RESPICE PROSPICE

UNIVERSITY of the
WESTERN CAPE

SOME NON-STANDARD STATISTICAL DEPENDENCE PROBLEMS

By

Alphonce Bere

SUBMITTED IN FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
UNIVERSITY OF THE WESTERN CAPE
CAPE TOWN, SOUTH AFRICA
MARCH 2016

To my late brother Ereck Watson Bere

Table of Contents

List of Figures

List of Tables

Abstract

The major result of this thesis is the development of a framework for the application of pair-mixtures of copulas to model asymmetric dependencies in bivariate data. The main motivation is the inadequacy of mixtures of bivariate Gaussian models which are commonly fitted to data. Mixtures of rotated single parameter Archimedean and Gaussian copulas are fitted to real data sets. The method of maximum likelihood is used for parameter estimation. Goodness-of-fit tests performed on the models giving the highest log-likelihood values show that the models fit the data well.

We use mixtures of univariate Gaussian models and mixtures of regression models to investigate the existence of bimodality in the distribution of the widths of autocorrelation functions in a sample of 119 gamma-ray bursts. Contrary to previous findings, our results do not reveal any evidence of bimodality. We extend a study by Genest et al. (2012) of the power and significance levels of tests of copula symmetry, to two copula models which have not been considered previously. Our results confirm that for small sample sizes, these tests fail to maintain their 5% significance level and that the Cramér−von Mises-type statistics are the most powerful.

Acknowledgements

I am immensely grateful to my supervisor Professor Chris Koen for accepting me as his PhD student, for the many insightful suggestions that he made, for his forthrightness and above all for the high level of professionalism and dedication he gave to my work.

Professor Christian Genest was always forthcoming with clarifications regarding copula modelling; that is much appreciated.

I also would like to acknowledge the financial support I got from the Cape Peninsula University of Technology and the University of Venda. My HODs in these two institutions also made it possible for me to have some time for research work. A very significant part of this work was completed during a month-long stay at the African Institute for Mathematical Sciences in 2014. I am grateful for their support.

I am also indebted beyond measure to my late elder brother Ereck Watson Bere who gave a lifelong commitment to the education of us, his siblings, after the passing away of our father in 1978 and to my parents who also made huge sacrifices in order to educate their children.

I am blessed with three kind hearted sisters Jane, Miriam and Ngoni, who have given me consistent support during the course of this project. I love them all.

I do not know if I will be able to pay back my lovely wife Martha and the kids Komborero, Admire, Ngoni and the ever-smiling Ereck (Jnr) for the many days that they endured without me while I was working on this project.

Finally, I wish to thank the following: Kudakwashe Ndhlukula, Sekai Shambira and Sebastian Mukumbira for their friendship and Dr. Gift Muchatibaya, Prof Stanford Shateyi, Dr. Jabulani Gumbo and Professor Winston Garira for their encouragement.

Thohoyandou, South Africa Alphonce Bere
July 31, 2015

Definitions of Astronomical terms used in the thesis

Gamma-ray Bursts (GRBs) : Flashes of high energy radiation associated with extremely energetic explosions that take place in distant galaxies.

Long duration GRBs: GRBs with a duration greater than two seconds.

Short duration GRBs: GRBs with a duration shorter than two seconds.

Light Curve of a GRB: Graph of radiation intensity as a function of time.

A pulsar: A highly magnetized, rotating neutron star that emits a beam of electromagnetic radiation.

Pulsar Period: Rate of rotation of the pulsar.

Period derivative: Rate of change of period.

Swift Mission: Consisted of a spacecraft called Swift, which was launched into orbit on November 20, 2004. The aim of the mission was to study gamma ray bursts.

Burst Alert Telescope (BAT): A telescope aboard Swift.

Gamma-ray burst monitor: An instrument (telescope) aboard the Fermi Gamma-ray spacecraft which is used for studying gamma-ray bursts. Fermi was launched on 11 June 2008.

Konus: A telescope aboard the WIND spacecraft, with the purpose of studying gamma-ray bursts. The WIND spacecraft was launched on 1 November 1994.

T_{90} : The central time interval over which 90 percent of the gamma-ray burst's energy is emitted.

Hardness ratio: The ratio of the two fluences (counts of photons) in two different energy bands, integrated over the time interval T_{90}.

Chapter 1

Introduction and Objectives

1.1 Introduction

In many cases, we want to determine if there is an association or dependence between two random variables. We may make important decisions based on knowledge about the presence or absence of statistical dependence. Some examples are given below.

- An electrical utility may decide to produce less power on a mild day based on the correlation between electricity demand and weather.
- An accurate model of dependence enables risk practitioners to price financial instruments fairly.
- Forecasting models are built utilizing the dependence between successive observations of a time series.
- If an accurate model can be found for the dependence between the concentrations of two minerals that are known to occur together in soil samples, then, the concentration of one of the minerals can be predicted on the basis of the concentration of the other. This is useful in situations where it is very expensive to determine the concentration of one of the minerals.

Identifying and modeling dependence is therefore a useful skill for any statistician.

There is a variety of statistical tools that can be used to assess the dependence between a pair of random variables (X, Y). The most commonly used is the Pearson correlation coefficient. Application of the Pearson correlation coefficient assumes a linear relationship between X and Y. A small size of the Pearson correlation coefficient does not not necessarily mean that the variables are not related; the coefficient sometimes fails to detect non-linear relationships.

Two rank-based correlation coefficients can sometimes overcome the above-mentioned limitation of the Pearson correlation coefficient. These are Kendall's rank correlation coefficient, usually denoted by τ and Spearman's rank correlation coefficient, commonly denoted by the symbol ρ. The coefficients τ and ρ are measures of monotone dependence. They are therefore less useful in situations where the dependence structure in the data is not monotone.

Copulas offer a more sophisticated measure of dependence compared to the correlation methods described above. A copula is a function that links univariate marginal distributions to their joint multivariate distribution.

Copula modelling offers many advantages. Copulas have the ability to model complex dependence structures, they allow for the modeling of the dependence structure between random variables independently of the marginal distributions and unlike the correlation approach, copulas also have the ability to capture the dependence between extreme events.

The notion of correlation is also widely applied in time series analysis where the focus is on the relationship between observations separated by a given time interval.

The autocorrelation function (ACF) is normally applied in this situation.

1.2 The nature of the problem

The main contribution of this work is in chapter 5 where we apply mixtures of rotated copulas to bivariate data. Our work is mainly motivated by attempts of previous researchers to model astrophysical phenomena using statistical methods including some of the statistical tools for assessing dependence described above. We have sought either to improve on the statistical techniques employed in the previous studies or to offer alternative methods. Below we summarize the studies from which we have derived the motivation for our work.

Borgonovo (2004), Borgonovo et al. (2007) and Vasquez and Kawai (2011) computed autocorrelation functions of long duration gamma-ray bursts. A characteristic timescale for the bursts can be defined in terms of the autocorrelation function width, commonly defined in the astronomy literature as the lag at which the autocorrelation has declined to a value of 0.5. The three studies mentioned above concluded that the widths of autocorrelation functions of gamma-ray burst light curves show a bimodal distribution.

Horváth et al. (2010) used bivariate Gaussian mixture models to model the dependence between durations (T_{90}) and the hardness ratios (which characterises the energy spectra) of a sample of gamma-ray bursts. The conclusion that came from their work was that the relationship between the duration and the hardness ratio can be modelled by a three-component bivariate Gaussian mixture model. From this, the authors conjectured that there are three physically distinct types of gamma-ray bursts in outer space.

Similarly, Lee et al. (2012) used the Gaussian mixture model to classify pulsars on the basis of the relationship between the pulsar period and the period derivative. It was concluded that a mixture of six bivariate Gaussians was needed to capture the relationship between the period and period derivatives.

The sample sizes considered in the studies of Borgonovo (2004), Borgonovo et al. (2007) and Vasquez and Kawai (2011), namely 16, 22 and 20 respectively, are rather small. Another shortcoming of these three studies is that no formal statistical techniques were employed. Only graphical evidence was used in arriving at the conclusion that the autocorrelation functions of gamma ray bursts exhibit a bimodal distribution. Graphical techniques do not offer an estimate of the margin of error in the analysis and as such, there is a need to employ more sophisticated statistical techniques to either prove or disprove the claimed bimodality.

With regards to the paper by Horváth et al. (2010), our preliminary analysis reveals that even if tests of hypotheses fail to reject the claimed three-component bivariate Gaussian mixture model, this model is not acceptable. The distribution of T_{90} values alone can be described by a three-component mixture model while that of hardness ratios alone is best described by a two-component model. A similar conclusion applies to the data modelled by Lee et al. (2012). Although a bivariate six-component model fits to the data, an analysis of the marginal data reveals that fewer than six components can adequately describe the marginal distribution of each variable. Our work on modelling using mixtures of rotated copulas was thus motivated by the failure of the mixtures of Gaussian models to adequately model the bivariate relations described above.

In arriving at the number of bivariate mixture components for the hardness ratio-duration data, Horváth et al. (2010) used the likelihood ratio test which compares the Gaussian-mixture likelihoods under the null and alternative hypothesis. Provided some regularity conditions (see for example Lehmann, 1998) are met, the likelihood ratio statistic is known to have an asymptotic chi-square distribution. Horváth et al. (2010), assumed and applied this chi-squared distribution for the likelihood ratio statistic but, in the context of mixture models, the regularity conditions required for the asymptotic chi-square distribution are not met; the parameters estimated under the alternative hypothesis are not identifiable under the null hypothesis. We also address this issue through obtaining the p-values of the likelihood ratio statistic by simulating under the null hypothesis. Furthermore, Horváth et al. (2010) made no attempt to actually test the final model for goodness-of-fit.

An important question is whether the three Gaussian clusters, identified by Horváth et al. (2010), or the six clusters in the Lee et al. (2012) data correspond to physically distinct classes of objects. In this study, we caution against the idea of associating the number of components in a statistical model with the number of distinct groups of astrophysical objects, by proposing alternative statistical models that adequately describe the two data sets in question.

In chapter 4, we digress slightly to discuss a related issue; evaluation of the power and significance levels of tests of copula symmetry. This follows earlier work by Genest et al. (2012). Our study uses different copula models from the ones studied in Genest et al. (2012).

The data sets used in this study have some unique features which hamper the application of the usual or "standard" statistical techniques. The light curves of gamma-ray bursts discussed in chapter 1 are highly non-stationary. This has necessitated an adjustment in the way the autocorrelation functions of the light curves are computed as will be seen in the next chapter. The lack of symmetry in other data sets has led us to use mixtures of rotated copulas instead of single copulas to model dependence.

1.3 Objectives of the study

In this study we suggest, develop, apply and evaluate alternative statistical dependence tools for modeling the data sets mentioned above.

The objectives of this research are:-

(i) to model the distribution of gamma-ray autocorrelation function widths using Gaussian mixtures of distributions and mixtures of regression models,

(ii) to further evaluate the fit of bivariate Gaussian mixture models to duration-hardness ratio data used in Horváth et al. (2010) and the period-period derivative data used in Lee et al. (2012),

(iii) to evaluate the power and significance levels of tests of copula symmetry, and,

(iv) to model the dependence in the data sets used in Horváth et al. (2010) and in Lee et al. (2012) using copulas, especially mixtures of rotated copulas, and to evaluate the goodness-of-fit of the copula models.

1.4 Significance of the study

Gamma-ray bursts are some of the most powerful explosions in the Universe. Both pulsars and gamma-ray bursts were discovered in the 1960s. To date, scientists all over the world grapple with questions regarding the origin and subdivision of members of these classes of objects. This research is an effort to contribute to this endeavour.

1.5 Layout of the thesis

Chapter 1 gives the introduction to the subjects to be covered. In chapter 2 we present the work on modelling the distribution of autocorrelation function widths of gamma ray bursts using mixtures of Gaussian distributions, mixtures of regression models and kernel density estimates. In chapter 3, we apply bivariate Gaussian mixture models to gamma-ray burst data and also to the pulsar data. Chapter 4 reports on a study of the power and significance levels of tests of copula symmetry. In chapter 5, we give a new approach to modelling the data sets in Horváth et al. (2010) and in Lee et al. (2012) based on mixtures of rotated copulas. Chapter 6 states the conclusions.

Chapter 2

Modelling using univariate mixtures of Gaussian distributions and mixtures of regressions

2.1 Gamma-ray bursts

Gamma-ray bursts (GRBs) are flashes of gamma rays associated with extremely energetic explosions that take place in distant galaxies. A typical burst lasts from ten milliseconds to several minutes. Bursts which last more that 2 seconds are classified as “long bursts” or “bursts of long duration” while those that last less than two seconds are referred to as “short bursts.” Figure 2.1, below shows a typical gamma-ray burst light curve - a graph of intensity as a function of time. The intensity is given by the number of photons received per unit area per unit time.

In figure 2.1, the actual burst starts roughly at time 237 seconds and ends at time 390 seconds, approximately. Before and after the burst, we observe the background/noise signal. The signal obtained during the burst is partly due to the source (GRB) and partly due to the background noise:

$$m = s + b \tag{2.1.1}$$

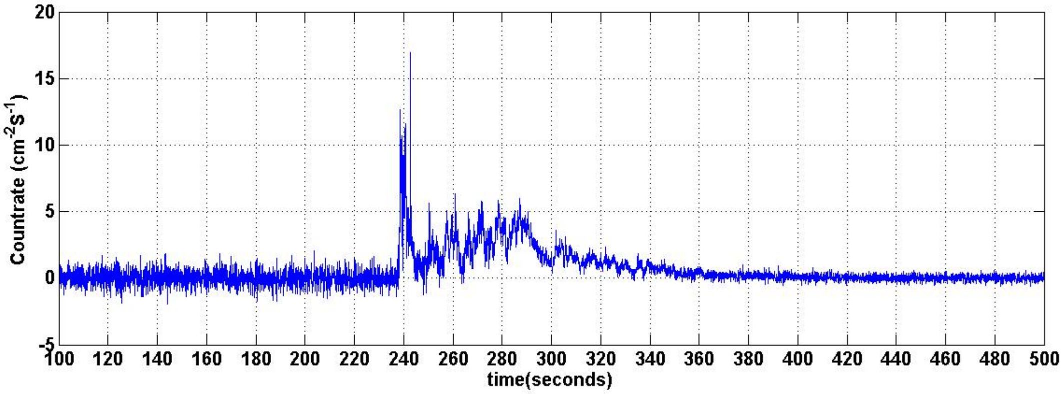


Figure 2.1: A typical gamma-ray burst light curve.

where s, b and m are respectively the source, background and total count rates. During the burst, we observe only the total photon counts; the source and background counts are not individually observable. In the next section, we describe the problem that motivated the work in this chapter.

2.2 The nature of the problem

There are a number of studies in the astrophysics literature focussed on the autocorrelation functions of gamma-ray bursts. Some of the recent publications in that area are Borgonovo (2004), Borgonovo et al. (2007) and Vasquez and Kawai (2011). The three studies cited above considered the distribution of autocorrelation function widths across long duration gamma-ray bursts and showed it to be bimodal, if corrected for relativistic effects. The autocorrelation function width referred to here is the time that it takes for the autocorrelation to decay to a value of 0.5.

In this chapter, we model the distribution of autocorrelation function widths of gamma-ray bursts using univariate mixtures of Gaussian distributions and mixtures of regressions. Firstly, we suggest an alternative, improved way of normalizing the gamma ray burst autocorrelation functions, i.e. computing autocorrelation at lag 0. We also suggest an alternative way of estimating autocorrelation function widths. Finally, we apply Gaussian mixture models and a mixture of regression models to address the question of whether the widths of autocorrelation functions is bimodal or not. The work summarized in this chapter was published in Koen and Bere (2012).

From a statistical point of view, two main shortcomings are evident in the studies cited above. In none of the three studies was a formal test of hypothesis employed to confirm the presence of two modes. Also, the studies of Borgonovo (2004) and Borgonovo et al. (2007) used data that were obtained from different sources; the Gamma-ray Burst Monitor (GRBM), the Burst and Transient Source Experiment, and Konus. This might have a bearing on the result because different instruments have different sensitivities, time resolutions and photon energy sensitivities.

The data used for the present study were obtained from a single source; the Burst Alert Telescope (BAT). Formal testing procedures are employed to address the question of whether the autocorrelation function widths have got a bimodal distribution or not.

2.3 Statistical tools used

In this section we describe the statistical concepts employed in this chapter. These include the autocorrelation function, the ”dip test” for bimodality, kernel density estimation, Gaussian mixture models, the likelihood ratio test, goodness-of-fit tests and mixtures of regressions.

2.3.1 Standard autocorrelation functions

The general autocorrelation function derives from the well-known Pearson correlation coefficient. The Pearson correlation coefficient is also applicable in a time series context where it can be used to quantify the linear dependence between values of the times series which are separated by a specified time difference (lag).

Let X_j be a time series for which the mean μ_j and variance σ_j^2 at any given time, j are known. Then the autocorrelation of X between time s and time t is given by

$$\rho(s,t) \quad = \quad \frac{E\left[(X_t - \mu_t)(X_s - \mu_s)\right]}{\sigma_t \sigma_s}$$

For stationary processes, the mean μ and variance σ^2 are independent of the time. For such processes, the correlation depends only on the time interval between a pair of values; not on their actual position in time. In that case the correlation can be expressed in terms of the time lag, τ

$$\rho(\tau) \quad = \quad \frac{E\left[(X_t - \mu)(X_{t+\tau} - \mu)\right]}{\sigma^2}$$

Given n observations $x_1, x_2, \ldots, x_n$ of a discrete time process X_j, we can compute the sample autocovariance c_k between two points which are a lag k apart using the simplified formula

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \overline{x})(x_{t+k} - \overline{x}) \tag{2.3.1}$$

The correlation coefficient at lag k is then computed as

$$r_k = \frac{c_k}{c_0}$$

for $k = 1, 2, \ldots, m$ where $m < n$. The sample autocorrelation function is a biased estimator of the population autocorelation function $\rho(k)$. However, as $n \to \infty$, $E(c_k) \to \rho(k)$ i.e. c_k is asymptotically unbiased for $\rho(k)$ (Chatfield, 2003).

The manner in which the autocorrelation function is computed in the work of Borgonovo (2007), Borgonovo et al. (2007) and Vasquez and Kawai (2011) differs slightly from the method described above. The reason for the difference is that the light curves of gamma-ray bursts are highly non-stationary. As a result, the autocorrelation function is not well defined.

The seemingly unorthodox way of computing the autocorrelation function is not a problem in the present context, since the "autocorrelation function" is only used as a means to calculate a statistic which characterizes a given gamma-ray burst. However, because of the non-stationarity, care has to be exercised in the computations.

2.3.2 Computation of autocorrelation functions of gamma-ray bursts

Following the practice in the astrophysics literature cited above, and in order to make it possible to compare our results with previous findings, we define the autocorrelation function at lag $l = k\triangle t$ as

$$\begin{aligned} A(k\triangle t) &= \frac{1}{n}\sum_{j=1}^{n-k} s_j s_{j+k}/A(0) \\ &= \frac{1}{n}\sum_{j=1}^{n-k}(m_j - b_j)(m_{j+k} - b_{j+k})/A(0), \end{aligned} \tag{2.3.2}$$

where s_j, b_j and m_j are defined in equation (2.1.1). The time interval between measurements (also referred to as the bin width) is $\triangle t$ and the duration of the burst is $n\triangle t$.

Equation (2.3.2) differs from the computational formula in the literature in that the upper limit to the summation in equation (2.3.2) is $n-k$, rather than n. The latter summation limit requires that data values be artificially defined at times $n+1, n+2, \ldots$. This contributes to greater uncertainty in the values of the autocorrelation function.

As pointed out earlier, neither the source nor background count rates are directly observable during the burst. The background level can however be estimated by for example, fitting a low-order polynomial to the pre- and post-burst light curves, and interpolating across the burst, giving estimates $\hat{b}_j$. For practical application, equation (2.3.2) is then replaced by

$$A(k\triangle t) \quad = \quad \frac{1}{n}\sum_{j=1}^{n-k}(m_j - \hat{b}_j)(m_{j+k} - \hat{b}_{j+k})/A(0) \tag{2.3.3}$$

In the current work, the background level was estimated by fitting a polynomial of order 2 to the pre- and post-burst data.

We propose the formula below for the variance of the source signal.

$$A(0) \quad = \quad \frac{1}{n}\sum_{j=1}^{n}\left[(m_j - \hat{b}_j)^2 - V\right] \tag{2.3.4}$$

where V is the variance of the background, assumed to be constant, independent of j. V can be estimated from the out-of-burst light curve using the formula

$$V \quad = \quad \frac{1}{L}\sum_{i=1}^{L}(b_i - \hat{b}_i)^2 \equiv \frac{1}{L}\sum_{1=1}^{L}(m_i - \hat{b}_i)^2 \tag{2.3.5}$$

where L is the number of pre- and post-burst measurements used.

In Borgonovo (2004) and Borgonovo et al. (2007), the analogue of (2.3.4) is

$$A(0) \;=\; \frac{1}{n}\sum_{j=1}^{n}\left[(m_j-\hat{b}_j)^2-m_j\right] \tag{2.3.6}$$

Inspection of figure 2.1 shows that, at least for this particular GRB, the background variance is *not* stationary. It does not appear to be feasible to model this change in variance. The incorrect assumption of homoscedasticity will affect the value of $A(0)$. A simple way of dealing with incorrect normalization is proposed in section 2.4.

Justification for (2.3.4) and (2.3.5) as opposed to (2.3.6) is as follows:

$$\begin{aligned}
E(m_j-\hat{b}_j)(m_{j+k}-\hat{b}_{j+k}) &= E(s_j+b_j-\hat{b}_j)(s_{j+k}+b_{j+k}-\hat{b}_{j+k})\\
&\approx Es_js_{j+k}+E(b_j-Eb_j)(b_{j+k}-Eb_{j+k})\\
&= Es_js_{j+k}+\mathrm{Cov}(b_j,b_{j+k}),
\end{aligned}$$

where it has been assumed that the source and background signals are uncorrelated, and also that $\hat{b}_j$ accurately estimates the mean background level at each time point $j\triangle t$ across the burst. For uncorrelated background noise,

$$\begin{aligned}
E(m_j-\hat{b}_j)(m_{j+k}-\hat{b}_{j+k}) &= Es_js_{j+k}+\mathrm{Cov}(b_j,b_{j+k})\\
&= Es_js_{j+k}+V\delta(k,0), \qquad (2.3.7)
\end{aligned}$$

where $\delta(k,0)$ is Kronecker's delta,

$$\delta(k,0) = \begin{cases} 0, & \text{if } k \neq 0 \\ 1, & \text{if } k = 0 \end{cases}$$

It follows that for $A(0)$ as defined in (2.3.4),

$$EA(0) = \frac{1}{n}\sum_j Es_j^2,$$

which is clearly the correct normalization in (2.3.2). For $A(0)$ as defined in (2.3.6), on the other hand,

$$E[(m_j - \hat{b_j})^2 - m_j] \;=\; Es_j^2 + [\mathrm{Var}(b_j) - Es_j - Eb_j].$$

If the background counts are Poisson distributed, as usually assumed, the first and last terms in the square brackets cancel out, but the term in Es_j remains.

2.3.3 The dip test for unimodality

The dip test devised by Hartigan and Hartigan (1985) tests the null hypothesis that a distribution is unimodal against the alternative that it has more than one mode. The test statistic is the maximum difference between the empirical distribution function and the unimodal distribution function that minimises that maximum difference. Large values of the test statistic indicate multimodality.

The reference distribution for calculating the dip statistic is the uniform, as a worst case unimodal distribution. In the R package "dip-test", p-values for the dip statistic are calculated by comparing the observed dip statistic with dip statistics for repeated samples of the same size from a uniform distribution.

There are other tests of bimodality in the literature. These include Silverman's test (Silverman, 1981), the excess mass test (Müller and Sawitski, 1991) and the Hall and York (2001) test. A power comparison of the latter tests is given in Xu et al. (2014), who unfortunately did not include the dip test in his study.

2.3.4 Kernel smoothing

Given a sample $x_1, x_2, \ldots, x_n$ of observations, the kernel estimator of the probability density function f at the point x is given by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2.3.8}$$

where K is a function satisfying $\int K(x)dx = 1$ which we call the kernel and h is a positive number, usually called the bandwidth or window width (Silverman, 1986). Some examples of kernel functions are

1. the uniform kernel,

$$K(u) = \frac{1}{2} I(|u| \leq 1),$$

2. the Epanechnikov kernel,

$$K(u) = \frac{3}{4}(1 - u^2) I(|u| \leq 1),$$

3. the triangular kernel,

$$K(u) = (1 - |u|) I(|u| \leq 1),$$

4. the Gaussian kernel,

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right),$$

and

5. the biweight kernel,

$$K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1).$$

Here $I(|u| \leq 1) = \begin{cases} 1, & |u| \leq 1 \\ 0, & \text{otherwise.} \end{cases}$

The quality of a kernel estimate in a particular context depends on the value of its bandwidth. If the bandwidth is too small, the resultant kernel estimate will be rough (under-smoothed). Too big a bandwidth results in the estimate missing essential details (over-smoothing).

2.3.5 Gaussian mixture models and the expectation maximization algorithm

Mixture models are usually applied when the population consists of sub-populations. As an example annual medical insurance claims may consist of claims from “sick” and “healthy” clients. A mixture model expresses the density function of a random variable/vector as a linear combination of a small number K of basis density functions:

$$f(\mathbf{x} \mid \Theta) \quad = \quad \sum_{j=1}^{K} \pi_j f_j(\mathbf{x} \mid \theta_j), \qquad (2.3.9)$$

where $\mathbf{x} = \{x_1, x_2, \ldots, x_d\} \in \mathbb{R}^d$ and $\Theta = \{\pi_1, \pi_2, \ldots, \pi_K, \theta_1, \theta_2, \ldots, \theta_K\}$ represents the parameters (Bilmes, 1998). The π_j’s are called mixing parameters while the θ_j are the parameters of the basis densities. It is possible to have more than one parameter for the basis density. The π_j's satisfy the constraints

$$\sum_{j=1}^{K} \pi_j \quad = \quad 1 \tag{2.3.10}$$

and

$$0 \leq \quad \pi_j \quad \leq 1. \tag{2.3.11}$$

In this thesis, we focus on Gaussian mixture models for which the component densities functions are given by

$$f_j(\mathbf{x}|\theta_j) \quad = \quad \frac{1}{(2\pi)^{d/2}(\det\Sigma_j)^{1/2}} e^{-1/2(\mathbf{x}-\mu_j)^T\Sigma_j^{-1}(\mathbf{x}-\mu_j)} \tag{2.3.12}$$

Here, $\mu_j \in \mathbb{R}^d$ is the mean vector for the j^{th} component and Σ_j is a $d \times d$ covariance matrix corresponding to the j^{th} component.

Given a density function $f(\mathbf{x}|\Theta)$ with parameter set Θ, and a data set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ of size n drawn from the corresponding distribution, if we can assume that the data vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are independent and identically distributed, then the resulting joint density for the sample will be

$$f(\mathcal{X}|\Theta) \quad = \quad \prod_{i=1}^{n} f(\mathbf{x}_i|\Theta) = \mathcal{L}(\Theta|\mathcal{X}) \tag{2.3.13}$$

In (2.3.13) $f(\mathcal{X}|\Theta)$ is the joint density function; a function of the data given the parameter(s), while $\mathcal{L}(\Theta|\mathcal{X})$ is the likelihood function which is a function of the parameters given the data. One of the methods that can be used for estimating the parameters of the distribution is the method of maximum likelihood.

The usual procedure is to set the derivative of log$\mathcal{L}(\Theta|\mathfrak{X})$ equal to zero and solve for the resultant equation for the parameters. However, in some cases alternative methods are more efficient. One such technique is the Expectation Maximization (EM) algorithm.

The EM algorithm was explained and given its name in a paper by Dempster et al. (1977). Each step of the algorithm is guaranteed to either retain the previous value of the likelihood or improve on it. A proof of this fact together with details on the derivation of the iterative formulae for the parameters, especially for Gaussian mixture models, can be found in Chen and Gupta (2010).

2.3.6 The likelihood ratio test

The likelihood ratio test is used to compare the fit of two models, a smaller (constrained) model and a more complex model. The smaller model must be nested within the larger model. The null hypothesis of the test is that the fit of the null (smaller) model is adequate. Rejection of the null hypothesis implies that the more complex model provides a significant data description improvement over the smaller model.

The likelihood ratio statistic is

$$\Lambda \ = \ -2ln\left(\frac{L_s}{L_c}\right). \tag{2.3.14}$$

Here L_s is the likelihood of the simpler model and L_c represents the likelihood of the more complex model.

According to a theorem attributed to Wilks (1938), under certain regularity conditions, the asymptotic distribution of Λ in equation (2.3.14) is approximately a chi-squared distribution with the degrees of freedom equal to the difference in the number of parameters for the two models under comparison. One of the required regularity conditions is that the parameters involved in L_s or L_c should not lie on the boundary of the parameter space. Full details on the necessary regularity conditions can be found in Lehmann (1998).

In the context of mixture models, where we test the null hypothesis of say, K components against an alternative hypothesis of $K+1$ components, the asymptotic chi-squared distribution does not hold because the likelihood computed under the null hypotheses of a K-component model implies that the mixture proportion π_{K+1} for the $(K+1)^{\text{th}}$ component is zero which is on the border of the proportion parameter space.

For this situation of mixture models, McLachlan (1987) suggested establishing the distribution of the likelihood ratio statistic by simulating under the null hypothesis. The steps of his algorithm are as outlined below:-

(i) Fit the two competing models and compute the value of Λ. Retain the estimated parameters $\hat{\Theta}_0$ for the model specified by H_o.

(ii) For a large integer M , and proceeding under H_o, repeat the following steps for $k = 1, 2, \ldots, M$:

(a) Generate a sample equal in size to the original data from $f(\mathbf{x}; \hat{\Theta}_o)$

(b) For each sample compute and retain the value of the test statistic $\Lambda^{(k)}$ after fitting the two competing models to the simulated data.

(iii) Compute approximate p-value for the test as

$$p = M^{-1} \sum_{k=1}^{M} I(\Lambda^{(k)} \geq \Lambda).$$

2.3.7 Goodness-of-fit tests for the normal distribution

Goodness-of-fit tests are tests for investigating if a sample of data is consistent with a specified population distribution.

Of particular interest is the situation where the parameters of the population distribution are unspecified, i.e. only the family of distributions is postulated.

In the sequel, tests will be required for testing the fit of the normal distribution to log-transformed autocorrelation function widths. The discussion in this section will therefore focus on goodness-of-fit tests for the normal distribution.

D'Agostino and Stephens (1986) classified goodness-of-fit tests for normality into five categories, chi-square type tests, empirical distribution function tests, moment tests, regression tests and miscellaneous tests.

These authors also attempted to give recommendations on which of these tests to apply in different situations. This after summarizing the results of a number of power studies where these different tests were applied to various non-normal populations. From the recommended tests we chose the Anderson-Darling and D'Agostino-Pearson tests. Brief descriptions of the two tests follow.

The Anderson-Darling statistic for testing if X comes from a distribution $F(x)$ uses the empirical distribution function $F_n(x)$:

$$A^2 = n\int_{-\infty}^{\infty} \frac{\{F_n(x) - F(x)\}^2}{[\{F(x)\}\{1 - F(x)\}]} dF(x) \tag{2.3.15}$$

Anderson and Darling (1954) derived the computational formula

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log z_{(i)} + \log\left\{1 - z_{(n+1-i).}\right\}\right] \tag{2.3.16}$$

Here $z_{(1)} < z_{(2)} < \cdots < z_{(n)}$ are the order statistics corresponding to $z_1 =$

$F(x_1), z_2 = F(x_2), \ldots, z_n = F(x_n)$. An alternative formula for A^2 is

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i-1) \log z_{(i)} + (2n+1-2i) \log \left\{ 1 - z_{(i)} \right\} \right] \tag{2.3.17}$$

The null hypothesis is rejected if the value of A^2 is large. In the case of a fully specified distribution F, the distribution theory of empirical distribution function statistics is well developed and tables which give the significance levels are available.

In cases where the null hypothesis does not specify the location or scale parameters, the exact distributions of empirical distribution function statistics are difficult to find. Luckily for quadratic statistics such as the Anderson-Darling statistic, the asymptotic distributions are known . It is also known that when the unknown parameters are estimated using appropriate methods, the distribution of the empirical distribution function statistics will not depend on the true values of the unknown parameters; instead they only depend on the family tested and on the sample size n (D'Agostino and Stephens, 1986). D'Agostino and Stephens (1986) suggested a modification of the Anderson-Darling statistic which makes it possible to compare the values of the statistic for finite sample size with asymptotic significance points. The modified statistic is

$$A^* = A^2 \left(1.0 + \frac{0.75}{n} + \frac{2.25}{n^2} \right).$$

Table 4.9 (page 127) in D'Agostino and Stephens (1986) gives formulae for approximating the p-values corresponding to the modified statistics.

The D'Agostino-Pearson test is an example of a moment test. Moment tests capitalize on the fact that for a random variable X which has a normal distribution with mean μ and variance σ^2, the third and fourth moments (about the mean) are respectively

$$\sqrt{\beta_1} = \frac{E(X-\mu)^3}{\sigma^3} = 0 \tag{2.3.18}$$

and

$$\beta_2 = \frac{E(X-\mu)^4}{\sigma^4} = 3 \tag{2.3.19}$$

Here, $\sqrt{\beta_1}$ indicates the skewness of the distribution while β_2 shows the kurtosis or peakedness of the distribution. Any substantial deviation of $\sqrt{\beta_1}$ and β_2 from the values indicated in equations (2.3.18) and (2.3.19) would indicate non-normality.

Given a random sample $x_1, x_2, \ldots, x_n$, estimates of $\sqrt{\beta_1}$ and β_2 can be obtained from

$$\sqrt{b_1} = \frac{m_3}{m_2^{\frac{3}{2}}} \tag{2.3.20}$$

and

$$b_2 = \frac{m_4}{m_2^2} \tag{2.3.21}$$

where the sample moments are

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k,\ k > 1 \tag{2.3.22}$$

and

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.3.23}$$

D'Agostino and Stephens (1986) outline various approaches for approximating the null distributions for $\sqrt{b_1}$ and b_2. We adopt the S_U approximation and the Anscombe and Glynn approximation respectively for $\sqrt{b_1}$ and b_2.

Steps of the Johnson S_U approximation for the null distribution of $\sqrt{b_1}$ are outlined below.

1. Compute $\sqrt{b_1}$ from sample data.

2. Compute

$$\begin{aligned} Y &= \sqrt{b_1}\left\{\frac{(n+1)(n+3)}{6(n-2)}\right\}^{\frac{1}{2}} && (2.3.24) \\ h_2 &= \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)} && (2.3.25) \\ W^2 &= -1+\{2(h_2-1)\}^{\frac{1}{2}} && (2.3.26) \\ q &= 1/\sqrt{\log W} && (2.3.27) \\ a &= \{2/(W^2-1)\}^{\frac{1}{2}} && (2.3.28) \end{aligned}$$

3. Compute

$$Z(\sqrt{b_1}) = q\log[Y/a + \left\{(Y/a)^2+1\right\}^{\frac{1}{2}}] \qquad (2.3.29)$$

Under normality, $Z(\sqrt{b_1})$ of (2.3.29) is approximately a standard normal random variable. The approximation is suitable for $n \geq 8$.

Steps of the Anscombe and Glynn approximation for the null distribution of b_2 are as outlined below.

1. Compute b_2 from sample data

2. Compute the mean and variance of b_2

$$E(b_2) = \frac{3(n-1)}{n+1} \qquad (2.3.30)$$

and

$$\mathrm{var}(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} \tag{2.3.31}$$

3. Compute the standardized value of b_2

$$x = \frac{b_2 - E(b_2)}{\sqrt{\mathrm{var}(b_2)}} \tag{2.3.32}$$

4. Compute the third standardized moment of b_2

$$\sqrt{\beta_1(b_2)} = \frac{6(n^2-5n+2)}{(n+7)(n+9)}\sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}} \tag{2.3.33}$$

5. Compute

$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}}\left[\frac{2}{\sqrt{\beta_1(b_2)}} + \sqrt{\left\{1 + \frac{4}{\beta_1(b_2)}\right\}}\right] \tag{2.3.34}$$

6. Compute

$$Z(b_2) = \left((1 - \frac{2}{9A}) - \left[\frac{1-(2/A)}{1 + x\sqrt{2/(A-4)}}\right]^{\frac{1}{3}} \right) \div \sqrt{2/(9A)} \tag{2.3.35}$$

Under H_0, $Z(b_2)$ of (2.3.35) is approximately a standard normal random variable. The Anscombe and Glynn approximation is suitable for $n \geq 20$.

D'Agostino and Pearson (1973) suggested the statistic

$$K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2) \tag{2.3.36}$$

where under H_0, $Z(\sqrt{b_1})$ and $Z(b_2)$ are independent standard normal random variables defined in equations (2.3.29) and (2.3.35). It then follows that if the null hypothesis of normality is true, K^2 is distributed as a chi-square random variable with two degrees of freedom.

2.3.8 Mixtures of regressions

Mixtures of regressions are mixture models that include covariates in the mixture formulation. They are useful in many disciplines. In the social sciences, they are known as latent class regressions and in machine learning they are referred to as hierarchical mixtures of experts.

A "mixture of linear regressions model" is a model where each $y_i,\ i = 1, 2, \ldots, n$ takes the form

$$y_i = \mathbf{x}_i^T \beta_j + \epsilon_{ij} \tag{2.3.37}$$

with probability $\pi_j,\ j = 1, 2, \ldots, K$. Here the number of predictor variables is p, y_i is the value of the response variable in the i^{th} observation, $\mathbf{x}_i^T$ $(i = 1, 2, \ldots, n)$ denotes the transpose of the vector of independent variables for the i^{th} observation, β_j $(j = 1, 2, \ldots, K)$ denotes the $p + 1-$dimensional vector of regressors for the j^{th} regression and π_j are the mixing probabilities satisfying

$$0 \leq \pi_j \leq 1 \text{ and } \sum_{j=1}^{K} \pi_j = 1.$$

The ϵ_{ij} are random errors which under the assumption of normality satisfy $\epsilon_{ij} \sim N(0, \sigma_j^2), i = 1, 2, \ldots, n; j = 1, 2, \ldots, K.$

The parameters

$$\Theta = (\pi_1, \pi_2, \ldots, \pi_K; \beta_1, \beta_2, \ldots, \beta_K, \sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2)$$

can be estimated by maximizing the log-likelihood

$$l(\Theta|x_1, \ldots, x_n; y_1, \ldots, y_n) \quad = \quad \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j \phi \left[\frac{y_i - \mathbf{x}_i^T \beta_j}{\sigma_j} \right] \right), \qquad (2.3.38)$$

where ϕ is the standard normal density function.

The expectation maximization algorithm is usually used for parameter estimation. Details can be found in e.g Turner (2000). Other methods for parameter estimation include the classification expectation maximization and the stochastic expectation maximization (Faria and Soromenho, 2010).

2.4 Results and discussion

A sample of 119 Swift gamma-ray bursts were available for analysis. The autocorrelation functions of the gamma-ray bursts were computed using equations (2.3.2), (2.3.4) and (2.3.5).

The normalization by $A(0)$ in (2.3.4) works reasonably well, but is not perfect. Figure (2.2) shows short lag autocorrelation functions for four gamma-ray bursts. Whereas the normalization is accurate for GRB100814A, the values of $A(0)$ seem to be slightly large for GRBs 091024 and 081008 and far too small for GRB 050904. The light curves of these GRBs are given in figure 2.3.

Our objective in this case is to determine the ACF width τ, defined as the lag at which the autocorrelation function decays to a value of 0.5. If the normalization is not correct then the estimated width τ will not be accurate; if $A(0)$ is too small τ will be overestimated, whereas if it is too large, τ will be underestimated. The remedy that

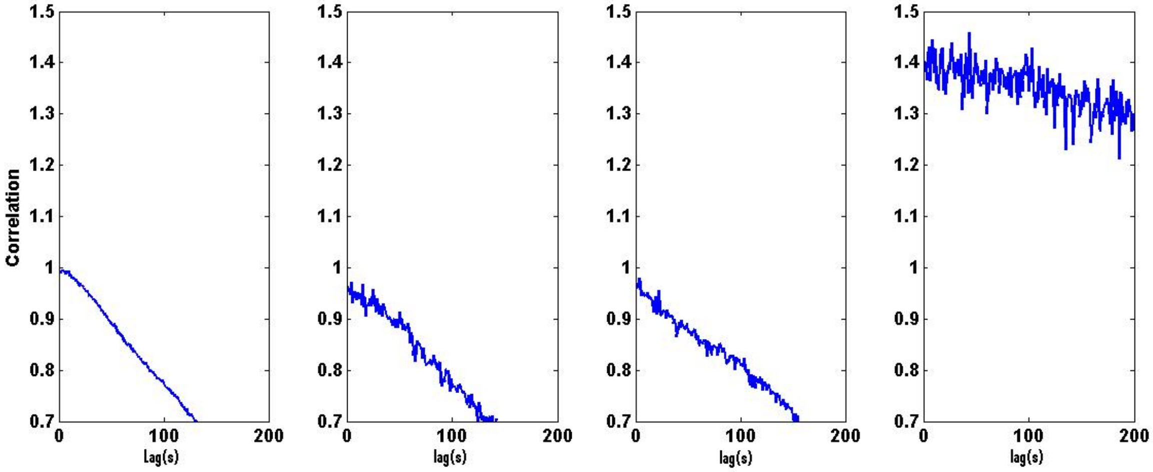


Figure 2.2: Short-lag ACFs for GRBs 100814A, 091024, 081008 and 050904.

we suggest for all such cases is to fit a low-order polynomial to the autocorrelation function values over small positive lags, extrapolate to find the value $A'(0)$ at zero lag and then normalize by this value. In practice, a quadratic was fitted to low-lag autocorrelation function values larger than 0.8. Examples of ACFs produced in this way are plotted in figure 2.4.

In Borgonovo (2004) and Borgonovo et al. (2007), the autocorrelation function width is computed on the basis of the assumption that the autocorrelation function decays in an approximately exponential fashion. A quadratic function $g(\tau)$ is then fitted to log$A(\tau)$ over the interval $0.4 \leq A(\tau) \leq 0.6$ and the lag τ_o such that $g(\tau_o) = \log(0.5)$ is noted.

The first and third panels of figure (2.4) show that this algorithm cannot be applied blindly since the respective secondary peaks near ~ 25 and ~ 27 also satisfy $0.4 \leq A(\tau) \leq 0.6$. It would also therefore seem useful to consider measures of autocorrelation function width which take the secondary peaks into consideration. We suggest use of

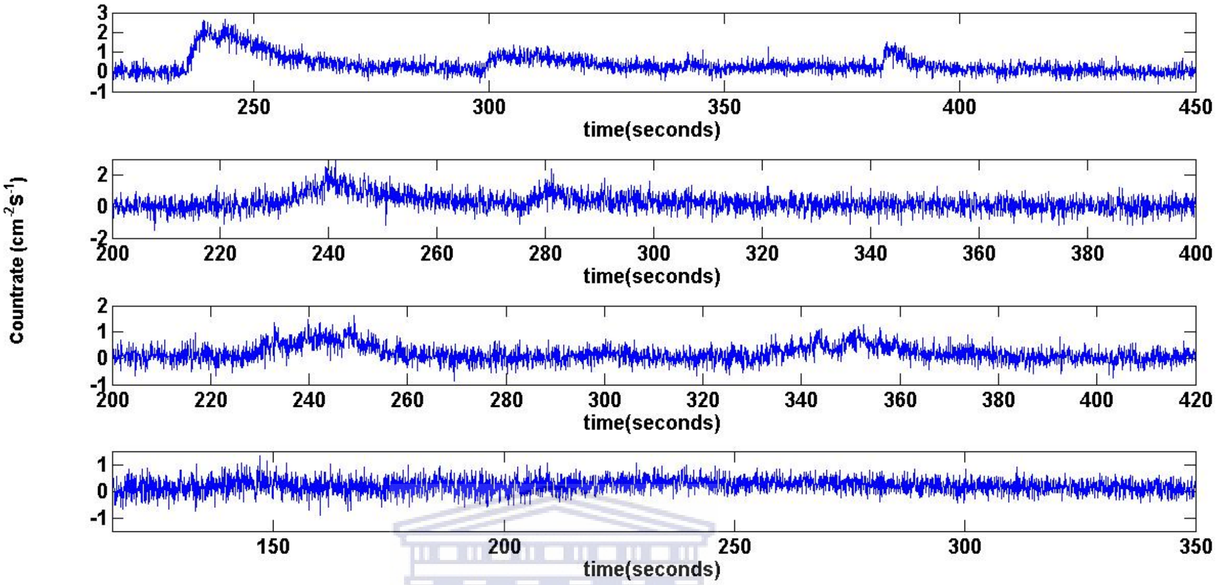


Figure 2.3: From top to bottom: light curves of GRBs 100814A, 091024, 081008 and 050904.

the mean $\tau_m \equiv \overline{\tau}$ of the lags such that $0.4 \leq A(\tau) \leq 0.6$ and this statistic will also be considered in the sequel. In the case of τ_o, if the ACFs have multiple maxima above 0.4, then only the small-lag part of the autocorrelation function is used in applying the algorithm in Borgonovo (2004) and Borgonovo et al. (2007).

Individual values of τ_o, and τ_m for different GRBs, of interest to astronomers, were published by Koen and Bere (2012). The discussion below focusses on whether the distributions are unimodal or not.

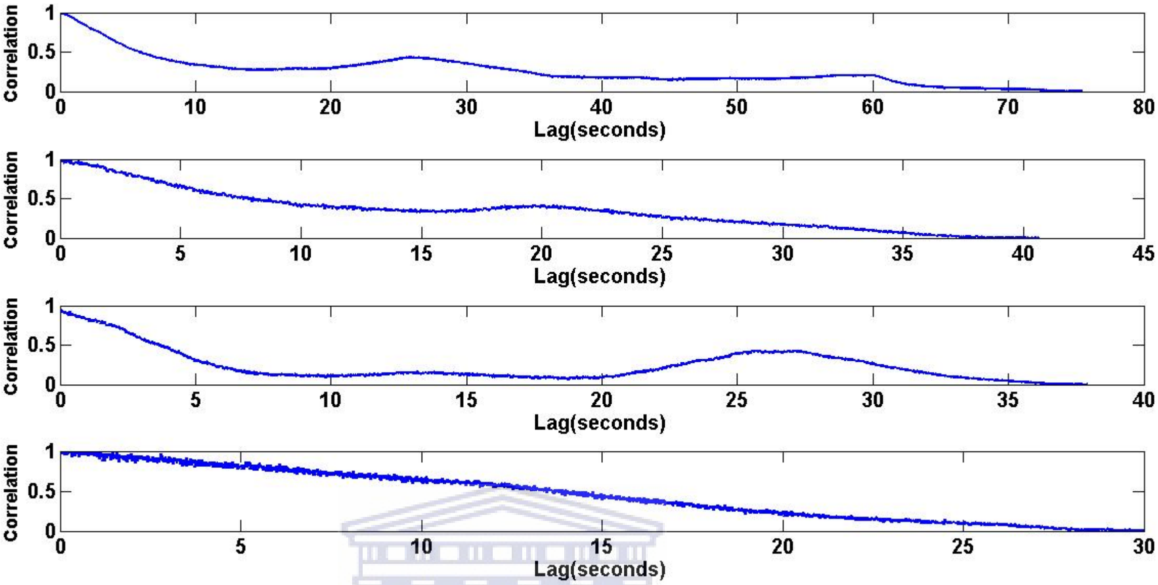


Figure 2.4: ACFs of the bursts in fig (2.3).

2.4.1 Histograms and kernel density estimates for log-transformed autocorrelation function widths

Figure (2.5) gives histograms of τ_o, τ_m and also autocorrelation function widths obtained by Borgonovo et al. (2007). The histograms reveal that the autocorrelation function width distributions are heavily skewed to the right. In the subsequent analysis, we work with logarithms (to base 10) of autocorrelation function widths as a way of reducing the influence of outlying observations.

Figure (2.6) shows histograms of logarithms (to base 10) of autocorrelation functions widths. There are suggestions of bimodality in figure (2.6) - there are dips in the histograms near $\log(\tau_o) \approx 0$ and $\log(\tau_m) \approx 0$. It is possible that these dips are a result of the choices of bin positions and/or random fluctuations.

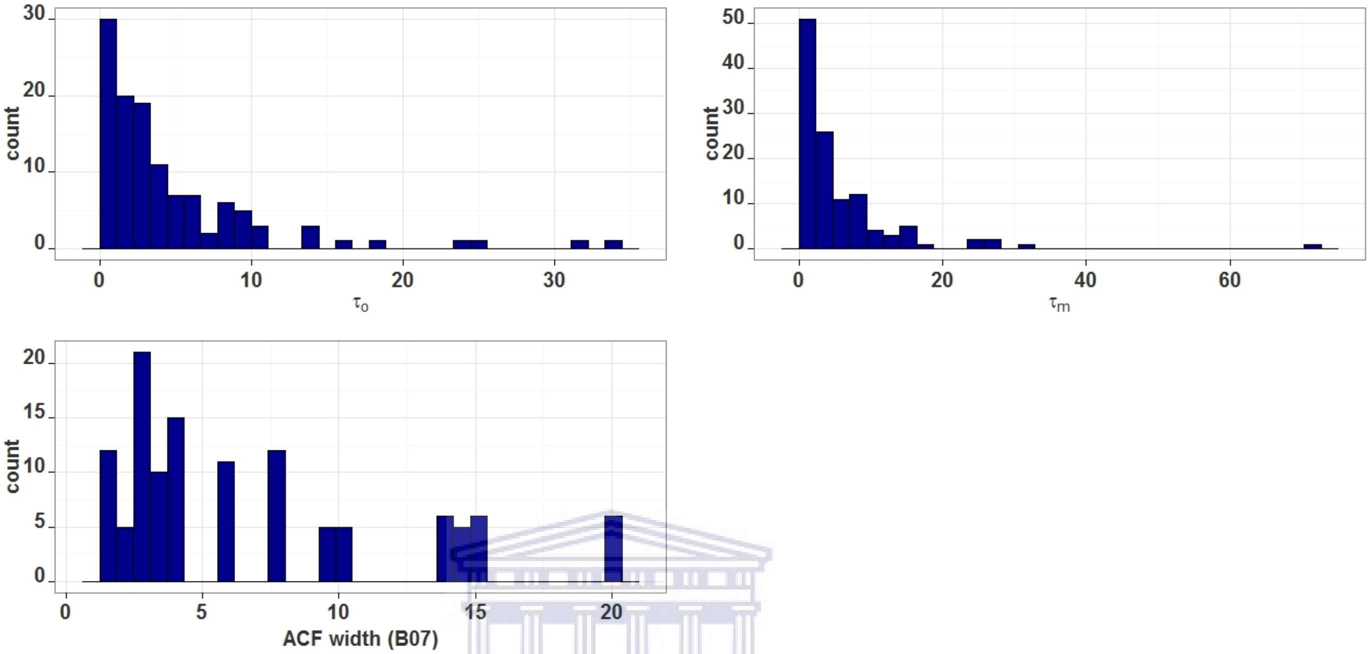


Figure 2.5: Histograms of autocorrelation function widths

Kernel smoothers provide more sophisticated estimators of the probability density functions of data than do histograms. Figure (2.7) gives plots of kernel density estimates for $\log(\tau_o)$ (on the left) and $\log(\tau_m)$ (on the right). We consider the Epanechnikov and Triangular kernels. "Normal scale" bandwidth values of $h = 2.34n^{-\frac{1}{5}}s$ and $h = 2.58n^{-\frac{1}{5}}s$ (see for example Wand and Jones, 1995, pages 60 and 178) were employed for the two kernels where s is an estimate of the spread of the data. The outlier resistant estimator

$$s = (x_{0.75} - x_{0.25})/1.34$$

was used. Here, $x_{0.75}$ and $x_{0.25}$ are the 75th and 25th percentiles of the distribution of autocorrelation function widths.

We also apply a density estimator proposed by Botev et al. (2010). The estimator

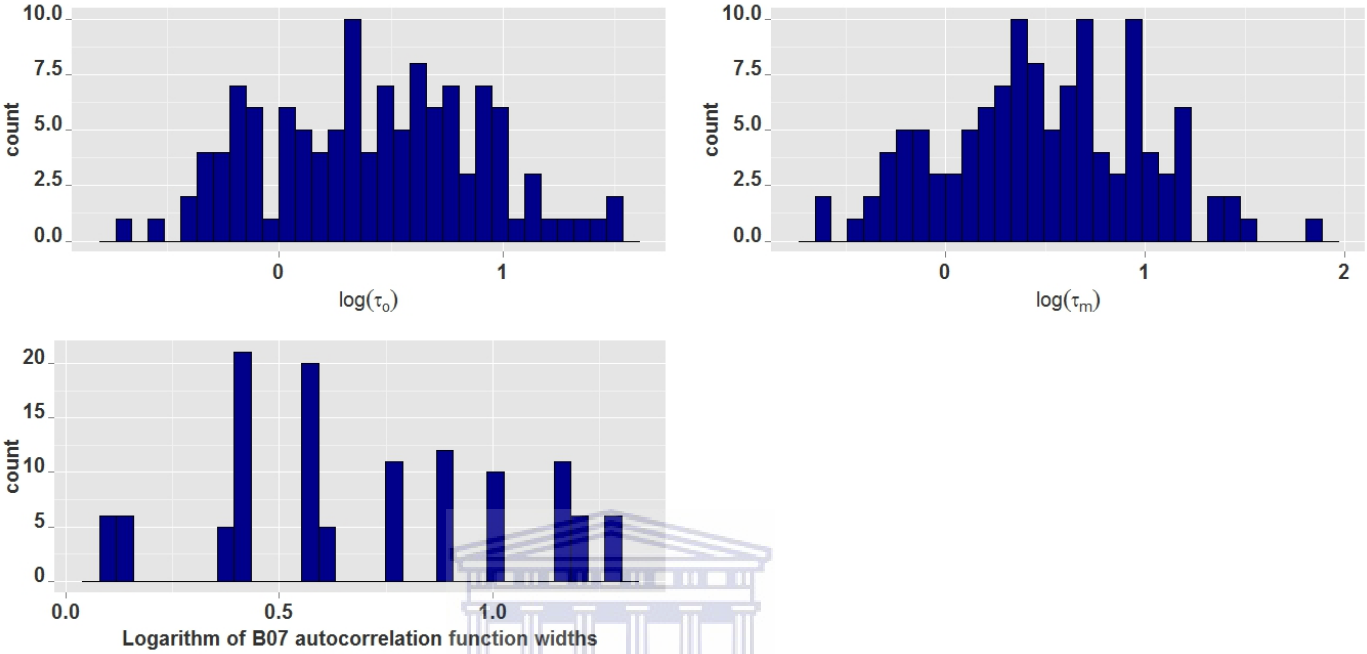


Figure 2.6: Histograms of logarithms of autocorrelation function widths

uses a non-parametric bandwidth and is good for multimodal data. Given a sample $\{\mathbf{x} = x_i,\ i = 1, \ldots, n\}$ and a bandwidth h, the kernel function K is of the form

$$K(x, x_i; h) = \sum_{j=-\infty}^{\infty} \left[\psi(x, 2j + x_i; h) + \psi(x, 2j - x_i; h)\right], \quad x \in [0, 1] \tag{2.4.1}$$

where ψ is defined as

$$\psi(x, x_i; h) = \frac{1}{\sqrt{2\pi h}} \exp\left(-\frac{(x - x_i)^2}{2h}\right). \tag{2.4.2}$$

For both $\log(\tau_o)$ and $\log(\tau_m)$, there is very good agreement between the three different kernel estimators. The estimated probability density function of $\log(\tau_m)$ is almost symmetrical and clearly unimodal. For $\log(\tau_o)$ the distribution has a slight bump. It is still necessary to carry out proper tests of hypothesis in respect of whether the autocorrelation function widths exhibit a bimodal distribution or not.

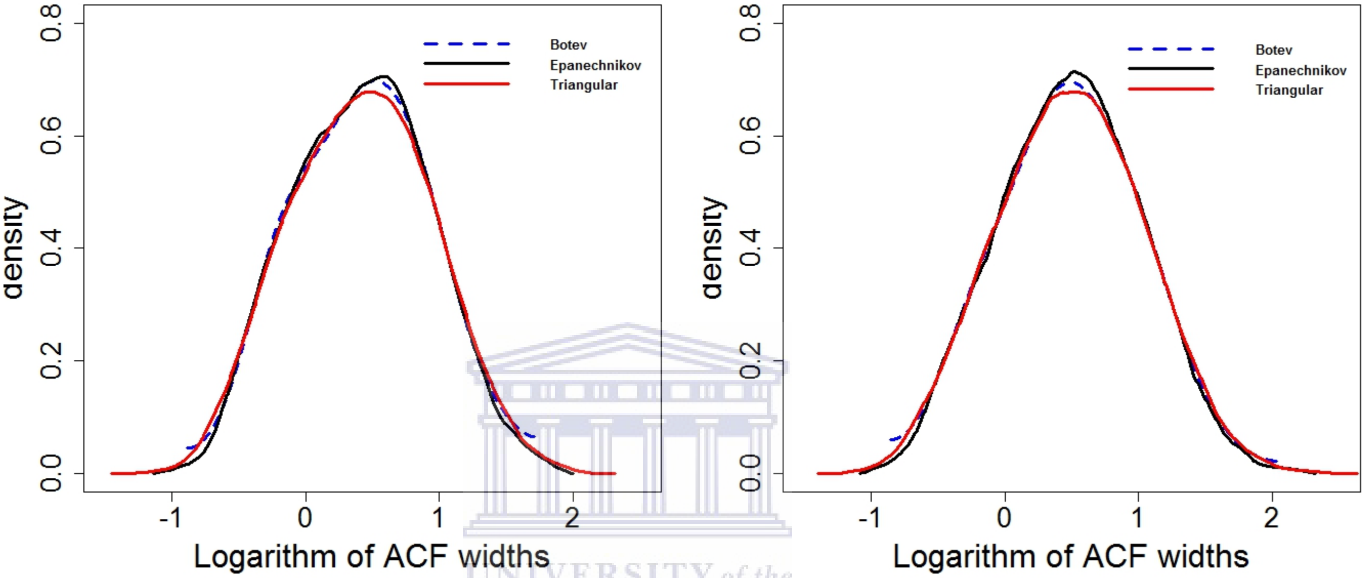


Figure 2.7: Kernel density estimates of logarithms of autocorrelation function widths τ_o (left) and τ_m (right).

2.4.2 The dip test

The dip statistics for the distributions of $\log(\tau_o)$ and $\log(\tau_m)$ are $D = 0.022$ and $D = 0.028$, respectively, with p-values 0.98 and 0.78. This indicates that there is no evidence for multimodality.

2.4.3 Modelling log-transformed autocorrelation function widths using mixtures of Gaussian distributions

Since bimodality of ACF widths may have important astrophysical implications, this point is explored further by explicit mixture modelling. We fitted the model

$$f(x;\mu,\sigma) \quad = \quad \pi_1 f(x;\mu_1,\sigma_1) + \pi_2 f(x;\mu_2,\sigma_2) \tag{2.4.3}$$

to the $\log(\tau_o)$ and $\log(\tau_m)$ data sets. Results obtained using MATLAB (version 2013a) are given in table (2.1) below. We also give the results for the 22 gamma-ray burst autocorrelation function widths reported in Borgonovo et al. (2007). There is very little, if any, correspondence between our results and those in Borgonovo et al (2007).

Data set	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\pi}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\pi}_2$
$\log(\tau_o)$	0.5	0.44	0.88	-0.22	0.083	0.12
$\log(\tau_m)$	0.54	0.47	0.93	-0.22	0.081	0.07
BO7	0.87	0.022	0.32	0.16	0.18	0.68

Table 2.1: Parameters estimated from fitting the Gaussian mixture model (2.4.3) to the logarithms of the ACF widths. Results for the logarithmically transformed B07 data are also given.

The next step was to investigate if there is a significant difference between the log-likelihood of the two component model (2.4.3) and that of a single Gaussian model. The likelihood ratio statistic

$$\Delta_{21} = -2(L_1 - L_2)$$

can be used used for this purpose. Here L_1 refers to log-likelihood of the single component Gaussian model and L_2 refers to the log-likelihood of a two-component Gaussian model.

Log-likelihoods, likelihood ratio statistics and p-values, based on a thousand simulations under the null hypothesis, are given in table (2.2). There is clearly no compelling evidence that the likelihood of the data significantly improves as we increase the number of components from one to two.

Data set	L_1	L_2	Δ_{21}	P-value
$\log(\tau_o)$	-81.08	-76.34	9.49	0.12
$\log(\tau_m)$	-85.52	-83.14	4.76	0.49

Table 2.2: Log-likelihood values, obtained after fitting single component and two-component Gaussian mixture models to the log(ACF) data, together with likelihood ratio statistics. p-values are based on $M = 1000$ replicates.

2.4.4 Testing if a single-Gaussian model fits adequately to the log-transformed autocorrelation function widths

The next step is to test if the single-Gaussian model fits the data adequately. As a first informal test, we examine normal quantile-quantile plots of the two data sets. These are given in figure (2.8) below. The plots do not show any serious deviation from normality since all the data points are fairly close to the straight line. This is followed by two formal tests of hypothesis namely the D'Agostino-Pearson test and the Anderson-Darling test.

P-values for the D'Agostino-Pearson statistic are based on the S_U transformation of $\sqrt{\beta_1}$ and the Anscombe and Glynn transformation of β_2. The results are summarized in table (2.3) below.

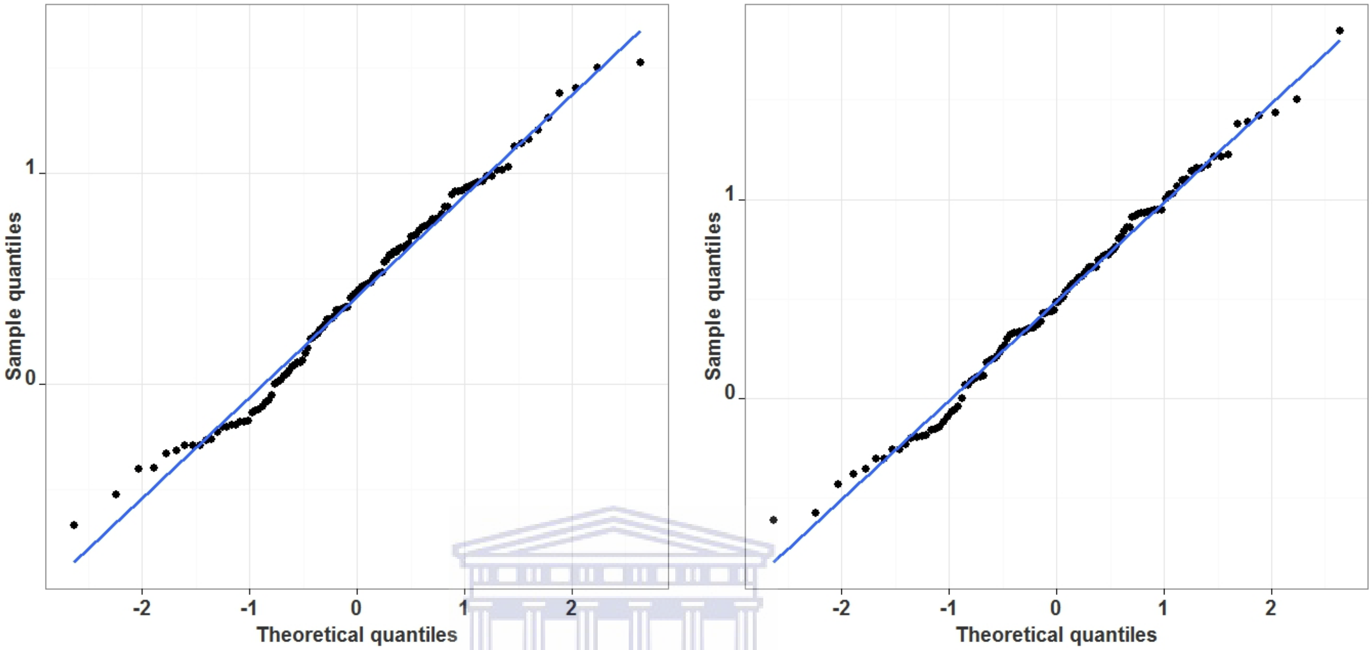


Figure 2.8: Quantile-quantile plots of logarithms of autocorrelation function widths for τ_o (left) and τ_m (right).

Data set	Anderson-Darling		D'Agostino-Pearson	
	Statistic	P-value	Statistic	P-value
$\log(\tau_o)$	0.4495	0.2767	2.1171	0.3470
$\log(\tau_m)$	0.2457	0.7586	0.9007	0.6374

Table 2.3: Values and p-values of the Anderson-Darling and D'Agostinho-Pearson statistics for testing if a normal distribution fits the $\log(\tau_o)$ and $\log(\tau_m)$ data.

The large p-values show that a single-Gaussian model is adequate for the data. The parameter values are $(\hat{\mu} = 0.42, \hat{\sigma} = 0.48)$ for $\log(\tau_o)$ and $(\hat{\mu} = 0.49, \hat{\sigma} = 0.50)$ for $\log(\tau_m)$. The implication is that τ_o and τ_m have got log-normal distributions.

2.4.5 Modelling log-transformed autocorrelation function widths using a mixture of regressions

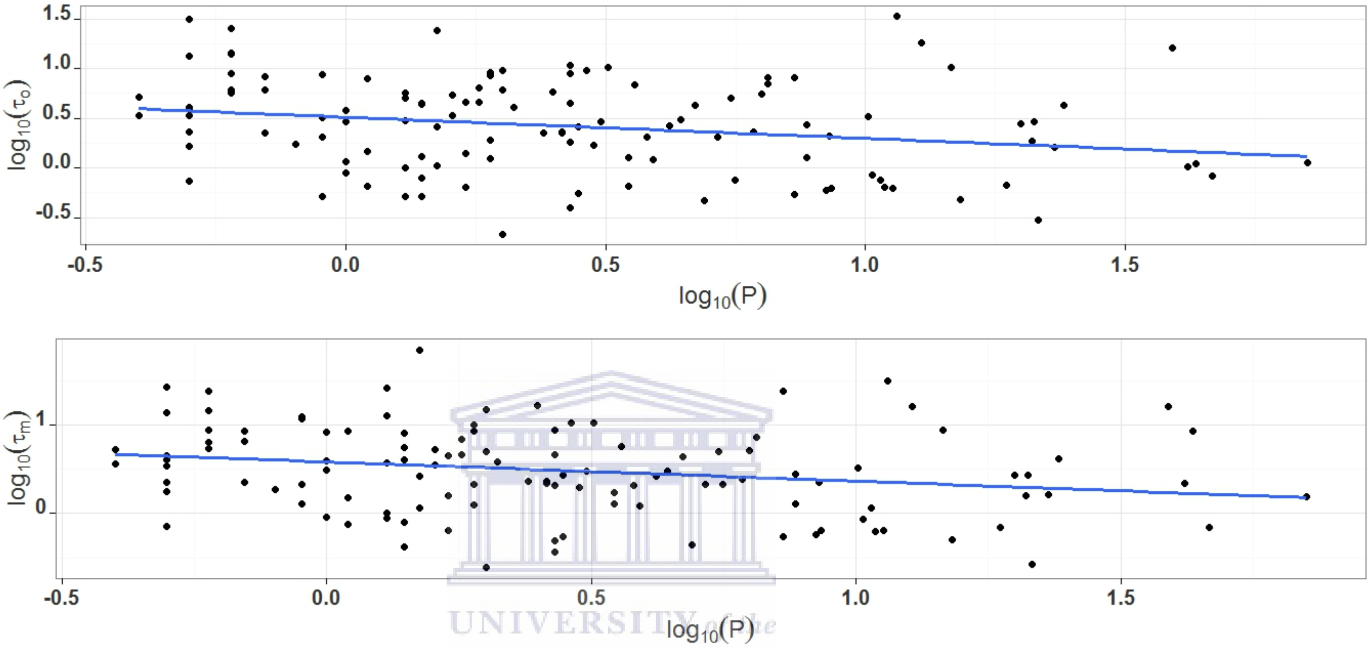


Figure 2.9: The dependence of autocorrelation function width on peak flux

Aside from the form of the evolution of the GRB radiation, other information can also be extracted from time series such as those in figure 2.1. One of the important measurables is the peak flux, i.e. the series maximum. Figure (2.9) shows that there is a significant correlation between autocorrelation function widths and peak flux of the gamma ray bursts. The regression lines are given by

$$\mathrm{Log}(\tau_o) = 0.50(0.056) - 0.21(0.081)\mathrm{Log}(P) \tag{2.4.4}$$

$$\mathrm{Log}(\tau_m) = 0.58(0.058) - 0.22(0.085)\mathrm{Log}(P) \tag{2.4.5}$$

The quantities in brackets are the standard errors of the parameter estimates. The respective F-statistics for the significance of the slopes are 6.814 (p-value= 0.0102) and 6.582 (p-value= 0.0116) indicating that the relationship between peak flux and the ACF widths is significant.

We set out to investigate the possibility that the data in figure (2.9) would better be modelled by a mixture of regressions. If this were the case, it could happen that subsets of autocorrelation function widths data selected on the basis of peak flux (e.g bright, high-flux bursts), would exhibit bimodality. The Gaussian likelihood ratio statistic was used to test the null hypothesis of a single linear regression versus the alternative of two regressions. The two competing likelihoods are

$$l_0 = \sum_{i=1}^{n} \log \left(\phi \left(\frac{y_i - \beta x_i}{\sigma} \right) \right)$$

and

$$l_1 = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{2} \pi_j \phi \left(\frac{y_i - \beta_j x_i}{\sigma_j} \right) \right).$$

The analysis was done in R using the package MIXREG. A parametric (Gaussian) bootstrap procedure with 1000 bootstrap replications was used to find the significance levels of the likelihood ratio statistic for the $\log(\tau_o)$ and $\log(\tau_m)$ data sets. The p-values were 0.54 and 0.92 respectively, implying that the gains from an extra regression are insignificant.

2.5 Conclusion

The objective of the work in this chapter was to prove or disprove the claim that widths of gamma-ray burst autocorrelation functions exhibit a bimodal distribution.

The contributions of our work are the following

(i) We suggested an alternative way of normalizing the gamma-ray burst autocorrelation function in the form of equation (2.3.4). This is not always adequate as demonstrated in figure (2.2). In such cases we have suggested the extrapolation of the autocorrelation function $A(l)$ from larger lags to $l = 0$ in order to determine $A(0)$.

(ii) We have suggested an alternative, more robust way of measuring the autocorrelation function width.

(iii) The following statistical techniques were employed in an effort to verify/disprove the claim that autocorrelation functions widths exhibit a bimodal distribution.

(a) The Dip Test of Bimodality.

(b) Gaussian mixture models with the number of components chosen on the basis of the likelihood ratio statistic whose p-values were obtained by simulation.

(c) The Anderson-Darling and D'Agostino-Pearson tests to confirm approximate unimodal normality.

(d) Mixtures of regressions to test if GRBs would exhibit bimodality if selected on the basis of their peak fluxes.

(iv) Contrary to findings in other studies, our analysis does not reveal any evidence of bimodality in the distribution of autocorrelation function widths.

2.6 Future work

In the current work, a visual inspection of the corresponding light curve was required to identify the starting point of each gamma-ray burst. This is obviously a very cumbersome method and hence it is necessary to come up with methods of automating this task.

An inspection of the light curves of gamma-ray bursts suggests that the pre- and post-burst data have constant means. Autocorrelation function plots also reveal that there is no serial autocorrelation in the pre- and post burst data. Informed by the two observations above we have tried to automate the process of identifying the start and end points of gamma-ray bursts using cumulative sums and the Box-Ljung test (Lung and Box, 1978). The Box-Ljung test is used to test if time series data are independently distributed.

The steps of our algorithm are as outlined below.

1. Identify the summit of the burst i.e. the position with highest intensity.

2. Moving to the left of the summit, take small windows of observations; and for each window, find the cumulative sum or apply the Box-Ljung test. The starting position of the burst is taken as the midpoint of the first window where the mean adjusted cumulative sum is zero or where the Box-Ljung statistic is non-significant.

3. The end position is identified by repeating step 2, moving to the right of the summit.

A preliminary investigation involving a small sample of gamma ray bursts shows that the results obtained from these algorithms compare very well with those obtained from visual inspection. A proper study would require realistic simulation of gamma-ray bursts with specified start and end points.

The work in this chapter involved use of univariate Gaussian mixture models. The next chapter contains applications of bivariate mixture modelling.

Chapter 3

Modelling using bivariate Gaussian mixture models

3.1 Introduction

In this chapter we apply bivariate Gaussian mixture models to model two data sets. The first data set comes from Horváth et al. (2010). The data are logarithms of the durations, T_{90} and the hardness ratios of 325 gamma-ray bursts observed by BAT on Swift (See the list of definitions for brief explanations of the meanings of these terms). These data were kindly supplied by Dr. Istvan Horváth. The results of the analysis that we performed on this data set are also reported in Koen and Bere (2012).

The second data set is from Lee et al. (2012). The data are observations of the spin periods (P) and period derivatives of a number of pulsars. The data were downloaded from http://www.atnf.csiro.au/research/pulsar/psrcat/. The data are plotted in figures (3.1) and (3.2).

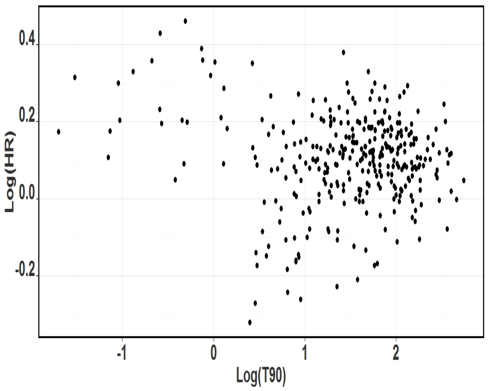


Figure 3.1: Scatter plot of GRB data.

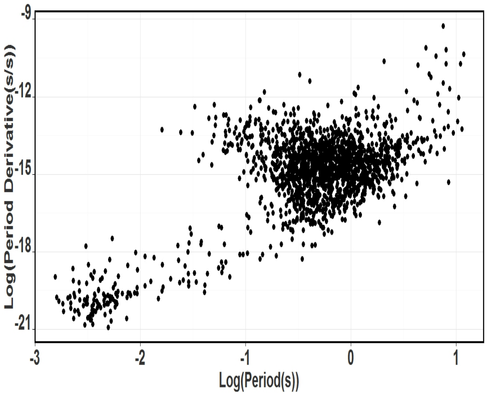


Figure 3.2: Scatter plot of pulsar data.

In arriving at the number of bivariate mixture components for the hardness ratio-duration data, Horváth et al. (2010), used the likelihood ratio test which compares the Gaussian-mixture likelihoods under the null and alternative hypothesis. It was assumed that the likelihood ratio statistic has an asymptotic chi-square distribution with the degrees of freedom being the difference in the number of parameters for the two competing models. As indicated before and as also discussed in Lehmann (1998), Andrews (2001), Miloslavsky and Vander Laan (2003), Lo (2005), and others, the regularity conditions required for the asymptotic chi-square distribution are not met; the parameters estimated under the alternative hypothesis are not identifiable under the null hypothesis and one of the mixing proportions lies on the boundary of the parameter space.

Furthermore, Horváth et al. (2010), identified the optimal number of components, but no attempt was made to actually test the final model for goodness-of-fit.

3.2 A goodness-of-fit test for bivariate data: the two-dimensional Kolmogorov-Smirnov test

The objective is to test if data come from a hypothesized bivariate distribution. An initial version of this test was developed by Peakock (1983). The test statistic T_n[1] is the largest difference between the empirical and theoretical cumulative probability distribution when all four possible ways to cumulate data following directions of the coordinate axes are considered. Peacock's test requires that both the empirical cumulative distribution and the cumulative distribution of the hypothesized model function be calculated in all $4n^2$ quadrants of the plane defined by

$$(x < X_i, y < Y_j), \quad (x < X_i, y > Y_j), \quad (x > X_i, y < Y_j), \quad (x > X_i, y > Y_j) \quad (i, j = 1, 2, \ldots, n)$$

for all possible combinations of the indices i and j.

Fasano and Franceschini (1987) proposed a simpler and faster to compute version of the Peacock test. Instead of considering all n^2 points $(X_i, Y_j, i, j = 1, \ldots, n)$ of the plane as suitable places to cumulate the data points and hypothesized distribution, Fasano and Franceschini suggested cumulating the data and model distribution in only the four quadrants of the plane defined by

$$(x < X_i, y < Y_i), \quad (x < X_i, y > Y_i), \quad (x > X_i, y < Y_i), \quad (x > X_i, y > Y_i) \quad (i = 1, 2, \ldots, n.)$$

In what follows, use will be made of the Fasano and Franceschini version of the two-dimensional Kolmogorov-Smirnov test. The exact distribution of the Fasano-Franceschini statistic is unknown. In practice, p-values for the test are obtained

[1] Statistics T_n are approximately proportional to $\frac{1}{\sqrt{n}}$ so sometimes use is made of the statistic $Z_n = \sqrt{n}T_n$.

through simulation.

3.3 Fitting of bivariate Gaussian mixture models to the GRB data

3.3.1 Likelihood ratio test for selecting the number of components in the mixture

Mixtures of two, three and four bivariate Gaussians were fitted to the paired ($\log T_{90}, \log HR$) data. The value of the likelihood ratio statistic comparing the three- and two component fits is $\Lambda_{32} = 35.86$, while $\Lambda_{43} = 13.54$. Percentage points of Λ_{32} were computed by simulating 1000 data sets of size 325 using parameters of the optimal two-component bivariate model fitted to the observations. The process was also carried out for Λ_{43} using parameters of the best fitting three-component model.The significance levels of the statistics are $p = 0.0030$ for Λ_{32} and $p = 0.6384$ for Λ_{43}. This confirms the conclusion of Horváth et al. (2010) that the best model is the one with three components.

The next task was to establish whether a three-component model represents an adequate representation of the data. This is done in three ways. Firstly we test whether the marginal distributions of the three-component bivariate Gaussian distribution fit to the separate duration and hardness ratio data. Next we try to establish and test for the number of univariate components in the individual duration/hardness ratio data. Finally we employ the two-dimensional Kolmogorov-Smirnov test to test the adequacy of the three-component bivariate Gaussian mixture model.

3.3.2 Testing if the marginal distributions of the fitted three-component model fit the durations and hardness ratio data

A goodness-of-fit test such as the Kolmogorov-Smirnov test can be used to establish whether the fitted marginal distribution gives a good representation of the empirical distribution. The distribution of the Kolmogorov-Smirnov statistic is known only in the case that the theoretical distribution is fully specified, i.e. none of the parameters need to be estimated. The same problem applies to many other test statistics. Use will therefore be made of the bootstrapping procedure described by Stute et al. (1993).

We assume here that the statistic T is based on the comparison of the empirical cumulative distribution function F_n and a partially specified theoretical cumulative distribution function $F(\boldsymbol{\theta})$. The steps of the bootstrapping procedure are outlined below.

(i) The theoretical probability density function (in the present case a mixture of Gaussians) depends on a number of unknown parameters. In the present context the parameters are means, variances and mixture proportions. Let $\boldsymbol{\theta}$ be the vector of unknown parameters. Estimate these and denote the estimate by $\hat{\boldsymbol{\theta}}$.

(ii) Compute the statistic of interest, $T_0 = T[F_n, F(\hat{\boldsymbol{\theta}})]$.

(iii) Compute a sample of the same size as the original data from the probability density function $f(\hat{\boldsymbol{\theta}})$ corresponding to $F(\hat{\boldsymbol{\theta}})$. Determine the empirical cumulative distribution F_{n*} of these data.

(iv) From the simulated sample, estimate the parameter values in exactly the same manner in which $\hat{\boldsymbol{\theta}}$ was estimated from the real observations. Let the vector of estimates be $\hat{\boldsymbol{\theta}}_*$.

(v) Calculate the statistic $T_* = T[F_{n*}, F(\hat{\boldsymbol{\theta}}_*)]$.

(vi) Repeat steps (iii)-(v) many (preferably a few thousand) times and the p-value will be the percentile of T_0 with respect to the collection of T_* values.

The procedure was carried out for the three component marginal distributions of the durations and hardness ratios using the Anderson-Darling statistic, $T = A^2$. A thousand simulated data sets were generated in each case. The Anderson-Darling statistics were $A^2 = 0.134$ ($p = 0.57$) and $A^2 = 0.416$ ($p = 0.002$) respectively, for durations and hardness ratios. It follows that the marginal distribution of the three component model provides an adequate description of durations, but not of the hardness ratio distribution.

3.3.3 Univariate tests for the number of components in the duration hardness ratio data

Univariate tests for the number of mixture components in the distribution of $\log T_{90}$ and log HR gave $\Lambda_{32} = 13.44$ with a significance level of $p = 0.045$, and $\Lambda_{32} = 2.79$ with $p = 0.48$ respectively, indicating three and two components respectively. The Anderson-Darling goodness-of-fit tests are not significant; the Anderson-Darling statistic for a three-Gaussian fit to the T_{90} data is 0.097 with a p-value of 0.88 and that for a two component fit to the hardness ratio data is 0.196 with a p-value of 0.42.

3.3.4 Testing if the three-component bivariate Gaussian mixture model provides a good fit

The two-dimensional Kolmogorov-Smirnov test was applied to the bivariate GRB data to ascertain if the three component mixture fits the data adequately. The significance level was determined by using the bootstrapping recipe described above. The value of the test statistic was $T_n = 0.0374$ and a thousand simulated data sets gave $p = 0.6134$. Contrary to what was found for the marginal distributions, this suggests that the three-component mixture is a good fit to the data. A brief discussion of this discrepancy follows in section 3.5.

3.4 Fitting of bivariate Gaussian mixture models to the pulsar data

3.4.1 Likelihood ratio test for fitting a mixture of up to six bivariate components

Lee et al. (2012) fit the data in figure 3.2 with a six-component bivariate mixture model. We fitted mixtures of two up to seven Gaussians to the data. Table 3.1 shows the change in log-likelihood that occurs each time the number of mixture components is increased by one. The corresponding p-values, obtained by simulation, are also given.

Statistic	p-value
$\Lambda_{32} = 295.2607$	$p_{32} = 0.0000$
$\Lambda_{43} = 91.5263$	$p_{43} = 0.0000$
$\Lambda_{54} = 88.0896$	$p_{54} = 0.0000$
$\Lambda_{65} = 46.0740$	$p_{65} = 0.0050$
$\Lambda_{76} = 24.6671$	$p_{76} = 0.2108$

Table 3.1: Values of the likelihood ratio statistics and the corresponding p-values.

We see that there is no significant change in the likelihood when we increase the number of components from six to seven. This confirms the conclusion of Lee et al. (2012) that a six-component bivariate Gaussian mixture model provides the best fit to the data. The next step is to try and establish the number of components in the margins.

3.4.2 Testing if the marginal distributions of the fitted six-component bivariate Gaussian mixture model fit the empirical data.

The Anderson-Darling test was used to ascertain if the margins of the fitted six-component bivariate distribution fitted the individual pulsar period or pulsar period derivative data. P-values were obtained using 1000 bootstraps of the test statistic. The values of the Anderson-Darling statistic were obtained as $A^2 = 0.3450$ ($p = 0.0000$) and $A^2 = 0.1949$ ($p = 0.02794$) for the period and period derivative data respectively. The conclusion is that the marginal distributions of the fitted 6-component bivariate Gaussian mixture model fit neither the period nor the period derivative data sets.

3.4.3 Use of the Anderson-Darling and likelihood ratio tests for determining the number of components of the period data.

Gaussian mixtures of from two up to seven components were fitted to the univariate pulsar period data. Values of the likelihood ratio statistic for comparing all models differing by one component were computed, together with their p-values. We also computed the values of the Anderson-Darling statistic for testing the fit of each of theses mixture models. The results are given in table (3.2) below.

Number of components	Anderson-Darling Statistic	P-value	Likelihood-ratio Statistic	P-value
6	0.0506	0.982	$\Lambda_{76} = 0.15$	0.8220
5	0.0662	0.9254	$\Lambda_{65} = 6.14$	0.3227
4	0.2924	0.1188	$\Lambda_{54} = 8.15$	0.2587
3	0.4223	0.0259	$\Lambda_{43} = 18.82$	0.00799
2	2.78	0.0000	$\Lambda_{32} = 80.47$	0.000

Table 3.2: Values of the likelihood ratio and Anderson-Darling statistics, together with the corresponding p-values, upon fitting mixtures of two up to seven Gaussians to the pulsar period data.

From the p-values in the table, we see that the Anderson-Darling test accepts anything between four and six components while the likelihood ratio test suggests that a four-component Gaussian mixture model will be adequate to describe the distribution of the period data.

3.4.4 Anderson-Darling and likelihood ratio tests for determining the number of components of the period derivative data

The analysis in the previous subsection was repeated for the period derivative data. The results are given in table (3.3) below.

Number of components	Anderson-Darling Statistic	P-value	Change in likelihood	P-value
6	0.0759	0.7300	$\Lambda_{76} = 0.25$	0.06993
5	0.1094	0.5205	$\Lambda_{65} = 8.75$	0.1389
4	0.1200	0.4525	$\Lambda_{54} = 12.17$	0.0400
3	0.3260	0.0939	$\Lambda_{43} = 9.946$	0.0549
2	3.83	0.0000	$\Lambda_{32} = 89.57$	0.000

Table 3.3: Values of the likelihood ratio and Anderson-Darling statistics, together with the corresponding p-values, upon fitting mixtures of two up to seven Gaussians to the pulsar period derivative data.

In this case, the Anderson-Darling test accepts anything between three and six components while the likelihood ratio test suggests that a four- or five-component Gaussian mixture model will be adequate to describe the distribution of the period derivative data.

3.4.5 Testing if the six-component mixture model provides a good fit using the bivariate Kolmogorov-Smirnov test

The value of the test statistic was $T_n = 0.0206$. A thousand simulated data sets gave a p-value of $p = 0.1637$. The result is again contrary to the results obtained from the analysis of the marginal distributions.

As was the case with the gamma-ray burst data, we see again that although the preferred model has six components, it does not give a very good fit in the margins. The distribution of the univariate period data can be adequately described by four Gaussian components while that of the period derivative data can be adequately described by five components.

3.5 Conclusion

Using simulated percentage points of the likelihood ratio statistic we have confirmed that the bivariate Gaussian mixture model with three components is the preferred model for the bivariate distribution of gamma-ray burst durations and hardness ratios considered in Horváth (2010). We have also confirmed that a bivariate mixture model with six components is the preferred model for the period-period derivative data considered in Lee et al. (2012). The bivariate Kolmogorov-Smirnov test also supports the adequacy of the fit of these models.

We also extended the analysis of the two data sets by investigating the number of components in the marginal distributions using the likelihood ratio and Anderson-Darling tests. The results show that the models above do not fit very well in the margins; the distribution of T_{90} values alone can be described by a three-component model while two-component mixture model is preferred for hardness ratios.

With regards to the data in Lee et al (2012), our results show that the distribution of the univariate period data can be adequately described by four Gaussian components while that of the period derivative data can be adequately described by five components.

The discrepancy between the multivariate Kolmogorov-Smirnov test result and the tests applied separately to the univariate distributions, is puzzling. Two pertinent issues, which will not be fully pursued here, are (i) power of the bivariate Kolmogorov-Smirnov test, and (ii) the overall significance level of the two univariate tests, bearing in mind the possible interaction between T_{90} and the hardness ratio. A study of the power of the Kolmogorov-Smirnov test could consist of the following steps.

1. Generate a large number of bivariate data sets where one of the margins consists of two components and the other margin is a mixture of three components. The two univariate data sets can be coupled by a copula.

2. Fit a three component mixture model to each data set.

3. For each data set use the Kolmogorov-Smirnov test to test if the three component model is acceptable.

4. Compute the percentage of rejections of the three component model.

We believe that copula models (see, for example, Genest and Favre, 2007 or Genest and Nešlehová, 2013) can do a better job of modelling the two data sets considered in this chapter. This idea will be explored in chapter 5 where we model the data sets using mixtures of copulas. In the next chapter we digress and look at the power and significance levels of tests of copula symmetry.

Chapter 4

Evaluating the power and significance levels of tests of symmetry for bivariate copulas.

4.1 Introduction

In chapter three, bivariate Gaussian mixture models were used to model the relationship between the following pairs of variables:-

(i) The natural logarithms of the hardness ratios and duration (T_{90}) values of gamma ray bursts as discussed in the paper by Horváth et al. (2010).

(ii) The period and period derivatives for the pulsar data discussed in Lee et al. (2012).

In both cases our analysis revealed that the bivariate Gaussian mixture models do not fit the data very well in the margins. We proposed that the data sets could possibly be modelled using models where the number of components in the X and Y margins differs, while the dependence pattern between the margins is captured by a copula (defined below).

There are many studies in the literature which have focussed on fitting copula models to data. Areas of application include agriculture (Larsen et al., 2013), hydrology (Nazemi and Elshorbagy, 2012), risk management (Embrechts et al., 2001) and marketing (Danaher and Smith, 2011).

Lack of symmetry (formally defined below) is quite evident in the scatter plots of the two astronomical data sets referred to above. This fact is formally verified in the next chapter. On the other hand, “most copula models used in practice are symmetric” (Genest et al, 2012). It is therefore prudent for researchers to test any data set for symmetry before attempting to fit symmetric copula models to it. Research on tests of copula symmetry started with the work of Jasson (2005) who proposed a test which is an adaptation of the chi-square test. The test statistic is therefore assumed to have an asymptotic chi-squared distribution under the null hypothesis.

Genest et al. (2012) argued that the test statistic cannot be assumed to be asymptotically distribution-free since its distribution depends on the underlying copula. They thus proposed three tests which are not distribution-free either but can be effectively implemented using the multiplier central limit theorem (van der Vaart and Wellner 1996). They gave a modified procedure for implementing the Jasson statistic, again based on the multiplier central limit theorem and also presented the results of a study of the power and significance levels of the proposed tests.

The study in Genest et al. (2012) covered a limited number of copula models. In this chapter, a similar study is extended to different copula models.

With specific regards to the tests discussed in this chapter, nothing is said in the literature about the role of the bandwidth parameter used for numerical estimation of the copula derivatives (required for the multiplier central limit theorem procedures).

Following ideas in Genest and Nešlehová, (2014), an investigation is carried out in an attempt to find a useful bandwidth for a given sample size.

Firstly we start by giving a formal definition of copulas and discuss those aspects of copulas that are relevant to the work in this and the next chapter. We then define copula symmetry and discuss the tests to be evaluated. After that we give the results of our own simulation study. The discussion will be limited to two dimensional copulas as applicable to this work.

4.2 Copulas

Copulas are joint cumulative distribution functions that describe dependencies among variables, independent of their marginal distributions (Joe, 1997). Copulas are "distribution functions whose one-dimensional margins are uniform" (Nelsen, 2006).

Formally, a two-dimensional copula is a function C from $[0,1]^2$ to $[0,1]$ with the following properties:

1. For every $u, v \in [0,1]$,

$$C(u,0) \;=\; C(0,v) = 0 \tag{4.2.1}$$

   and

$$C(u,1) \;=\; u \text{ and } C(1,v) = v \tag{4.2.2}$$

2. For every $u_1, u_2, v_1, v_2 \in [0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \quad \geq \quad 0 \tag{4.2.3}$$

If we consider a copula as a function which assigns points in $[0,1]^2$ to a point in $[0,1]$, then equation (4.2.3) says that the number assigned to each rectangle $[u_1, u_2] \times [v_1, v_2]$ in $[0,1]^2$ must be non-negative.

A theorem in Sklar (1959), is central to the theory and application of copulas. We state the theorem below as it is given in Nelsen (2006).

Theorem 4.2.1. *Let F be a joint distribution function with margins F_1 and F_2. Then there exists a copula C such that for all x, y in the real plane,*

$$F(x,y) \quad = \quad C(F_1(x), F_2(y)) \tag{4.2.4}$$

If F_1 and F_2 are continuous, then C is unique; otherwise C is uniquely determined on rangeF_1 $\times$ rangeF_2. Conversely, if C is a copula and F_1 and F_2 are distribution functions, then the function F defined by equation (4.2.4) is a joint distribution function with margins F_1 and F_2.

The major advantage of copulas is that they allow us to separate the marginal distributions from the dependence structure and model these separately. Furthermore, with copula modelling, it is quite possible for the marginal distributions to belong to different parametric distributions or even to be represented by their empirical estimates.

Another advantage of copulas which has given them popularity in the financial services industry especially in insurance is their ability to model tail dependence or dependency in the extreme values. Not all copulas allow for tail dependence though.

4.3 Other measures of dependence

Besides copulas, other measures of dependence include Kendall's τ and Spearman's ρ. These two are rank based dependence measures which better capture the existence of monotonic, but not necessarily linear dependence, compared to the Pearson correlation coefficient. Just like copulas these two measures are scale invariant. Furthermore, both Kendall's τ and Spearman's ρ can be expressed in terms of the copula function. Nelsen (2006) gives the formulae and proofs. It follows then that copula parameters can be expressed in terms of Kendall's τ or Spearman's ρ and vice-versa.

4.4 Families of copulas

The most popular families of copulas are elliptical copulas which are associated with elliptical distributions; Archimedean copulas which are defined by strictly decreasing functions called generators; extreme value copulas which are applicable to the modelling of the dependence structure between rare events; and survival copulas which are associated with survival functions. There are many other copulas which do not belong to these broad categories.

4.5 Symmetrical copulas

A two-dimensional copula C is termed as being symmetric or exchangeable if for any $u, v \in [0, 1]$,

$$C(u, v) \quad = \quad C(v, u) \tag{4.5.1}$$

i.e. the value that the copula function takes does not change when we switch the arguments.

4.6 Tests of copula symmetry

We start off by describing the tests proposed by Genest et al. (2012) then we give a description of the test proposed by Jasson (2005), together with its modification as proposed by Genest et al. (2012).

4.6.1 The tests proposed by Genest et al. (2012)

We assume that the pairs of observations (x_i, y_i) $i = 1, 2, \ldots, n$, are a random sample from the variables X and Y which are characterized by a bivariate distribution H with continuous margins F_1 and F_2. It follows from Sklar's theorem that there exists a copula C coupling H to the marginal cumulative distribution functions F_1 and F_2.

We define an estimator C_n of C by

$$C_n(u, v) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left(U_i \le u, V_i \le v \right) \tag{4.6.1}$$

where

$$(U_i, V_i) \;=\; (F_1(X_i), F_2(Y_i))$$

Usually the pair (U_i, V_i) is unknown and is estimated by

$$(\hat{U}_i, \hat{V}_i) \;=\; (F_{1n}(X_i), F_{2n}(Y_i)),$$

where

$$F_{1n}(X_i) \;=\; \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(X_j \leq X_i) = \frac{R_i}{n} \tag{4.6.2}$$

$$F_{2n}(Y_i) \;=\; \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(Y_j \leq X_i) = \frac{S_i}{n}. \tag{4.6.3}$$

Here, R_i stands for the rank of X_i among $X_1, X_2, \ldots, X_n$ and S_i stands for the rank of Y_i among $Y_1, Y_2, \ldots, Y_n$. Substituting $\hat{U}_i$ and $\hat{V}_i$ into equation (4.6.1) gives the empirical copula,

$$\hat{C}_n(u, v) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\hat{U}_i \leq u, \hat{V}_i \leq v). \tag{4.6.4}$$

To test the hypothesis of exchangeability as specified in equation (4.5.1), Genest et al. (2012) proposed three tests which compare the values of the empirical copula $\hat{C}_n$ at (u, v) and (v, u) for all choices of $u, v \in [0, 1]$. The rationale behind the tests is that we would expect $\hat{C}_n(u, v)$ to be close to $\hat{C}_n(v, u)$ whenever the null hypothesis of exchangeability holds. The three test statistics are

$$\begin{aligned} R_n &= \int_0^1 \int_0^1 \left\{ \hat{C}_n(u,v) - \hat{C}_n(v,u) \right\}^2 dvdu \\ S_n &= \int_0^1 \int_0^1 \left\{ \hat{C}_n(u,v) - \hat{C}_n(v,u) \right\}^2 d\hat{C}_n(u,v) \qquad (4.6.5) \\ T_n &= \sup_{(u,v)\in[0,1]^2} |\hat{C}_n(u,v) - \hat{C}_n(v,u)| \end{aligned}$$

Genest et al. (2012) show that the expressions for R_n, S_n and T_n are equivalent to the expressions given below which are easier to compute.

$$\begin{aligned} R_n &= \frac{1}{n^2}\mathbf{1}^T\mathbf{A}\mathbf{1} \\ S_n &= \frac{1}{n^3}\sum_{i=1}^{n}\mathbf{1}^T\mathbf{B}_k\mathbf{1} \qquad (4.6.6) \\ T_n &= \max_{i,j\in\{1,\ldots,n\}} \left| \hat{C}_n\left(\frac{i}{n},\frac{j}{n}\right) - \hat{C}_n\left(\frac{j}{n},\frac{i}{n}\right) \right| \end{aligned}$$

In equations (4.6.6) $\mathbf{1}$ is an $n \times 1$ vector of $1's$ and $\mathbf{A}, \mathbf{B}_1, \ldots, \mathbf{B}_n$ are $n \times n$ matrices with entry at position (i, j) given by

$$\begin{aligned} A_{ij} &= 2(1-\hat{U}_i \vee \hat{U}_j)(1-\hat{V}_i \vee \hat{V}_j) - 2(1-\hat{U}_i \vee \hat{V}_j)(1-\hat{U}_j \vee \hat{V}_i) \\ B_{kij} &= \mathbb{I}(\hat{U}_i \vee \hat{U}_j \leq \hat{U}_k, \hat{V}_i \vee \hat{V}_j \leq \hat{V}_k) - \mathbb{I}(\hat{U}_i \vee \hat{V}_j \leq \hat{U}_k, \hat{V}_i \vee \hat{U}_j \leq \hat{V}_k) \\ &\quad -\mathbb{I}(\hat{U}_i \vee \hat{V}_j \leq \hat{V}_k, \hat{V}_i \vee \hat{U}_j \leq \hat{U}_k) + \mathbb{I}(\hat{U}_i \vee \hat{U}_j \leq \hat{V}_k, \hat{V}_i \vee \hat{V}_j \leq \hat{U}_k), \end{aligned}$$

where $a \vee b = \max(a, b)$.

4.6.2 Asymptotic behaviour of the tests

The limiting behaviour of the statistics R_n, S_n and T_n is derived from the asymptotic behaviour of the empirical copula process $\hat{\mathbb{C}}_n$ defined for all $(u, v) \in [0, 1]^2$ by

$$\hat{\mathbb{C}}_n(u,v) \quad = \quad \sqrt{n}\left\{\hat{C}_n(u,v) - C(u,v).\right\} \tag{4.6.7}$$

For all $(u,v) \in [0,1]^2$ we can also define the symmetrised empirical process, $\hat{\mathbb{H}}_n$ as

$$\hat{\mathbb{H}}_n \quad = \quad \sqrt{n}\left\{\hat{C}_n(u,v) - \hat{C}_n(v,u)\right\} \tag{4.6.8}$$

If $C(u,v) = C(v,u)$ holds for all $(u,v) \in [0,1]^2$,

$$\hat{\mathbb{H}}_n \quad = \quad \hat{\mathbb{C}}_n(u,v) - \hat{\mathbb{C}}_n(v,u) \tag{4.6.9}$$

It follows that the statistics defined in equations (4.6.5) can be expressed in terms of $\hat{\mathbb{H}}_n$ as reflected in equations (4.6.10) below.

It is shown in Genest et al. (2012) that if the null hypothesis of exchangeability stated in equation (4.5.1) holds, and the copula C is "regular" i.e. the partial derivatives

$$\dot{C}_{11}(u,v) = \frac{\partial}{\partial u}C(u,v) \text{ and } \dot{C}_{12}(u,v) = \frac{\partial}{\partial v}C(u,v)$$

exist and are continuous respectively on the sets $\{(u,v) \in [0,1]^2 : 0 < u < 1\}$ and $\{(u,v) \in [0,1]^2 : 0 < v < 1\}$, then as $n \to \infty$,

$$\begin{aligned}
nR_n &= \int_0^1 \int_0^1 \left\{\hat{\mathbb{H}}_n(u,v)\right\}^2 dvdu \rightsquigarrow \mathbb{H}_R = \int_0^1 \int_0^1 \left\{\hat{\mathbb{H}}(u,v)\right\}^2 dvdu, \\
nS_n &= \int_0^1 \int_0^1 \left\{\hat{\mathbb{H}}_n(u,v)\right\}^2 d\hat{C}_n(u,v) \rightsquigarrow \mathbb{H}_S = \int_0^1 \int_0^1 \left\{\hat{\mathbb{H}}(u,v)\right\}^2 dC(u,v) \\
\sqrt{n}T_n &= \sup_{(u,v)\in[0,1]^2} |\hat{\mathbb{H}}_n(u,v)| \rightsquigarrow \mathbb{H}_T = \sup_{(u,v)\in[0,1]^2} |\hat{\mathbb{H}}(u,v)|
\end{aligned} \tag{4.6.10}$$

Here $\rightsquigarrow$ denotes weak convergence and, $\hat{\mathbb{H}}$ is a Gaussian random field defined for all $(u,v) \in [0,1]^2$ by

$$\hat{\mathbb{H}}(u,v) \;=\; \mathbb{H}(u,v) - \dot{C}_{11}(u,v)\mathbb{H}(u,1) - \dot{C}_{12}(u,v)\mathbb{H}(1,v), \qquad (4.6.11)$$

in terms of a centred Gaussian random field $\mathbb{H}$ with covariance function given by

$$\Gamma_{\mathbb{H}}(u,v,s,t) = \text{cov}\left\{\mathbb{H}(u,v), \mathbb{H}(s,t)\right\} \;=\; 2\left\{\Gamma_{\mathbb{C}}(u,v,s,t) - \Gamma_{\mathbb{C}}(u,v,t,s)\right\}$$

for each $u, v, s, t \in [0,1]$. Here $\Gamma_{\mathbb{C}}(u,v,s,t) = \text{cov}\left\{\mathbb{C}(u,v), \mathbb{C}(s,t)\right\}$.

It follows from equations (4.6.10) and (4.6.11)that the asymptotic null distributions of the statistics $nR_n,\ nS_n$ and $\sqrt{n}T_n$ depend on the underlying form of the copula which is unknown in practice. It is therefore impossible to compute p-values using simulation.

One way to get around this problem is to generate bootstrap replicates of the limiting distributions of these statistics using the multiplier central limit theorem (see for example, van der Vaart and Wellner, 1996).

The steps of the “multiplier” procedure are outlined below.

(i.) Compute the statistic R_n, S_n or T_n.

(ii.) Define P_n at any $u, v \in [0,1]$ as the $n \times 1$ vector with i^{th} component

$$P_{in} \;=\; \mathbb{I}(\hat{U}_i \leq u, \hat{V}_i \leq v) - \mathbb{I}(\hat{U}_i \leq v, \hat{V}_i \leq u)$$

(iii.) For all $u, v \in [0,1]$ estimate

$$\widehat{\dot{C}}_{11}(u,v) \;=\; \frac{\hat{C}_n(u + l_n, v) - \hat{C}_n(u - l_n, v)}{2l_n} \qquad (4.6.12)$$

$$\widehat{\dot{C}}_{12}(u,v) = \frac{\hat{C}_n(u,v+l_n) - \hat{C}_n(u,v-l_n)}{2l_n} \tag{4.6.13}$$

(iv.) Fix a bandwidth $l_n \in (0, 0.5)$. Typically $l_n = b_n/\sqrt{n}$ (Genest and Nešlehovă, 2014) for some small integer-valued b_n. For each $h \in \{1, 2, \ldots, M\}$, where M is a large integer, repeat the following steps.

(a) Draw a vector $\kappa^{(h)} = (\kappa_1^{(h)}, \ldots, \kappa_n^{(h)})$ of independent non-negative random variables with unit mean and variance. A normal distribution with unit mean and variance, or an exponential distribution with unit mean is usually used to this end. Set

$$\overline{\kappa}_n^{(h)} = \frac{1}{n}\left(\kappa_1^{(h)} + \cdots + \kappa_n^{(h)}\right)$$

and

$$\Xi_n^{(h)} = \left(\frac{\kappa_1^{(h)}}{\overline{\kappa}_n^{(h)}} - 1, \ldots, \frac{\kappa_n^{(h)}}{\overline{\kappa}_n^{(h)}} - 1\right)$$

(b) Define the bootstrap replicate $\hat{\mathbb{H}}_n^{(h)}$ of $\hat{\mathbb{H}}$ at any $u, v \in [0,1]$ by

$$\hat{\mathbb{H}}_n^{(h)} = n^{-\frac{1}{2}}\Xi_n^{(h)}\left\{P_n(u,v) - \widehat{\dot{C}}_{11}(u,v)P_n(u,1) - \widehat{\dot{C}}_{12}(u,v)P_n(1,v)\right\} \tag{4.6.14}$$

(c) Compute the bootstrap replicate of the appropriate test statistic:

$$\begin{aligned} R_n^{(h)} &= \int_0^1 \int_0^1 \left\{ \hat{\mathbb{H}}_n^{(h)}(u,v) \right\}^2 dvdu, \\ S_n^{(h)} &= \int_0^1 \int_0^1 \left\{ \hat{\mathbb{H}}_n^{(h)}(u,v) \right\}^2 d\hat{C}_n(u,v), \qquad (4.6.15) \\ T_n^{(h)} &= \sup_{(u,v)\in[0,1]^2} |\hat{\mathbb{H}}_n^{(h)}(u,v)|. \end{aligned}$$

(v.) Compute the approximate respective p-values of R_n, S_n and T_n as

$$\frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(R_n^{(h)} > R_n), \quad \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(S_n^{(h)} > S_n) \text{ and } \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(T_n^{(h)} > T_n)$$

As proposed by Genest et al. (2012), the partial derivatives (4.6.12) and (4.6.13) were one sided at the boundary points of $[0, 1]$. More recently, Genest and Nešlehová (2014, 2013) replaced these with two-sided forms. If the bandwidth is sufficiently small, there should not be much difference in the results obtained using the two schemes.

For computational convenience, we let $\hat{\mathbb{H}}_n^{(h)} = \frac{1}{\sqrt{n}}\Xi_n^{(h)} Q_n$ for each $h \in \{1, \ldots, M\}$, where for all $u, v \in [0, 1]^2$,

$$Q_n(u,v) = P_n(u,v) - \hat{\dot{C}}_{11} P_n(u,1) - \hat{\dot{C}}_{12}(u,v) P_n(1,v)$$

$\hat{C}_n$ is a discrete distribution function, hence $S_n^{(h)}$ in the set of equations (4.6.15) can be computed as

$$S_n^{(h)} = n^{-3} \sum_{i=1}^{n} \left\{ \Xi_n^{(h)} Q_n(\hat{U}_i, \hat{V}_i \right\}^2 .$$

A numerical approximation involving an $N \times N$ grid is used to obtain rough estimates of $R_n^{(h)}$ and $T_n^{(h)}$ as,

$$\begin{aligned}
\hat{R}_n^{(h)} &\approx \frac{1}{nN^2} \sum_{k=1}^{n} \sum_{l=1}^{n} \left\{ \hat{\mathbb{H}}_n^{(h)} \left(\frac{k}{N}, \frac{l}{N} \right) \right\}^2 \\
&= \frac{1}{n^2 N^2} \sum_{k=1}^{n} \sum_{l=1}^{n} \left\{ \Xi_n^{(h)} Q_n \left(\frac{k}{N}, \frac{l}{N} \right) \right\}^2, \\
\hat{T}_n^{(h)} &\approx n^{-\frac{1}{2}} \max_{k,l \in \{1,\dots,N\}} \left| \hat{\mathbb{H}}_n^{(h)} \left(\frac{k}{N}, \frac{l}{N} \right) \right| \qquad (4.6.16) \\
&= \frac{1}{n} \max_{k,l \in \{1,\dots,N\}} \left| \Xi_n^{(h)} Q_n \left(\frac{k}{N}, \frac{l}{N} \right) \right|
\end{aligned}$$

The following result proven in Genest et al. (2012) ensures the validity of the multiplier method described above.

Proposition 4.6.1. *Let C be a regular symmetric copula. If*

$$\lim_{n \to \infty} l_n = 0, \quad \inf_{n \in \mathbb{N}} \sqrt{n} l_n > 0,$$

then for arbitrary $M \in \mathbb{N}$, the sequence $(\hat{\mathbb{H}}_n, \hat{\mathbb{H}}_n^{(1)}, \dots, \hat{\mathbb{H}}_n^{(M)})$ converges weakly as $n \to \infty$, to $(\hat{\mathbb{H}}, \hat{\mathbb{H}}^{(1)}, \dots, \hat{\mathbb{H}}^{(M)})$, where $\hat{\mathbb{H}}, \hat{\mathbb{H}}^{(1)}, \dots, \hat{\mathbb{H}}^{(M)}$ are independent copies of $\hat{\mathbb{H}}$.

4.6.3 Jasson's test

Given a sample of bivariate observations, the test involves partitioning the set $[0,1]^2$ into squares of width $1/L$ for some integer $L > 2$ and using the scaled ranks of the observations $(\hat{U}_1, \hat{V}_1), \dots, (\hat{U}_n, \hat{V}_n)$ to construct a contingency table by counting how many of them fall in each of the squares.

The idea of the test is that if the data are symmetric, the counts in the cells (i, j)

and (j, i) would roughly be the same. The test therefore involves comparison of the proportions of observations in the cells (i, j) and (j, i).

To be specific, for a random pair (U, V) with copula distribution C, and for $k, l \in \{1, \ldots, L,\}$ we define

$$p^L_{kl}(C) \quad = \quad P\left\{(U, V) \in \left(\frac{k-1}{L}, \frac{k}{L}\right] \times \left(\frac{l-1}{l}, \frac{l}{L}\right]\right\}. \tag{4.6.17}$$

When C is symmetric, then $p^L_{kl}(C) = p^L_{lk}(C)$. Jasson (2005) therefore proposed the test statistic

$$J_n \quad = \quad \sum_{k<l} J_{n,(k,l)} \tag{4.6.18}$$

where

$$J_{n,(k,l)} \quad = \quad \sqrt{n}\frac{W^L_{n,(k,l)}}{\{p^L_{kl}(\hat{C}_n) + p^L_{lk}(\hat{C}_n)\}^{\frac{1}{2}}} = \sqrt{n}\frac{p^L_{kl}(\hat{C}_n) - p^L_{lk}(\hat{C}_n)}{\{p^L_{kl}(\hat{C}_n) + p^L_{lk}(\hat{C}_n)\}^{\frac{1}{2}}} \tag{4.6.19}$$

It should be noted that in equation (4.6.18), the focus is restricted to pairs (k, l) where $1 \leq k < l \leq L$ because $W^L_{n,(k,l)} = -W^L_{n,(l,k)}$ and also, $W^L_{n,(k,k)} = 0$ for all $k, l \in \{1, \ldots, L\}$.

In Jasson (2005) the asymptotic distribution of J_n was assumed to be a chi-squared distribution with $(L-1)(L-2)/2$ degrees of freedom. However, Genest at al. (2012) argued that the statistic J_n is not distribution-free; the asymptotic distribution depends on the underlying copula. They proposed an alternative test statistic, based on bootstrap replicates of $\hat{\mathbb{H}}$. The steps involved in the computation of the statistic are outlined below.

1. Let M be a large integer, as before. For all $h \in \{1, 2, \ldots, M\}$ and $k < l$, set

$$\begin{aligned} \sqrt{n} W_{n,(k,l)}^{L(h)} &= \hat{\mathbb{H}}_n^{(h)}\left(\frac{k}{L}, \frac{l}{L}\right) - \hat{\mathbb{H}}_n^{(h)}\left(\frac{k-1}{L}, \frac{l}{L}\right) \\ &- \hat{\mathbb{H}}_n^{(h)}\left(\frac{k}{L}, \frac{l-1}{L}\right) + \hat{\mathbb{H}}_n^{(h)}\left(\frac{k-1}{L}, \frac{l-1}{L}\right), \end{aligned} \qquad (4.6.20)$$

   where $\hat{\mathbb{H}}_n^{(h)}$ is as defined in equation (4.6.14).

2. For each $h \in \{1, 2, \ldots, M\}$, write

$$\sqrt{n}\mathbf{W}_n^{L(h)} = \left(\sqrt{n} W_{n,(1,2)}^{L(h)}, \ldots, \sqrt{n} W_{n,(L-1,L)}^{L(h)}\right)^{\top}. \qquad (4.6.21)$$

   It then follows from the proposition (4.6.1) that when the null hypothesis of exchangeability is correct, the vectors $\sqrt{n}\mathbf{W}_n^{L(1)}, \ldots, \sqrt{n}\mathbf{W}_n^{L(M)}$ are asymptotically independent copies of $\mathbf{W}^L$ where $\mathbf{W}^L$ is a centred Gaussian vector. Furthermore, the empirical covariance matrix based on $\mathbf{W}_n^{L(1)}, \ldots, \mathbf{W}_n^{L(M)}$ provides a consistent estimate $\hat{\sum}_L$ of the covariance matrix $\sum_L$ of $\mathbf{W}^L$

3. The statistic proposed in Genest et al. (2012) is

$$J_n^L = (\mathbf{W}_n^L)^{\top} \hat{\sum}_L^{-1} \mathbf{W}_n^L, \qquad (4.6.22)$$

   with an asymptotic χ_v^2 null distribution where $v = \text{rank}(\sum_L)$ degrees of freedom. Genest et al. (2012) speculated that for most classical copula models, $\sum_L$ is of full rank, therefore, $v = (L-1)(L-2)/2$.

4.7 Results on the power and significance levels of the tests

A combination of R (version 3.1.2) and MATLAB (2013A) programs was used for computation. It was possible to run a few R commands in MATLAB using the statconnDCOM link (Baier and Neuwirth, 2007). This link was used in those situations where MATLAB functions were not readily available to carry out certain tasks. Examples of such tasks include generation of values of the Joe copula and conversion from Kendall's τ to the copula parameter.

4.7.1 Significance levels of the tests

The first step was to carry out an investigation of a suitable bandwidth l_n for estimating the partial derivatives (4.6.12) and (4.6.13). Following Genest and Nešlehová (2014) and Quessy and Bahraoui (2013), we experimented with $l_n = \frac{b_n}{\sqrt{n}}$, $b_n = 1, 2, 3, 4$. The Clayton copula

$$C_\theta(u,v) \;\; = \;\; (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}},\ \theta \in (0, \infty) \qquad (4.7.1)$$

was used for this purpose. This choice was made so that our results could be compared to those of Genest et al. (2012) who also considered this copula. Sample sizes of $n = 100$ and $n = 250$ were considered.

Following Genest et al. (2012), the copula parameters chosen corresponded to values of Kendal's τ equal to $0.25, 0.5$ and 0.75. The different values of τ were selected to demonstrate the effect of the degree of dependence on the significance levels. The random variables $(\kappa_1^{(h)}, \kappa_2^{(h)}, \ldots, \kappa_n^{(h)})$ were taken to be normal random variables with unit mean and variance. Genest et al. (2012) used exponential random variables with a mean of 1.

There is ambiguity in Genest et al. (2012) regarding the value of N used in approximation (4.6.16). For $n = 100$, the heading of their table 1 suggests that $N = 50$ was used. However, in the paragraph immediately below the table, it is claimed that N was given by $N = n/5$. This would imply $N = 20$ for $n = 100$. It was thus deemed prudent to start by experimenting with both $N = 20$ and $N = 50$. All tests were carried out at 5% level of significance.

Tables (4.1) and (4.2) give the results on the number of rejections (out of a thousand simulations) of the null hypothesis of symmetry using the statistics R_n, S_n and T_n with $N = 50$ and $N = 20$ respectively, in approximation (4.6.16). Results for J_n^L, $L = 3, 4, 5, 6$ are also given in table 4.2. (The computation of the Jasson statistics does not involve N, hence we only give the single set of results).

b_n	τ	R_n	S_n	T_n
1	0.25	1.6	2.2	0.6
	0.5	2.2	1.7	1.6
	0.75	0.2	1.3	1.9
2	0.25	1.4	1.1	2.5
	0.5	1.3	2.2	2.9
	0.75	0.0	0.1	2.2
3	0.25	2.2	2.3	3.6
	0.5	0.7	1.7	2.5
	0.75	0.0	0.1	0.2
4	0.25	1.3	1.4	1.2
	0.5	0.3	0.9	0.7
	0.75	0.0	0.0	0.1

Table 4.1: The percentage of rejections of the null hypothesis of symmetry for samples of size $n = 100$ from the Clayton copula. For each of 1000 samples, H_0 was tested at the 5% level, using $M = 250$ bootstrap replications. $N = 50$ was used in approximation 4.6.16.

It is clear in from table (4.1) that for tests R_n, S_n and T_n the significance levels are far below the anticipated 5% when $N = 50$. There is an improvement with $N = 20$ as can be seen in table 4.2.

b_n	τ	R_n	S_n	T_n	J_n^3	J_n^4	J_n^5	J_n^6
1	0.25	2.2(2.0)	2.4(3.1)	3.6(3.9)	4.1(3.6)	4.5(4.6)	5.2(5.7)	7.4(7.3)
	0.5	1.6(1.5)	1.8(2.0)	4.4(6.1)	3.8(3.8)	3.4(3.5)	3.9(4.9)	5.0(6.1)
	0.75	0.3(0.5)	0.9(2.0)	7.5(5.7)	0.0(0.0)	5.2(5.5)	3.8(9.9)	21.0(9.4)
2	0.25	1.7	1.7	5.7	4.9	4.6	6.0	8.0
	0.5	1.0	1.4	7.2	4.7	3.7	4.2	5.3
	0.75	0.0	0.3	9.4	0.0	2.7	6.4	18.9
3	0.25	1.9	2.2	5.5	5.4	6.6	7.0	8.8
	0.5	0.5	1.1	5.3	4.5	5.4	4.3	4.8
	0.75	0.0	0.0	5.4	0.3	0.0	2.0	7.0
4	0.25	1.0	1.9	5.2	4.0	3.1	6.0	9.5
	0.5	0.2	0.3	5.6	4.5	2.5	2.3	4.1
	0.75	0.0	0.0	0.0	0.0	0.0	3.0	1.4

Table 4.2: The percentage of rejections of the null hypothesis of symmetry for samples of size $n = 100$ from the Clayton copula. For each of 1000 samples, H_0 was tested at the 5% level, using $M = 250$ bootstrap replications. For the statistics R_n, S_n and T_n, $N = 20$ was used in approximation 4.6.16. Results obtained by Genest et al. (2012) are given in brackets.

From table (4.2), we can see that the best results (in terms of being close to the 5% nominal level) are obtained when $b_n = 1$. Except for the discrepancy in the significance levels of J_n^5 and J_n^6 when $\tau = 0.75$, our results reasonably match the results in Genest et al. (2012) which are given in brackets.

For $\tau = 0.75$, pairs of ranks of the data tend to be concentrated along the line $y = x$. When $L = 5, 6$ we have very fine partitions of $[0, 1]^2$ and hence a high chance that a significant number of the partitions will be having zero entries. The end result will be that the variance-covariance matrix based on the vectors

$$\mathbf{W}_n^{L(1)}, \mathbf{W}_n^{L(2)}, \ldots, \mathbf{W}_n^{L(M)}$$

will be nearly singular and the corresponding inverse will be unstable.

Genest et al. (2012) also mentioned the sparseness of pairs of ranks of observations in the vicinity of $[0, 1]$ and $[1, 0]$ for $\tau = 0.75$. More so, our simulations also revealed many instances of $\tau = 0.75$ and $L = 6$ where the software would report that the variance-covariance matrix was either "singular to working precision" or "close to singular or badly scaled" and therefore the results were not accurate. In such cases the value of the Jasson statistic would be non-real (either complex or ill-defined). This seems to contradict the claim in Genest et al (2012) that "From numerical experimentation it seems Σ_L is of full rank for many classical (copula) models".

Our computer programs were coded in such a way that for each set of parameters, generation of random samples would continue until a real value of the test statistic was obtained. It is not clear how Genest et al. (2012) dealt with the samples where the Jasson statistic was ill-defined. This might explain why their significance levels for $\tau = 0.75$ and $L = 5, 6$ are lower than ours. We note that our approach leads to biased results in the sense that only samples with well-conditioned Σ_L are retained. It is not obvious how to remedy this.

Table (4.3) gives the results for $n = 250$. There is no single value of b_n which gives good results for all the five statistics. For the statistics R_n and S_n, the results are closest to 5% when $b_n = 3$. The case $\tau = 0.75$ aside, all the Jasson statistics and T_n give the best results when $b_n = 1$. Using these values our results reasonably match those obtained by Genest et al. (2012) which are given in brackets. In what follows we therefore use $N = 20$ in approximation (4.6.16) and $b_n = 1$ for all the statistics when $n = 100$. For $n = 250$, we use $N = 50$, $b_n = 3$ for R_n and S_n and $b_n = 1$ for T_n and the Jasson statistics. These are the values of b_n which result in significance levels which are nearest to 0.05 in tables (4.2) and (4.3). The values of b_n are chosen to ensure that all tests are as close as possible to the nominal level and at the same time are not too liberal, i.e. they do not reject the null hypothesis too often (Genest and Nešlehová, 2014).

In addition to the Clayton copula, the following bivariate copulas were also considered:-

(i) The Frank copula, given for all $(u, v) \in (0, 1]^2$ by

$$C_\theta(u, v) = -\frac{1}{\theta} \ln \left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right), \quad \theta \in (-\infty, \infty)\backslash\{0\}$$

(ii) The Joe copula, given for all $(u, v) \in (0, 1]^2$ by

$$C_\theta(u, v) = 1 - \left[(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta\right]^{\frac{1}{\theta}}, \quad \theta \in [1, \infty).$$

The results on the significance levels of the tests are given in table (4.4).

b_n	τ	R_n	S_n	T_n	J_n^3	J_n^4	J_n^5	J_n^6
1	0.25	3.3	3.2	3.7(4.0)	4.8(4.0)	3.4(3.1)	4.3(4.0)	5.1(6.1)
	0.5	2.3	2.9	4.4(4.0)	5.1(5.1)	2.9(2.7)	3.2(3.3)	4.4(4.1)
	0.75	0.7	2.4	6.0(4.0)	0.0(0.1)	1.1(4.2)	9.5(7.0)	21.7(9.9)
2	0.25	3.4	3.4	6.4	4.7	4.9	4.9	6.7
	0.5	3.0	2.8	7.0	5.6	3.5	3.5	3.7
	0.75	1.0	2.0	7.8	0.0	0.2	0.9	1.9
3	0.25	4.0(4.1)	3.5(4.2)	6.9	4.5	4.7	5.9	8.1
	0.5	2.7(2.1)	4.1(3.4)	6.9	4.7	4.3	4.1	3.7
	0.75	0.5(1.0)	1.7(2.7)	6.3	0.0	0.1	0.9	2.6
4	0.25	1.8	3.0	7.6	5.0	4.9	7.0	6.1
	0.5	1.6	3.1	5.4	4.9	4.3	4.1	3.8
	0.75	0	0.3	6.6	0.0	0.1	1.0	1.3

Table 4.3: The percentage of rejections of the null hypothesis of symmetry for samples of size $n = 250$ from the Clayton copula. For each of 1000 samples, H_0 was tested at the 5% level, using $M = 250$ bootstrap replications. For the statistics R_n, S_n and T_n, $N = 50$ was used in approximation 4.6.16. Results obtained by Genest et al. (2012) are given in brackets.

copula	n	τ	R_n	S_n	T_n	J_n^3	J_n^4	J_n^5	J_n^6
Frank	100	0.25	1.7	1.9	3.7	3.6	4.4	5.2	6.4
		0.5	1.6	1.8	3.8	4.0	2.9	3.5	4.0
		0.75	0.5	1.6	6.6	0.0	4.1	1.4	40.9
Joe	100	0.25	2.0	2.0	3.0	4.1	3.3	6.0	6.6
		0.5	1.9	2.7	3.7	2.7	3.5	4.3	2.9
		0.75	0.3	1.1	7.3	0.0	2.6	6.2	13.6
Frank	250	0.25	3.6	4.1	3.6	4.3	3.9	3.9	5.1
		0.5	2.3	2.9	4.2	4.3	4.3	4.9	5.0
		0.75	0.5	2.6	5.2	0.0	2.1	7.5	17.0
Joe	250	0.25	4.1	4.0	3.8	5.0	5.2	5.7	6.2
		0.5	3.1	3.2	4.4	4.3	4.4	4.2	4.2
		0.75	0.6	2.7	5.7	0.0	0.1	1.9	2.6

Table 4.4: The percentage of rejections of the null hypothesis of symmetry for samples of size $n = 100$ and $n = 250$ from the Frank and Joe copulas. For each of 1000 samples, H_0 was tested at the 5% level, using $M = 250$ bootstrap replications. $N = 20$ and $N = 50$ were used for sample sizes $n = 100$ and $n = 250$ respectively, in approximation 4.6.16.

When $n = 100$ the significance levels of the statistics R_n and S_n are much lower than the anticipated 5%. The statistics T_n, J_n^3 and J_n^4 have slightly better significance levels though they tend to perform poorly when $\tau = 0.75$. This agrees with Genest et al. (2012) who observed that for $n = 100$, the statistics T_n and J_n^4 had better performance as compared to the other statistics. J_n^5 and J_n^6 tend to be either too liberal or too conservative when $n = 100$.

As obtained in Genest at al. (2012) there is an improvement in the significance levels when n is increased to 250 though there are discrepancies when $\tau = 0.75$.

4.7.2 Power of the tests

To assess the power of the tests, the Frank and the Joe Copula were made asymmetric by Khoudraji's device (Genest et al. 1998, Liebscher 2008). An asymmetric version of a copula C is given for $u, v \in [0,1]^2$ by

$$K_\eta(u,v) = u^\eta C(u^{1-\eta}, v), \quad \eta \in (0,1). \tag{4.7.2}$$

The full algorithm for sampling from an asymmetric Archimedean copula is given in Mai and Scherer (2012). We will use the symbols K_η^F and K_η^J to denote the asymmetric versions of the Frank and Joe Copula respectively. As an alternative to the use of Khoudraji's device use could have been made of Liouville Copulas (Liebscher 2008, McNeil and Nešlehová 2010).

Table (4.5) below gives the power of the tests for $n = 100$. The results for $n = 250$ are given in table (4.6). The pattern of the results is partly summarized in figure (4.1) which gives the power of the tests for $\tau = 0.7$, $\eta \in \{0.1, 0.2, \ldots, 0.5\}$.

copula	τ	η	R_n	S_n	T_n	J_n^3	J_n^4	J_n^5	J_n^6
K_η^F	0.5	0.25	11.6	11.1	10.5	8.6	9.0	7.8	11.2
		0.5	20.7	20.6	18.1	10.6	10.6	11.2	13.4
		0.75	9.6	9.6	9.9	8.0	5.2	5.1	7.2
	0.7	0.1	7.7	7.5	8.6	6.1	8.4	8.6	11.3
		0.2	30.0	26.7	19.6	17.4	17.2	15.5	16.7
		0.3	54.8	53.5	37.6	29.0	27.1	29.2	30.2
		0.4	68.4	67.6	52.5	38.3	41.1	36.1	38.1
		0.5	77.5	78.8	58.9	38.2	37.4	33.9	39.9
	0.9	0.25	72.2	76.5	51.1	36.5	49.4	51.5	52.1
		0.5	99.9	99.9	96.2	50.3	84.4	92.0	95.0
		0.75	94.2	99.6	91.6	0.2	4.2	50.9	85.5
K_η^J	0.5	0.25	37.1	38.8	27.0	17.0	18.3	18.9	22.7
		0.5	20.7	20.7	25.7	17.3	15.6	15.2	20.8
		0.75	10.6	7.7	11.7	7.9	7.5	6.1	8.3
	0.7	0.25	72.0	74.0	45.1	32.8	38.5	40.0	43.0
		0.50	93.5	94.6	75.7	45.5	51.3	48.9	59.5
		0.75	49.4	53.6	37.8	14.8	12.8	13.3	19.3
	0.9	0.25	78.4	86.4	56.0	40.0	55.5	58.0	61.7
		0.5	99.4	99.4	96.6	48.3	84.2	93.0	97.2
		0.75	93.9	99.5	92.0	0.7	6.8	46.9	82.6

Table 4.5: Power of the tests of copula symmetry based on R_n, S_n, T_n and J_n^L with $L \in \{3, 4, 5, 6\}$, as estimated from 1000 samples of size $n = 100$ from the Frank and Joe copulas. $N = 20$ was used in approximation 4.6.16. All tests were conducted at 5% level of significance and $b_n = 1$ and $M = 250$ bootstrap replicates were used for all the tests.

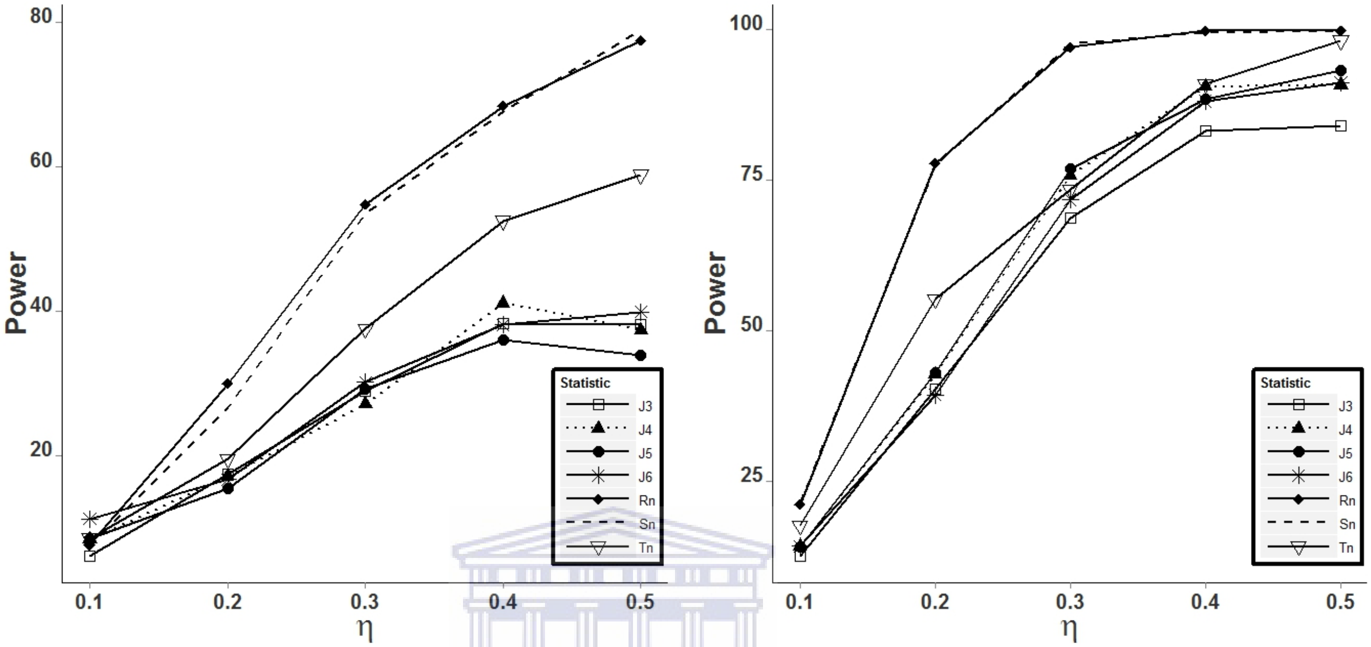


Figure 4.1: Power of the tests of copula symmetry based on R_n, S_n, T_n and J_n^L with $L \in \{3, 4, 5, 6\}$, as estimated from 1000 samples of size $n = 100$ (left panel) and $n = 250$ (right panel) from the Frank copula ($\tau = 0.7$), using $M = 250$ bootstrap replicates.

The following conclusions can be made from figure (4.1), tables (4.5) and (4.6).

1. The Cramér−von Mises statistics R_n and S_n are the most powerful, followed by T_n. Genest et al. (2012) concluded that the Statistic S_n is the most powerful.

2. The Jasson type statistics are the least powerful. Their hierarchy in terms of power is not easy to tell. Genest et al. (2012) concluded that the power increases with the value of L.

3. As expected the power increases with sample size and with the value of τ. This is in agreement with the results in Genest et al. (2012).

4. For a fixed value of τ, the maximum power is achieved when $\eta = 0.5$. This is not

surprising; Genest et al. (2012) used an asymmetry index due to Nelsen (2007) to show that for a given copula, Khoudraji's device gives maximum asymmetry when $\eta = 0.5$.

copula	τ	η	R_n	S_n	T_n	J_n^3	J_n^4	J_n^5	J_n^6
K_η^F	0.5	0.25	40.8	40.2	21.7	21.1	20.0	19.2	18.2
		0.5	66.2	65.5	43.4	30.8	32.2	31.8	25.4
		0.75	37.2	35.2	22.1	17.7	15.0	12.6	12.3
	0.7	0.1	21.2	21.5	17.7	12.6	14.2	14.1	14.3
		0.2	77.8	77.6	55.3	40.4	42.8	43.0	39.4
		0.3	97.1	97.7	73.5	68.8	75.8	76.9	71.8
		0.4	99.8	99.6	91.0	83.2	90.5	88.4	88.1
		0.5	99.8	99.7	98.2	84.0	90.8	93.2	91.2
	0.9	0.25	100.0	100.0	100.0	78.7	96.2	97.2	96.8
		0.5	100.0	100.0	100.0	96.9	100.0	100.0	100.0
		0.75	100.0	100.0	100.0	10.8	43.1	83.1	98.2
K_η^J	0.5	0.25	100.0	100.0	60.4	40.8	49.0	55.7	54.9
		0.5	91.1	90.6	60.9	47.9	51.3	53.2	51.4
		0.75	100.0	100.0	19.7	19.2	18.3	16.7	15.9
	0.7	0.25	99.4	99.7	89.2	71.3	88.3	91.4	90.5
		0.50	100.0	100.0	100.0	92.8	97.1	97.7	98.2
		0.75	98.4	98.9	82.0	60.6	55.4	52.1	45.8
	0.9	0.25	100.0	100.0	97.2	78.0	96.1	98.8	98.4
		0.5	100.0	100.0	100	96.4	100.0	100.0	100.0
		0.75	100.0	100.0	99.7	13.1	47.0	81.0	95.1

Table 4.6: Power of the tests of copula symmetry based on R_n, S_n, T_n and J_n^L with $L \in \{3, 4, 5, 6\}$, as estimated from 1000 samples of size $n = 250$ from the Frank and Joe copulas. $N = 50$ was used in approximation 4.6.16. All tests were conducted at 5% level of significance.$M = 250$ bootstrap replicates and $b_n = 1$ was used for all the tests except R_n and S_n for which $b_n = 3$ was used.

4.8 Application of the tests of copula symmetry to the GRB and pulsar data sets

We also applied the tests discussed in this chapter to the GRB data set and to a two-sample partition of the pulsar data which was obtained using the K-means clustering algorithm. For the GRB data ($n = 325$) and the larger pulsar data set ($n = 1595$), $l_n = \frac{3}{\sqrt{n}}$ was used for R_n and S_n and $l_n = \frac{1}{\sqrt{n}}$ was used for T_n and the Jasson statistics i.e. we chose the recommended bandwidth for $n = 250$, the largest sample size for which a suitable bandwidth had previously been established. Similarly, for the smaller pulsar data ($n = 164$), a bandwidth of $l_n = \frac{1}{\sqrt{n}}$ which had been found to be suitable for $n = 100$ was used for all the statistics.

The results are given in table (4.7). We give the value of each statistic together with the corresponding p-value (in brackets). We can see that the statistics R_n, S_n and T_n classify all the data sets as asymmetric. All the tests confirm asymmetry in the larger pulsar data set. For the GRB data, the statistics J_n^3 and J_n^4 do not reject the null hypothesis of exchangeability while the other two Jasson statistics lead to rejection of this null hypothesis. For the smaller pulsar data set, J_n^3 and J_n^5 lead to rejection of the null hypothesis while the other two Jasson statistics lead to non-rejection.

Overall the results in table (4.7) suggest lack of exchangeability in all the three data sets. It will therefore be a futile exercise to attempt to fit symmetric copula models to these data sets.

Statistic	GRB data	Smaller pulsar	Larger pulsar
	set	data set	data set
R_n	$5.1 \times 10^{-4}(0.00)$	$3.2 \times 10^{-4}(0.00)$	$5.1 \times 10^{-4}(0.00)$
S_n	$4.3 \times 10^{-4}(0.00)$	$4.4 \times 10^{-4}(0.00)$	$5.2 \times 10^{-4}(0.00)$
T_n	$5.8 \times 10^{-2}(0.00)$	$6.1 \times 10^{-2}(0.04)$	$2.4 \times 10^{-2}(0.01)$
J_n^3	$0.7(0.40)$	$4.0(4.55 \times 10^{-2})$	$8.9(2.85 \times 10^{-3})$
J_n^4	$1.9(0.60)$	$4.9(0.18)$	$37(4.60 \times 10^{-8})$
J_n^5	$16.22(1.26 \times 10^{-2})$	$18(6.23 \times 10^{-3})$	$71(2.55 \times 10^{-13})$
J_n^6	$29(1.25 \times 10^{-3})$	$16(0.10)$	$110(0.00)$

Table 4.7: Values of the tests statistics together with the corresponding p-values (in brackets) upon applying the tests of copula symmetry R_n, S_n, T_n and J_n^L with $L \in \{3, 4, 5, 6\}$, to the GRB data set and to partitions of the pulsar data

4.9 Conclusion

We have extended the study of the power and significance levels of tests of copula symmetry/exchangeability which was initially conducted by Genest at al. (2012). The Clayton copula was used to determine a good bandwidth l_n for estimating copula derivatives and the appropriate value of N for obtaining approximate values for the bootstrap replicates of the statistics R_n and T_n. The preferred values of l_n and N in this case were the ones resulting in tests whose significance levels were close to the nominal 5% level. The results show that for samples size 250, the statistics R_n and S_n have their significance levels closest to the nominal 5% when $l_n = 3/\sqrt{n}$; while for $n = 100$ the significance levels are closest to 0.05 when $l_n = 1/\sqrt{n}$. The rest of the tests give the best results with $l_n = 1/\sqrt{n}$. Nothing is said in Genest et al. (2012) about the issue of the bandwidth, making it very difficult for their results to be reproducible.

The Frank and Joe Copula were used in the actual study of the significance levels. Khoudraji's device was used to obtain asymmetric versions of these two copulas which were used in the power study.

Our results show that for $n = 100$, the majority of the tests have significance levels below the expected 5% nominal level. There is a slight improvement in the significance levels when the sample size increases to 250. This agrees with Genest et al. (2012).

With regards to power, our results show that the two Cramḗr−von Mises type statistics R_n and S_n are the most powerful followed by T_n. The Jasson type statistics are the least powerful. The power of the tests increases with sample size and also with the value of τ. The power also increases with the value of the parameter η used in Khoudraji's technique up to a peak value at $\eta = 0.5$. All this is in agreement with previous findings. Contrary to the Genest et al. (2012) who concluded that the power of the Jasson tests increases with finer partitions of the $[0, 1]^2$ grid, our results do not show any hierarchy in terms of the power of these tests.

Our study is very much limited, covering only two copula models, two sample sizes and integer values between 1 and 4 for the numerator of l_n. It was not possible to widen the scope of this study because of the computer run-time involved.

Chapter 5

Modelling dependence using mixtures of copulas

5.1 Introduction

In chapter three, bivariate Gaussian mixture models were used to model the relationship between the following pairs of variables:-

(i) The natural logarithms of the hardness ratios and duration (T_{90}) values of gamma ray bursts as discussed in the paper by Horváth et al. (2010).

(ii) The logarithms of periods and period derivatives for the pulsars discussed in Lee et al. (2012).

In this chapter we propose a new approach to modelling the same data based on mixtures of rotated copulas. It is well known that appropriately weighted mixtures of copulas are also copulas (see for example Nelsen, 2006, page 14). It follows that mixtures of copulas retain the usual benefits derived from copula modelling. For example, unlike the bivariate Gaussian models employed in chapter three, copula mixture modelling allows the relaxation of the requirement that the mixture components be bivariate normal and also facilitates separate modelling of the marginal distributions and the dependence structure.

Mixtures of copulas have the added advantage that they can model a variety of distribution shapes, that are totally different from the shapes captured by the individual constituent copulas. For example if we combine a Gaussian copula (which does not capture tail dependence) and a Clayton copula, we get a copula structure which is able to capture lower tail dependence. This flexibility can be further enhanced by the inclusion of rotated copulas in the mixtures.

There are a lot of previous research studies where mixtures of copulas have been applied. The applications are mainly in the financial disciplines; modelling dependencies between variables such as stocks in different markets, stock prices and volumes, prices of different commodities such as gold and oil, or exchange rates of different countries.

Hu (2006) uses mixtures of preselected copulas to model the dependence between four stock market indices from different parts of the world. The copulas used in the mixture are the Gaussian copula, the Gumbel copula and the Gumbel survival copula which is just a rotation of the Gumbel copula through an angle of 180°. The Gaussian copula is chosen because of its use in previous studies (see for example Li, 2000) involving financial data. The choice of the other two copulas is motivated by the desire to investigate the existence of left and right tail dependence in the data. A chi-squared test is used for testing the fit of the copula models. To our knowledge, the properties of this test such as power and the ability to maintain the intended significance level are not documented for this context.

Rodriguez (2007) also applies mixtures of copulas to model the dependence between pairs of daily returns for East Asian and Latin American stocks. He uses "regime-switching copulas" where the dependence pattern is allowed to vary with the variances (volatility) in the marginal distributions. The objective is to determine if the dependence pattern varies with marginal volatility. The models considered are two- and three-component mixtures involving the Clayton, Gumbel and Frank copulas. Wang et al. (2013) also use a "dependence-switching copula" to model the association between stock returns. Their model switches between an equally weighted mixture of the Clayton copula and the survival Clayton copula and another equally weighted mixture of Clayton copulas rotated by 90 and 270 degrees respectively. The former mixture is meant to capture positive dependence between the stock and foreign exchange markets while the latter is meant to capture negative dependence between the markets.

Other studies in which mixtures of copulas are applied to financial data include Li and Liang (2005), Hong et al. (2007), Wang (2008), Ning and Wirjanto (2009), Trivedi and Zimmer (2009), Dias and Embrechets (2010) and Arakelian and Karlis (2014). Li and Liang (2005) consider mixtures of Gaussian copulas only. Hong et al. (2007) apply a mixture of the Gaussian and Clayton copulas while Ning and Wirjanto (2009) use two- and three component mixtures involving the Frank, Clayton, Gumbel and survival Clayton copulas. Trivedi and Zimmer (2009) apply two- and three-component mixtures involving the Gaussian, Gumbel and Clayton copulas, Dias and Embrechts (2010) use a mixture of the Clayton and survival Clayton copula while Arakelian and Karlis (2014) apply two-component mixture copulas resulting from all the possible combinations of the Gaussian, Gumbel, Frank, Joe and Frank copulas.

The work of Wang (2008) is interesting in the sense that it partially addresses a long-standing question in copula modelling; how to select the appropriate copula functions. In Wang (2008), the components for a copula mixture are selected through a penalized maximum likelihood estimation algorithm. The method starts with a "working model"; a mixture of say three or four candidate copulas. The working models considered are three- and four-component mixtures involving the Clayton, Gaussian, Gumbel, survival Gumbel and Frank copulas. Component copulas with small weights are then removed by a thresholding rule given by the penalty function. The method produces promising results with synthetic data sets. The question of how to choose the constituent copulas of the "working model" remains open though.

There are also other studies where mixtures of copulas are applied to non-financial data. Tewari et al. (2011) proposes the use of Gaussian Mixture Copula Models (GMCMs) for clustering data with non-Gaussian components. The models are applied to synthetic data sets and also in image segmentation. The R-package GMCM (Bilgrau et al., 2015) was developed for estimation of parameters in Gaussian mixture copula models. Ghosh et al. (2011) apply six copulas (the Gaussian, Student-t, Gumbel, Clayton, Frank and Kernel)and the fifteen possible combinations of two different (copulas) to the pricing of crop insurance based on the crop yield-price joint distribution. Vrac et al. (2012) use mixtures of copulas to partition a sample of (cumulative) distribution functions into clusters. Wu (2014) applies mixtures of copulas to model the joint distribution of the age and usage of cars. Only two copula mixtures are considered; a convex combination of the Gumbel copula rotated 180° and another Gumbel copula rotated by 270°, and the combination of two Gumbel copulas rotated by 180°. The motivation behind Wu's work is slightly different from ours though; he uses rotated copulas in order to produce asymmetry.

A recent study by Kosmidis and Karlis (2015) uses simulated data sets to illustrate the failure of Gaussian mixture models to capture tail dependence and advocates the use of mixtures of copulas. The authors illustrate their proposed method using four-component mixtures built from all possible permutations of two Clayton copulas which are intended to model lower tail dependence and two Gumbel copulas which are expected to capture lower tail dependence. The use of rotated copulas is also suggested in the study. For example, the authors show that the Gumbel copulas can be replaced by the survival Clayton copula which can also capture upper tail dependence. An interesting contribution of this study is a formulation where the angle of rotation of each constituent copula is among the parameters used for maximizing the likelihood. The advantages are two-fold. Firstly this makes it possible to rotate the copulas through any angle between 0 and 360°. Also, as the authors illustrate, the choice of copulas to go into the mixture is simplified; instead of working with many copulas, different rotations of the same copula can be used to capture the shape of the data and the tail dependence pattern.

In the majority of the papers cited above, the choice of copulas which make up the mixture model is limited to a few copulas. The choice is usually guided by features of the data such as asymmetry and tail dependence. Also, only a few studies cited above have utilized rotated copulas. In this study, we present a more general and systematic method of constructing mixtures of copulas which makes it possible to use many candidate copula models and utilizes all four standard rotations of the copulas.

The models presented in the studies cited above and those proposed in this thesis are not easily applicable in multivariate settings beyond two dimensions. Simple parametric copula models are not flexible enough to uncover complex dependence structures of higher dimensional data (Weiß and Scheffer, 2015), and have other restrictions such as parameter restrictions (Kim et al., 2013). Vine copulas (see for example Mai and Scherer, 2012, chapter 5) have been developed to model complex multivariate data sets. Vine copulas are graphical models that allow us to represent a $d-$dimensional multivariate density using $d(d-1)/2$ bivariate copula densities (sometimes also referred to as "pair copulas") in a hierarchical manner.(Kim et al., 2013).

There are two popular types of vines; C-vines where one needs to specify in advance the relationships between one specific variable and the others, and D-vines where one starts off by specifying pairings of the variables.

Some recent research studies have extended the idea of a mixture model to vine copulas. For example, Kim et al. (2013) use mixtures of D-vine copulas to uncover complex and hidden patterns in simulated and real data sets. Weiß and Scheffer (2015) propose the use of mixtures of copulas as pair copulas in multi-dimensional vine copula models to minimize the risk of misspecifying a vine model. Vine copula models and their variants are not considered in this thesis because the data sets discussed herein are all bivariate.

Very few of the studies cited above used formal copula goodness-of-fit tests. In most cases the best copula mixture models are arrived at on the basis of log-likelihood values or information criteria such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). There is no guarantee that the model with the lowest AIC or BIC will fit the data. In this project goodness-of-fit testing for the copula mixture models is performed using the Cram$\acute{e}$r$-$von Mises test. In the next section we give a description of the model used. The discussion is limited to the bivariate case for the reason that the application data sets are all bivariate.

5.2 Model formulation

Given pairs of observations (x_{1i}, x_{2i}) $i = 1, 2, \ldots, n$, of the variables X_1 and X_2, we wish to model the joint distribution of X_1 and X_2, $H(x_1, x_2)$. In the present work, X_1 could be the logarithm of the duration (T_{90}) while X_2 will be the logarithm of the hardness ratio or, for the pulsar data set, X_1 could be the logarithm of the period, while X_2 is the logarithm of the period derivative.

According to Sklar's theorem, there exists a copula C such that for all (x_1, x_2) in $\mathbb{R}^2$,

$$H(x_1, x_2; \mathbf{\Theta}) \;\; = \;\; C(F_1(x_1; \boldsymbol{b}_1), F_2(x_2; \boldsymbol{b}_2); \mathbf{\Phi}). \qquad (5.2.1)$$

In (5.2.1), $F_1(x_1; \boldsymbol{b}_1)$ and $F_2(x_2; \boldsymbol{b}_2)$ are the respective marginal cumulative distributions of the variables X_1 and X_2 and $\mathbf{\Theta} = \{\boldsymbol{b}_j, j = 1, 2; \mathbf{\Phi}\}$. The $\boldsymbol{b}_j s$ are the parameter vectors of the marginal distributions and $\mathbf{\Phi}$ is the parameter vector of the copula C.

We assume that $H(x_1, x_2)$ can be decomposed into a mixture of two bivariate joint distributions as

$$H(x_1, x_2; \mathbf{\Theta}) \;\; = \;\; \sum_{k=1}^{2} p_k H_k(x_1, x_2; \mathbf{\Theta}_k). \qquad (5.2.2)$$

Here $\mathbf{\Theta} = \{p_k, \mathbf{\Theta}_k; k = 1, 2\}$ represents the parameter vector of the joint distribution

H and $\mathbf{\Theta}_k$ is the parameter vector of each component. The p_k are the proportions, satisfying $p_1 + p_2 = 1$.

Similarly to equation (5.2.1), we can express each component $H_k(x_1, x_2; \mathbf{\Theta}_k)$ as

$$H_k(x_1, x_2; \mathbf{\Theta}_k) \;=\; C_k\left(F_1(x_1; \boldsymbol{b}_1^k), F_2(x_2; \boldsymbol{b}_2^k); \mathbf{\Phi}_k\right), k = 1, 2. \tag{5.2.3}$$

Here $\mathbf{\Theta}_k = \left\{\boldsymbol{b}_j^k, j = 1, 2; \mathbf{\Phi}_k, k = 1, 2\right\}$ is the parameter vector of the copula C_k. The $\boldsymbol{b}_j^k$'s are the parameter vectors of the marginal distributions while the $\mathbf{\Phi}_k$'s represent the parameter vectors of the k^{th} copula. This means that (5.2.2) can be expressed as

$$H(x_1, x_2; \mathbf{\Theta}) \;=\; \sum_{k=1}^{2} p_k C_k\left(F_1(x_1; \boldsymbol{b}_1^k), F_2(x_2; \boldsymbol{b}_2^k); \mathbf{\Phi}_k\right) \tag{5.2.4}$$

The number of components in the copula mixture can be different from the number of components in the marginal distributions. Such is the case with the GRB data set discussed in Horváth et al. (2010) and the pulsar data set in Lee et al. (2012). To accommodate this scenario we drop the superscript k on the $\boldsymbol{b}_j s$ in equation (5.2.4) yielding

$$H(x_1, x_2; \mathbf{\Theta}) \;=\; \sum_{k=1}^{2} p_k C_k\left(F_1(x_1; \boldsymbol{b}_1), F_2(x_2; \boldsymbol{b}_2); \mathbf{\Phi}_k\right) \tag{5.2.5}$$

To be able to compute the likelihood, we need the joint density

$$\begin{aligned} h(x_1, x_2; \mathbf{\Theta}) &= \frac{\partial^2}{\partial x_1 \partial x_2} H(x_1, x_2; \mathbf{\Theta}) \qquad (5.2.6) \\ &= \sum_{k=1}^{2} p_k \left\{\prod_{j=1}^{2} \frac{d}{dx_j} F_j(x_j; \boldsymbol{b}_j)\right\} \frac{\partial^2}{\partial F_1 \partial F_2} C_k\left(F_1(x_1; \boldsymbol{b}_1), F_2(x_2; \boldsymbol{b}_2); \mathbf{\Phi}_k\right) \\ &= \sum_{k=1}^{2} p_k \left\{\prod_{j=1}^{2} f_j(x_j; \boldsymbol{b}_j)\right\} \frac{\partial^2}{\partial F_1 \partial F_2} C_k\left(F_1(x_1; \boldsymbol{b}_1), F_2(x_2; \boldsymbol{b}_2); \mathbf{\Phi}_k\right). \end{aligned}$$

The term $\frac{\partial^2}{\partial F_1 \partial F_2} C_k \left(F_1(x_1; \boldsymbol{b}_1), F_2(x_2; \boldsymbol{b}_2); \boldsymbol{\Phi}_k\right)$ in (5.2.6) represents the k^{th} copula density which we denote here by c_k. We thus rewrite (5.2.6) as

$$h(x_1, x_2; \boldsymbol{\Theta}) = \sum_{k=1}^{2} p_k \left\{ \prod_{j=1}^{2} f_j(x_j; \boldsymbol{b}_j) \right\} c_k \left(F_1(x_1; \boldsymbol{b}_1), F_2(x_2; \boldsymbol{b}_2); \boldsymbol{\Phi}_k\right). \quad (5.2.7)$$

5.3 Estimation of parameters

Given a sample of observations $(x_{1i}, x_{2i}), i = 1, 2, \ldots, n$, we seek to identify the marginal distributions $F_1(x_1; \boldsymbol{b}_1)$ and $F_2(x_2; \boldsymbol{b}_2)$ and to find the values of the parameters $p_k, \boldsymbol{\Phi}_k$ and $\boldsymbol{b}_j$ $k = 1, 2,$ $j = 1, 2$ which maximize the likelihood

$$\begin{aligned} L(p_k, \boldsymbol{\Theta}_k) = & \prod_{i=1}^{n} \sum_{k=1}^{2} p_k \left\{ \prod_{j=1}^{2} f_j(x_{ji}; \boldsymbol{b}_j) \right\} \\ & \times c_k \left(F_1(x_{1i}; \boldsymbol{b}_1), F_2(x_{2i}; \boldsymbol{b}_2); \boldsymbol{\Phi}_k\right). \end{aligned} \quad (5.3.1)$$

The task of parameter estimation can be simplified by adopting a semi-parametric approach where the marginal distributions and the corresponding densities are estimated using non-parametric methods and a parametric estimation method such as maximum likelihood estimation is used to estimate the mixing proportions and copula parameters. In practice, the marginal distributions are replaced by

$$\begin{aligned} F_{n1}(x_{1i}) = \hat{U}_i &= \frac{1}{n+1} \sum_{j=1}^{n} I(x_{1j} \leq x_{1i}) = \frac{R_i}{n+1} \\ F_{n2}(x_{2i}) = \hat{V}_i &= \frac{1}{n+1} \sum_{j=1}^{n} I(x_{2j} \leq x_{2i}) = \frac{S_i}{n+1}. \end{aligned} \quad (5.3.2)$$

Here, R_i stands for the rank of x_{1i} among $x_{11}, x_{12}, \ldots, x_{1n}$ and S_i stands for the rank of x_{2i} among $x_{21}, x_{22}, \ldots, x_{2n}$.

The non-standard normalization constant $\frac{1}{n+1}$ is preferred instead of the classical $\frac{1}{n}$ because F_n and G_n later serve as arguments in pseudo-likelihoods such as (5.3.4) below that can take infinite values when given 1 as one of the arguments.

With the semi-parametric approach described above, maximizing the likelihood expression (5.3.1) becomes equivalent to maximizing

$$L'(p_k, \boldsymbol{\Theta}_k) \quad = \quad \prod_{i=1}^{n} \left[\sum_{k=1}^{2} p_k c_k \left(\frac{R_i}{n+1}, \frac{S_i}{n+1}; \boldsymbol{\Phi}_k \right) \right]. \tag{5.3.3}$$

In this thesis, the marginal distributions are estimated using a non-parametric method. The term $\prod_{j=1}^{2} f_j(x_{ji}; \boldsymbol{b}_j)$ which appears in equation (5.3.1) has therefore been omitted from equation (5.3.3) because the parameters of the marginal density functions are not going to be estimated.

It is usually simpler to maximize the log-likelihood,

$$l(p_k, \boldsymbol{\Theta}_k) \quad = \quad \sum_{i=1}^{n} \log \left[\sum_{k=1}^{2} p_k c_k \left(\frac{R_i}{n+1}, \frac{S_i}{n+1}; \boldsymbol{\Phi}_k \right) \right]. \tag{5.3.4}$$

In practice, the estimation of the parameters p_k and $\boldsymbol{\Phi}_k$ of equation (5.3.4) is performed using the maximum likelihood algorithms.

5.4 Applications of mixtures of copula models

We apply the mixture of copula models to the natural logarithms of the hardness ratios and durations (T_{90}) of gamma ray bursts as discussed in the paper by Horváth et al. (2010) and to the logarithms of pulsar periods and period derivatives discussed in Lee et al. (2012).

5.4.1 The data sets.

We consider two partitions of the pulsar data set as given in figure (5.1) below. The first partition (left panel), which was made using the k-means clustering algorithm, divides the data set of 1759 pairs of observations into a smaller group of 164 pairs of observations and a larger group of size 1595. The second partition (right panel), which was created to investigate the sensitivity of the results to a small change in the partitions, has 150 observations in the smaller data set and 1609 observations in the larger data set. The results reported in the tables and figures correspond to the partition that was created using the k-means clustering algorithm. The results for the other partition are given in appendix B.

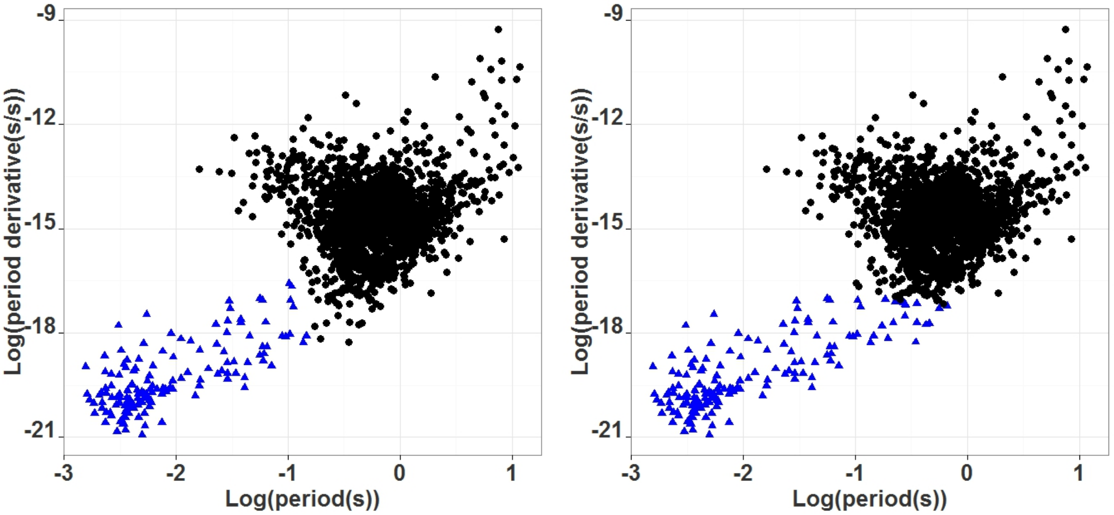


Figure 5.1: Scatter plots of log(pulsar period) versus log(period derivative)

5.4.2 The copula models considered

The copula models considered in this study include the 22 Archimedean copulas described in section 4.2 of the book by Nelsen (2006) together with one-parameter elliptical copulas implemented in the r-package fCopulae. These are the normal, Cauchy, logistic and Laplace copulas.

Apart from the bivariate copula families listed above, we do also consider rotations of these copulas by 90, 180 and 270 degrees. Given $0 \leq u, v \leq 1$, and the copula densities $c(u, v)$, the densities of rotated copulas are given by the following equations (see Mai and Scherer, 2012, page 207):-

$$\begin{aligned} c_{90}(u,v) &= c(1-u,v) \qquad (5.4.1) \\ c_{180}(u,v) &= c(1-u,1-v) \\ c_{270}(u,v) &= c(u,1-v) \end{aligned}$$

The corresponding cumulative distribution functions can be obtained by integration of the densities as:-

$$\begin{aligned} C_{90}(u,v) &= v - C(1-u,v) \qquad (5.4.2) \\ C_{180}(u,v) &= C(1-u,1-v) + u + v - 1 \\ C_{270}(u,v) &= u - C(u,1-v) \end{aligned}$$

The 26 canonical copula forms together with 3 rotations of each gives a total of 104 candidate copulas. R programs were developed to successively fit all the 5356 possible pairings of the candidate copulas to the data sets. These mixtures were fitted to each partition of the pulsar data set and also to the gamma-ray burst data. For each data set, the five copula pairs with the highest log-likelihood values were retained for goodness-of-fit testing. In this particular case it was not necessary to consider an information criterion since all the copulas considered had only one parameter to be estimated, i.e. the number of parameters was the same for all models and comparing likelihoods was therefore equivalent to comparing information criteria.

Evaluation of copula density and cumulative distribution functions and random number generation for these copula families was done with the help of the R package fCopulae, (version 3011.81; Rmetrics Core Team, 2013).

5.4.3 Parameter estimation

Estimation of copula parameters together with the mixing proportions was done with the aid of the R package maxLik, version 1.2-4 (Toomet and Henningsen, 2011). In the sequel, we give results obtained after fitting mixtures of copulas, as described above, to the two pulsar data sets and also to the Horváth et al. (2010) GRB data.

5.4.4 Results for larger pulsar data set

Table (5.1) gives the results obtained after fitting pairs of copulas to the larger pulsar data set. We report on the families of the constituent copulas, their rotations, the corresponding parameter estimates and standard errors and also an estimate of the proportion of the first copula and its standard error. The log-likelihood values and $p-$values of the Cramér$-$von Mises statistic for copula goodness-of-fit testing are given in the last two columns. A brief discussion of copula goodness-of-fit testing is given in the sequel.

The results given in table (5.1) are for the five copulas with the highest log-likelihood values. The top 5 likelihood values ranged from 72.19 to 89.20. All the parameters reported here differ significantly from zero; the largest p-value was 1.1×10^{-3}. It is striking to note that the five copula pairs with the highest log-likelihood values are each a combination of copula number 16 in Nelsen (2006) section 4.2 with another copula.

As a first step towards testing for goodness-of-fit, a scatter plot of the original data, transformed to copula scale, is juxtaposed with plots of equal samples sizes drawn from the five copula pairs with the highest log-likelihood values. Other researchers (see for example, Genest and Favre, 2007) superimpose the plot derived from the original data on a plot of a very large sample (say 10000 values) drawn from the fitted copula. The sample from the fitted copula is made large in order to cater for sampling variability. We have not adopted this procedure here for the reason that the scatter plot is hereby only serving the purpose of a qualitative procedure which will be followed by a formal goodness-of-fit test. Figure (5.2) shows the resultant plots. None of the five plots suggests any deviation from the pattern depicted in the first plot, which was derived from the actual data.

Copula 1						Copula 2				log-	$p-$value of
Family[1]	rotation	$\hat{\Phi}_1$	S.E($\hat{\Phi}_1$)	$\hat{p}_1$	S.E($\hat{p}_1$)	Family	rotation	$\hat{\Phi}_2$	S.E($\hat{\Phi}_2$)	likelihood	Cram$\acute{e}$r$-$von Mises statistic
16	270	0.150	0.026	0.609	0.043	16	180	0.151	0.046	89.20	0.740
16	270	0.316	0.056	0.523	0.053	20	180	0.200	0.020	87.00	0.788
16	270	0.290	0.052	0.523	0.057	6	0	1.331	0.061	84.58	0.810
16	270	0.338	0.064	0.539	0.054	1	180	0.513	0.076	84.26	0.785
16	270	0.323	0.068	0.540	0.057	4	0	1.238	0.047	78.27	0.767

Table 5.1: Parameter estimates, standard errors and angles of rotation for the five copula pairs giving the highest log-likelihood values when fitted to the larger pulsar data set. Log-likelihood values and $p-$values of the Cram$\acute{e}$r$-$von statistic are given in the last two columns.

[1] Family refers to the copula family number in Nelsen, 2006, section 4.2

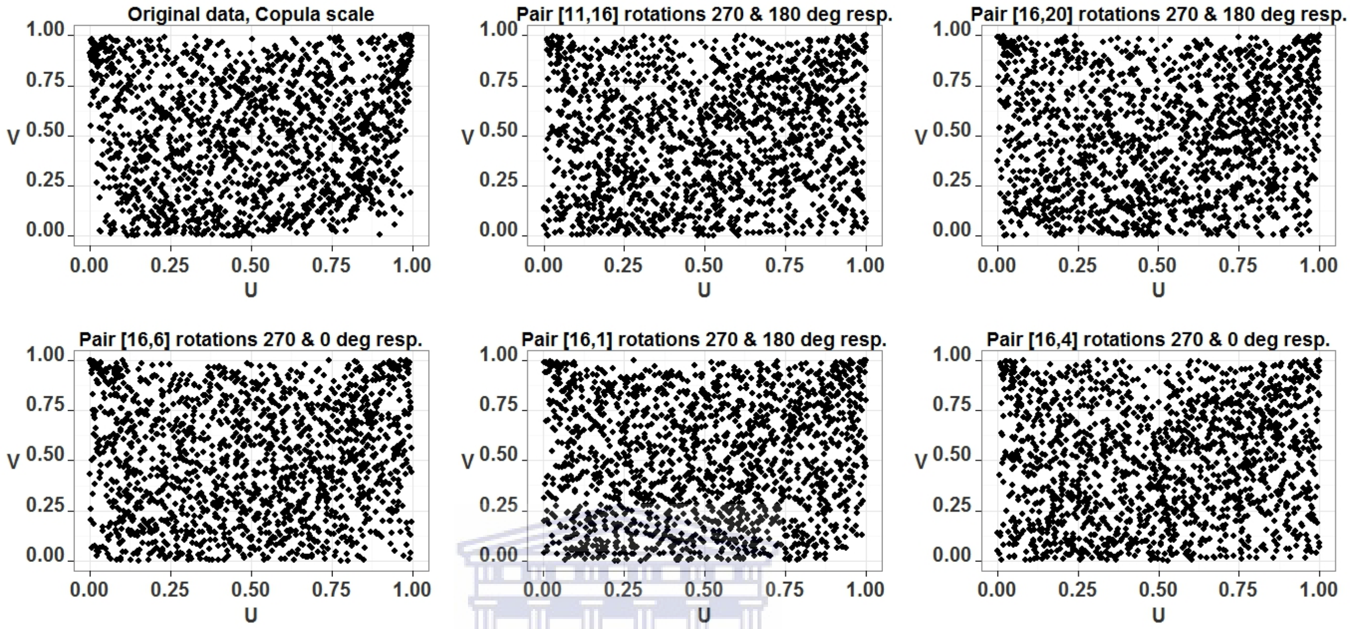


Figure 5.2: Scatter plot of the original data (copula scale) together with the corresponding plots of simulated data derived from the best five copula mixture models.

The scatter plots in figure (5.2) provide a graphical check of the goodness-of-fit of the dependency structure only, i.e. the copula function taken in isolation. Figure (5.3) below was constructed in an effort to qualitatively assess the goodness-of-fit of the complete bivariate model, i.e. the copula together with the marginal distributions. The margins of the random pairs (U_i, V_i) from each of the estimated pair-copula models were transformed back into the units of the original data using the inverse of the empirical marginal distributions $F_{n1}(x)$ and $F_{n2}(x)$ of equation (5.3.2). This task is hampered by the fact that the inverse empirical cumulative distributions are not in the form of closed formulae. Use was therefore made of the R-package logspline (version 2.1.5) to estimate the cumulative distribution functions using spline functions. Figure (5.3) shows that all the five models capture the pattern of the original data fairly well.

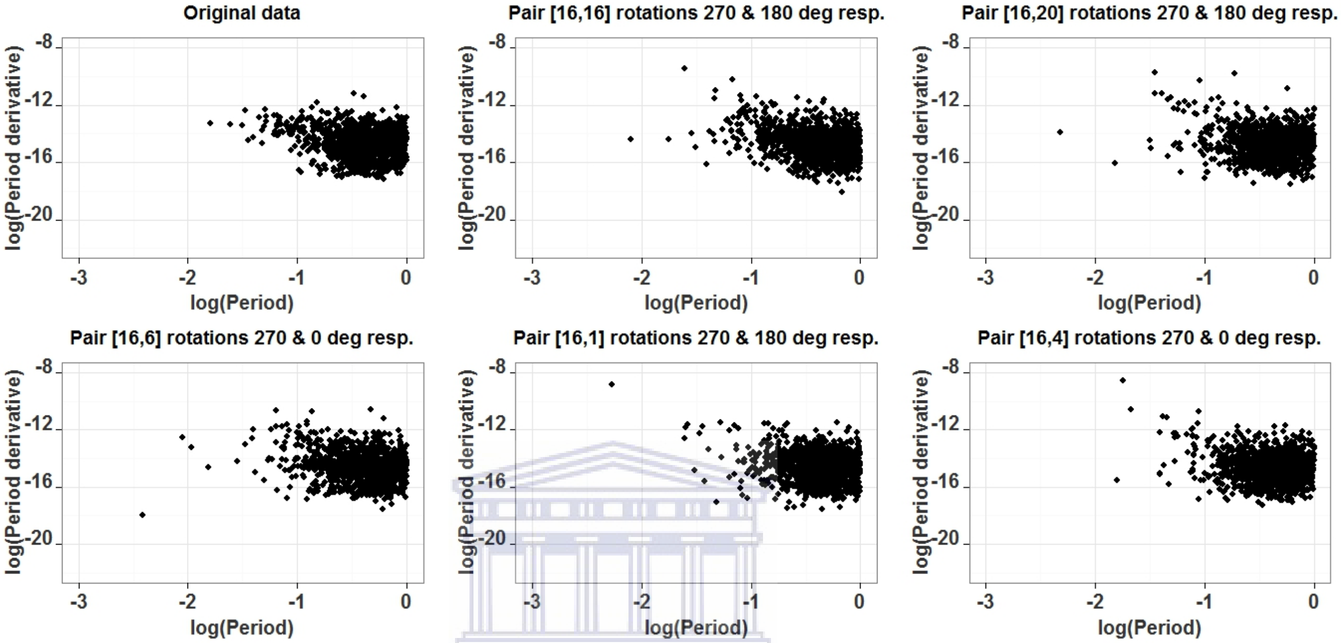


Figure 5.3: Same as in fig (5.2) upon transforming back to the scale of the original data using the inverse empirical distributions of period and period derivative values.

The next step was to employ the Cramér−von Mises test to determine if the chosen mixture models fitted the data. Besides the Cramér−von Mises test, there are many other goodness-of-fit tests for copula models. Details can be found in Genest et al. (2009) and Berg (2009). Results of simulation studies carried out in these two studies reveal that the Cramér−von Mises test is the most powerful among the blanket goodness-of-fit tests for copula models (although see below).

The Cramér−von Mises statistic S_n compares the empirical copula $C_n(u, v)$ as defined in the previous chapter to its parametric estimate under the null hypothesis C_{θ_n} :

$$\begin{aligned} S_n &= \int_{[0,1]^2} n\left\{C_n(u,v) - C_{\theta_n}(u,v)\right\}^2 dC_n(u,v) \\ &\approx \sum_{i=1}^{n} \left\{C_n\left(\hat{U}_i, \hat{V}_i\right) - C_{\theta_n}\left(\hat{U}_i, \hat{V}_i\right)\right\}^2 \end{aligned} \tag{5.4.3}$$

Here $\hat{U}_i$ and $\hat{V}_i$ are as defined in equations (5.3.2). Large values of the statistic indicate lack of fit. The Cram$\acute{e}$r$-$von Mises statistic is not distribution free; the limiting distribution depends on the underlying copula type and also on the unknown value(s) of the parameter(s). P-values are therefore approximated by simulation under the null hypothesis. Details of the simulation procedure can be found in Genest et al. (2009). Alternatively the multiplier techniques described in the previous chapter can be used to approximate the p-values. More details about the use of multiplier techniques for goodness-of-fit testing for copula models can be found in Kojadinovic et al. (2011) and Kojadinovic and Yan (2011). The multiplier techniques are computationally less demanding than the parametric bootstrap because their application does not require repeated simulation and estimation of parameters for the hypothesized copula. They were however not applied in the current work.

Genest and Nešlehová, (2013) proposed an Anderson-Darling-type statistic for copula goodness-of-fit testing, with $p-$values determined using the multiplier techniques. A simulation study performed by these authors showed that this test is more powerful than the Cram$\acute{e}$r$-$von Mises test. The present study however used the Cram$\acute{e}$r$-$von Mises test because it was considered the best at the time the study was conducted.

The $p-$values for the Cram$\acute{e}$r$-$von Mises statistics are given in table 5.1. The values of the test statistic ranged from 0.059 to 0.074. The p-values all exceed 0.70

indicating that all the models fit very well to the data.

5.4.5 Results for Smaller pulsar data set

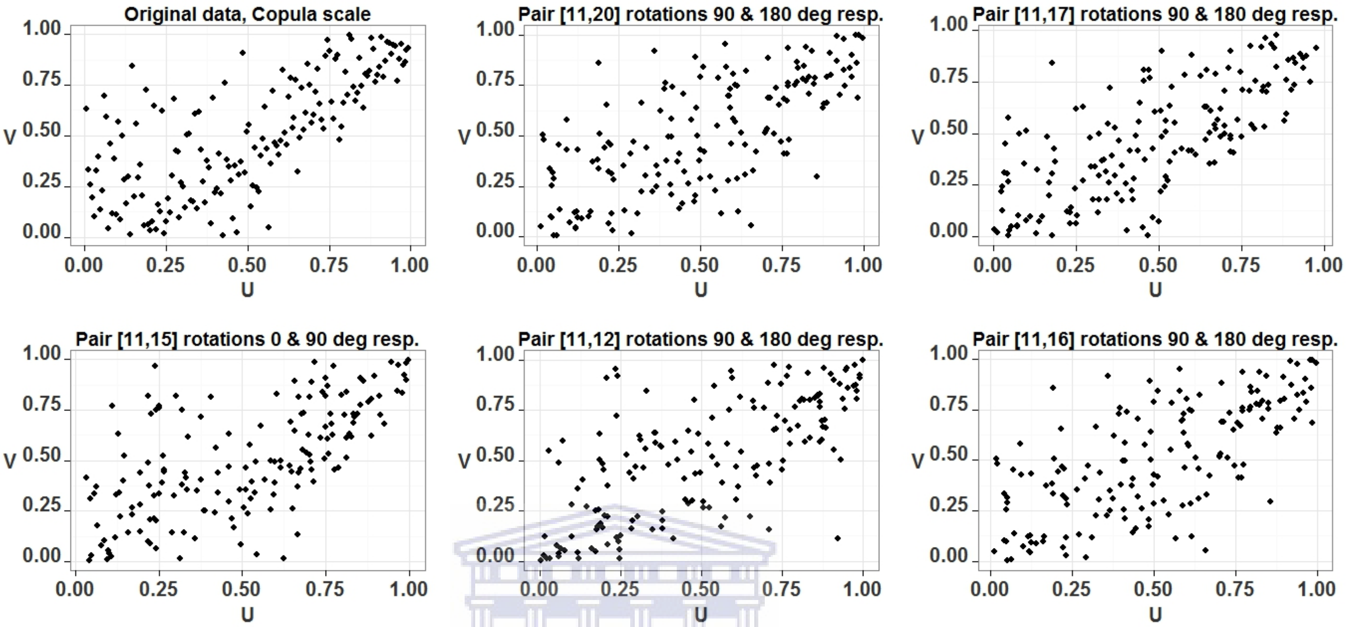
Table (5.2) reports the five copula pairs that gave the highest values of the log-likelihood when fitted to the smaller pulsar data set. The results are dominated by combinations of copula number 11 rotated 90 degrees, together with other copulas. The log-likelihood values range from 45 to 50.

Plots serving as qualitative goodness-of-fit tests are given in figures (5.4) and (5.5) below. Again the plots show that the models are capable of reproducing the pattern of the original data which is given in the first subplot of each figure. The last column of table (5.2) also supports this claim, with goodness-of-fit $p-$values ranging from 0.71 to 0.95 indicating that all the models fit the data well.

Copula 1						Copula 2				log-	$p-$value of
Family[2]	rotation	$\hat{\Phi}_1$	S.E($\hat{\Phi}_1$)	$\hat{p}_1$	S.E($\hat{p}_1$)	Family	rotation	$\hat{\Phi}_2$	S.E($\hat{\Phi}_2$)	likelihood	Cramér$-$von Mises statistic
11	90	0.482	0.001	0.624	0.092	20	180	0.655	0.077	67.544	0.94
11	90	0.482	0.001	0.457	0.138	17	180	9.260	3.071	66.58	0.92
11	90	0.448	0.002	0.492	0.126	15	0	2.285	0.091	65.46	0.82
11	90	0.482	0.001	0.627	0.122	12	180	1.100	0.179	64.06	0.80
11	90	0.482	0.001	0.676	0.094	16	180	9.115	0.355	64.04	0.71

Table 5.2: Parameter estimates, standard errors and angles of rotation for the five copula pairs giving the highest log-likelihood values when fitted to the smaller pulsar data set. Log-likelihood values and $p-$values of the Cramér$-$von statistic are given in the last two columns.

[2] Family refers to the copula family number in Nelsen, 2006, section 4.2


Original data, Copula scale

Pair [11,20] rotations 90 & 180 deg resp.

Pair [11,17] rotations 90 & 180 deg resp.

Pair [11,15] rotations 0 & 90 deg resp.

Pair [11,12] rotations 90 & 180 deg resp.

Pair [11,16] rotations 90 & 180 deg resp.

Figure 5.4: Scatter plot of the original data (copula scale) together with the corresponding plots of simulated data derived from the best five copula mixture models.

Mixtures of copula models were also fitted to the data sets depicted in the alternative partition in the right hand panel of figure (5.1) above. The results, which are given in appendix B, do not differ much from those of the first partition, at least in terms of the two copula families selected. For the larger data set, the best five models for the first partition are among the top eight models for the second partition. For the smaller data set, three of the best models are common to both partitions.

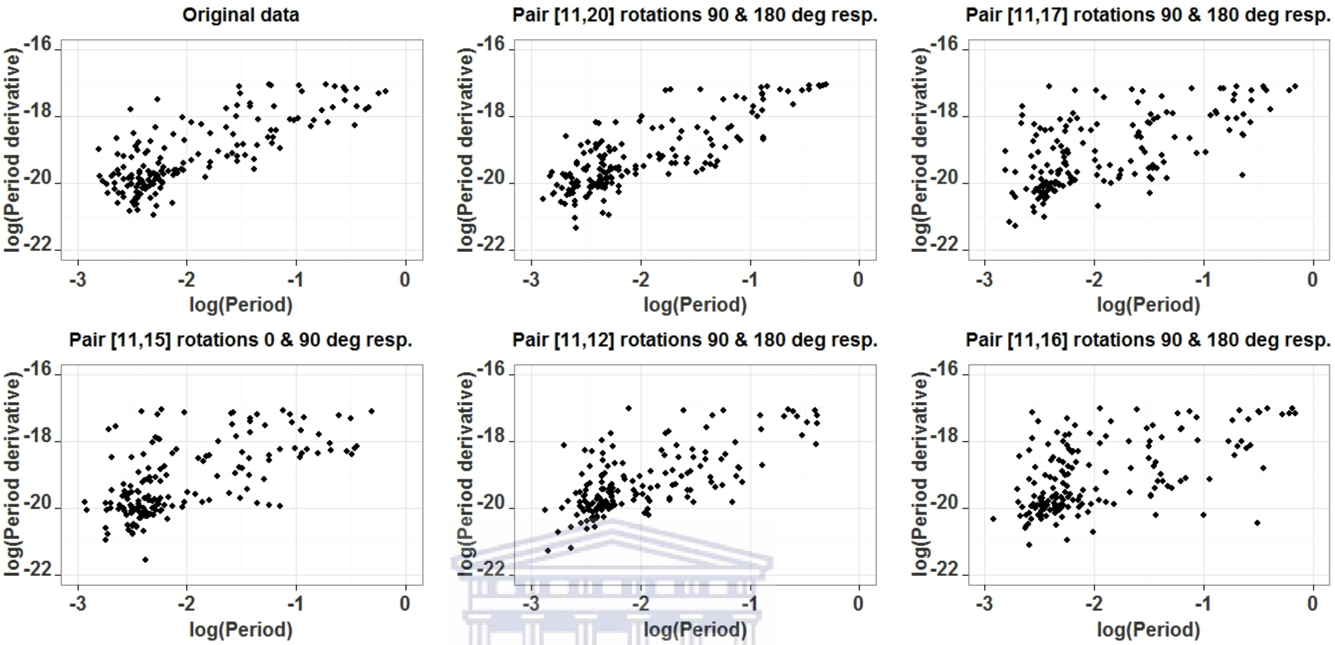


Figure 5.5: Same as in fig (5.4) upon transforming back to the scale of the original data using the inverse empirical distributions of period and period derivative values.

5.4.6 Results for the Horváth et al. (2010) GRB data set

Finally, in table (5.3) we list the five copula pairs giving the highest likelihood values when fitted to the GRB data. The log-likelihood values range between 14 and 20. Again, figures (5.6) and (5.7) reveal a fairly good fit and this is confirmed in the last column in table (5.3) where the p-values for the Cramér−von Mises statistic range from 0.77 to 0.94.

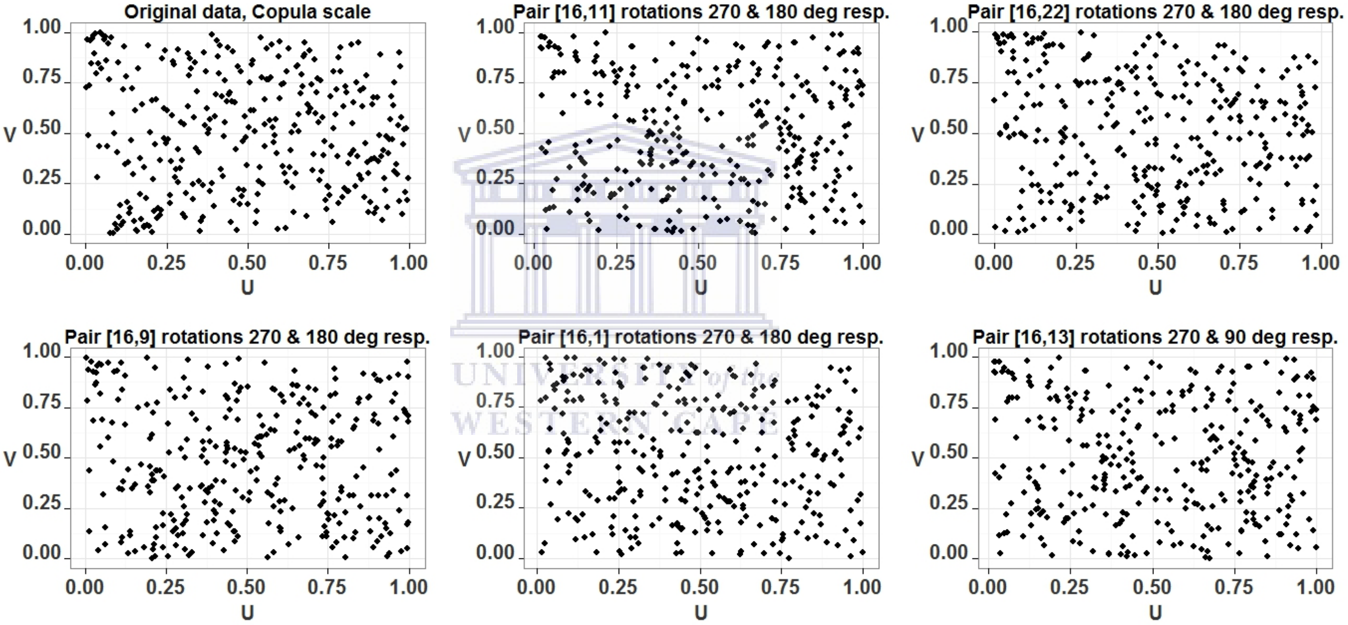


Figure 5.6: Scatter plot of the original data (copula scale) together with the corresponding plots of simulated data derived from the best five copula mixture models.

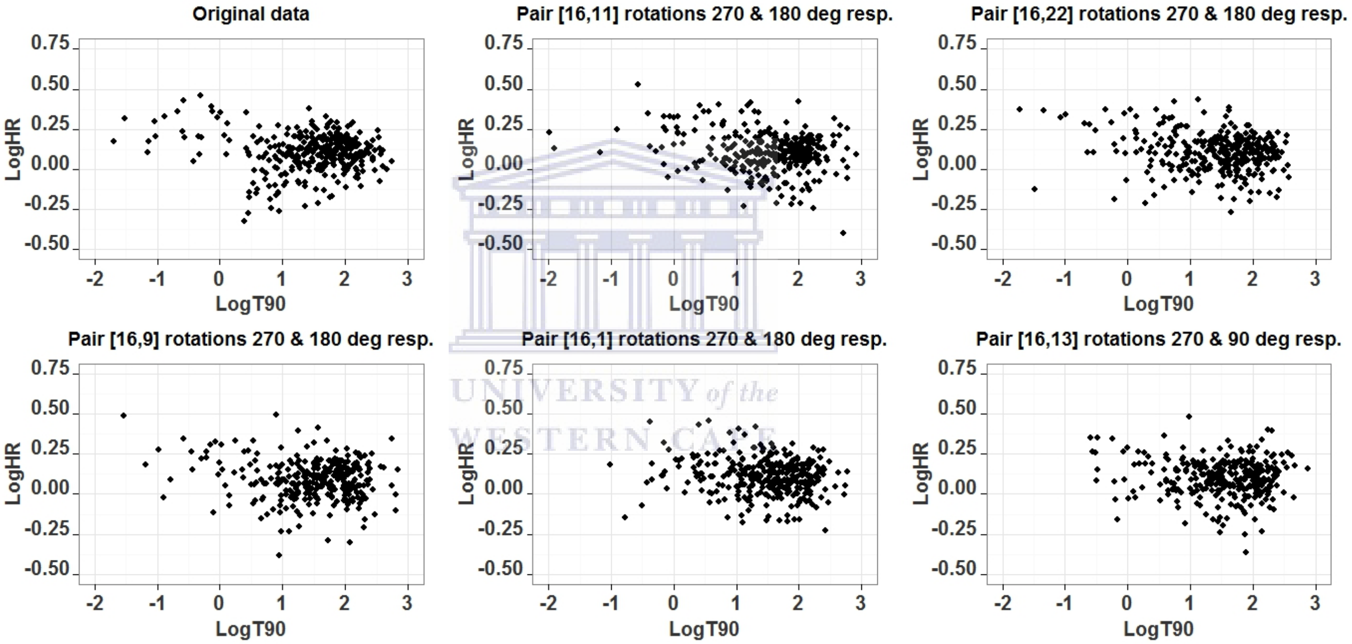


Figure 5.7: Same as in fig (5.6) upon transforming back to the scale of the original data using the inverse empirical distributions of duration and hardness ratio values.

Copula 1						Copula 2				log-	$p-$value of
Family[3]	rotation	$\hat{\Phi}_1$	S.E($\hat{\Phi}_1$)	$\hat{p}_1$	S.E($\hat{p}_1$)	Family	rotation	$\hat{\Phi}_2$	S.E($\hat{\Phi}_2$)	likelihood	Cramér$-$von Mises statistic
16	270	0.093	0.041	0.507	0.124	11	180	0.226	0.037	19.24	0.91
16	270	0.092	0.041	0.514	0.125	22	180	0.533	0.103	19.00	0.93
16	270	0.099	0.045	0.533	0.129	9	180	0.566	0.213	17.81	0.90
16	270	0.103	0.045	0.507	0.118	1	180	-0.338	0.017	19.64	0.94
16	270	0.156	0.060	0.700	0.134	13	90	1.780	0.877	14.31	0.77

Table 5.3: Parameter estimates, standard errors and angles of rotation for the five copula pairs giving the highest log-likelihood values when fitted to the GRB data set. Log-likelihood values and $p-$values of the Cramér$-$von statistic are given in the last two columns.

[3] Family refers to the copula family number in Nelsen, 2006, section 4.2

5.5 Summary

In the present chapter we have introduced the idea of modelling data using mixtures of copulas including rotated copulas. The models are applied to the Period-Period derivative data discussed in Lee et al. (2012) and also to the data discussed in Horváth et al. (2010) using mixtures of the single parameter Archimedean and Elliptical copulas offered in the statistical package f-Copula (Rmetrics Core Team, 2013).

A semi-parametric estimation procedure is employed where the empirical cumulative distribution functions are used to transform the marginal data to copula scale, then the method of maximum likelihood is used to estimate the mixing proportions and copula parameters.

Five models giving the highest likelihood values are selected for each data set and subjected to goodness-of-fit testing. P-values of the Cramér−von Mises statistic based on simulating under the null hypothesis confirm that all the selected models fit the data adequately. This is also confirmed informally using graphical techniques.

If we were to deduce the number of physical classes of gamma-ray bursts and pulsars based on our copula results, we would arrive at two classes for gamma-ray bursts and four classes for pulsars. This is clearly in contradiction to the findings of Horváth et al. (2010) and Lee et al. (2012) and in our view, this contradiction indicates that the number of components in a statistical model that fits data may not necessarily be the same as the number of distinct classes of the physical object under consideration. Our work also indirectly offers a way of generating asymmetric copulas.

One limitation of the models we proposed is that they are not easily applicable to data whose dimension exceeds two.

Chapter 6

Conclusion

This thesis sought to improve on the statistical analyses presented in a number of previous studies of astrophysical phenomena. The main contribution is the modelling of the pulsar and gamma-ray burst data sets using mixtures of rotated bivariate copulas.

The studies of Borgonovo (2004), Borgonovo et al. (2007) and Vasquez and Kawai (2011) which focussed on the distribution of the widths of autocorrelation functions of gamma-ray bursts each used very small samples of about 20 gamma ray bursts. No formal statistical tests were employed in any of these papers.

Our work involved a large sample of 119 gamma-ray bursts the data of which were all collected from the same instrument. This helps to reduce non-uniformity. We suggested an alternative way of normalizing the gamma-ray burst autocorrelation function. Where the normalization does not perform well, we have suggested the extrapolation of the autocorrelation function $A(l)$ from larger lags to $l = 0$ in order to determine $A(0)$. We have also suggested an alternative, more robust way of measuring the autocorrelation function width.

Several statistical techniques were employed in an effort to verify/disprove the claimed bimodality in the distribution of the widths of autocorrelation functions of gamma-ray bursts. These include kernel density estimates, the dip test of bimodality and univariate Gaussian mixtures models. The number of components in the Gaussian mixture models was arrived at using a likelihood ratio test whose $p-$values were obtained by simulating under the null hypothesis. The Anderson-Darling test and the D'Agostino-Pearson test were used for goodness-of-fit testing. Mixtures of regression models were employed to investigate the possibility that the gamma-ray burst autocorrelation function widths could reveal bimodality according to their peak fluxes.

Contrary to findings in other studies, our analysis does not reveal any evidence of bimodality in the distribution of autocorrelation function widths, although there is evidence for slight asymmetry in the distribution.

In chapter 3, we used simulated percentage points of the likelihood ratio statistic to confirm that a bivariate Gaussian mixture model with three components is the preferred model for the joint distribution of gamma-ray burst durations and hardness ratios considered in Horváth (2010). We also confirmed that a bivariate mixture model with six components is the best model for the period-period derivative data considered in Lee et al. (2012). The bivariate Kolmogorov-Smirnov test was used to test the fit of these models.

We extended the analysis of these two data sets by investigating the number of components in the marginal distributions, using likelihood ratio and Anderson-Darling tests. The results show that the models above do not fit very well in the margins; the distribution of T_{90} values alone can be described by a three-component model while a two-component mixture model is preferred for hardness ratios.

With regards to the data in Lee et al (2012), our results are that the distribution of the univariate period data can be adequately described by four Gaussian components while that of the period derivative data can be modelled by five components.

In chapter 4 we extended the study of the power and significance levels of tests of copula symmetry/exchangeability which was initially conducted by Genest at al. (2012) by using two copula models that had not been considered in the earlier study. We used Khoudraji's device to obtain asymmetric versions of these two copulas which were used in the power study. The study began with an investigation of an effective bandwidth parameter l_n for estimating the partial copula derivatives. Our results show that for sample size 250, the statistics R_n and S_n have their significance levels closest to the nominal 5% when $l_n = 3/\sqrt{n}$; while for $n = 100$ they work best with $l_n = 1/\sqrt{n}$. The rest of the tests give the best results with $l_n = 1/\sqrt{n}$. For $n = 100$, the majority of the tests have significance levels below the expected 5% nominal level. There is a slight improvement in the significance levels when the sample size increases to 250. This agrees with previously obtained results.

With regards to power, the two Cramér−von Mises statistics R_n and S_n are the most powerful followed, by T_n. The Jasson type statistics are the least powerful. The power of all the tests increases with sample size and also with the correlation τ and the parameter η used in Khoudraji's technique up to a peak value at $\eta = 0.5$. All this is in agreement with previous findings. Contrary to Genest et al. (2012) who concluded that the power of the Jasson tests increases with finer partitions of the $[0, 1]^2$ grid, our results do not show any clear hierarchy in terms of the power of the Jasson tests.

Our work also brought out the difficulties associated with the Jasson statistics i.e.

with finer partitions of $[0,1]^2$ the variance-covariance matrix derived from bootstrap vectors is not always non-singular, leading to undefined values of the test statistic.

The main contribution is in chapter 5 where we introduced the idea of modelling asymmetric bivariate dependence structures using mixtures of two rotated copulas. These models are applied to the pulsar period-period derivative data discussed in Lee et al. (2012) and also to the data discussed in Horváth et al. (2010) using mixtures of single parameter rotated Archimedean copulas. Estimation of parameters is performed using the method of maximum likelihood. The models giving the highest log-likelihood values were selected for each data set and subjected to goodness-of-fit testing. Qualitative (graphical) goodness-of-fit tests and the Cram$\acute{e}$r$-$von Mises test confirmed the fit of the chosen models. Through this work we indirectly offer a way of generating asymmetric copulas. Also, a comparison of our results with earlier results casts doubt on the previously held view that the number of mixtures components in a statistical model reflects the number of physical classes of the objects from which the data are derived.

We note in passing that a satisfactory parametric model of the GRB data consists of marginal Gaussian mixture models with respectively three and two components while the dependence structure is well described by a mixture of two one-parameter rotated copulas. Similar considerations apply to the pulsar data analysed by Lee et al. (2012).

Appendices

Appendix A

MATLAB and R functions coded

Function	Purpose	Requires
pvalr.m	Likelihood ratio test to test if it is worthwhile to increase number of univariate mixture components by 1	
andr.m	Anderson-Darling test for fitting a mixture of Gaussians to univariate data	
pulspval.m	Fits a mixture of up to 7 components to the Lee data and gives the p-values for the change in log-likelihood	
pval.m	Anderson-Darling test to determine if the margins of a distribution fit univariate data	
ks2dg.m	Bivariate K-S test	
jaspt.m	Jasson tests	jasqfc.m
jasptn.m	R_n and S_n statistics	Tnqev.m
Tnqev.m	R_n and S_n statistics	Aikij.m Ind.m Pin.m
Tnqlb.R	T_n statistic	

Function	Purpose	Requires
jasptpow.m	Power of Jasson statistics	jasqfc.m
Swift logav.m	Computes the ACF width for each GRB	
mixcops.r	fits mixtures of rotated copulas to bivariate data and selects pairs with highest loglikelihood	
gof.R	Goodness-of-fit testing for mixtures of rotated copulas	
Hovgraphs.R	Goodness-of-fit plots for copula mixture models	Atnfgraphs2.R

Appendix B

Results for a second partition of the pulsar data set

Copula 1						Copula 2				log-	$p-$value of
Family[1]	rotation	$\hat{\Phi}_1$	S.E($\hat{\Phi}_1$)	$\hat{p}_1$	S.E($\hat{p}_1$)	Family	rotation	$\hat{\Phi}_2$	S.E($\hat{\Phi}_2$)	likelihood	Cramér$-$von Mises statistic
16	270	0.1447	0.022	0.65	0.041	16	180	0.1421	0.041	92.92	0.56
16	270	0.2810	0.050	0.5361	0.054	20	180	0.2015	0.020	90.60	0.74
16	270	0.3014	0.057	0.5524	0.055	1	180	0.5193	0.078	87.89	0.79
4	0	1.2399	0.047	0.4507	0.057	16	270	0.2897	0.052	82.33	0.63
1	270	0.7450	0.151	0.2911	0.047	1	90	-0.3112	0.008	80.28	0.67
11	90	0.2201	0.016	0.6400	0.055	1	270	0.6890	0.130	75.92	0.57
22	90	0.5283	0.034	0.6249	0.053	1	270	0.6793	0.122	74.16	0.61
11	90	0.1992	0.014	0.7528	0.049	16	180	0.033	0.011	72.97	0.33

Table B.1: Parameter estimates, standard errors and angles of rotation for the eight copula pairs giving the highest log-likelihood values when fitted to the larger pulsar data set. Log-likelihood values and $p-$values of the Cramér$-$von statistic are given in the last two columns

[1] Family refers to the copula family number in Nelsen, 2006, section 4.2

Copula 1						Copula 2				log-	$p-$value of
Family[2]	rotation	$\hat{\Phi}_1$	S.E($\hat{\Phi}_1$)	$\hat{p}_1$	S.E($\hat{p}_1$)	Family	rotation	$\hat{\Phi}_2$	S.E($\hat{\Phi}_2$)	likelihood	Cramér$-$von Mises statistic
11	90	0.4250	0.005	0.4398	0.141	15	0	2.2271	0.030	49.99	0.92
11	90	0.4250	0.005	0.4712	0.141	1	180	1.5953	0.340	47.46	0.95
11	90	0.3982	0.008	0.3917	0.176	13	180	4.3275	0.644	46.71	0.91
13	180	3.9216	0.568	0.6889	00.220	16	270	0.0222	0.011	45.31	0.96
11	90	0.4262	0.005	0.3851	0.163	17	180	7.7142	1.559	45.63	0.90
11	90	0.3982	0.006	0.5172	0.192	12	180	1.10023	0.158	45.03	0.75

Table B.2: Parameter estimates, standard errors and angles of rotation for the six copula pairs giving the highest log-likelihood values when fitted to the smaller pulsar data set. Log-likelihood values and $p-$values of the Cramér$-$von statistic are given in the last two columns

[2] Family refers to the copula family number in Nelsen, 2006, section 4.2

Bibliography

[1] T. W. Anderson and D. A. Darling, *A test of goodness of fit*, Journal of the American Statistical Association **49(268)** (1954), 765–769.

[2] D. W. K. Andrews, *Testing when a parameter is on the boundary of the maintained hypothesis*, Econometrica **69** (2001), 683–734.

[3] V. Arakelian and D. Karlis, *Clustering dependencies via mixtures of copulas*, Communications in Statistics-Simulation and Computation **43** (2014), 1644–1661.

[4] T. Baier and E. Neuwirth, *Excell::COM::R*, Journal of Computational Statistics **22(1)** (2007), 91–108.

[5] D. Berg, *Copula goodness-of-fit testing: An overview and power comparison*, European Journal of Finance **15** (2009), 675–701.

[6] J. A. Bilmes, *A gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models* , Tech. Report TR-97-021, University of California, Berkeley, Department of Electrical Engineering and Computer Science, April 1998.

[7] L. Borgonovo, *Bimodal distribution of the autocorrelation function in gamma-ray bursts*, Astronomy and Astrophysics **418** (2004), 487–493.

[8] L. Borgonovo, F. Frontera, C. Guidorzi, E. Montanari, L. Vetere, and P. Soffitta, *On the temporal variability classes found in long gamma-ray bursts with known redshift*, Astronomy and Astrophysics **465** (2007), 765–775.

[9] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, *Kernel density estimation via diffusion*, Annals of Statistics **38(5)** (2010), 2916–2957.

[10] C. Chatfield, *The Analysis of Time Series, An Introduction*, 6 ed., Chapman and Hall, London, 2003.

[11] Y. Chen and M. R. Gupta, *EM Demystified: An Expectation-Maximization Tutorial* , Tech. Report UWEETR-2010-0002, University of Washington, Department of Electrical Engineering, February 2010.

[12] R. B. D'Agostino and E. S. Pearson, *Testing for departures from Normality. empirical results for the distribution of b_2 and $\sqrt{b_1}$*, Biometrika **60** (1973), 613–622.

[13] R. B. D'Agostino and M. A. Stephens, *Goodness-of-fit Techniques*, Marcel Dekker, Inc., New York, 1986.

[14] P. J. Danaher and M. S. Smith, *Modeling multivariate distributions using copulas: Applications in marketing*, Marketing Science **30** (2011), 4–21.

[15] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, Series B **39 (1)** (1977), 1–38.

[16] A. Dias and P. Embrechts, *Modelling exchange dependence dynamics at different time horizons*, Journal of International Money and Finance **29** (2010), 1687–1705.

[17] P. Embrechts, F. Lindskog, and A. McNeil, *Modelling Dependence with Copulas and Applications to Risk Management*, Tech. Report CH-8092 Zürich, ETH Zürich, Department of Mathematics, September 2001.

[18] S. Faria and G. Soromenho, *Fitting mixtures of linear regressions*, Journal of Statistical Computation and Simulation **80(2)** (2010), 201–225.

[19] G. Fasano and A. Franceschini, *A multidimensional version of the Kolmogorov-Smirnov test*, Monthly Notices Of the Royal Astronomical Society **225** (1987), 155–170.

[20] N. I. Fisher and P. Switzer, *Chi-plots for assessing dependence*, Biometrika **72(2)** (1985), 253–265.

[21] ______, *Graphical assessment of dependence: Is a picture worth 100 tests?*, The American Statistician **55(3)** (2001), 233–239.

[22] C. Genest, K. Ghoudi, and L. P. Rivest, *Discussion of "Understanding relationships using copulas", by E. W. Frees and E. A. Valdez*, North American Actuarial Journal **2** (1998), 143–149.

[23] C. Genest and J. C. Boies, *Detecting dependence with Kendall plots*, The American Statistician **57(4)** (2003), 275–284.

[24] C. Genest and A. C. Favre, *Everything you wanted to know about copula modeling but were afraid to ask*, Journal of Hydrological Engineering **12(4)** (2007), 347–368.

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

[26] C. Genest, J. G. Nešlehová, and J. Quessy, *Tests of symmetry for bivariate copulas*, Annals of the Institute of Mathematical Statistics **64(4)** (2012), 811–834.

[27] C. Genest, W. Huang, and J. Dufour, *A regularized goodness-of-fit test for copulas*, Journal de la Société Française de Statistique **154(1)** (2013), 64–77.

[28] C. Genest and J. G. Nešlehová, *Copulas and Copula Models*, Encyclopedia of Environmetrics, Wiley Online Library, 2013.

[29] C. Genest and J. G. Nešlehová, *On tests of radial symmetry for bivariate copulas*, Statistical Papers **55** (2014), 1107–1119.

[30] S. Gosh, J. D. Woodard, and D. Vedenov, *Efficient estimation of copula mixture models: An application to the rating of crop revenue insurance*, Paper presented at the Agricultural and Applied Economics Association (AAEA) Annual Meeting, Pittsburgh, Pennsylvania, July 24-26, 2011.

[31] P. Hall and M. York, *On the calibration of Silverman's test for multimodality*, Statistica Sinica **11** (2001), 515–536.

[32] J. A. Hartigan and P. M. Hartigan, *The dip test of unimodality*, Annals of Statistics **13** (1985), 70–84.

[33] A. Henningsen and O. Toomet, *Maxlik: A package for maximum likelihood estimation in R*, Computational Statistics **26(3)** (2011), 443–458.

[34] Y. Hong, J. Tu, and G. Zhou, *Asymmetries in stock returns: Statistical tests and economic evaluation*, The Review of Financial Studies **20(5)** (2007), 1547–1581.

[35] I. Horváth, Z. Bagoly, L.G. Balázs, A. de Ugarte Postigo, P. Veres, and A. Mészáros, *Detailed classification of Swift's gamma ray bursts*, Applied Physics Journal **713** (2010), 552–557.

[36] L. Hu, *Dependence patterns across financial markets: a mixed copula approach*, Applied Financial Economics **16** (2006), 717–729.

[37] S. Jasson, *L'asymétrie de la dépendence, quel impact sur la tarification?*, Tech. report, AXA Group Risk Management, Paris, France, April 2005.

[38] H. Joe, *Multivariate Models and Dependence Concepts*, Chapman and Hall/CRC, Florida, 1997.

[39] D. Kim, J. M. Kim, S. M. Liao, and Y. S. Jung, *Mixture of D-vine copulas for modelling dependence*, Computational Statistics and Data Analysis **64** (2013), 1–19.

[40] C. Koen and A. Bere, *On multiple classes of gamma-ray bursts, as deduced from autocorrelation functions or bivariate duration/hardness ratio distributions*, Monthly Notices of the Royal Astronomical Society **420** (2012), 405–415.

[41] I. Kojadinovic and J. Yan, *A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems*, Statistical Computation **21** (2011), 17–30.

[42] I. Kojadinovic, J. Yan, and M. Holmes, *Fast large-sample goodness-of-fit tests for copulas*, Statistica Sinica **21** (2011), 841–871.

[43] I. Kosmidis and D. Karlis, *Model-based clustering using copulas with applications*, Statistics and Computing, in press, 2015.

[44] R. Larsen, J.W. Mjelde, D. Klinefelter, and J. Wolfley, *The use of copulas in explaining crop yield dependence structures for use in geographic diversification*, Agricultural Finance Review **73** (2013), 469–492.

[45] K. J. Lee, L. Guillemot, Y. L. Yue, M. Krammer, and D.J. Champion, *Application of the Gaussian mixture model in pulsar astronomy. Pulsar classification and candidates ranking for the Fermi 2fgl catalog*, Monthly Notices of the Royal Astronomical Society **424** (2012), 2832–2840.

[46] E. L. Lehmann, *Elements of Large-Sample Theory*, Springer, New York, 1998.

[47] D. X. Li, *On default correlation: A copula function approach*, Journal of Fixed Income **9(4)** (2000), 43–54.

[48] E. Liebscher, *Construction of asymmetric multivariate copulas*, Journal of Multivariate Analysis **102(4)** (2008), 869–870.

[49] G. M. Ljung and G. E. P. Box, *On a measure of a lack of fit in time series models*, Biometrika **65(2)** (1978), 297–303.

[50] Y. Lo, *Likelihood ratio tests of the number of components in a normal mixture with unequal variances*, Statistics and Probability Letters **71** (2005), 225–235.

[51] J. Mai and M. Scherer, *Simulating Copulas. Stochastic Models, Sampling Algorithms, and Applications*, Imperial College Press, London, 2012.

[52] H. B. Mann and A. Wald, *On stochastic limit and order relationships*, The Annals of Mathematical Statistics **14(3)** (1943), 217–226.

[53] A. J. McNeil and J. Nešlehová, *From Archimedean to Liouville copulas*, Journal of Multivariate Analysis **101(8)** (2010), 1772–1790.

[54] D. W. Müller and G. Sawitski, *Excess mass estimates and tests for modality*, Journal of the American Statistical Association **86** (1991), 738–746.

[55] M. Miloslavsky and M. J. van der Laan, *Fitting of mixtures with unspecified number of components using cross validation distance estimate*, Computational Statistics and Data Analysis **41** (2003), 413–428.

[56] R. B. Nelsen, *An Introduction to Copulas*, Springer, London, 2006.

[57] ______, *Extremes of nonexchangeability*, Statistical papers **48** (2007), 329–336.

[58] C. Ning and T. S. Wirjanto, *Extreme return-volume dependence in east-Asian stock markets: A copula approach*, Finance Research Letters **6** (2009), 202–209.

[59] Z. Ouyang, H. Liao, and X. Yang, *Modeling dependence based on mixture copulas and its application in risk management*, Applied Mathematics-A Journal of Chinese Universities **24(4)** (2009), 393–401.

[60] J. A. Peacock, *Two-dimensional goodness-of-fit testing in astronomy*, Monthly Notices of the Royal Astronomical Society **202** (1983), 615–627.

[61] J. Quessy and T. Bahraoui, *Graphical and formal statistical tools for assessing the symmetry of bivariate copulas.*, The Canadian Journal of Statistics **41(4)** (2013), 637–656.

[62] ______, *Graphical and formal statistical tools for the symmetry of bivariate copulas*, Canadian Journal of Statistics **41(4)** (2013), 637–656.

[63] J. C. Rodriguez, *Measuring financial contagion: A copula approach*, Journal of Empirical Finance **14** (2007), 401–423.

[64] B. W. Silverman, *Using kernel density estimates to investigate multimodality*, Journal of the Royal Statistical Society, Series B **43** (1981), 97–99.

[65] ______, *Density Estimation for Statistics and Data Analysis* , Chapman and Hall/CRC, London, 1986.

[66] W. Stute, W. Gonzalez Manteiga, and M. P. Quindimil, *Bootstrap based goodness-of-fit tests*, Metrika **40** (1993), 243–256.

[67] Rmetrics Core Team, *R package fcopulae*, 2013, R package version 3011.81.

[68] A. Tewari, M. J. Giering, and A. Raghunathan, *Parametric characterization of multimodal distributions with non-Gaussian modes*, Proceedings of the 11th IEEE International Conference on Data Mining Workshops, Vancouver, Canada, 2011.

[69] P. K. Trivedi and D. M. Zimmer, *Pitfalls in modelling dependence structures: Explorations with copulas*, The Methodology and Practice of Econometrics (J. L. Castle and N. Shephard, eds.), Oxford University Press, Oxford, 2009.

[70] T. R. Turner, *Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions*, Journal of the Royal Statistical Society: Series C (Applied Statistics) **49(3)** (2000), 371–384.

[71] W. van der Vaart and J.A Wellner, *Weak convergence and empirical processes*, Springer Series in Statistics, New York, 1996.

[72] N. Vasquez and N. Kawai, *Pulse characterization of long gamma-ray bursts with known redshift*, Physica E **43** (2011), 689–691.

[73] M. Vrac, L. Billard, E. Diday, and A. Chédin, *Copula analysis of mixture models*, Computational Statistics **27(3)** (2012), 427–457.

[74] M. P. Wand and M.C. Jones, *Kernel Smoothing*, Chapman and Hall, New York, 1995.

[75] X. Wang, *Selection of mixed copulas and finite mixture models with applications in finance*, Ph.D. thesis, The University of North Carolina at Charlote, 2008.

[76] Y. C. Wang, J. L. Wu, and Y. H. Lai, *A revisit to the dependence structure between stock and foreign exchange markets: A dependence-switching copula approach*, Journal of Banking and Finance **37** (2013), 1706–1719.

[77] G. N. F. Weiß and M. Scheffer, *Mixture pair-copula-constructions*, Journal of Banking and Finance **54** (2014), 175–191.

[78] S. S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, The Annals of Mathematical Statistics **9** (1938), 60–62.

[79] S. Wu, *Construction of asymmetric copulas and its application in two-dimensional reliability modelling*, European Journal of Operations Research **238(2)** (2014), 476–485.

[80] L. Xu, E. J. Bedrick, T. Hanson, and C. Restrepo, *A comparison of statistical tools for identifying modality in body mass distributions*, Journal of Data Science **12** (2014), 175–196.