin, O., Okhrin, Y. and Schmid, W. (2013a). On the structure and estimation of hierarchical Archimedean copulas, *Journal of Econometrics* **173**(2): 189–204.

Okhrin, O., Okhrin, Y. and Schmid, W. (2013b). Properties of hierarchical Archimedean copulas, *Statistics & Risk Modeling* **30**(1): 21–54.

Okhrin, O. and Ristig, A. (2014a). Hierarchical Archimedean copulae: The HAC package, *Journal of Statistical Software* **58**(4): 1–20.

Okhrin, O. and Ristig, A. (2014b). Hierarchical archimedean copulae: The HAC package, *Journal of Statistical Software* **58**(4).

Page, E. S. (1954). Continuous inspection schemes, *Biometrika* **41**(1/2): 100–115.

Patton, A., Fan, Y. and Chen, X. (2004). Simple tests for models of dependence between multiple financial time series, with applications to u.s. equity returns and exchange rates, *Working paper* .

Patton, A. J. (2012). A review of copula models for economic time series, *Journal of Multivariate Analysis* **110**: 4–18.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations: Iii. the analysis of variance test, *Biometrika* **29**(3/4): 322–335.

Qiu, P. (2017). Some perspectives on nonparametric statistical process control, **to appear in**, *Journal of Quality Technology* .

Qiu, P. and Hawkins, D. (2001). A rank-based multivariate cusum procedure, *Technometrics* **43**(2): 120–132.

Qiu, P. and Hawkins, D. (2003). A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions, *Journal of the Royal Statistical Society: Series D (The Statistician)* **52**(2): 151–164.

Rachev, S., Stoyanov, S. and Fabozzi, F. (2008). *Advanced stochastic models, risk assessment, and portfolio optimization: The ideal risk, uncertainty, and performance measures*, John Wiley & Sons.

Radulovic, J.-D. F. D. and Wegkamp, M. (2004). Weak convergence of empirical copula processes, *Bernoulli* **10**(5): 847–860.

Rezapour, M. (2015). On the construction of nested Archimedean copulas for *d*-monotone generators, *Statistics & Probability Letters* **101**(0): 21–32.

Rizzo, M. L. and Székely, G. J. (2016). Package ‘energy’, *R User’s Manual* .

Roberts, S. (1959). Control chart tests based on geometric moving averages, *Technometrics* **1**(3): 239–250.

Rosenblatt, M. (1952). Remarks on a multivariate transformation, *Annals of Mathematical Statistics* **23**: 470–472.

Ross, G. J. et al. (2013). Parametric and nonparametric sequential change detection in R: The cpm package, *Journal of Statistical Software* **78**.

Savu, C. and Trede, M. (2010a). Hierarchies of Archimedean copulas, *Quantitative Finance* **10**(3): 295–304.

Savu, C. and Trede, M. (2010b). Hierarchies of Archimedean copulas, *Quantitative Finance* **10**(3): 295–304.

Scaillet, O. (2007). Kernel-based goodness-of-fit tests for copulas with fixed smoothing parameters, *Journal of Multivariate Analysis* **98**(3): 533–543.

Schloegl, L. and O’Kane, D. (2005). A note on the large homogeneous portfolio approximation with the student-t copula, *Finance Stochastics* **9**: 577–584.

Schönbucher, P. (2002). Taken to the limit: Simple and not-so-simple loan loss distribution, *Working Paper* .

Schönbucher, P. and Schubert, D. (2000). Copula-dependent default risk in intensity models, *Working paper* .

Schreyer, M., Paulin, R. and Trutschnig, W. (2017). On the exact region determined by kendall's τ and spearman's ρ, *to appear in: Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Segers, J. and Uyttendaele, N. (2014). Nonparametric estimation of the tree structure of a nested Archimedean copula, *Computational Statistics & Data Analysis* **72**: 190–204.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, Vol. 162, John Wiley & Sons.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*, ASQ Quality Press.

Shewhart, W. A. and Deming, W. E. (1939). *Statistical method from the viewpoint of quality control*, Courier Corporation.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data, *Biometrics* **51**(4): 1384–1399.

Sims, C. A. (1980). Macroeconomics and reality, *Econometrica: Journal of the Econometric Society* pp. 1–48.

Sklar, A. (1959). Fonctions de répartition à n dimension et leurs marges, *Publications de l'Institut de Statistique de l'Université de Paris* **8**: 299–231.

Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension, *InterStat* **5**.

Székely, G. J. and Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method, *Journal of Classification* **22**(2): 151–183.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: statistics based on distances, *Signal Processing* **143**: 1249–1272.

Markit$^{\text{TM}}$ (2008). Markit credit indices: A primer., *Technical Report,* .
URL: *https://www.markit.com/news/Credit%20Indices%20Primer.pdf*

Wang, W. and Wells, M. (2000). Model selection and semiparametric inference for bivariate failure-time data, *Journal of the American Statistical Association* **95**(449): 62–76.

Whelan, N. (2004). Sampling from Archimedean copulas, *Quantitative Finance* **4**(3): 339–352.

Woodall, W. H. and Montgomery, D. C. (2014). Some current directions in the theory and application of statistical process monitoring, *Journal of Quality Technology* **46**(1): 78.

Xu, Y. F. (2017). *EnergyOnlineCPM: Distribution free multivariate control chart based on energy test.*
URL: *https://cran.r-project.org/web/packages/EnergyOnlineCPM/index.html*

Zech, G. and Aslan, B. (2003). A multivariate two-sample test based on the concept of minimum energy, *Proc. Statistical Problems in Particle Physics, Astrophysics, and Cosmology* pp. 8–11.

Zeileis, A., Leisch, F., Kleiber, C. and Hornik, K. (2005). Monitoring structural change in dynamic econometric models, *Journal of Applied Econometrics* **20**(1): 99–121.

Zhang, S., Okhrin, O., Zhou, Q. M. and Song, P. X.-K. (2016). Goodness-of-fit test for specification of semiparametric copula dependence models, *Journal of Econometrics* **193**(1): 215–233.

Zhou, M., Zi, X., Geng, W. and Li, Z. (2015). A distribution-free multivariate change-point model for statistical process control, *Communications in Statistics: Simulation and Computation* **44**: 1975–1987.

Zhu, W., Wang, C.-W. and Tan, K. S. (2016). Structure and estimation of Lévy subordinated hierarchical Archimedean copulas (LSHAC): Theory and empirical tests, *Journal of Banking & Finance* .

Zou, C. and Tsung, F. (2011). A multivariate sign EWMA control chart, *Technometrics* **53**: 84–97.

Zou, C., Wang, Z. and Tsung, F. (2012). A spatial rank-based multivariate ewma control chart, *Naval Research Logistics (NRL)* **59**(2): 91–110.

Declaration of Authorship

I hereby confirm that I have authored this dissertation and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Yafei Xu

Berlin, 20. October, 2017