Copula models for epidemiological research and practice

Eirini Koutoumanou

THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY COLLEGE LONDON

Great Ormond Street Institute of Child Health

2019

Declaration

I, Eirini Koutoumanou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

……………………………………

Abstract

Investigating associations between random variables (rvs) is one of many topics in the heart of statistical science. Graphical displays show emerging patterns between rvs, and the strength of their association is conventionally quantified via correlation coefficients. When two or more of these rvs are thought of as outcomes, their association is governed by a joint probability distribution function (pdf). When the joint pdf is bivariate normal, scalar correlation coefficients will produce a satisfactory summary of the association, otherwise alternative measures are needed.

Local dependence functions, together with their corresponding graphical displays, quantify and show how the strength of the association varies across the span of the data. Additionally, the multivariate distribution function can be explicitly formulated and explored.

Copulas model joint distributions of varying shapes by combining the separate (univariate) marginal cumulative distribution functions of each rv under a specified correlation structure. Copula models can be used to analyse complex relationships and incorporate covariates into their parameters. Therefore, they offer increased flexibility in modelling dependence between rvs.

Copula models may also be used to construct bivariate analogues of centiles, an application for which few references are available in the literature though it is of particular interest for many paediatric applications. Population centiles are widely used to highlight children or adults who have unusual univariate outcomes. Whilst the methodology for the construction of univariate centiles is well established there has been very little work in the area of bivariate analogues of centiles where two outcomes

are jointly considered. Conditional models can increase the efficiency of centile analogues in detection of individuals who require some form of intervention. Such adjustments can be readily incorporated into the modelling of the marginal distributions and of the dependence parameter within the copula model.

Impact statement

The methods and results presented in this thesis have the potential to improve patients' lives by allowing clinicians and researchers to better explore bivariate associations between random variables, e.g. clinical outcomes.

Such improvements can be achieved via firstly having a better understanding of the limitations of correlation coefficients. Secondly, by appreciating the enhancements local dependence functions can bring as well as alternative graphical displays that complement the conventional scatterplot. Thirdly, via multivariate distribution functions that can capture a wide range of association patterns and strengths between two or more random variables.

Copula models are a flexible tool that can be used for the construction of multivariate distribution functions and allow the exploration of varying relationships across the range of the two rvs. Informing healthcare professionals, including paediatricians, about these models will advance their understanding of multivariate relationships and enable them to construct more in-depth and flexible models for associations of variables and their joint characteristics.

Moreover, healthcare professionals often need to evaluate patients' individual or combined results and classify them according to whether they fall inside or outside a normal range of values (normal range is defined as a range within which the vast majority of the population lies). When two or more individual results are available, it would be beneficial to be able to classify patients according to whether they fall within a multivariate normal range or not, rather than evaluating multiple univariate ranges. Using a multivariate normal range is expected to reduce the number of false positive results, i.e. patients falsely being classified as unusual according to multiple individual tests which might in fact be within the normal range when jointly considered. The use of a multivariate range will also enable the identification of unusual cases that might have been missed based on univariate centiles alone, resulting in the identification of hidden extremes.

This thesis reviews the empirical exploration of bivariate associations, including conditional models, and showcases a new classification method for extreme values. This new method has the potential to become a conventional tool for everyday use in a healthcare setting that will enable its users to make better informed choices regarding unusual observations within a sample of subjects.

The applicability of the results from this thesis are not limited to healthcare professions. Greater flexibility in the investigation of bivariate associations and the ability to identify unusual observations in a multivariate setting is likely to have wider applicability.

Table of Contents

List of Figures

List of Tables

Abbreviations and Notation

BAC	bivariate analogue of centile
BCPEo	Box Cox Power Exponential original
BIC	Bayesian information criterion
CBAC	copula bivariate analogue of centiles
CCBA	conditional copula bivariate analogue of centiles
cdf, F	cumulative distribution function
CH	convex hull
df	degrees of freedom
EWMA	Exponentially Weight Moving Average
GIG	Generalised Inverted Gamma
ldf, γ	local dependence function
ldm	local dependence map
MSITAR	Multivariate Superimposition by Translation and Rotation
pdf, f	probability distribution function
rv	random variable
sd	standard deviation
SHASH	Sinh-Arcsinh
TD	tail dependence

m	number of dimensions/rvs
n	sample size
μ	mean
σ	population standard deviation
r	Pearson’s correlation coefficient
ρ	Spearman Correlation coefficient
τ	Kendall’s tau
λ_U	upper tail dependence
λ_L	lower tail dependence
C	copula function
θ	copula dependence parameter

Personal statement

(omitted)

Acknowledgments

This research was supported by the NIHR Great Ormond Street Hospital Biomedical Research Centre. The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health.

At the time of my registration as a PhD student and for few years thereafter, I was part of the ICH Centre for Paediatric Epidemiology and Biostatistics, MRC Centre of Epidemiology for Child Health. The Centre partially funded my studies and I am grateful for their support.

Overview

This thesis is structured in six Chapters and focuses on bivariate associations with discussion on appropriate extensions to multivariate dimensions of the methods and work presented here in the final Chapter.

The first half of Chapter 1 provides a description of conventional approaches to exploring associations (i.e. bivariate distributions and scalar measures of association). The second half of Chapter 1 introduces the notion of copulas and lists their fundamental properties. Several copula functions are shown in detail and each is graphically displayed to facilitate understanding of the variety of association patterns copula models can capture. Chapter 1 closes with a summary table of the copulas introduced and a flowchart diagram of the step-by-step copula fitting procedure.

Chapter 2 discusses local dependence and explores in detail various ways this can be measured and displayed; via the local dependence function, local dependence maps, chi-plots.

Chapter 3 starts with a literature review on the topic of bivariate centiles/tolerance regions. The concept of Bivariate Analogues of Centiles (BACs) is introduced and these are then applied to simulated datasets to test their efficiency and robustness. Joint, bivariate outliers are compared to values that univariately fall outside the normal range. The process is used to highlight potentially false extremes as well as hidden extremes, i.e. subjects whose univariate characteristics would respectively falsely flag them as extreme and miss them.

Chapter 4 combines copulas and BACs. This results in copula BACs (CBACs) and conditional copula BACs (CCBAC). Each of these extensions of BACs produce relevant and realistic results regarding joint extremes as they take into account the marginal distribution of each response variable with covariate adjustment where necessary.

Chapter 5 applies the techniques presented in Chapters 2 to 4 to analyse a large dataset comprising all live-births recorded in Mexico City in 2017. The exploration of this dataset starts with scalar coefficient measures, local association measures, followed by conditional copula modelling. The results of the BAC algorithm are presented, and these are also extended to CBACs and CCBACs.

The final Chapter draws conclusions on the analyses and results presented. Gaps in the literature are discussed alongside ideas for future work.

All analysis (plots, numerical results, etc.) in this thesis have been produced in R, The Comprehensive R Archive Network [1], version 3.6.1, unless otherwise stated. Where R libraries have been used, these are cited within the relevant sections.

1. Exploring relationships

Researchers often investigate the joint behaviour of several numerical response variables, such as size and shape of an object; speed, emissions, and noise of a car; right and left parts of the body; weight and height of children.

When exploring bivariate relationships, the tools used to understand and quantify the association of interest include probability distribution functions and measurements of association. This Chapter provides an overview of these two tools and also introduces copulas to further explore the association between random variables.

1.1 Probability distribution functions

A probability distribution function (pdf) describes the behaviour of a random variable (rv) across the entire range of possible values. More specifically, the pdf is a function whose value at any given datapoint can be interpreted as a relative likelihood that the random variable would equal that datapoint. All distributions require a set of parameters in order to be perfectly defined. These take the form of at least one of the following four parameters: location (μ), scale/dispersion (σ), skewness (ν) and kurtosis (τ). For example, for the Normal distribution μ is the mean and σ the standard deviation (sd) and has $\nu = 0$ and $\tau = 3 \cdot \sigma^4$, whilst the Gamma distribution is defined in terms of scale and shape parameters, with μ, σ, ν, τ being functions of them.

Results can be extended to multidimensional settings for dependence between two or more variables $(Y_1, Y_2, \cdots, Y_m)$, where bivariate pdfs, applicable to just two dependent variables, are most commonly seen in real-life applications.

1.1.1 Marginals and joint probability distribution functions

Let Y_1 be a continuous univariate rv. The distribution function of Y_1 in an interval (a, b) is given by the integral of Y_1's probability density function, f_1, over the interval:

$$\Pr(a \leq Y_1 \leq \mathrm{b}) = \int_a^b f_1(y_1)\,\partial y_1$$

F_1 is the cumulative distribution function (cdf) of Y_1 and provides the probability of its values falling in the interval $[-\infty, y_1]$, i.e. values less than or equal to y_1:

$$F_1(y_1) = \Pr(Y_1 \leq y_1) = \int_{-\infty}^{y_1} f_1(u)\,\partial \mathrm{u}$$

A pdf can be defined via the cdf as follows:

$$f_1(y_1) = \frac{\partial}{\partial y_1} F_1(y_1)$$

The joint pdf and cdf for rvs Y_1 and Y_2 are respectively given below, along with the function that connects the two:

$$\Pr(a_1 \leq Y_1 \leq b_1, a_2 \leq Y_2 \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f_{12}(y_1, y_2)\,\partial y_2\,\partial y_1$$

$$F_{12}(y_1, y_2) = \Pr(Y_1 \leq y_1, Y_2 \leq y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f_{12}(u_1, u_2)\,\partial u_2\,\partial u_1$$

$$f_{12}(y_1, y_2) = \frac{\partial^2 F_{12}(y_1, y_2)}{\partial y_1 \partial y_2}$$

The marginal pdf for each individual variable is given by:

$$f_1(y_1) = \int_{y_2} f(y_1, u_2)\partial u_2 \text{ and } f_2(y_2) = \int_{y_1} f(u_1, y_2)\partial u_1$$

If there are m $(m > 2)$ rvs, $Y_1 \ldots Y_m$, then the bivariate marginal pdf of $Y_1$and Y_2 is given by:

$$f_{12}(y_1, y_2) = \int_{y_3} \int_{y_4} \cdots \int_{y_m} f(y_1, y_2, u_3 \cdots, u_m) \partial u_3 \cdots \partial u_m$$

Any m-dimensional joint cdf F of continuous rvs Y_i, $i = 1 \ldots m$, with univariate margins $F_1, F_2, \ldots, F_m$, is bounded below and above by the Fréchet-Hoeffding [2,3] lower and upper bounds, F_L and F_U, respectively, defined as:

$$F_{L_{1 \ldots m}}(y_1, y_2, \ldots y_m) = \max\left[\sum_{i=1}^{m} F_i - m + 1, 0\right]$$

$$F_{U_{1 \ldots m}}(y_1, y_2, \ldots y_m) = \min[\, F_1, F_2, \ldots, F_m]$$

Two of the most common multivariate distribution functions, the Normal and the t, are presented in the following two sections.

1.1.1.1 The bivariate Normal distribution

The Normal distribution has a bell-shaped and symmetrical shape. It peaks at the centre of the curve and its values tail off evenly on either side of the centre (mean).

If μ and σ^2 denote the mean and the variance of a Normally distributed rv Y_1, $Y_1 \sim N(\mu, \sigma^2)$, the univariate Normal pdf of Y_1 is given by:

$$f_1(y_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_1 - \mu)^2}{2\sigma^2}\right)$$

The multivariate Normal pdf for an m-dimensional set of univariate rvs $Y_1, Y_2, \cdots, Y_m$ is given by:

$$f_{1\ldots m}(y_1, y_2, \ldots, y_m) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{m/2}} \exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right)$$

where y and μ represent, respectively, the m-dimensional vector of rvs $Y_1, Y_2, \cdots, Y_m$ and the m-dimensional vector of univariate means:

$$\mu = (\mu_1, \mu_2, \cdots, \mu_m)$$

Σ is the $m \times m$ covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1m}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{m1}^2 & \cdots & \sigma_m^2 \end{bmatrix}$$

where σ_{ij} (for $i, j = 1, \ldots, m$) is the covariance between rvs Y_i and Y_j and $\sigma_i{}^2$ is the variance of the i-th rv.

The pdf of a bivariate Normal distribution can be written as follows:

$$f_{12}(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-r_{12}^2)}} \exp\left\{\frac{1}{2(1-r_{12}^2)}\left[\frac{(y_1-\mu_1)^2}{\sigma_1^2} - 2r_{12}\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

Where $r_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ is Pearson's correlation coefficient [4], r (it measures linear dependence between two continuous rvs, more details about r follow in Section 1.2.1).

Figure 1-1 and Figure 1-2 below present the pdf and contour plots of 2 bivariate Normal distributions of a pair of random variables (y_1, y_2), where $\mu_1 = \mu_2 = 0$, $\sigma_{11}^2 = \sigma_{22}^2 = 10$ with varying correlation r_{12}.

Contour plots slice the probability density function horizontally in small areas and show how dense the function is within each sliced area (25 levels are displayed in Figure 1-1 and Figure 1-2). The z-axis of the bivariate 3D-pdf represents the bivariate probability density and the values of the pair (y_1, y_2) are represented on the two horizontal axes respectively, as annotated on the plot.

Figure 1-1: Bivariate Normal distribution, $r_{12} = -0.9$

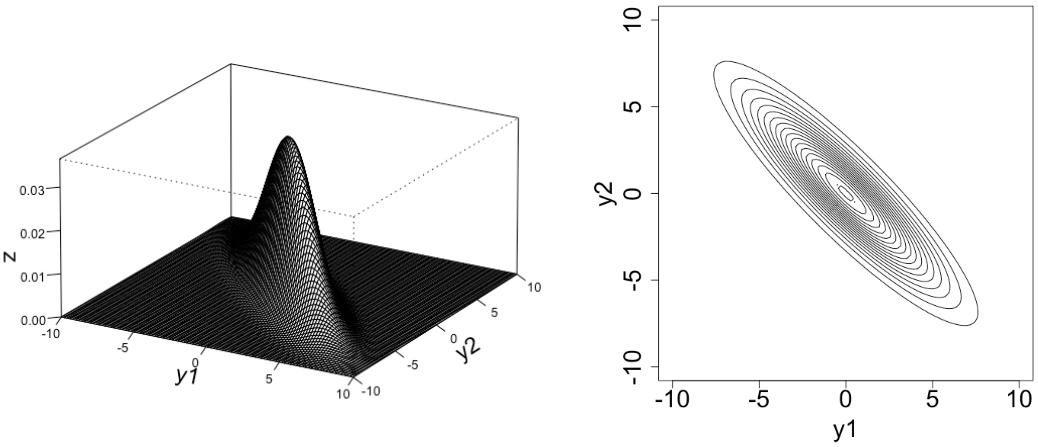
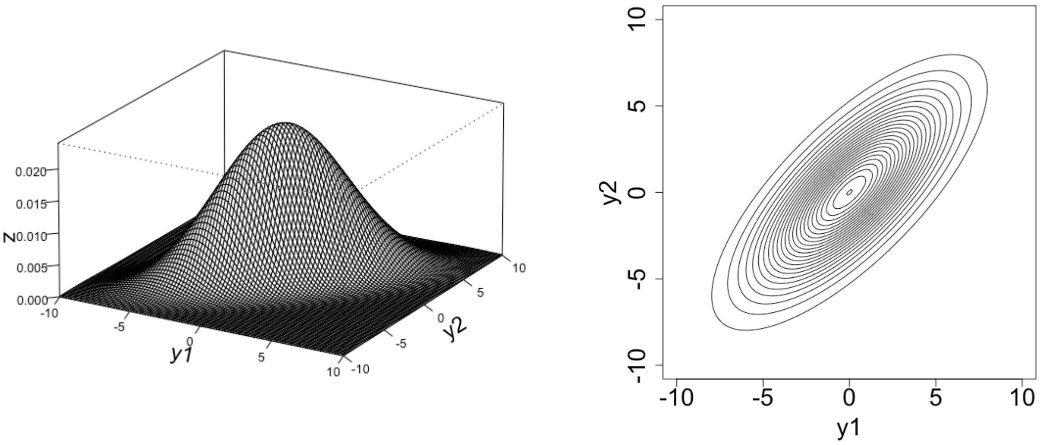


Figure 1-2: Bivariate Normal distribution, $r_{12} = 0.75$



The main characteristics of the bivariate Normal distribution [5] are constant correlation, radial symmetry, and Normally distributed conditional and marginal distributions for any dimensions smaller than m. In situations where the above are appropriate assumptions of the bivariate association between the rvs under investigation, the bivariate Normal can be a good model choice.

However, there are many examples, where at least one of these features will not be evident. For example, non-constant correlation is often evident in the medical field where it is common to find distributions with a few patients producing extreme responses, i.e. most healthy children's visual acuity in the right and left eyes would be expected to be scattered around a 'normal' centre depending on their age group, but visually impaired (but otherwise healthy) children may distort the shape of this distribution.

Normality tests are often used to evaluate evidence (in the form of p-values) for or against the assumption of a Normally distributed population from which a rv has come from [6]. However, the key question at this stage is whether a given data set approximates the Normal distribution well, but significance tests for normality answer the alternative question of whether the population that the sample was derived from could be Normally distributed.

Multivariate skewness and kurtosis indices were introduced with corresponding statistical tests by Mardia [7,8] and can assist in quantifying the shape of the distribution under investigation. The notation generally used is β_1 for skewness and β_2 for kurtosis, and their values indicate the extent of departure from multivariate normality. Values of β_1 close to 0 and to $m(m+2)$ for β_2, correspond to bivariate (or higher) Normal distributions, where m denotes the dimensionality of the outcome ($m=2$ for bivariate scenarios).

The multivariate t distribution provides an alternative with heavier tails and hence has wider applicability (the Normal distribution is a special case of the t).

1.1.1.2 The bivariate t distribution

The t distribution [9,10] is also symmetric and bell-shaped but has heavier tails than the Normal distribution. If X is Normally distributed with mean μ and variance σ^2 (sample estimate s), then $t_1 = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ is a t-distributed rv, $t_1 \sim t(\mathrm{df})$, where $df = n - 1$ and whose values can range within $(-\infty, \infty)$, with pdf:

$$f_1(t_1) = \frac{\Gamma\left(\frac{\mathrm{df}+1}{2}\right)}{\sqrt{\mathrm{df}\cdot\pi}\,\Gamma\left(\frac{\mathrm{df}}{2}\right)}\left(1+\frac{\mathrm{t_1}^2}{\mathrm{df}}\right)^{-\frac{\mathrm{df}+1}{2}}$$

Γ is the gamma function and df denotes the degrees of freedom which must be positive but not necessarily an integer.

The number of degrees of freedom, df, is a shape parameter and determines how strong/dominant the peak of the curve is around the centre. As it increases, the t distribution approaches the standard Normal distribution. The mean of the t distribution is 0 for $\mathrm{df} > 1$ and the variance for $\mathrm{df} > 2$, is given by:

$$\mathrm{Var}(t_1) = \frac{\mathrm{df}}{\mathrm{df}-2}$$

The pdf of the bivariate t distribution [11] is given by:

$$f_{12}(t_1, t_2) = \frac{\Gamma\left(\frac{\mathrm{df}+2}{2}\right)}{\mathrm{df}\cdot\pi\cdot\Gamma\left(\frac{\mathrm{df}}{2}\right)}\frac{1}{|R|^{1/2}}\left(1+\frac{\underline{t}'R^{-1}\underline{t}}{\mathrm{df}}\right)^{-\frac{\mathrm{df}+2}{2}}$$

With:

- $\underline{t} = (t_1, t_2)$ representing the bivariate vector of variables t_1 and t_2
- the correlation matrix, R is a positive-definite real 2×2 matrix, where the off-diagonal elements contain the correlation between t_1 and t_2 (notice that this is not the covariance matrix as it was for the Normal distribution). The covariance is given by [12]: $\frac{\mathrm{df}}{\mathrm{df}-2}R$ for $\mathrm{df} > 2$.

Similarly, to the Normal distribution above, bivariate t pdfs can be drawn along with contour plots. In fact, the bivariate t plots look very similar to the equivalent Normal distribution plots with the only difference seen in the heavier tails of the t distribution.

Summary

The end of this review of the bivariate Normal and t distributions highlights the importance of best practice when variables do not comply with the characteristics of either of these distributions, which so often is the case in the medical and other fields. Copulas are multivariate distribution functions that allow great flexibility in the shape/form of the association that is being investigated. Copula functions are described in detail in section 1.3 and section 1.4 and their benefits as means of constructing bivariate distributions are demonstrated. Firstly, however, the idea and principles of scalar dependency, including correlation coefficients need to be explored (Section 1.2) as these play a significant role in the understanding of copulas.

1.2 Scalar dependence coefficients

There are several quantities that measure the association/dependence between two continuous random variables [13,14] in the numerical scale (i.e. scalar coefficients). Three of the most commonly used ones will be considered here: linear correlation, concordance (or rank correlation) and tail dependence.

Scalar coefficients summarise the relationships between pairs of variables (Y_1 and Y_2) in one number. They assume that the strength of each pairwise relationship condenses to a single parameter/numerical value. Similarly, tail dependence measurements are scalar parameters that focus on describing the tails of the marginal distribution functions of the variables whose association is being explored.

A “good” scalar measure of association between any two pairs of bivariate observations (y_{1i}, y_{2i}) and (y_{1j}, y_{2j}) from rvs Y_1 and Y_2, as defined by Gibbons and Chakraborti [15] is one that satisfies the following criteria:

i. The association measure equals 1 if the relationship is direct and perfect (perfect concordance) such that: $y_{1i} < y_{1j}$ when $y_{2i} < y_{2j}$ or $y_{1i} > y_{1j}$ when $y_{2i} > y_{2j}$

ii. The measure equals -1 if the relationship is indirect and perfect (perfect discordance) such that: $y_{1i} < y_{1j}$ when $y_{2i} > y_{2j}$ or $y_{1i} > y_{1j}$ when $y_{2i} < y_{2j}$

iii. If neither criterion (i) nor (ii) is true for all pairs, the measure lies between the two extremes, -1 and 1

iv. The measure equals 0 if y_1 and y_2 are independent

v. The measure for y_1 and y_2 is the same as for y_2 and y_1 or $-y_1$ and $-y_2$ or $-y_2$ and $-y_1$

vi. The measure for $-y_1$ and y_2 or y_1 and $-y_2$ is the negative of the measure for y_1 and y_2

vii. The measure should be invariant under all transformations of y_1 and y_2 for which order is preserved

The next three sections briefly introduce each of the three scalar association measurements mentioned earlier (linear correlation, concordance (or rank correlation) and tail dependence) and each section concludes with several graphical displays for illustration. Via this overview the pros and cons of each scalar measurement will be highlighted and the need for an alternative will emerge. Association measures should be expected to yield more than just a single number in order to successfully represent complicated changes in magnitude and direction of correlation across the range of two or more rvs. Chapter 2 explores in detail alternatives to scalar measures.

1.2.1 Scalar linear correlation coefficient

By far, the most commonly used dependence measure is the Pearson's linear correlation coefficient [4], r, which quantifies linear dependence, by "averaging out" the linear association between two rvs (where $\cong$ denotes the approximation of a population parameter by a sample estimate):

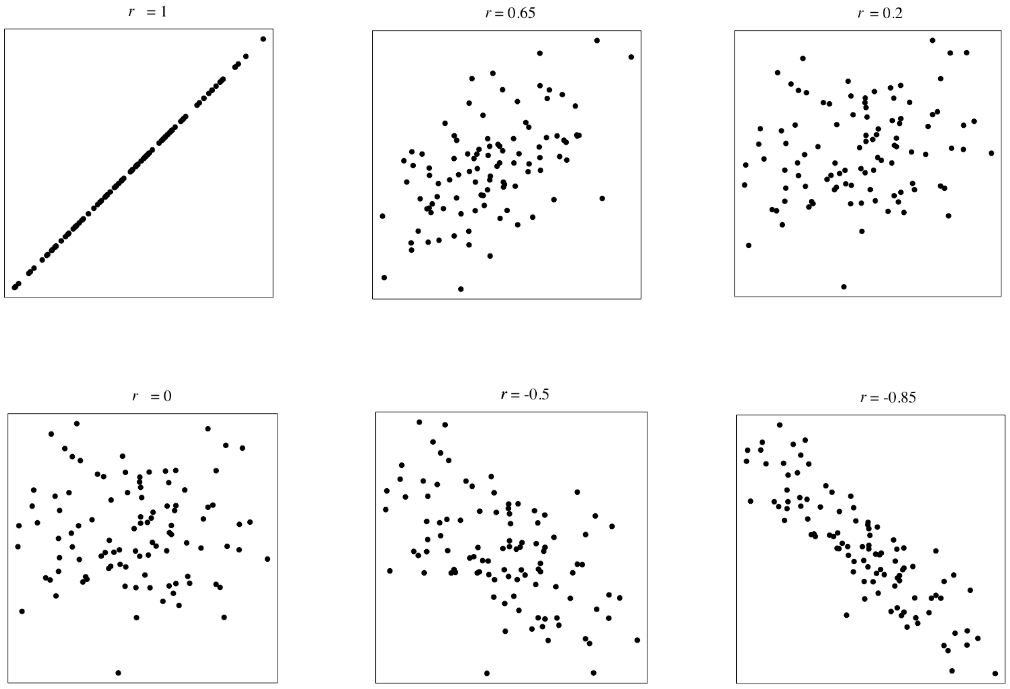
$$r_{12} = r(y_1, y_2) = \frac{\mathrm{Cov}(y_1, y_2)}{\sigma_1 \sigma_2} \cong \frac{\sum_i (y_{1i} - \overline{y_1})\,(y_{2i} - \overline{y_2})}{\sqrt{\sum_i (y_{1i} - \overline{y_1})^2}\sqrt{\sum_i (y_{2i} - \overline{y_2})^2}}$$

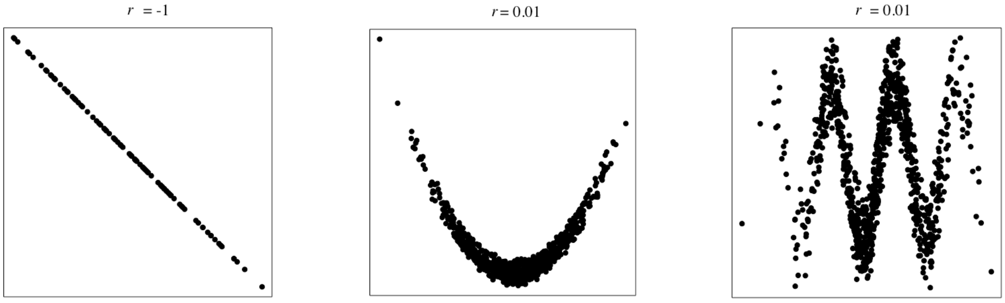
- $\mathrm{Cov}(y_1, y_2)$, $\sigma_1$and σ_2 denote the covariance between rvs y_1 and y_2 and the standard deviation of each, respectively.
- r satisfies the first six general properties of association measures seen earlier.
- However, it is invariant only with respect to linear transformations of y_1 and y_2, e.g. $r_{12} = r(y_1 + a, y_2) = r(y_1 + b, cy_2 - d)$ where a, b, c and d are real numerical values, but not for all order-preserving transformations.

Pearson's r is a reflection of the proximity of the data to a straight line, so it only detects linear relationships between the rvs being investigated. Non-linear associations cannot be validly explored with Pearson's coefficient. Additionally, if the bivariate distribution between the two rvs is Normal, it then provides a complete description of the dependence structure between them, as a straight line cutting through a bivariate Normal shape is definitely a good summary of the linear association seen. Moreover, if the pair (y_1, y_2) follows a bivariate Normal distribution, then $r_{12} = 0$ implies that the two variables are independent; the latter is not necessarily true when the pair of rvs does not follow a bivariate Normal distribution.

The first seven graphs in Figure 1-3 show examples of a Bivariate Normal distribution with different values for r, including its two most “extreme” values of $+1$ and -1. The last two graphs show two cases where there is a very well-defined non-linear association between the two rvs and Pearson’s r is unable to capture this.

Figure 1-3: Examples of r values for 100 simulated Bivariate Normal cases





The necessary prerequisites of linear structure and bivariate Normality of the data can prove too restrictive in many real-life scenarios. Correlation measures for non-linear associations are discussed in detail in Barbour [16] and two of the most common such measures are described in the following sections.

1.2.2 Scalar concordance correlation coefficients

The Spearman's ρ [17] and Kendall's τ [18] are two alternative distribution-free scalar dependence measures, referred to as rank correlation measures. The formulae for each are presented below:

$$\rho_{12} \cong \frac{\sum_i (y_{r1i}-\overline{y_{r1}})\,(y_{r2i}-\overline{y_{r2}})}{\sqrt{\sum_i (y_{r1i}-\overline{y_{r1}})^2\,\sum_i (y_{r2i}-\overline{y_{r2}})^2}}$$

(Spearman's)

If there are tied ranks, where y_{r1i} and y_{r2i} are the ranked values of the original data y_1 and y_2.

$$\rho_{12} \cong 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$$

(Spearman's)

A simpler calculation that can be used if there are no tied ranks, where d is the difference in paired ranks (i.e. ranks of y_{1i} and y_{2i})

$\tau_{12} \cong \frac{n_c - n_d}{\binom{n_c + n_d}{2}} = \frac{2(n_c - n_d)}{n(n-1)}$

(Kendall's)

n_c and n_d are the number of concordant and discordant pairs, respectively

Two pairs (y_{1i}, y_{2i}) and (y_{1j}, y_{2j}) are concordant if $y_{1i} > y_{1j}$ and $y_{2i} > y_{2j}$ or if $y_{1i} < y_{1j}$ and $y_{1i} < y_{2j}$. The pairs are said to be discordant if $y_{1i} > y_{1j}$ and $y_{1i} < y_{2j}$ or if $y_{1i} < y_{1j}$ and $y_{2i} > y_{2j}$.

Spearman's ρ and Kendall's τ satisfy all criteria seen in the earlier section. In contrast with Pearson's r, they satisfy the seventh criterion as ranks are preserved under all order-preserving transformations, hence ρ and τ will remain constant for such transformations.

Finally, note that all the correlation measurements introduced are specific applications of the generalised correlation coefficient that was first discussed by Daniels in 1944 [19]. The generalised correlation coefficient (ignoring standardisation) of ordered rvs Y_{1i} and Y_{2j} for $i, j = 1 \dots n$ is given by:

$$\Gamma = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{ij}$$

Where a_{ij} and b_{ij} are scores for every pair of (y_{1i}, y_{1j}) and (y_{2i}, y_{2j}) observations, respectively. The following choices for a_{ij} and b_{ij} lead respectively to Pearson's r, Spearman's ρ and Kendall's τ:

$$a_{ij} = (y_{1i}, y_{1j}) \quad \text{and} \qquad b_{ij} = y_{2i} - y_{2j}$$

$$a_{ij} = \text{rank}(y_{1i}) - \text{rank}(y_{1j}) \quad \text{and} \qquad b_{ij} = \text{rank}(y_{2i}) - \text{rank}(y_{2j})$$

$$a_{ij} = \text{sign}(y_{1i} - y_{1j}) \quad \text{and} \qquad b_{ij} = \text{sign}\big(y_{2i} - y_{2j}\big)$$

Figure 1-4: Simulated data with scalar coefficients

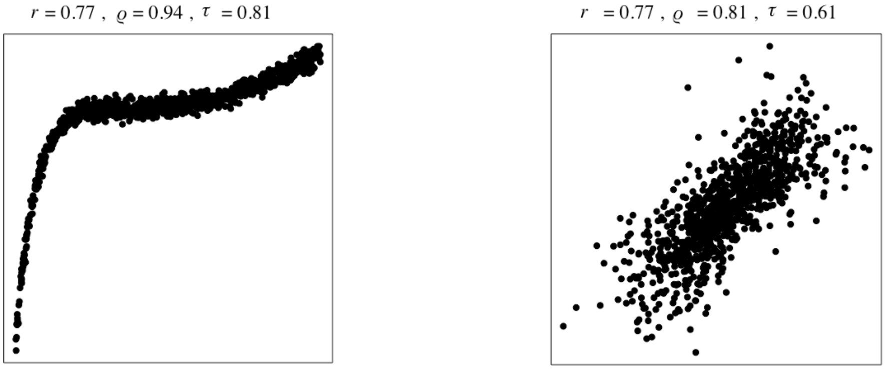


Figure 1-4 shows that the three scalar coefficients differ depending on the association pattern evident in the data. In the first example, Spearman's and Kendall's are larger than Pearson's as they are not influenced by the curvature of the pattern. In the second example, Pearson's is affected by outliers. All three measures result in no more than a single number to summarise dependency, which can be deemed inadequate in many cases, as illustrated by the examples above.

1.2.3 Scalar tail dependence (TD) coefficient

The concept of bivariate tail dependence (λ) [20,21] relates to the dependence in extreme values at the tails of a bivariate distribution. It measures dependence in the upper-quadrant tail or lower-quadrant tail of a bivariate distribution. As previously, let $f_{12}(y_1, y_2)$ denote the bivariate pdf of variables Y_1 and Y_2 and $f_1(y_1)$ and $f_2(y_2)$ the two marginals. Tail dependence (TD) is defined as the probability a value of one of the marginals exceeds a high/low threshold (t') under the condition that the other marginal has already exceeded that threshold.

The upper tail dependence coefficient, λ_U, is defined [5], if the following limit exists, as:

$$\lambda_U = \lim_{t\prime \to 1^-} \mathrm{Pr}\,(f_1(y_1) > t'|(f_2(y_2) > t')$$

If $\lambda_U > 0$, Y_1 and Y_2 are said to be upper-tail dependent and upper-tail independent if $\lambda_U = 0$.

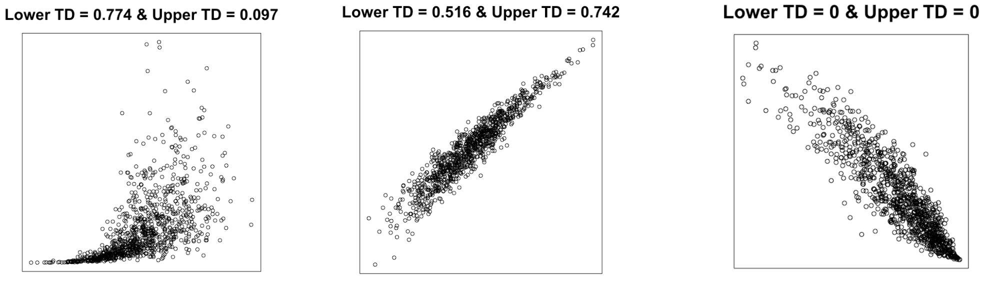
Similarly, the lower tail dependence coefficient, λ_L, is given by:

$$\lambda_L = \lim_{t\prime \to 0^+} \mathrm{Pr}\,(f_1(y_1) \leq t'|(f_2(y_2) \leq t')$$

If $\lambda_L > 0$, Y_1 and Y_2 are said to be lower-tail dependent and lower-tail independent if $\lambda_L = 0$.

Figure 1-5 shows examples of various tail dependence scenarios. For negative correlation, lower and upper tail dependence is zero as the two rvs do not span small or large values at the same time, i.e. likelihood of the x axis to be below a certain threshold whilst values on the y axis are also below this threshold is 0.

Figure 1-5: Scenarios of varying tail dependence



TD does not satisfy the criteria for a “good” measure of association presented at the beginning of section 1.2 as it only measures association at the tail of the bivariate distribution, as opposed to any pair of values of y_1 and y_2 and it can only range from 0 (no tail dependence) to 1 (perfect tail dependence), as opposed to from -1 to 1.

Summary

Scalar dependence measurements do not adequately capture non-constant correlation (mixture of positive, negative and zero associations) across the defined range of the two variables of interest or just in their tails. Scalar correlation measures assume that correlation remains constant along the range of the two rvs. However, varying association structures do not always comply with this assumption, i.e. patterns of association that change in strength across the range of observed values are sometimes more realistic. Out of the 5 measures, the TD represents a local concept of correlation by simply narrowing down the range of values it focuses on. This concept will be further explored in Chapter 2 via the introduction of the local dependence function.

Finally, there is clearly a need for an improved way for exploring bivariate associations that reflect realistic data patterns. Multivariate distribution functions provide an answer to this objective and are described in the next section.

1.3 Copulas

For multivariate distributions to be constructed, the univariate distribution of each of the variables of interest should be fully defined. Many applications and methodologies assume that the joint behaviour of several variables is best described via a multivariate Normal distribution. However, this is often unrealistic. Copulas provide flexible means of multivariate distribution construction that goes beyond the conventional multivariate Normal or t distributions.

Copulas or copula distributions or copula models are multivariate probability distribution functions (bivariate in their simplest form) that describe the joint behaviour of two or more rvs; in other words, they provide a comprehensive model of the dependence structure between several variables.

The term originates from the Latin word "cōpula"; from *co*- which means together and *apere* which means to fasten [22]. It is used to describe anything (e.g. a word, an object), that connects, ties, or bonds elements together.

Mathematically, copulas are joint cumulative distribution functions generated from given marginals. In other words, they couple multivariate distribution functions to their marginal distribution functions.

This thesis focuses on continuous random variables (rvs), unless otherwise stated. For continuous rvs $Y_i,\ i = 1 \dots m$, if F is an $m-$dimensional cumulative distribution function (cdf) with one-dimensional margins $F_1, F_2, \dots, F_m$, then there exists an m-dimensional copula function C such that:

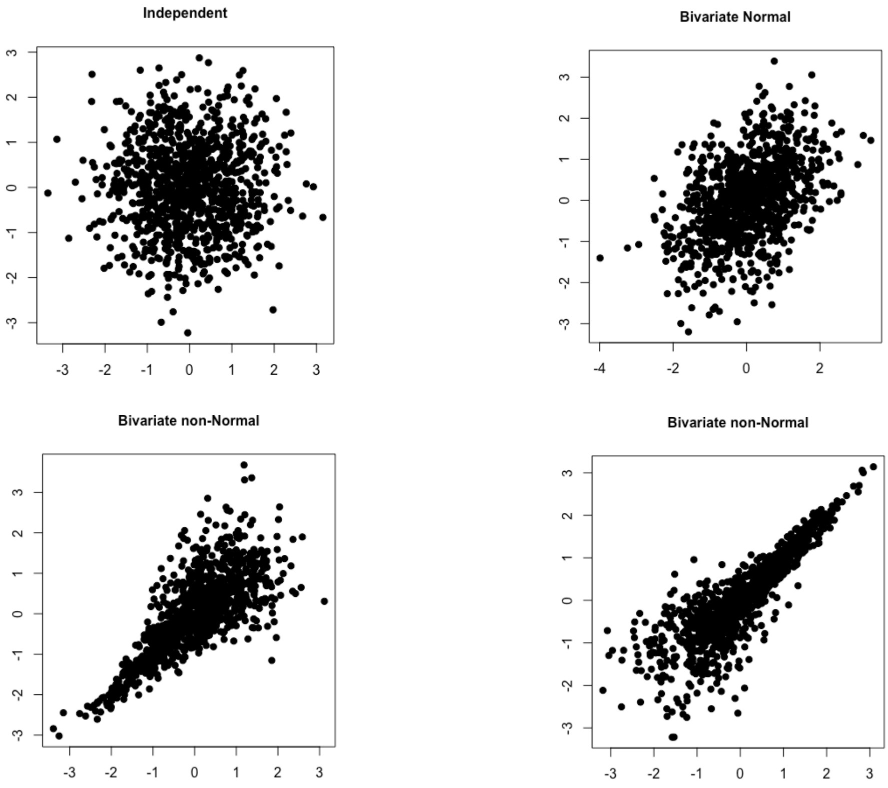
$$F_{1,2,\dots,m}(y_1, y_2, \dots, y_m\ ; \theta) = C(F_1(y_1), F_2(y_2), \dots, F_m(y_m); \theta)$$

Where $F_i(y_i) = F_{y_i}(y_i)$, for $i = 1, \dots, m$ and θ is the dependence or copula parameter, which governs the degree of association between the marginals (more details about θ will follow in Section 1.3.3).

*The univariate marginals of the response variables are **coupled** via the copula function C, which combined with the copula parameter (θ) leads to the **copula distribution**.*

The bivariate case, $m = 2$, has attracted special attention, with the best-known bivariate copula function being the bivariate Normal distribution $F(y_1, y_2)$ with Normal margins, F_1 and F_2. However, within the copula framework, two Normal marginals can also lead to a non-Normal bivariate distribution. To illustrate this, Figure 1-6 shows 4 sets of data whose marginals are Normally distributed with mean 0 and standard deviation 1 (i.e. N$(0,1)$). They are clearly different though with respect to their joint distribution. The first graph shows complete independence and Kendall's rank correlation τ equals 0. For the remaining three graphs τ equals 0.5. The "Bivariate Normal" graph is an example where the joint distribution of the Normal marginals is also Normal, whereas for the remaining two ("Bivariate non-Normal") the Normal marginals join to produce non-Normal bivariate probability density functions, skewed on the lower and upper tails respectively. Copulas provide the means via which an explicit exploration of dependence patterns between rvs such as the ones seen in Figure 1-6 is undertaken.

Figure 1-6: Examples of bivariate distributions with Normal marginals



Copulas first appeared in the works of Hoeffding [23] and Fréchet [2]. Sklar was the first to use the term “copulas” to denote these functions in an article published in 1959 [24]. In 1990, Dall’Aglio organised the first conference devoted to the idea of marginal distributions and the way they join together, entitled “Probability distributions with given marginals” [25].

By the end of the 1990s, the notion of copulas was increasingly popular with two text books becoming standard references in this area. In 1997, Joe [14] published a book on multivariate models and in 1999 Nelsen [26] published the first edition of an introductory text on copulas, followed by a second edition in 2006 [27]. The main reason for this increased interest was the realisation of the advantages copula models offer in research fields as diverse as finance and hydrology. This led to appropriate computational developments which assisted copulas’ further applicability and ease of implementation. In 2007, Schweizer [28] noted that:

“The ‘era of i.i.d.’ is over: and when dependence is taken seriously, copulas naturally come into play. It remains for the statistical community at large to recognise this fact. And when every statistics text contains a section or Chapter on copulas, the subject will have come of age.”

Twelve years later, I believe there is still work to be done in the recognition of the benefits of copula multivariate analysis; this can be partly achieved by enabling researchers to extend familiar univariate methods to their multivariate equivalents, an area which I think lacks behind. This thesis aims to fill in one of these gaps by producing an equivalent to univariate centiles.

In a nutshell, copulas provide increased flexibility in capturing dependence between rvs to supplement the commonly used independence models (perhaps with non-Normal marginals) and multivariate Normal models.

The remaining sections of this Chapter describe theoretical aspects of copulas (i.e. Sklar’s theorem, copula parameter and copula types).

1.3.1 Sklar’s theorem

Abe Sklar in 1959 [24] published the following definition of copulas:

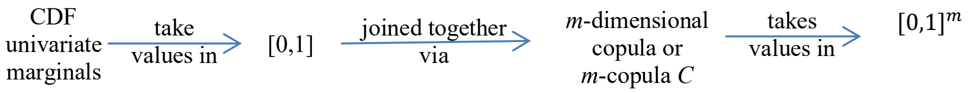
Let H be an m −dimensional cdf with marginals $F_i(y_i), i = 1, \dots, m$. Then there exists a copula $C: [0,1]^m \rightarrow [0,1]$ such that for all rvs Y_i in [-∞, +∞]:

$$H(y_1, y_2, \dots, y_m) = C\big(F_1(y_1), F_2(y_2), \dots, F_m(y_m)\big) \qquad \textbf{(Eq. 1-1)}$$

If $F_1, F_2, \dots, F_m$ are all continuous, then C is unique. Conversely, if C is an $m-$copula and $F_1, F_2, \dots, F_m$ are distribution functions, then the function H defined in (Eq. 1-1) is an $m-$dimensional distribution function with margins $F_1, F_2, \dots, F_m$. A proof of this theorem can be found in Schweizer and Sklar [29].

This thesis will focus on continuous copulas and all C functions that follow will be considered as continuous unless otherwise stated.

Researchers often have information about marginal distributions of individual variables but know little about their joint behaviour. Copulas can be used to piece together joint marginal distributions or express a multivariate distribution in terms of its marginals.



The function C can be obtained as:

$$C(u_1, u_2, u_3, \dots, u_m) = H\left(F_1^{-1}(y_1), F_2^{-1}(y_2), F_3^{-1}(y_3), \dots, F_m^{-1}(y_m)\right)$$

Where $F_i^{-1}(y_i)$, also known as the quantile function of y_i, denotes the inverse of F_i . Thus, copulas are essentially transformations of rvs $Y_1, \dots, Y_m$ into another set of variables $U_1, \dots, U_m$ whose margins are uniform on the unit range $[0,1]$, i.e. $U_i = F_{Y_i}(Y_i) \sim U(0,1)$. Such strictly increasing transformation (known as the probability integral transform [30]) is invariant with regards to the association between Y_is, i.e. the dependence structure is preserved amongst the new components, U_i.

For example, in the bivariate case:

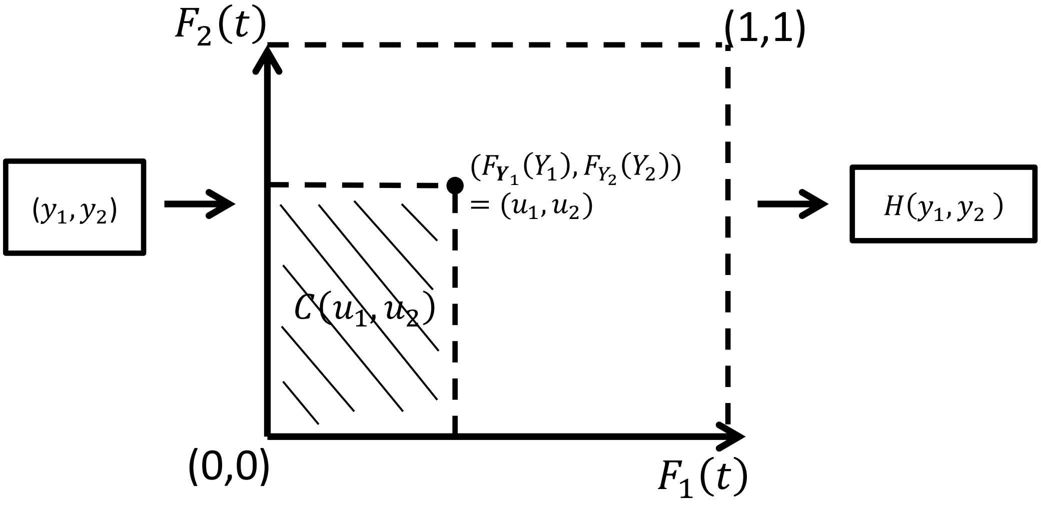
$$\Pr(U_1 \leq u_1) = \Pr\left(F_{Y_1}(Y_1) \leq u_1\right) = \Pr\left(Y_1 \leq F_{Y_1}^{-1}(u_1)\right) = F_{Y_1}(F_{Y_1}^{-1}(u_1)) = u_1$$

$$C(u_1, u_2) = \Pr(U_1 \leq u_1, U_2 \leq u_2) = \Pr\left(Y_1 \leq F_{Y_1}^{-1}(u_1), Y_2 \leq F_{Y_2}^{-1}(u_2)\right)$$

$$= H\left(F_{Y_1}^{-1}(u_1), F_{Y_2}^{-1}(u_2)\right)$$

Finally, Figure 1-7 below represents all the relationships described above from the perspective of the $[0,1]^2$ plane defined by the two uniform marginal distributions, where t denotes any real number.

Figure 1-7: The marginal and copula distributions in the $[\mathbf{0},\mathbf{1}] \times [\mathbf{0},\mathbf{1}]$ plane



The copula represents the area of the $[0, u_1] \times [u_2, 0]$ rectangle.

The joint distribution is expressed in terms of its respective marginal distributions and a function C that binds them together. The copula C depends on θ – this is the parameter of the copula function (known as dependence or copula parameter), which governs the degree of association between the marginals. It may be multivariate, though it is often defined as a scalar. More details about the copula parameter θ follow in Section 1.3.3.

To summarise, copula analysis involves specifying univariate marginals for each rv along with a copula function that binds them together. A copula can incorporate various forms of dependence structures regardless of the form of the marginals. Choosing the right copula function to capture the underlying dependence structure becomes the pivotal problem in many applications.

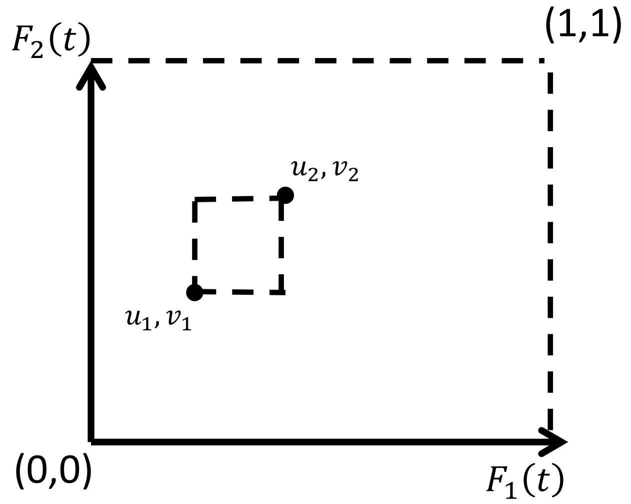
1.3.2 Copula properties

Bivariate copulas are written below in terms of standard uniform rvs U_1 and U_2 such that $C(u_1, u_2)$: $[0,1]^2 \rightarrow [0,1]$. Any copula function must satisfy the following properties [26].

- For every u_1, u_2 in $[0,1]$, $C(u_1, 0) = 0 = C(0, u_2)$, $C(u_1, 1) = u_1$ and $C(1, u_2) = u_2$.

- For every u_1, u_2, v_1, v_2 in $[0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$, $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$. In other words, the volume of the rectangle in the unit cube, defined by the univariate marginals on $[0,1] \times [0,1]$ is positive. In terms of rectangles, similar to Figure 1-7 seen earlier, this property translates to the volume of the rectangle in Figure 1-8 being positive.

Figure 1-8: Copula properties – positive volume



- If the copula is a product of two marginals, then this implies that the variables are independent and separate estimation of each marginal is appropriate.

- Suppose T_1 and T_2 are non-decreasing continuous functions of u_1 and u_2, then the random vector $(T_1(u_1), T_2(u_2))$ has the same copula C as u_1 and u_2. Hence, provided a marginal distribution for each rv can be specified, copulas do not require data transformations of any kind to perform/fit well.

- Any copula C is bounded by copulas M, W such that $\forall$ u_1, u_2 in [0,1]:

  $W(u_1, u_2) \leq C(u_1, u_2) \leq M(u_1, u_2)$ and

  $\max(u_1 + u_2 - 1{,}0) \leq C(u_1, u_2) \leq \min(u_1, u_2)$

  M and W are called the Fréchet-Hoeffding bounds for C. Any m-dimensional joint cdf F of continuous rvs $Y_i,\ i = 1 \dots m$, with univariate margins $F_1, F_2, \dots, F_m$, is bounded below and above by the Fréchet-Hoeffding [2,3].

- The lower bound $W(u_1, u_2)$ corresponds to perfect negative dependence.

- The upper bound $M(u_1, u_2)$ describes perfect positive dependence.

1.3.3 Dependence measures and the copula parameter

A copula is fully defined via its dimensions, dependence parameter and marginals. The aim is to find a copula that best describes the dependence structure between observed variables

The dimension of a copula is that of its characteristics and is equal to the number of variables whose dependence is being investigated. Hence, a 2-dimensional copula is a bivariate distribution function; a 3-dimensional copula is a trivariate distribution function and so on.

The dependence parameter, θ, accounts for the strength of the relationship under investigation.

A key consideration for the choice of the right copula is the ability of the model to capture the dependence between the variables of interest.

The scalar correlation measures introduced earlier can be expressed in terms of the y_1 and y_2 univariate margins, $F_1(y_1)$, $F_2(y_2)$ and bivariate distribution, $F_{12}(y_1, y_2)$ [31–33] as follows:

Pearson's: $r_{12} = \frac{1}{\sigma_1 \sigma_2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [F_{12}(y_1, y_2) - F_1(y_1) F_2(y_2)] \partial y_1 \partial y_2$

Spearman's: $\rho_{12} = 12 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [F_{12}(y_1, y_2) - F_1(y_1) F_2(y_2)] \partial F_1(y_1) \partial F_2(y_2)$

Kendall's: $\tau_{12} = 4 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F_{12}(y_1, y_2) \partial F_{12}(y_1, y_2) - 1$

Considering $u_1 = F_1(y_1)$ and $u_2 = F_2(y_2)$ as defined in section 1.3.1 $(U_i = F_{Y_i}(Y_i) \sim U(0{,}1))$, the above equations can be re-written as follows, where C is the copula function of (y_1, y_2).:

Pearson's: $r_{12} = \frac{1}{\sigma_1\sigma_2}\int_0^1\int_0^1 [C(u_1,u_2) - u_1u_2]\partial F_1^{-1}(u_1)\,\partial F_2^{-1}(u_2)$

Spearman's: $\rho_{12} = 12\int_0^1\int_0^1 [C(u_1,u_2) - u_1u_2]\partial u_1 \partial u_2$

Kendall's: $\tau_{12} = 4\int_0^1\int_0^1 C(u_1,u_2)\,\partial C(u_1,u_2) - 1$

(Eq. 1-2)

These equations clearly show that these scalar coefficients are functions of C over specified ranges and each can be calculated from the copula function. Hence, copulas are also meaningful in scenarios where the dependence structure is well described by a scalar correlation coefficient.

Recall that a bivariate copula function C is formed by two univariate marginals and the dependence parameter θ. The marginals are often specified according to observed data, hence from the formulae above it is clear that θ is the only element directly associated with the scalar correlation coefficients. The exact relationship binding each of these with θ depends on the functional form of C and hence will differ for different types of copula. More specifically, the relationship between τ and the θ plays an integral part in the definition of certain classes of copulas. This is because τ is the only one of the correlation coefficients above that depends just on the copula function and not on the marginals themselves [27]. For some copulas, as seen in Table 1-1 towards the end of this Chapter, θ can be expressed as a function of Kendall's τ. But for others, the equation connecting τ and θ might not yield an exact solution due to the mathematical formulations involved, but it is always possible to numerically approximate θ from τ.

1.4 Copula families

Having specified the marginal distributions of each variable, an appropriate copula function should be selected that best captures the dependence structure of the data.

A large number of copulas have been proposed, each imposing a different dependence structure on the data. Some of these are grouped together in copula families as their functions fall under the same general formula/rules and have the same properties. Hence, there are copulas that are stand-alone, not belonging to a family and there are others that fall under a specific categorisation of copulas, i.e. a family of copulas.

An extensive description of bivariate copulas is given by Joe [14] and Nelsen [26]. Here, the focus is on only some of them and at the end of the section, a table will summarise the dependence features these copulas can capture along with additional copulas that have not been described here.

There are no definite rules about which copula type is right for a certain data set; selecting a copula to fit specific data is an important but difficult problem. The true data generation mechanism is often unknown and hence it is possible that several, or none, of the possible copulas may fit the data reasonably well. A maximum likelihood method can be used to compare candidate copulas and select the optimum as that with the highest likelihood based on the Bayesian Information Criterion (BIC) and/or Akaike's Information Criterion (AIC) [34,35] values.

Each copula introduced in this section is followed by graphical examples of the respective bivariate density and contour plots of u_1 and u_2. Contour plots slice the pdf horizontally in small areas and show how dense the function is within each sliced area; 25 levels are displayed in all the graphs of this Chapter. For the remainder of this Chapter alone, the density levels of each contour will also be superimposed over the contour graph for easier interpretation. Variables u_1 and u_2 are displayed on the x and y axes of the contour plots. For the 3d pdf plot the horizontal axes represent rvs u_1 and u_2 and the y axis represents the density function. Kendall's τ and θ are also presented. Both marginal distributions of u_1 and u_2 are set to Normal$(0,1)$.

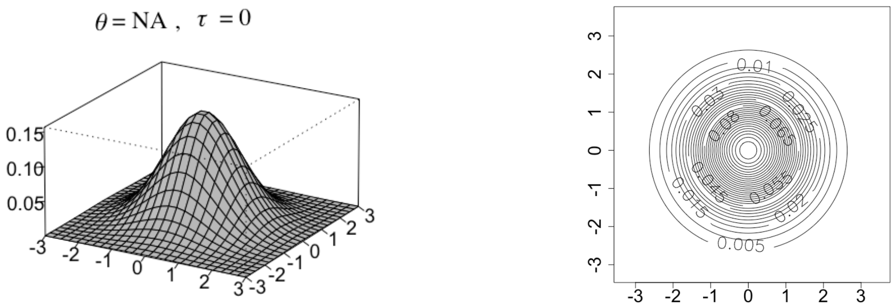
1.4.1 Independent copula

This is the simplest copula and it is also known as the product copula:

$$C(u_1, u_2) = u_1 \cdot u_2$$

This copula has independent marginals and analysing these variables separately will yield the same results as their joint analysis. There is no dependence structure between the rvs of interest, hence the resulting copula has no dependence parameter and Kendall's $\tau = 0$. The bivariate density on the left-hand side of Figure 1-9 peaks at 0 for both variables (pre-set univariate mean for each) and spreads along the x and y axes with $\mathrm{SD} = 1$ (also pre-set). The contour plot on the right-hand side shows the density at selected levels, where the numbers denote the density within each encircled area.

Figure 1-9: Independent Copula



1.4.2 Farlie-Gumbel-Morgenstern copula

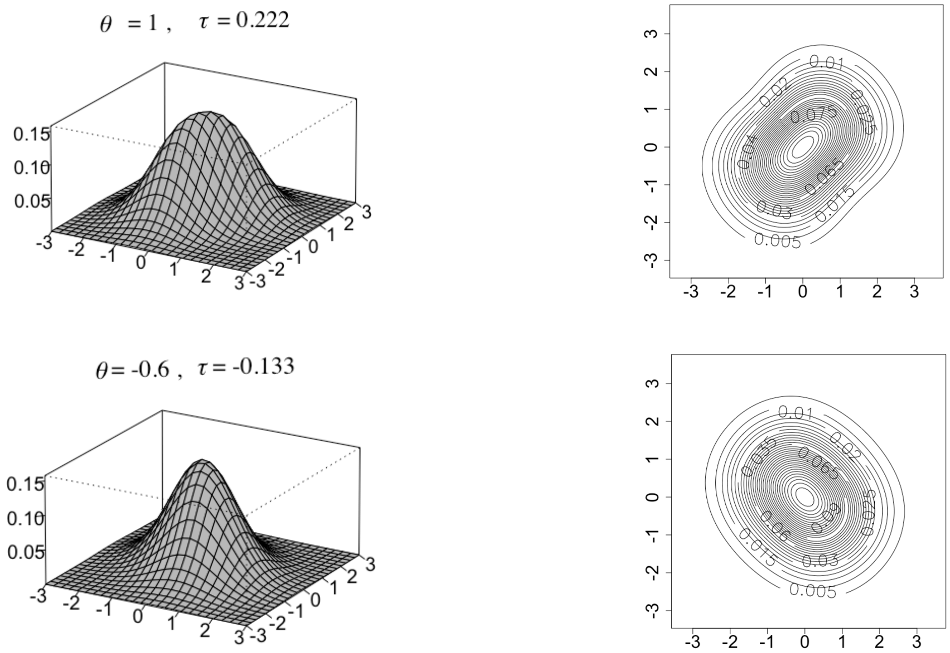
The Farlie-Gumbel-Morgenstern (FGM) copula [36–38] is a computationally simple copula:

$$C(u_1, u_2; \theta) = u_1 u_2 \big(1 + \theta(1 - u_1)(1 - u_2)\big)\ ;\ -1\ \leq \theta \leq 1$$

The association between θ and τ follows from (Eq. 1-2): $\tau = \frac{2\theta}{9}$. When applying this copula to a dataset, an initial value for θ can be easily decided upon by inverting the above formula for an observed τ, i.e. $\theta = 9\tau/2$.

Positive and negative θ values correspond to positive and negative dependence respectively. The independent copula is given by FGM with $\theta =\ 0$. FGM copulas can only model relatively weak dependence as illustrated in Figure 1-10 for $\theta = 1$ and $\theta = -0.6$, which correspond to τ values of $2/9$ and $-1.2/9$.

Figure 1-10: FGM Copula



$\theta\ = 1\ ,\quad \tau = 0.222$

$\theta = -0.6\ ,\quad \tau = -0.133$

1.4.3 Archimedean copulas

The two previous copulas (independent and FGM) are single, stand-alone types of copulas. Archimedean copulas form a particularly popular and important family of copulas[39]. They are easily constructed (e.g. additivity property mentioned below), are capable of capturing a wide range of dependencies (displayed via the three types below) and have convenient statistical properties (listed below). A 2-dimensional copula C is called Archimedean if it has the following property:

$$C(u_1, u_2; \theta) = \varphi^{[-1]}[\varphi(u_1; \theta) + \varphi(u_2; \theta)] \qquad \text{(Eq. 1-3)}$$

Archimedean copulas are characterised by a single function, φ, called the generator function, which is unique to each copula and satisfies the following properties:

- $\varphi: [0,1] \to [0, \infty]$
- $\varphi(0) = \infty$, $\varphi(1) = 0$
- $\varphi^{[-1]}$: pseudo-inverse of φ
- φ is a continuous, additive, strictly decreasing $(\varphi'(t) < 0)$ and convex $(\varphi''(t) \geq 0)$ function

For example, if $\varphi(t) = 1 - t$, for $t \in [0,1]$; then $\varphi^{[-1]}(t) = \max(1 - t, 0)$, from (Eq. 1-3) for any $u_1, u_2 \in [0,1]$: $C(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$, which is the lower Fréchet-Hoeffding bound in the bivariate case (discussed earlier in section 1.3.2). In other words, the lower bound of any copula, if it exists, is an Archimedean copula itself.

One of the main advantages of Archimedean copulas is that the introduction of an extra dimension/rv (in other words, a marginal distribution u_3 describing a 3rd rv) can be done additively, by including the generator $\varphi(u_3; \theta)$:

$$C(u_1, u_2, u_3; \theta) = \varphi[\varphi^{-1}(u_1; \theta) + \varphi^{-1}(u_2; \theta) + \varphi^{-1}(u_3; \theta)]$$

More properties of this family of copulas are:

- $C(u_1, u_2) = C(u_2, u_1),\ \forall\, u_1, u_2 \in [0,1]$ – commutative
- $C(C(u_1, u_2), u_3) = C\big(u_1, C(u_2, u_3)\big),\ \forall\, u_1, u_2, u_3 \in [0,1]$ – associative
- $C(u_1, u_2) \leq C(u_3, u_4),\ \forall\, u_1 \leq u_3\, , u_2 \leq u_4 \in [0,1]$ – order preserving

For any Archimedean copula and some marginal distribution t (defined in $[0,1]$), τ is given by [40]:

$$\tau = 4 \iint_{(0,1)^2} C(u_1, u_2) \partial C(u_1, u_2) - 1 = 1 + 4 \int_0^1 \frac{\varphi(\mathrm{t})}{\varphi'(\mathrm{t})} \partial \mathrm{t} \qquad \textbf{(Eq. 1-4)}$$

In the sections that follow, three Archimedean copulas (Gumbel, Frank and Clayton) are described and examples presented. The generator function differs for each of them as well as the range within which θ is defined.

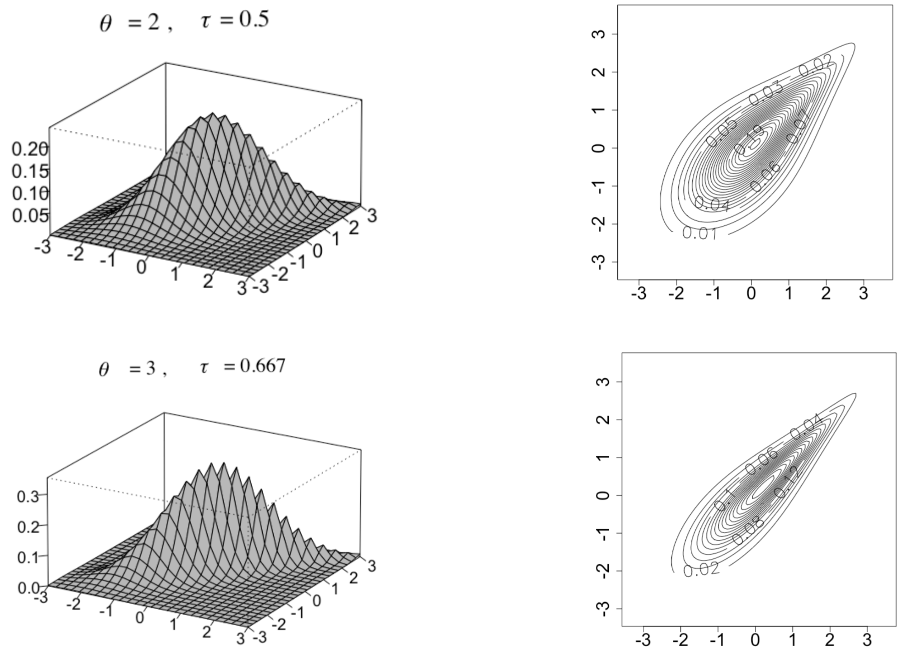
1.4.3.1 Gumbel

For the Gumbel copula [37,41] θ is restricted in $[1, \infty)$. Values of 1 and ∞ correspond, respectively, to independence and Fréchet-Hoeffding upper bound, but this copula does not attain the Fréchet-Hoeffding lower bound for any value of θ (hence cannot cope with negative dependence).

$$C(u_1, u_2; \theta) = \exp\left(-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta\right]^{1/\theta}\right)$$

The generator function of the Gumbel copula [42] is given by $\varphi(t) = (-\ln t)^\theta,\ \ t\ \in (0,1]$, and τ is equal to $1 - \theta^{-1}$. The Gumbel copula ($\theta = 1/(1-\tau)$) does not allow negative dependence ($0 \leq \tau \leq 1$). It can capture skewed associations and it measures more precisely upper than lower tail dependence. It is therefore suited not only for positively correlated rvs, but also for rvs whose high values are more strongly correlated than low values (Figure 1-11).

Figure 1-11: Gumbel Copula



1.4.3.2 Frank

For the Frank copula [43] θ may take any real value $(-\infty, \infty)$ apart from 0. Values of $-\infty$ and +∞ correspond to the lower and upper bounds, respectively. The Frank copula function is given by:

$$C(u_1, u_2; \theta) = -\frac{1}{\theta} \ln\left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right)$$

It is constructed using the following generator function φ:

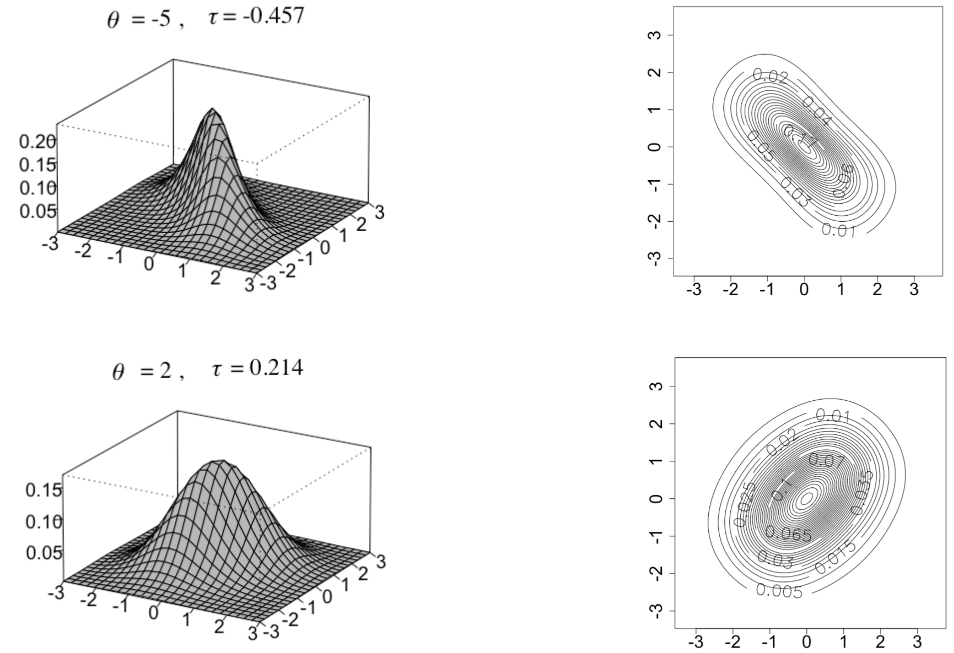
$$\varphi(t) = -\ln\left(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right), \text{ where } t \in [0, \infty)$$

Kendall's τ for the Frank generator is given below:

$$\tau = 1 - \frac{4}{\theta}\left[\frac{1}{\theta}\int_0^\theta \frac{t}{e^t - 1}\partial t + \frac{\theta - 2}{2}\right]$$

For this copula, there is no exact solution for θ in terms of τ. The Frank copula permits negative dependence between marginals, dependence is symmetric on both tails and both its lower and upper bounds can be reached. However, it has been noted that tail dependence tends to be relatively weak and the strongest dependence is centred in the middle of the distribution. In conclusion, Frank copulas are suitable when either strong negative or positive association is observed with weak tail dependence. Notice the subtle difference between the first graph of Figure 1-10 and the last graph of Figure 1-12, whose τ's are almost identical.

Figure 1-12: Frank Copula


$\theta = -5$, $\tau = -0.457$

$\theta = 2$, $\tau = 0.214$

1.4.3.3 Clayton copula

The function of the Clayton copula [44], also referred to as Cook and Johnson [45], was originally studied by Kimaldorf et al [46,47] and is given below:

$$C(u_1, u_2) = \left[\max\left(u_1{}^{-\theta} + u_2{}^{-\theta} - 1{,}0\right)\right]^{-1/\theta}$$

The dependence parameter, θ, is restricted in the region $[-1, \infty)$, whilst excluding 0. As θ approaches zero, the marginals become independent. As θ approaches infinity, the copula attains its upper bound. The distribution tends to the lower Fréchet-Hoeffding bound as θ approaches 1, but does not attain it for no value of θ (i.e. not comprehensive).

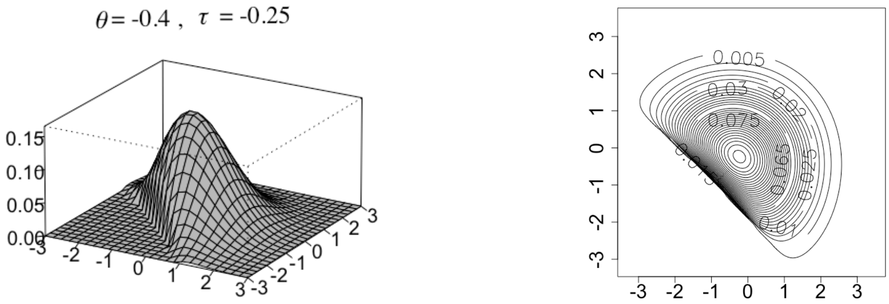
The Clayton copula can account for negative dependence (Figure 1-13). It generates asymmetric dependence and lower tail dependence, but relatively weak upper tail dependence.

Its generator function and Kendall's τ are given, respectively, by:

$$\varphi(t) = \frac{1}{\theta}\left(t^{-\theta} - 1\right) \text{ and } \tau = \frac{\theta}{\theta + 2}$$

The simplicity of the relationship between τ and θ for the Clayton copula is a great example of scenarios where a reliable initial value for θ can be obtained directly from the data when trying to identify the best copula.

Figure 1-13: Clayton copula



1.4.4 Elliptical copulas

Ellipses are curves such that the sum of the distances from two fixed points (called foci) for every point on the curve is constant. Elliptical distributions have contour shapes of ellipses. Two well-known examples of elliptical distributions are the bivariate Normal and t.

Elliptically-contoured distributions were introduced by Kelker [48] and widely discussed by Fang [49]. Let Ψ_m be a class of functions $\psi(t)\colon [0, \infty] \rightarrow \mathrm{R}$, where R represents a set of real numbers, such that function $\psi(\sum_{i=1}^{m} t_i^2)$ is an m-dimensional characteristic function[*] for all $t \in \mathrm{R}^m$ [50]. A rv Y has a m-multivariate elliptical distribution, written as $Y{\sim}E_m(\mu, \Sigma, \psi)$ or $Y{\sim}EC_m(\mu, \Sigma, \psi)$, if its characteristic function can be expressed as:

$$\varphi_Y(t) = \exp(it'\mu)\psi\left(\frac{1}{2}t'\Sigma t\right)$$

for some m-long column-vector μ, $m \times m$ positive definite matrix Σ and for some function $\psi(t) \in \Psi_m$ which is called the characteristic generator. The parameter μ is a location parameter and the Σ matrix determines the scale and the correlation of the rvs.

If $Y{\sim}E_m(\mu, \Sigma, \psi)$ exists, then Y has a density $f_Y(y)$ that takes the following form:

$$f_Y(y) = \frac{c_m}{\sqrt{|\Sigma|}} g_m\left(\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

Where c_m is a Normalising constant and $g_m(\cdot)$ is called the density generator [49].

For two components, Y_1 and Y_2, of an m-dimensional rv Y it has been shown [51,52] that Kendall's τ is given by:

$$\tau_{12} = \frac{2}{\pi} \arcsin r_{12}$$

[*] The characteristic function of any real-valued rv completely defines its probability distribution and determines the variable's behaviour and properties. It provides an alternative route to analytical results compared to working directly with pdfs or cdfs.

Hence, Kendall's τ only depends on Pearson's r and neither the characteristics generator nor the shape of the distribution affects the rank correlation [53].

The Normal and t distributions, which comply with the elliptical distribution pdf seen above, are examples of elliptical distributions, with respective density generators and Normalising constants given below:

$$g(u) = \exp(-u)\ ; \quad c_m = (2\pi)^{-\frac{m}{2}},$$

$$g_m(u) = \left(1 + \frac{u}{k_p}\right)^{-p}\ ; \quad c_m = \frac{\Gamma(p)}{\Gamma\left(p - \frac{m}{2}\right)}(2\pi k_p)^{-m/2},$$

Where the parameter $p > m/2$ and k_p is some constant that may depend on p. For example, in the bivariate case where $m = 2$ and df is the degrees of freedom, assuming $p = (\mathrm{df} + 2)/2$ and $k_p = \mathrm{df}/2$, the resulting distribution is the multivariate student t (identical to the one seen in Section 1.1.1.2).

Elliptical copulas describe such joint distribution functions. More specifically, if F is the cdf of an elliptical distribution and F^{-1} its inverse function, the elliptical copulas determined by F are:

$$C(u_1, u_2) = F[\mathrm{F}^{-1}(u_1), \mathrm{F}^{-1}(u_2)]$$

The corresponding Normal and t copulas are studied in the next two sections. These copulas are appropriate when radial symmetry (i.e. equivalent upper and lower tail dependence) is evident between the two random variables of interest. This constraint is not necessarily present in the more flexible Archimedean copulas. Both copulas have a correlation matrix, inherited from the elliptical distributions, which determines the dependence structure between the rvs, hence related to θ.

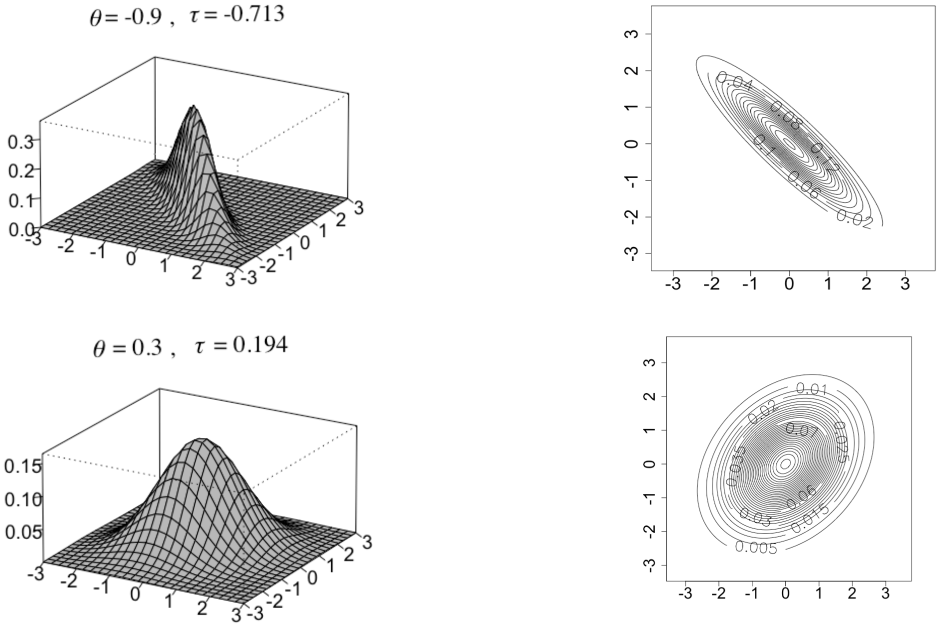
1.4.4.1 Gaussian (Normal) copula

The form of the (bivariate) Normal copula [54] is given by:

$$C(u_1, u_2; \theta) = \Phi\left(F_1{}^{-1}(u_1), F_2{}^{-1}(u_2); \theta\right)$$

Where Φ is the cdf for the bivariate standard Normal distribution, $N(0,1)$ and $F_1{}^{-1}(u_1)$, $F_2{}^{-1}(u_2)$ are the inverse of the two univariate marginals. In this copula case, the parameter θ is equal to $r.$ If the marginals are Normally distributed and the Normal copula is used to describe the association, the resulting joint distribution is multivariate Normal (Figure 1-14). This copula assigns equal degrees of positive and negative dependence to the joint distribution whilst assuming no tail dependence. As the θ approaches -1 and 1, the Normal copula attains the Fréchet lower and upper bound, respectively.

Figure 1-14: Normal Elliptical Copula



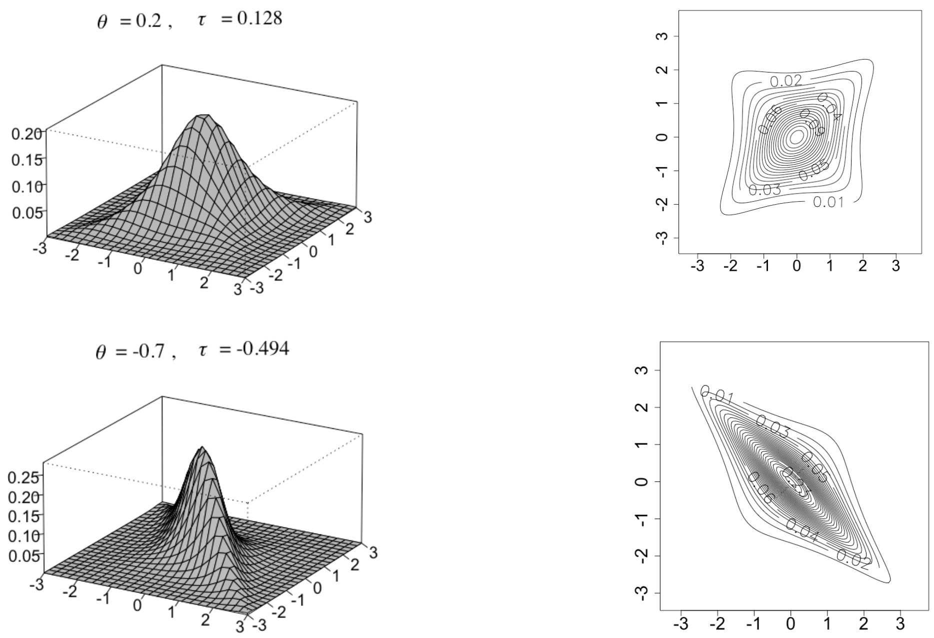
1.4.4.2 Student's t copula

The copula parameter θ is a vector of two parameters, r and df (degrees of freedom), where the latter controls the heaviness of the tails [49]. Its function is given by:

$$C(u_1, u_2; \theta) = T_{\mathrm{df}}\left(F_1{}^{-1}(u_1), F_2{}^{-1}(u_2); \theta\right)$$

where T_{df} and $F_i{}^{-1}(u_i)(i = 1{,}2)$ are the bivariate Student's t cdf and the inverse of the univariate marginals, respectively, with $\theta = r$. As df increases the Student's t copula converges to the Gaussian copula, hence decreasing the probability of tail events. This copula has more points in the tails than the Normal one and a star-like shape. Figure 1-15 shows examples of varying r (labelled as θ) with $\mathrm{df} = 2$ (called Cauchy copula [55]).

Figure 1-15: *t* Elliptical Copulas (df=2)



1.4.5 Extreme value copulas

In cases where researchers' interest is in joint extreme events, extreme value copulas can be a good choice for modelling of the dependence structure between exceptional events [56–58]. Extreme value copulas' biggest advantage is the fact that they are not symmetric and can, of course, account for tail dependence.

A bivariate copula $C(u_1, u_2)$ is an extreme value copula if there exists a copula C_O such that:

$$C(u_1, u_2) = \lim_{n\to\infty} C_O(u_1^{1/n}, u_2^{1/n})^n$$

In other words, the family of extreme-value copulas arises at the limits of ordinary copulas as the sample size n tends to infinity.

Alternatively, a copula $C(u_1, u_2)$ is an extreme value copula if and only if:

$$\begin{aligned} C(u_1, u_2) &= \exp\left(\ln(u_1 u_2)\, A\left(\frac{\ln u_1}{\ln(u_1 u_2)}; \theta\right)\right) \\ &= (u_1 u_2)^{A\left(\frac{\ln u_1}{\ln(u_1 u_2)}; \theta\right)} \end{aligned}$$

for $(u_1, u_2) \in [0,1]^2$ where $A: [0,1] \to [\frac{1}{2}, 1]$ is an appropriately chosen convex function, which satisfies the following conditions: $A(0) = A(1) = 1,\ \max(t, 1-t) \leq A(t) \leq 1$ for all $t \in [0,1]$. A is Pickand's dependence function [59]. The upper bound of function A corresponds to the independent copula, section 1.4.1 and the lower bound corresponds to the comonotone copula, $C(u_1, u_2) = \max(u_1, u_2)$.

If $A(t; \theta) = \left(t^\theta + (1-t)^\theta\right)^{1/\theta}, \theta \geq 1$, then the equation above yields the Gumbel copula. Hence, the Gumbel copula as seen earlier in section 1.4.3.1 is the only

Archimedean copula that is also an extreme value copula. There are no other Archimedean copulas with this property [42].

Kendall's τ and Spearman's ρ (unless there is independence) are given by:

$$\tau = 4\int_0^1 \frac{t(1-t)}{A(t)}\partial A'(t)$$

$$\rho = 12\int_0^1 \frac{1}{(1+A(t))^2}\partial t - 3$$

In the context of extremes, it is natural to also study the coefficient of upper and lower tail dependence.

$$\begin{aligned}\lambda_U &= \lim_{u\uparrow 1} \Pr(U > u|V > v)\\ &= 2\left(1 - A\left(\frac{1}{2}\right)\right)\\ \lambda_L &= \lim_{u\downarrow 0} \Pr(U \le u|V \le v)\\ &= \lim_{u\downarrow 0} u^{(2A(1/2)-1)}\end{aligned}$$

1.4.5.1 t-EV copula

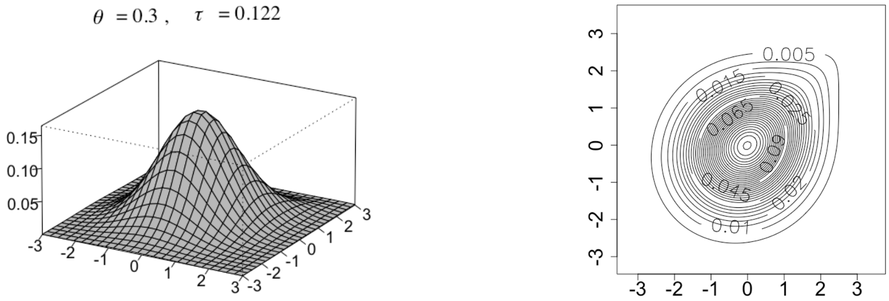
In section 1.4.4.2, the bivariate t-copula was explored. It is shown [56] that this copula also falls in the domain of the extreme-value copula C with Pickand's dependence function A equal to:

$$A(w) = wt_{\mathrm{df}+1}(z_w) + (1-w)t_{\mathrm{df}+1}(z_{1-w})$$

$$z_w = (1+\mathrm{df})^{1/2}\left[\left\{\frac{w}{1-w}\right\}^{1/\mathrm{df}} - r\right](1-r^2)^{-1/2}$$

where $w \in [0,1]$ and t_{df} represents the distribution function of the univariate t-distribution with df degrees of freedom (Figure 1-16).

Figure 1-16: t-extreme value copula



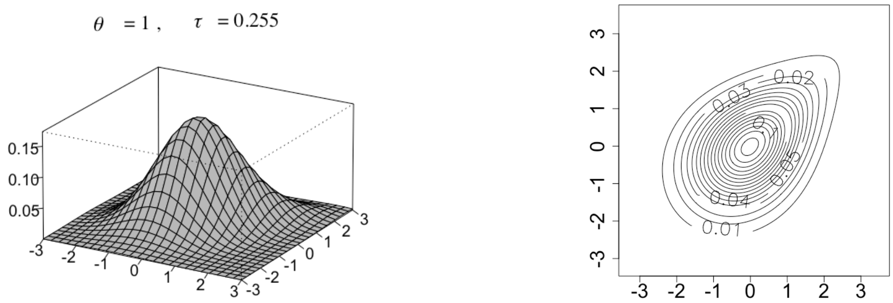
1.4.5.2 Hüsler-Reiss copula

The Hüsler-Reiss [56,60] copula C (Figure 1-17) is the bivariate extreme copula with Pickand's dependence function:

$$A(w) = (1 - w) \cdot \Phi\left(\lambda + \frac{1}{2\lambda}\ln\frac{1 - w}{w}\right) + w \cdot \Phi\left(\lambda + \frac{1}{2\lambda}\ln\frac{1 - w}{w}\right)$$

for $w \in [0,1]$, with Φ representing the standard Normal cumulative distribution function. The parameter λ measures the degree of dependence, going from independence ($\lambda \to \infty$) to complete dependence ($\lambda = 0$).

Figure 1-17: Hüsler-Reiss copula



1.5 Conclusions

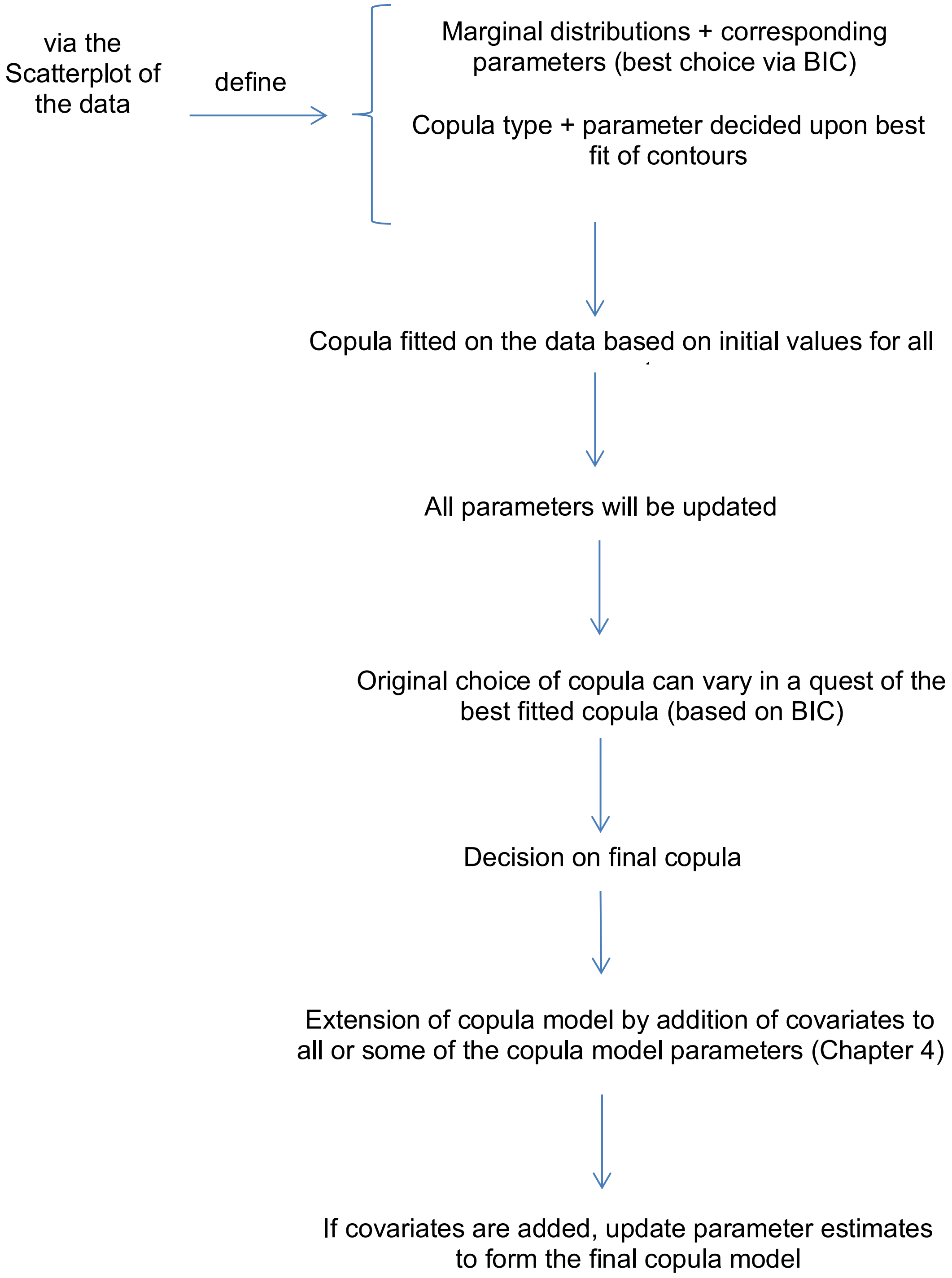
The review provided in this Chapter outlines the fundamental aspects of bivariate associations and forms the basis of the rest of the thesis. Chapter 2 will provide information on several alternative methods that can be used to investigate bivariate associations in a more local scale, i.e. in smaller neighbourhoods around bivariate points.

The flowchart in Figure 1-18 aims to provide a diagrammatical representation and summary of copula analysis as described in the earlier sections of this Chapter.

Table 1-1 summarises some of the dependence characteristics of the copulas seen thus far as well as some new ones, not explicitly defined here.

Copula models will be used in the coming Chapters of this thesis to enhance the exploration of local dependence as well as of the overall association pattern under investigation.

Figure 1-18: Summative flowchart of copula modelling



via the Scatterplot of the data

define

Marginal distributions + corresponding parameters (best choice via BIC)

Copula type + parameter decided upon best fit of contours

Copula fitted on the data based on initial values for all

All parameters will be updated

Original choice of copula can vary in a quest of the best fitted copula (based on BIC)

Decision on final copula

Extension of copula model by addition of covariates to all or some of the copula model parameters (Chapter 4)

If covariates are added, update parameter estimates to form the final copula model

Table 1-1: Summary table of two dimensional copulas

Copula	θ range	Kendall's τ	τ range	Dependence structure	Tail dependence
Independent	NA	0	NA	Independent	NA
FGM	$[-1,1]$	$\theta(2/9)$	$[-0.22,0.22]$	Symmetric Weak overall	Weak/Independent
	Archimedean family				
Frank	$(-\infty,\infty)$ excl 0	$1-\frac{4}{\theta}\left[1-\frac{1}{\theta}\int_0^\theta \frac{t}{e^t-1}\partial t\right]$	$[-1,1]$ excl 0	Symmetric Strong centre	Weak/Independent
Gumbel	$[1,\infty)$	$1-\theta^{-1}$	$[0,1]$	Asymmetric	Upper
Clayton	$[-1,\infty)$ excl 0	$\theta(\theta+2)^{-1}$	$(-1,1)$ excl 0	Asymmetric	Lower

Ali-Mikhail-Haq*	$[-1,1]$	$1+2\left[\frac{\frac{-1}{6\theta}-[(\theta-1)^2\ln(1-\theta)]}{3\theta^2}\right]$	$[-0.18,0.33]$	Asymmetric	Lower
Joe*	$[1,\infty)$	$1+\frac{4}{\theta}\int_{t=0}^{1}\frac{(\ln(1-t^\theta))(1-t^\theta)}{t^{\theta-1}}$	$(0,1]$	Asymmetric	Upper
	Elliptical family				
Normal	$[-1,1]$	$(2/\pi)\arcsin\theta$	$[-1,1]$	Symmetric	Weak/Independent
t	$[-1,1]$	$(2/\pi)\arcsin\theta$	$[-1,1]$	Symmetric	Upper and lower
	Extreme value family				
t-EV	$[-1,1)$	no closed form	$[0,1]$	Asymmetric	Upper
Hüsler-Reiss	$[0,\infty)$	no closed form	$[0,1]$	Asymmetric	Upper
* not presented in detail in this thesis					

2. Local dependence

The three scalar measures of dependence described in Chapter 1 use one (linear, rank correlation) or two (tail dependence) numerical values to summarise the association of two rvs, thought of as constant along the range of their bivariate relationship.

However, a single scalar dependence measure will not always reflect the dependence between a pair of continuous variables and will not convey the true dependence structure [61]. When the dependency structure is not constant (the most common scenario), coefficients that evaluate dependence locally, i.e. in smaller areas across the ranges of the rvs, should be used instead. The tail dependence coefficient partially addresses this by concentrating on the tails of the bivariate association but remains quite limited.

The following sections focus on 2 local dependence functions: the local dependence map and the chi-plot [62,63]. Details of a specific application with Beta marginals of these measures is explored in section 2.3.

2.1 Local dependence function and local dependence map

Local dependence (LD) [64] measures the correlation between Y_1 and Y_2 in a neighbourhood of any point (y_1, y_2) in the domain of the bivariate density function.

The local dependence function (LDF), γ was introduced by Holland and Wang and focuses on the association of Y_1 and Y_2 in smaller areas of their range (localisation) rather than their entire range, i.e. global dependence structure. Pearson's correlation coefficient, as seen earlier, is defined as follows:

$$
\begin{aligned}
r_{12} &= \frac{\mathrm{Cov}(y_1, y_2)}{\sigma_1 \sigma_2} \\
&= \frac{E(Y_1 Y_2) - E(Y_1)E(Y_2)}{\left(E\left({Y_1}^2\right) - E(Y_1)^2\right)^{1/2} \left(E\left({Y_2}^2\right) - E(Y_2)^2\right)^{1/2}}
\end{aligned}
$$

Localisation can be achieved with the use of kernel methods [65], where the correlation of Y_1 and Y_2 is calculated conditional on Y_1 and Y_2 being in the neighbourhood of a point (y_{10}, y_{20}). So, the formula above changes as follows to incorporate the indicator function $w_o(Y_1, Y_2)$:

$$
\begin{aligned}
&r_{12}(y_{10}, y_{20}, h_1, h_2) = \\
&= \frac{E(w_o(Y_1, Y_2)Y_1 Y_2) - E(w_o(Y_1, Y_2)Y_1)E(w_o(Y_1, Y_2)Y_2)}{\left(E\left(w_o(Y_1, Y_2){Y_1}^2\right) - E(w_o(Y_1, Y_2)Y_1)^2\right)^{\frac{1}{2}} \left(E\left(w_o(Y_1, Y_2){Y_2}^2\right) - E(w_o(Y_1, Y_2)Y_2)^2\right)^{\frac{1}{2}}}
\end{aligned}
$$

Where

$$
w_o(Y_1, Y_2) = \begin{cases} 1, & \text{if } (Y_1, Y_2) \in [y_{10} \pm h_1, y_{20} \pm h_2] \\ 0, & \text{otherwise} \end{cases}
$$

is a weight function and h_1, h_2 are smoothing parameters, $h_1 \to 0, h_2 \to 0$. If $w_o(Y_1, Y_2) = 1$ for all h_1 and h_2, the conventional equation of the Pearson's r is recovered.

Applying the above notion of localisation on all possible pairs of (y_{10}, y_{20}) and following the proof provided by Jones [65], the LDF is given by:

$$
\gamma(y_1, y_2) = \frac{\partial^2 \ln f_{12}(y_1, y_2)}{\partial y_1 \partial y_2}
$$

$$= \frac{1}{f_{12}(y_1, y_2)} \left\{ f_{12}^{11}(y_1, y_2) - \frac{f_{12}^{10}(y_1, y_2)\, f_{12}^{01}(y_1, y_2)}{f_{12}(y_1, y_2)} \right\}$$

Where $f^{ij}(y_1, y_2) = \frac{\partial f(y_1,y_2)}{\partial y_1^i \partial y_2^j}$.

The interpretation of positive and negative values of γ correspond to positive and negative dependence in the same way as positive and negative values of Pearson's r do. Values equal to 0 correspond to global (i.e. across all values) independence between y_1 and y_2.

In the case of the bivariate Normal distribution, a global association measure provides a very good estimation of the overall dependency between the rvs; hence the local dependence function would be constant. In fact, the local dependence function is constant if and only if the conditional distribution has an exponential family with its canonical parameter being a linear function [66]. In the bivariate Normal case the LDF is given by:

$$\gamma(y_1, y_2) = \frac{r}{(1 - r^2)}$$

As per Jones [65] *"the local dependence function can be used to show how dependency can be measured when both the degree and the direction of the dependence is different in different regions of the plane"*.

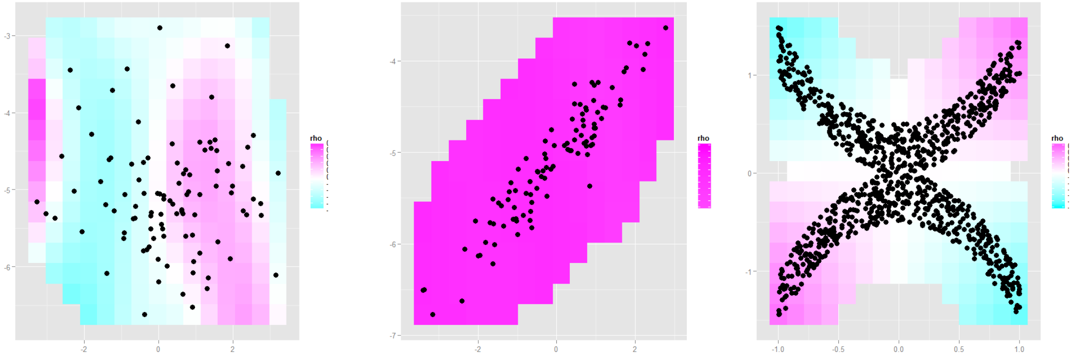
However, the LDF local dependence function can provide too detailed exploration of the association, which is contrary to the scalar coefficients that often average out too much of the dependence structure, hence neither of them is ideal.

Local dependence maps [67] are a compromise between the two; simplifying the estimated local dependence structure by identifying regions of (significant) positive, (non-significant) zero and (significant) negative local dependence. In essence, the local dependence map provides permutational tests of significance, i.e. tests based on permutational arguments as opposed to asymptotic distributional assumptions.

The graphs in Figure 2-1 present the LD map of 3 simulated example data sets (using the `localgauss R` library [68], version 0.35). The palette of colours presented on the side of each graph indicates varying levels of local dependence as estimated by the Pearson's correlation coefficient.

Figure 2-1: Local dependence maps

Black dots represent simulated data for three different association patterns



The graphical representation of the LDF is very informative and an improvement to the scalar coefficients. Association is evaluated at small neighbourhoods of pairs of (y_1, y_2) values and this can vary from strong/weak negative to strong/weak positive correlation. The last example of Figure 2-1 is a very good representation of the change in direction and strength of the correlation. The association at the central part of the scatterplot is virtually non-existent whilst this changes towards each of the four tails. However, it is essentially a tool only for measuring associations and extensions to more than two dimensions and/or incorporation of additional predictor variables would be challenging. Finally, the marginal distributions of the rvs have not been accounted for and they could potentially add a great deal of assistance in the exploration of the association.

2.2 Chi-plots

The chi-plot [62,63] is a data-driven transformation for bivariate observations which consists of plotting a rank-based measure of local dependence versus a function of the distance between each point and the median-centre of the dataset.

The chi-plot is a rank-based graphical tool and has characteristic patterns depending on whether the variables are independent, have some degree of monotone relationship or have more complex dependence structure. It is also a well-suited tool for identifying dependencies in the tails of bivariate distributions.

The chi-plot is a scatterplot of pairs (λ_i, χ_i); λ_i is a measure of the distance of point (y_{1i}, y_{2i}) from the centre of the data set (i.e. vector of the two medians). A positive λ_i value means that both Y_1 and Y_2 are large relative to their respective medians (or both small), i.e. positively correlated, whereas a negative value corresponds to y_{1i} and y_{2i} being on opposite sides of their respective medians, i.e. negatively correlated. When the data are a random bivariate sample from independent continuous marginals, then $\lambda_i \sim U[-1,1]$.

χ_i is a correlation measurement between the dichotomised Y_1 and Y_2 values at point (y_{1i}, y_{2i}), i.e. below-above/lower-greater than this point as evaluated from multiple directions, and its interpretation reduces to the local Pearson correlation coefficient. It is equal to 1 (-1) for all sample cut points when Y_2 is a strictly increasing (decreasing) function of Y_1. It is similar to the χ^2 statistic for testing independence in the 2×2 table generated by the cut-point (y_{1i}, y_{2i}). In other words, χ_i approximates the failure of the bivariate distribution function to factorise into a product of marginal distribution functions at the sample argument (y_{1i}, y_{2i}). It is equal to a scaled transformation of the difference $H_i - F_i G_i$, where H, F and G are the empirical bivariate and univariate distributions of y_1 and y_2, respectively. Under independence, $H_i = F_i G_i$, hence $\chi_i = 0$.

All values of λ_i and χ_i are based on ranked values of the data. Also, they lie in the interval $[-1,1]$, hence the chi-plot is drawn on a $[-1,1] \times [-1,1]$ plane. Points for which $|\lambda_i| > 4\left(\frac{1}{n-1} - 0.5\right)^2$ are not plotted on the chi-plot [63] hence only non-extreme values are displayed. The plot is approximately horizontal under independence $\left(\chi_i \sim N\left(0, \frac{1}{n}\right) \text{and } \lambda_i \sim U[-1,1]\right)$. In other words, the chi-plot measures dependence locally and draws it against the distance of the data point to the data centre.

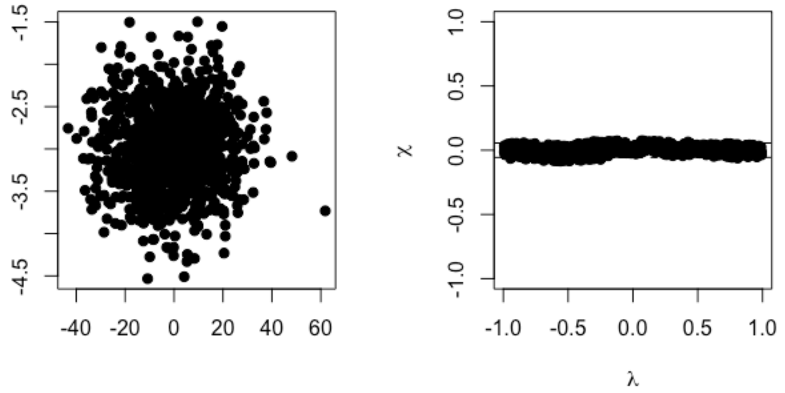
The chi-plot can be interpreted depending on the area its points are scattered amongst 5 possible sections of the graph between λ_i and χ_i, i.e. horizontal line, left and right areas above and below the horizontal line. Figure 2-2 contains examples of simulated data sets with varying dependence structures that aim to show the change in the patterns of the chi-plot in each of the aforementioned 5 areas. Each pair of graphs shows the scatterplot of the simulated data on the left and the corresponding chi-plot on the right:

- Figure 2-2 (i): the horizontal line → this is the line of independence and is usually plotted on the graph along with the “control limits” which define (non-parametrically, i.e. via Monte-Carlo simulations) the range within which 95% of the observation lie under independence. Any scatter around the horizontal line and its 95% region is due to sample variability, while deviations from the horizontal line correspond to positive and negative departures from independence.

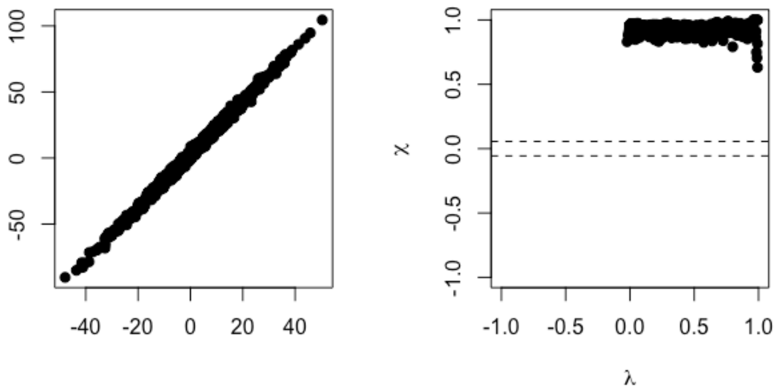
- Figure 2-2 (ii): positive χ and positive λ values → positive dependence ($\chi > 0$) in the lower left and upper right corner and $\lambda > 0$, i.e. same direction of distance from the bivariate centre for both rvs, i.e. all pairs of Y_1 and Y_2 are both above or below the median; positive association is uniform throughout.
- Figure 2-2 (iii): negative χ and negative λ values → indicate negative dependence ($\chi < 0$) in the upper left and lower right corner of the data where Y_1 is high/low and Y_2 low/high, respectively. Negative association is evident for the most part of this example with some deviations for those pairs of values in the middle of the graph where both Y_1 and Y_2 are on the same side of the median
- Figure 2-2 (iv) shows the chi-plot of a U-shape where a mixture of types of associations are displayed and are colour coordinated to show the direct correspondence of the points on the scatterplot and the chi-plot, summarising all the points relating to the different types of association described above.

Figure 2-2: Chi-plot simulated examples

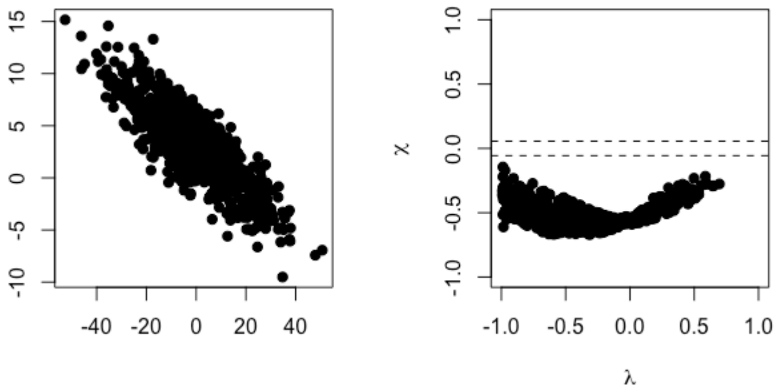
(i) Independent association



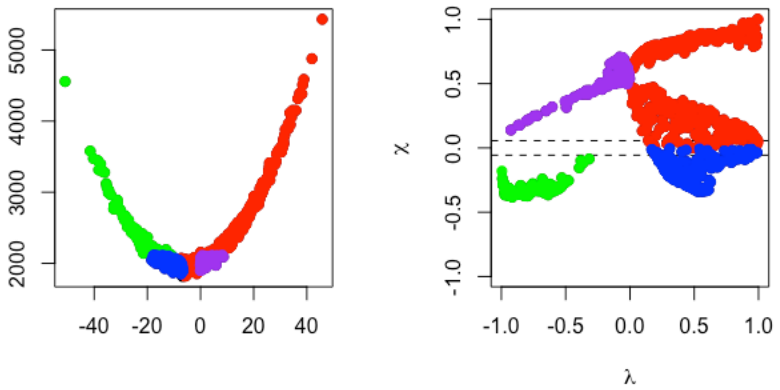
(ii) Positive linear association



(iii) Negative linear association



(iv) Quadratic association



Summary

This section has provided an overview of a variety of methods that enable the exploration of bivariate associations in more detail compared to a scalar coefficient. The LDF is a localised version of the correlation coefficient and via appropriate maps it provides a meaningful alternative to scalar measures. The chi plot is based on transformations of the data (and their ranks) and its ultimate goal is to provide a more explicit outcome regarding the association between rvs.

There is not one single best graph that can quantify every single element of dependence between rvs in its entirety. The dependency plots shown may prove useful in deciding, for example, the parameter of the copula function and/or the copula type itself. In the remaining section of this Chapter, the dependency measures will be applied to simulated data from a variety of copula types with specified marginals and in Chapter 5 they will be utilised in the modelling of characteristics of live births in Mexico.

2.3 Local dependence in bivariate copula models with Beta marginals

Bivariate Beta models provide an interesting framework in which to explore the role of the LDF in revealing bivariate structures between rvs bounded in $[0,1]$, since the Beta distribution can produce probability density functions (pdfs) in many shapes – U- and J-shaped, symmetric, and even uniform. There are many fields of application for such joint models, typically involving proportions, e.g. mathematics and language exam marks of students (proportions correct), the percentiles of height and weight, or the proportions of household income spent on food and heating.

There are several models for bivariate Beta distributions: some are derived from transformations of three standard [69], non-central and five [70] Gamma-distributed rvs; others arise from the relations between the Beta, F and skew-t distributions [71,72]. Transformations of Gamma densities impose constraints on the data-generating mechanism and when such constraints are not desired alternative processes are required. Such alternative process is provided by the class of bivariate distributions with Beta marginals constructed via copula functions.

Some work on this has been done by Gupta [73] where the LDF formula for the FGM and AMH copulae with Beta marginals are presented. I did not locate equivalent results for other copula functions, therefore I further extended these results by working out the expressions of the LDF for 3 additional (to the FGM and AMH) bivariate copulas with Beta marginals, Frank, Gumbel and Joe. These results were published in 2017 in a peer-reviewed journal [74] found in Appendix 3: Publications. The paper also includes an application of the resulting LDF expressions on student exam marks. More specifically, this involved joint modelling of the marks of students from a theoretical set of statistics questions and the marks from a statistics task performed on the Statistics Package for Social Science (SPSS) (this application is not presented here). The computational software Mathematica (version 10) [75] and the MathStatica extension [76] were used to derive the analytical expressions of the LDFs. Each of the following sections is accompanied by illustrations of the copula density, the LDF and chi plot with varying parameters.

2.3.1 Copula-defined bivariate distributions with Beta marginals

Let Y_1, Y_2 be univariate random variables each with a univariate Beta distribution with shape parameters $a_i, b_i \geq 0$, $i = 1,2$ respectively:

$$f_i(x_i) = \frac{{x_i}^{a_i - 1}(1-x_i)^{b_i - 1}}{B(a_i, b_i)} \text{ and } F_i(x_i) = \frac{B(x_i, a_i, b_i)}{B(a_i, b_i)}$$

Where $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1} \partial t$ denotes the incomplete beta function, and $B(a, b) = B(1, a, b)$. The survival function is denoted as $\bar{F}(x) = 1 - F(x)$.

In each copula section below, the closed form of the LDF is presented along with the contours of the copula density function (on the left), the LDF map and the chi plot for different marginal and dependence parameters. The parameters of the bivariate distribution (a_i, b_i and θ) were randomly chosen and differ between each copula for the first pair of graphs (Figure 2-3, Figure 2-5 and Figure 2-7), but are the same for the second set of graphs (Figure 2-4, Figure 2-6 and Figure 2-8). The figures illustrate how the same copula function can lead to markedly different bivariate structures for varying marginal and dependence parameters as well as how the same parameters can lead to markedly different bivariate associations for differing copula functions. In these examples, the contours have been drawn at the following density levels: 0, 0.2, 1, 2, 3, 4, 5, 10, 15 and 100. These represent density levels on the left-hand side graph and local dependence levels on the right. For both graphs, the x and y axes represent the two rvs, Y_1 and Y_2 respectively.

2.3.1.1 Frank copula

The density of the Frank copula (introduced in Section 1.4.3.2) with Beta marginals can be written as follows and this corresponds to a five-parameter family of bivariate distributions with Beta marginals (for $i = 1, 2$), where $\theta \in [-\infty, +\infty, \backslash\{0\}]$:

$$f(y_1, y_2) = \theta\left(e^{\theta} - 1\right)\left[\prod_i f_i(y_i; a_i, b_i)\right] \frac{e^{\theta\left(1+\sum F_i(y_i)\right)}}{\left(\sum_i e^{\theta\left(1+F_i(y_i)\right)} - e^{\theta \sum F_i(y_i)} - e^{\theta}\right)^2}$$

The LDF of this copula is equal to:

$$\gamma(y_1, y_2) = 2\theta^2\left(e^{\theta} - 1\right)\left[\prod_i f_i(y_i; a_i, b_i)\right] \frac{e^{\theta\left(1+\sum F_i(y_i)\right)}}{\left(\sum_i e^{\theta\left(1+F_i(x_y)\right)} - e^{\theta \sum F_i(y_i)} - e^{\theta}\right)^2}$$

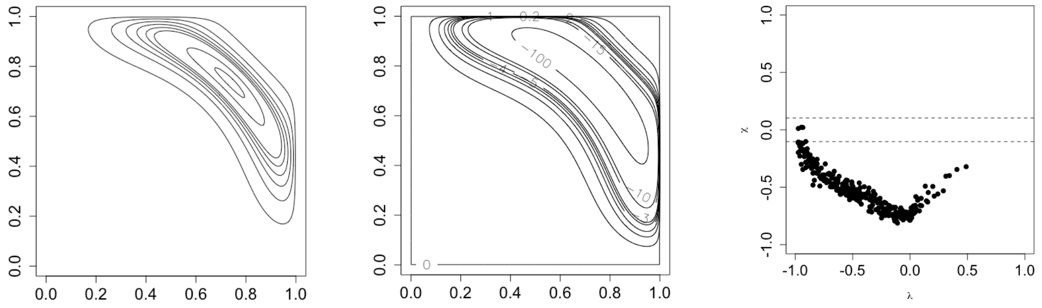
$$= 2\theta f(y_1, y_2)$$

Notice that the latter has the same sign as θ all over the unit square.

Figure 2-3 presents the pdf and LDF contours of a Frank copula (notice the different shape of the Frank copula contours compared to Figure 1-12 driven by different marginals/parameters) and the chi plot.

Figure 2-3: Frank copula PDF and LDF

Copula Parameters: $a_1 = 5, b_1 = 2, a_2 = 5, b_2 = 2, \theta = -12$

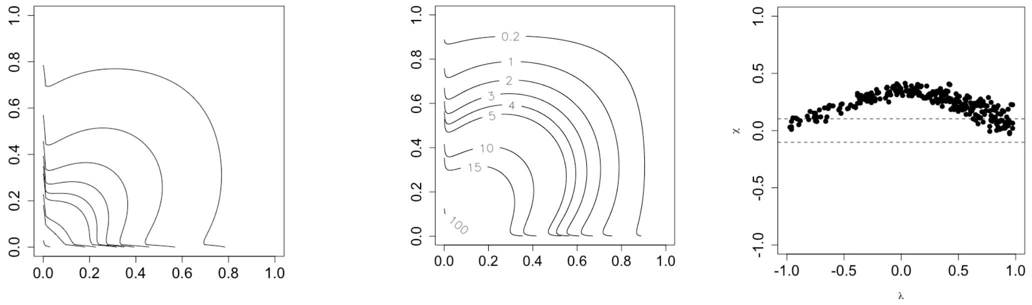


Both the pdf and the LDF follow the same general pattern with the only difference being the fact that the local dependence function expands more widely to accommodate for pairs of x_1 and x_2 that have low bivariate density but are locally associated with respect to the whole range of values. Notice the negative values of the LDF contours, as a result of the negative dependence parameter.

The following set of graphs (Figure 2-4) show a Frank copula with different parameters. Notice that the general shape of the LDF is again very similar to that of the pdf, i.e. local dependence values mirror very well the density values around the same areas of the bivariate relationship.

Figure 2-4: Frank copula PDF and LDF (same parameters, different copula)

Copula Parameters: $\boldsymbol{a_1 = 0.8, b_1 = 3, a_2 = 0.8, b_2 = 3, \theta = 3}$



2.3.1.2 Gumbel copula

The Gumbel copula is defined as follows, where $\theta \in [1, \infty]$:

$$F(x_1, x_2) = \exp\left(-\left[(-\ln F_1(x_1))^{\theta} + (-\ln F_2(x_2))^{\theta}\right]^{1/\theta}\right)$$

The density of the copula when the two marginals are Beta distributions and its LDF can be written as (for $i = 1{,}2$):

$$f(x_1, x_2) = \prod_i f_i(x_i; a_i, b_i) \prod_i F_i^{-1}(x_i; a_i, b_i) \prod_i {L_{x_i}}^{\theta-1}$$

$$\cdot \left(\theta - 1 + \sum_i {L_{x_i}}^{\theta}\right)^{\frac{1}{\theta}} \left(\sum_i {L_{x_i}}^{\theta}\right)^{\frac{1}{\theta}-2} \exp\left(-\sum_i {L_{x_i}}^{\theta}\right)^{\frac{1}{\theta}}$$

where $L_{x_i} = -\ln(F_i(x_i))$, and

$$\gamma(x_1,x_2) = (\theta-1)\prod_i f_i(x_i;a_i,b_i)\prod_i F_i^{-1}(x_i;a_i,b_i)\left(\sum_i L_{x_i}{}^{\theta}\right)^{\frac{1}{\theta}-2}\prod_i L_{x_i}{}^{\theta-1}$$

$$\cdot\left\{\theta(2\theta-1)\left[(\theta-1)\big(2+5\theta(\theta-1)\big)+\left(\sum_i L_{x_i}{}^{\theta}\right)^2\right]+\left(\sum_i L_{x_i}{}^{\theta}\right)^3\right\}$$

In contrast with the previous copula, there is no linear relationship between the dependence parameter θ and the LDF for this copula.

The pdf graph of the copula shown in Figure 2-5 may give the impression of a similar positive association for most of their joint range. However, the LDF graph provides a more thorough view of the dependence. The correlation is maximal across the main positive diagonal whilst it decreases rather quickly off diagonal and becomes minimal along the $(0,1)$ and $(1,0)$ ranges of the x and y axes.

Figure 2-5: Gumbel copula PDF and LDF

Copula Parameters: $\boldsymbol{a_1=5, b_1=2,\ a_2=5, b_2=2, \theta=2}$

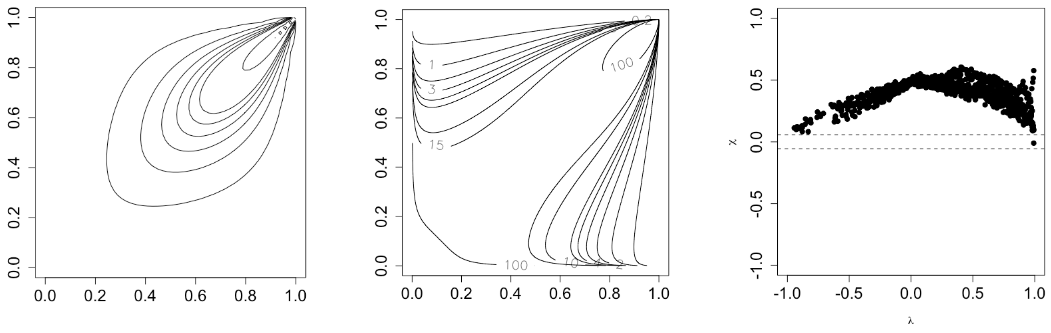
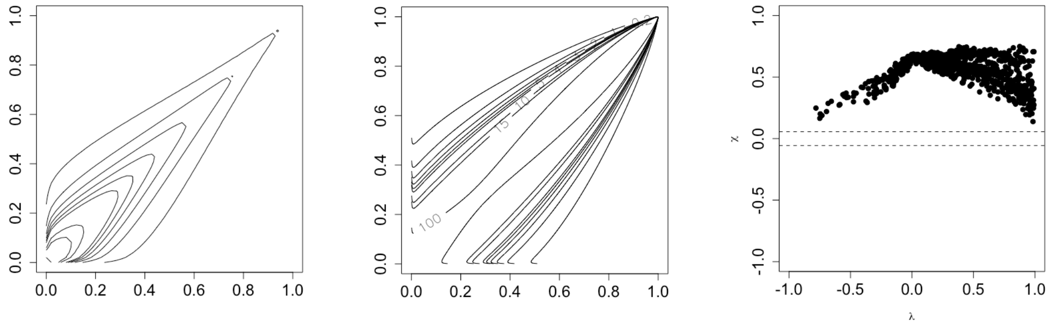


Figure 2-6 shows another example of the density and dependence functions of a Gumbel copula, with same parameters as the example in Figure 2-4. The pdf emphasises the high density at the bottom left corner of the bivariate distribution, whereas the LDF shows a strong correlation along the central part of the distribution.

Figure 2-6: Gumbel copula PDF and LDF (same parameters, different copula)

Copula Parameters: $\boldsymbol{a_1 = 0.8, b_1 = 3, a_2 = 0.8, b_2 = 3, \theta = 3}$



2.3.1.3 Joe copula

The Joe copula with Beta marginals yields the following bivariate distribution, where $\theta \in [1, \infty)$:

$$F(x_1, x_2) = 1 - \left[(1 - F_1(x_1))^\theta + (1 - F_2(x_2))^\theta - (1 - F_1(x_1))^\theta (1 - F_2(x_2))^\theta\right]^{1/\theta}$$

The density of the copula when the two marginals are Beta distributions and its LDF can be written, respectively, as follows:

$$f(x_1, x_2) = f(x_1) f(x_2) \bar{F}(x_1)^{\theta-1} \bar{F}(x_2)^{\theta-1} \times \left[\bar{F}(x_1)^\theta - \bar{F}(x_2)^\theta \left(\bar{\bar{\bar{F}}}_{x_1}\right)\right]^{\frac{1}{\theta}-2} \left[\theta - \bar{\bar{\bar{F}}}_{x_1} \bar{\bar{\bar{F}}}_{x_2}\right]$$

Where $\bar{\bar{\bar{F}}}_{x_i} = \bar{F}(x_i)^\theta - 1$ for $i = 1, 2$ and:

$$\gamma(x_1, x_2) = f(x_1) f(x_2) \theta(\theta - 1)[F(x_1) F(x_2)]^{\theta-1}$$

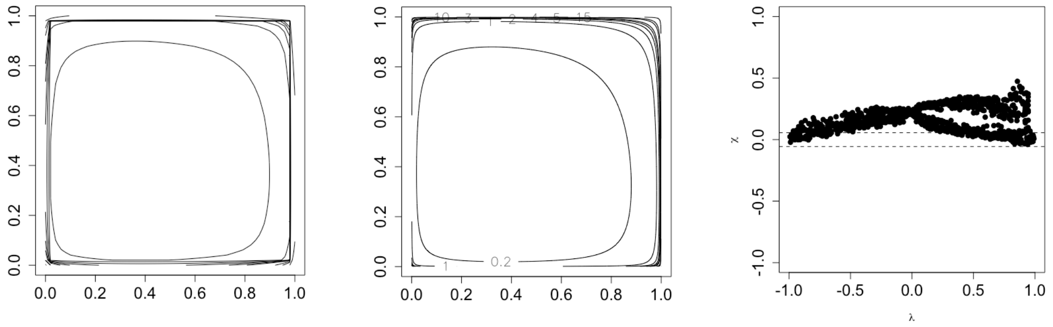
$$\times \frac{-2\theta^2 - \bar{\bar{\bar{F}}}_{x_1}{}^2 \bar{\bar{\bar{F}}}_{x_2}{}^2 + \theta \bar{\bar{\bar{F}}}_{x_1} \bar{\bar{\bar{F}}}_{x_2}\left[3 - \bar{F}(x_1)^\theta + \bar{\bar{\bar{F}}}_{x_1} \bar{\bar{\bar{F}}}_{x_2}\right]}{\left\{\left[\bar{F}(x_1)^\theta - \bar{F}(x_2)^\theta \bar{\bar{\bar{F}}}_{x_1}\right]\left[\theta - \bar{\bar{\bar{F}}}_{x_1} \bar{\bar{\bar{F}}}_{x_2}\right]\right\}^2}$$

As with the Gumbel copula, the overall sign of the LDF cannot be determined directly from the sign of the θ parameter.

In Figure 2-7, density is highest at the top right corner whilst the remaining corners have lower peaks. The highest levels of local dependence are found in similar locations on the LDF graph too, i.e. all corners, with the top right corner having the highest level of local dependence, i.e. large values of both X_1 and X_2 rvs are most highly correlated.

Figure 2-7: Joe copula PDF and LDF

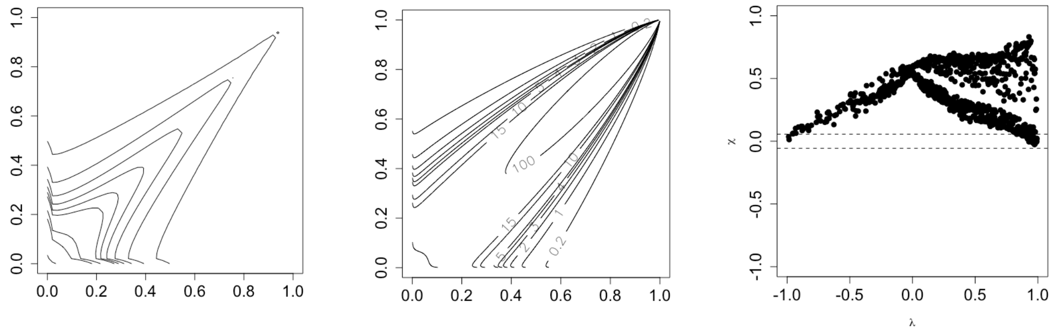
Copula Parameters: $\boldsymbol{a_1 = 0.5, b_1 = 0.1, a_2 = 0.5, b_2 = 0.1, \theta = 1.5}$



The bivariate copula shown in Figure 2-8 has the same parameters as the second example of the previous two copulas, Frank (Figure 2-4) and Gumbel (Figure 2-6). This illustrates the flexibility of various copula models to capture changing bivariate associations/structures given the same parameters. This is mirrored in the LDF graphs too, capturing local dependence at all sets of neighbouring points. The LDF in Figure 2-8 is an example of how the same amount of local dependence is found at two different areas of the bivariate plot, i.e. for low values of X_1 and X_2 and for values of X_1 and X_2 within the range of $(0.4, 1)$.

Figure 2-8: Joe copula PDF and LDF (same parameters, different copula)

Copula Parameters: $a_1 = 0.8, b_1 = 3, a_2 = 0.8, b_2 = 3, \theta = 3$



2.3.1.4 More examples and results

For completeness, examples of the FGM and AMH copulas based on the expressions presented by Gupta [73] are presented in Figure 2-9 and Figure 2-10 respectively. Notice that, although the general shapes of the two densities are similar, the LDF's are markedly different. The LDF of this FGM model has the highest correlation for x_1 values higher than 0.4 and for low values of x_2 and similarly for x_2 values higher than 0.4 for low values of x_1, whilst this pattern was not evident on the density graph. In contrast, the AMH copula reaches its highest local dependence close to the bottom left corner of the graph, where the highest density is also observed.

Figure 2-9: FGM copula PDF and LDF

Copula Parameters: $a_1 = 1, b_1 = 2, a_2 = 1, b_2 = 2, \theta = 0.75$

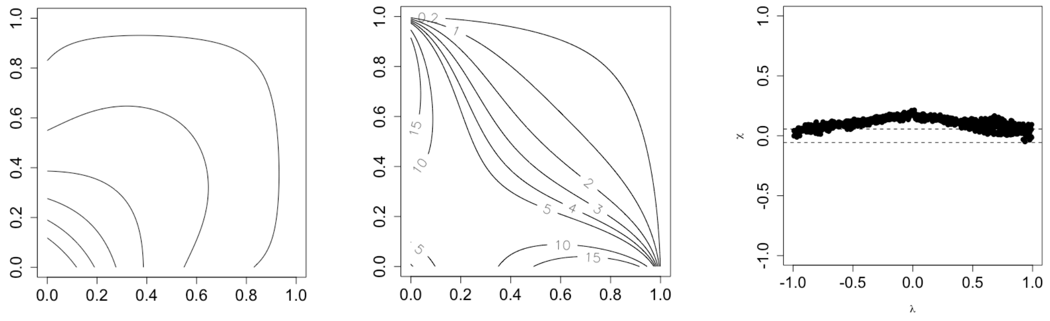
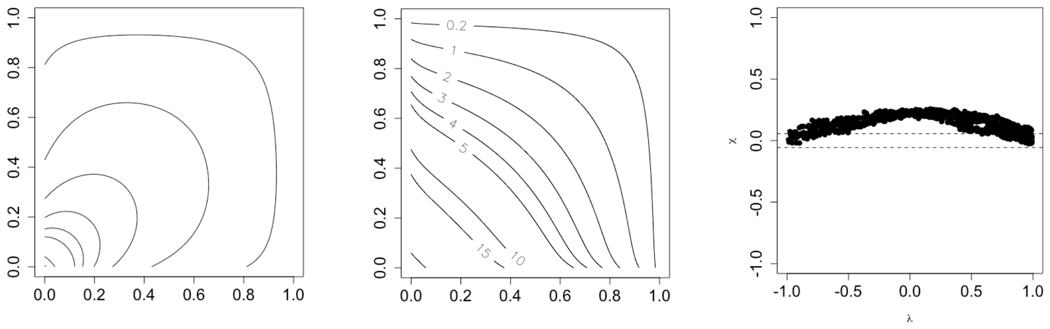


Figure 2-10: AMH copula PDF and LDF

Copula Parameters: $a_1 = 1, b_1 = 2, a_2 = 1, b_2 = 2, \theta = 0.75$



An interesting result for the FGM copula expresses Pearson's r for this bivariate model as a function of its five parameters. Since this bivariate distribution is constructed as an FGM copula its joint moments reduce to the product of univariate integrals in each variable [77,78] hence:

$$r = \theta \prod_{i=1}^{2} \left[\frac{\dfrac{2B(2a_i + 1, b_i)\ {}_3F_2\left(\begin{matrix} a_i, 2a_i + 1, 1 - b_i \\ a_i + 1, 2a_i + b_i + 1 \end{matrix} \middle| 1 \right)}{B(a_i, b_i)B(a_i + 1, b_i)} - 1}{\sqrt{\dfrac{a_i b_i}{a_i + b_i + 1}}} \right]$$

Where ${}_3F_2$ is the generalized regularized hypergeometric function [72]. It has been shown [79] that the absolute value of the correlation coefficient for any FGM copula is less than or equal to $1/3$. It is easy to see that this bound is reached for this bivariate distribution, e.g. if $a_1 = b_1 = a_2 = b_2 = 1$ then $r = \theta/3$. If all the four parameters of the univariate Beta distributions in the copula are equal, r takes values between $\theta/4$ (all 0) and $\theta/3$ (all 1) and then decreases very little to stay just above $0.318\,\theta$ for parameter values much larger than 1. It was only possible to extract this result for the FGM copula due to the ease of the calculation of the bivariate moments.

2.4 Conclusions

It is important for researchers to realise that correlation coefficients do not always convey the relationship between numerical variables in the best possible way. Summarising the entire correlation structure in a single constant value does not account for changes in the strength of the association across the joint range of Y_1 and Y_2. The local dependence function overcomes this problem by producing a detailed graphical display of the association between Y_1 and Y_2 across all values. The LDF can produce distinctly different graphs for bivariate density functions that are similar, as demonstrated in the contrasting Figure 2-9 and Figure 2-10.

The flexibility of individual copula types to cover a wide range of bivariate distributions with Beta marginals is emphasised via the use of different parameter sets. I have presented three LDF formulae, which can be easily programmed to produce LDF graphs for bivariate scenarios with Beta marginals. The same principles could be applied to marginals other than univariate Beta, for example skew-normal distributions [80].

The LDF along with the chi plot comprise a selection of alternative avenues for the exploration of bivariate associations.

In the next Chapter, my focus is on the identification of joint outliers and how this can be informed via the use of copula models and enhanced with the inclusion of covariates as and where appropriate.

3. Bivariate Analogues of Centiles (BAC) for convex bivariate pdfs

Population centiles are widely used in medicine to highlight individuals who have unusual outcomes [81]. For example, height and weight centiles are used to identify children who may have growth defects or require intervention; centile charts for respiratory function examine the reduction in function of the lung attributable to diseases such as cystic fibrosis; lab-based tests of body fluids are commonly interpreted with reference to standards provided by the test's manufacturer. Often this procedure is an important aspect of screening programmes which try to identify subjects with a particular phenotype or biomedical marker correlated with a disorder. These individuals may benefit from further investigation or direct preventive action to prevent disability or to improve their quality of life [82].

It is also common for a patient to undergo a battery of tests and each of these is referred to a separate test-specific centile chart with several tests/centile charts considered jointly. An individual may be within the normal range on one or more tests but outside the normal range for others. If referral is made on the basis of at least one test lying below the 5th population centile (a common scenario), then substantially more than 5% of patients may be referred leading to increased need for further, potentially invasive and/or expensive, investigations [83].

Where correlations exist between the outcomes of tests and more specifically when the magnitude of the correlation is not constant across the bivariate range, considering the multivariate relationship between the outcomes should improve the accuracy with which unusual individuals are identified. Hence application of these models may decrease the number of false referrals and highlight the number of hidden extremes, potentially improving allocation of resources and reducing patient anxiety [84].

Most commonly, scatterplots are the first point of call for researchers investigating bivariate relationships. The pattern produced in a simple scatterplot can be very informative and here I present a way of interpreting this pattern/shape to produce inferences in the form of centiles on a bivariate scale without making any distributional assumptions about the data. The term bivariate analogue of centile(s) (BAC(s)) will be used in the remainder of this Chapter and thesis to refer to an analogue of centiles on the bivariate scale. The reason behind using the word 'analogue' as opposed to simply 'bivariate centiles' is the fact that there is more than one way of ordering multivariate data and my proposed way is just one of many. The proposed method is focused on an ordering procedure based on convex hull peeling, which will be discussed in detail later in this Chapter.

My proposal is non-parametric and has proven to work fast in the examples presented in this thesis (and beyond), compared to other existing `R` functions (more details are presented in section 3.6). Its results are also very interpretable for researchers (percentile coverage within a bivariate range) as they can be seen as a direct analogue of univariate centiles, which are very well understood.

In this Chapter, I focus on the association between two continuous outcomes. Section 3.1 presents the results of a literature review on the topic of bivariate centiles. Convex hulls are defined in section 3.2 followed by the rationale behind the idea of a sequence of convex hulls/convex hull peeling in section 3.3 which will formulate the definition of

BACs. Sections 3.4 and 3.5 present two simulation examples, one based on a Gumbel copula and another based on a banana-shaped distribution, which is an example of a mixture of 3 uncorrelated bivariate Normal distributions [85–87]. More specifically, the proposed BACs are used to identify subjects who would have been considered to be within the normal range via the use of two separate univariate centile charts but are flagged up as unusual via this bivariate centile method. In Section 3.6 results from existing `R` libraries (`geometry`, `cxhull`, `depth` and `depthProc`) are presented and, where appropriate, their results are compared with my proposed BAC algorithm in terms of each library's flexibility and interpretability of bivariate analogues of centiles. Chapter 3 closes with an overview of the presented results and how these relate to gaps that are still pending (section 3.7).

3.1 Literature review

An online search was conducted using "The Web of Science" platform [88] in August 2019. Even though the focus of this Chapter is the application of centiles in bivariate scenarios, the search conducted included both "bivariate" and "multivariate" terms to ensure that all advances in this field are covered and commented on.

More specifically, the aim was to capture phrases such as bi/multi-variate reference regions/ ranges/ ellipsoids/ intervals/ charts/ models, and bi/ multi-variate tolerance regions/ ranges/ intervals. Tolerance regions are regions within which a specified proportion of a distribution lies with a fixed probability [89]. The following terms were also identified as relevant during the preliminary literature search: "trivariate reference range/region" and "dimensional reference region/range". The final list of search terms is shown below:

- bivariate centile* (0)
- bivariate reference* (9)
- bivariate tolerance* (10)
- multivariate centile* (1)
- multivariate reference* (20)
- multivariate tolerance* (20)
- trivariate reference* (2 + 1 repeat)
- *dimens* reference region* (1 + 1 repeat)
- *dimens* reference range* (1 repeat)
- *non-parametric tolerance* (3)

(The asterisk * symbol dictates that the space before or after the asterisk is flexible and can vary, hence capturing different beginnings and endings of words.)

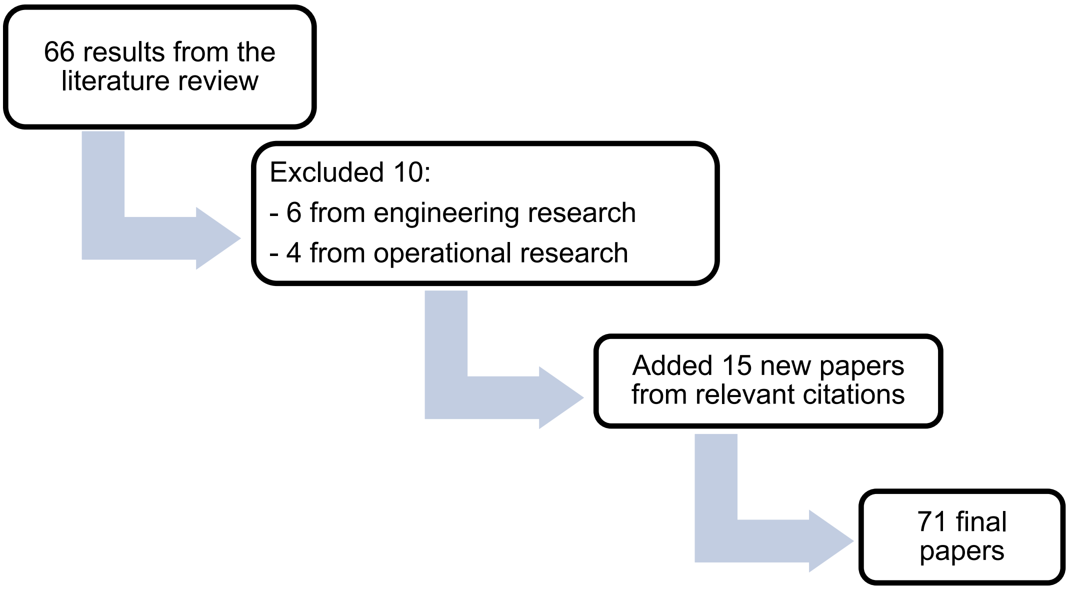
The outcome of this search resulted in 66 unique papers. The numbers in the brackets above indicate the frequency of each term (additional terms, as per the list at the previous paragraph, not included resulted in no publications). No uses of the term 'bivariate centile' were identified and just one paper used the term 'multivariate centile'. Nine results matched the *bivariate reference* term and another 10 the *bivariate tolerance*. These increased to 20 each with the use of the term *multivariate* instead of bivariate. Two papers were identified when the search phrase included the word *trivariate* and another one for the term *dimensional*. Finally, the last term revealed an extra 3 publications about bivariate and multivariate non-parametric *tolerance limits/ ranges*.

Several (10) of the results were not relevant to our search as they referred to engineering, operational design or financial terms, with no connection to the rationale of this Chapter; some of these terms were: bivariate tolerance design/ model/ signals/

techniques. After their removal, there was a remainder of 56 papers for further exploration.

Several additional papers (15) cited in some of the publications above, but not using any of the exact search terms listed earlier, were also identified as relevant. These were notably older papers (9 published prior to 1956) whose titles were more generic and did not explicitly mention the dimensions of the data in question (i.e. multivariate, bivariate; which were the focus of this search). These were added to the list of results, bringing the total number of papers reviewed to 71, as seen in the flowchart in Figure 3-1.

Figure 3-1: Flowchart of literature review results


66 results from the literature review
Excluded 10:
- 6 from engineering research
- 4 from operational research
Added 15 new papers from relevant citations
71 final papers

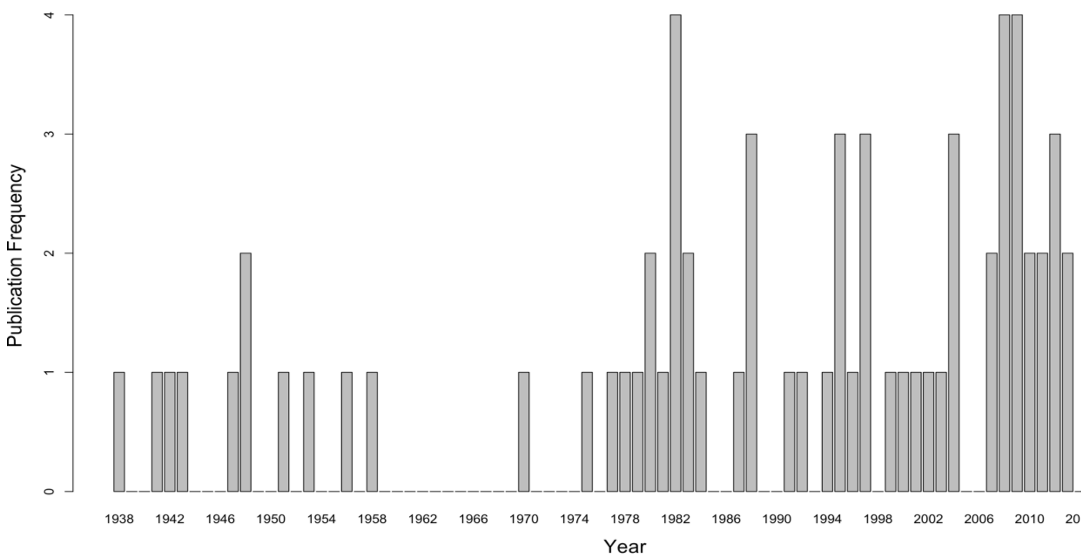
These 71 results spanned a variety of research fields including general statistical/computational methodology, neurology, endocrinology, nutrition, biomedicine, forensic sciences and more. Fields with more than 2 papers accounted for are summarised in Table 3-1.

Table 3-1: Papers by science field

Science Field	Papers
Statistics & Probability	29
Medical & Laboratory Technology	9
Endocrinology & Metabolism	5
Nutrition & Dietetics	5
Mathematics & Computational Biology	5
Medicine & Experimental Research	4
Medical Informatics	4

The first publication in this field appears in 1938, followed by minimal publication activity until 1956. A book by Irwin Gutmann [90] in 1970 breaks a 14 year-long silent publishing gap and, quite fittingly, it comes to summarise the theoretical developments in the field of (mainly) univariate and multivariate tolerance regions from a frequentist as well as Bayesian point of view. Since 1975, there is publication activity almost every 1 to 2 years adding up to 33 distinct years of 59 publications (Figure 3-2). The number of papers varied between 1 and 4 with the last two decades having the highest per year publications rate (17 during the 2000's and 13 since 2010).

Figure 3-2: Publications per year



In more detail, the inception of the idea of non-parametric tolerance limits came from Thompson in 1938 [91] and Wilks in 1941 [92] and 1942 [93]. Wilks proved that for continuous variables, the percentage of a given tolerance range has a Beta distribution ("distribution of the coverage") and was independent of the distribution of the variable of interest but instead was a function of the specific order statistic chosen and the sample size. In 1943, Wald [94] extended this idea to multivariate scenarios based on successive elimination of multivariate points and in 1947 Tukey [95] published an additional extension introducing the term 'statistically equivalent blocks'. The latter report forms the oldest paper of the literature search I conducted based on the terms shown earlier and comes as part of a series of 3 heavily technical publications from John Wilder Tukey and colleagues at the Annals of Mathematical Statistics. The first dealt with non-parametric order statistics [96] whilst the last explored tolerance regions for discrete continuous data [97]. The 2nd paper of this series proves particularly interesting to this thesis as it presents and quotes, for the first time, a polygon as the desired tolerance region, i.e. a bivariate analogue to a centile. The means of producing such polygon are entirely geographical with instructions from the author to the reader to draw lines crossing the most south-westerly, north-westerly, northerly, easterly, southerly and westerly points. Interestingly, our proposal has a direct association to Tukey's idea where lines are drawn through existing data points to form the edges of the polygon/centile.

More mathematical foundations for the computation of non-parametric multivariate tolerance regions/centiles were presented by Tukey in 1948 [98] as well as other authors between 1951 and 1956 [99–103]. As mentioned earlier, Guttman [90] publishes the first book dedicated to (mainly) univariate and multivariate tolerance regions in 1970, which contains comprehensive details regarding the publications mentioned in the previous two paragraphs.

Five years later (1975), an abstract for the International Symposium of Prospective Biology [104] in French refers to the application of multivariate reference ranges to the blood ionogram with no further details of the methodology used. No future citations have been identified for this work. In 1977, a publication in German by Abt [105] is claimed (as it has not been possible to obtain this publication) to present a method for the construction of scale-independent, non-parametric multivariate tolerance regions; we will see more from Abt in the coming years, hence mentioned in the next few paragraphs.

In 1978 and 1979, two papers are published quoting trivariate reference regions in the field of Clinical Chemistry [106] and Mineral and Electrolyte Metabolism [107], respectively, with one author in common. The earlier paper is the first *application* found in the literature of higher than two-dimensional reference regions. The authors explore the relationship between 3 outcomes simultaneously with the aim of identifying outliers in their sample and calculate Mahalanobis' distance [108] based on the assumption of a trivariate Normal distribution. In this given example, one of the outcomes was logged to comply with the Normal distribution and the final reference region result took the form of an ellipsoid shown on a 3-dimensional printout. The 1979 paper has not been obtainable.

Whilst dwelling further into the results of this literature review, it soon became evident that the use of Mahalanobis' distance (MD) [108], D^2, is one of the most commonly used methods for the calculation of bivariate or higher dimensional reference/tolerance ranges [109,110,119–125,111–118] to date, including the most recent, 2018, results of the literature review. So far there have not been any relevant publications in 2019.

MD is the distance of each multivariate point (Y_i) from the multivariate mean (μ), whilst accounting for the interrelationships of the variables via their covariance (Σ_i, positive definite variance-covariance matrix). In other words, it is the sum of the squared normalised distance of each observation from the centre of the distribution. Once all distances are calculated and ordered the desired multivariate reference value can be identified.

$$D^2 = (y_i - \mu)^T \Sigma^{-1} \left(y_i - \mu^{-}\ \right)$$

where T denotes the transpose.

MD is very restrictive in its applications as the variables (or their transformations) should follow a multi/bi-Normal distribution. This is an assumption often unrealistic for data in the medical and other fields. Hence, the need for non-parametric alternatives.

The first publication of the 1980's [126] is an abstract by Kauerz et al at the Joint Congress of the Scandinavian and German Societies of Clinical Chemistry. The abstract mentions that 95% dispersion ellipsoids were presented based on the assumption of normally distributed data.

In the same year, the term “multivariate reference range” is quoted at the proceedings of another conference [127], organised by the Joint American and Canadian Society of Clinical Chemistry. Even though I was not able to identify additional details about this presentation/ discussion, the main author, James C Boyd, returns two years later, in 1982 (and in 2004) with yet further ideas in the field of multivariate reference ranges (both discussed below).

In 1981, Abt joins with Ackermann for a paper in German in the field of Medicine [128], which was unfortunately currently unobtainable. Abt publishes again on his own in 1982 [129] in English where he presented the results of his proposed ‘parallelogram method’ for the construction of tolerance regions that can be of irregular multiplanar shape, but also scale independent, i.e. the regions are invariant with respect to linear transformations.

Boyd and Lacher in 1982 [84] dealt with multivariate reference ranges (20 dimensional in fact) and their benefits in comparison to 20-fold univariate tests. Their methods were solely based on the assumption of Normal distribution via the Mahalanobis’ distance statistic. Their results supported the assumption of less falsely abnormal reported cases when the multidimensional range is explored but they acknowledged the limitation of this application to datasets that are not Normally distributed.

Ackermann publishes with a different team of authors in 1982 in a German Paediatrics journal [130] and also presented at the 18th workshop of Pediatric Research in Germany [131]. According to the workshop abstracts, the authors discussed the range of applicability of bivariate tolerance regions, with particular interest in diagnosing subclinical rickets based on children located outside the bivariate tolerance region for alkaline phosphatase and 25-hydroxycholecalciferol. Additional methodological information of the applications of tolerance regions in this specific field of application has not been detected online.

In 1983, a brief article by D.L. Massart [132] recommends to clinical chemists to consider the use of multivariate reference regions in combination with univariate results as per Boyd's 1982 paper. Also, in the same year, Ackermann [133] proposed a new construction technique, which unfortunately has not been located.

In 1984, Ackermann and Abt [134] published their findings regarding sample size calculations for multivariate tolerance regions. Their paper essentially contained extensive tables for sample size determination for the construction of specific limits, inner and outer, of non-parametric multivariate tolerance intervals based on the results presented by Tukey in 1947 and Abt in 1982. From a statistical point of view, the procedures cited so far and the developments discussed in the above publications, are characterised by arbitrariness; they depend on auxiliary ordering functions [135] (like any other non-parametric method) and are not necessarily asymptotically minimal with respect to a chosen indexing class [136], i.e. produce tolerance regions that do not converge to a minimal index (this could be volume/density/probability/etc) when compared to another region from the same sample.

During the end of the '80s and start of the '90s centuries, more publications are seen based on MD ([109,110,119–125,111–118] cited earlier too) and in 1995 a new approach to multidimensional data is brought into the field of tolerance regions. Principal Component Analysis (PCA) [137] does not depend on distributional assumptions and is routinely employed to reduce the dimensionality of the data but the construction of the reference ranges is based on the assumption of normality [138,139].

Another conference abstract follows in 1996 from Northern Europe [140] producing an "innovative graphical method" for the identification of patients with unusual/interesting clinical results (i.e. body fluid overload). A similar team of authors published just over a year later a detailed paper on their findings about their proposed graphical method, the resistance-reactance (RXc) graph [141]. Their resulting bivariate ellipsoids are fully dependent on the assumption of the normal distribution and the specific formulae they are based on is presented in another publication from one of the authors [89]. A short commentary of this work followed the next year [142], with few more papers following in the coming years [143–147], but in none of these publications did the authors make a reference to non-normally distributed data.

At the end of the 1990s, a paper relating to reference ranges for longitudinal data was published [148] and a new computer program was introduced for the construction of multivariate reference models [149] (no longer available). Both proposed techniques were completely dependent on the assumption of the normal distribution. The former calculated centiles based on the minor and major component of principal component analysis and the authors concluded that "the centiles are no longer in units which are meaningful to the practitioner and hence their use becomes mechanistic, without the possibility for interaction". The authors of the latter paper mentioned in their conclusions that "the multivariate reference model can be helpful in some cases but must be seen as an addition to the univariate reference interval and not a replacement". In this thesis, I will aim to argue the case that they are both useful and address different questions, so they could potentially be used in isolation of each other.

In 1999, a paper from an Italian team of researchers, Capitani et al [150], utilised some of the early 20th century results of our literature search seen earlier (Wilks, Wald, etc) to propose the construction of tolerance hexagons by drawing parallel and non-parallel lines around the regression line between the two outcome variables. As the authors mention, if there is weak correlation or no dependence at all, it would be better for the tolerance region not to be based on the regression line between the two variables. However, using a regression line to summarise the relationship between two outcome rvs does not make the most of the shape of the cloud of points formed when plotting the two rvs against each other.

The start of the 21st century sees a paper from the forensic sciences mentioning bivariate tolerance regions [151] in their attempt to re-certify the National Institute of Standards and Technology Standard Reference Material on DNA quality assurance methods. They produce several graphical displays of tolerance ellipses, which are all based on the assumption of the bivariate normal distribution.

From 2000 onwards, there were several other publications based on MD and normally distributed data (as cited earlier on page 103), while less than two handfuls of papers prove interesting for our research and a couple of those are very relevant to the rest of this Chapter. The latter two are Bucchianico et al in 2001 [135] and Li et al in 2008 [152]. Both papers refer to convex hulls and several of their results will be discussed in section 3.3.

In 2003, Petersen [81] describes two new 2-dimensional reference charts, the bivariate reference and directional percentile charts. The non-parametric estimator of the former is a result of the empirical distribution functions of Y_1 and Y_2. It produces very rugged lines due to the discrete nature of the indicator function, I, involved in the calculation of the empirical distribution functions of Y_1, $S_1(\cdot)$ and Y_2, $S_2(\cdot)$:

$$\hat{S}_1(y) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Y_{1i}>y\}} \quad \text{and} \quad \hat{S}_{2|1}(y|y_1) = \frac{1}{m}\sum_{M} I_{\{Y_{2i}>y\}}$$

Where: $M = \{i | Y_{1i>y_1}\}$ is the selection of data indices that have a value higher than y and $m = \#\{i|Y_{1i} > y_1\}$ is the total number of those indices.

The directional percentile chart estimates a reference curve in the plane such that, in any direction from the centre of the distribution (estimated medians of Y_1 and Y_2), a certain part of the mass of the distribution is outside the curve. The non-parametric estimator of the two-dimensional reference curves is based on conditional quantile estimation [153,154] and uses kernel functions.

In 2008, Amin et al [155] extended a previously presented method, MaxMin Chart to multivariate scenarios and utilised information from an Exponentially Weighted Moving Average (EWMA) control chart to obtain smoothed tolerance limits. These take the form of the minimum and maximum observations of a range with a desired percentage coverage. This procedure depends on the multivariate normal distribution and the right choice of a smoothing parameter.

Petersen returns in 2009 [156] with an extension of his earlier 2003 work on unconditional non-parametric estimators of the bivariate reference curve (discussed earlier). It now accounts for covariates which can make the range of tolerance intervals more relevant and realistic to real-life situations. The analysis is extended by the non-parametric estimator being conditioned on specific values of the explanatory variable and, as previously, a kernel estimator of the conditional quantile is minimised. However, as the author concludes "…the (proposed) bivariate reference does not give a curve that demarcates 95% of the total probability mass (to the right or above) and 5% (to the left or below). … If one was to insist on giving a boundary region of probability content 95%, this could be achieved by identifying the particular bivariate percentile with a pre-specified proportion of the reference population outside. This is a topic of further research." Notice the use of the term bivariate percentile from Petersen in a similar context to how it is used here.

Wellek in 2011 [157] proposes the use of rectangular half-spaces whose edges determine univariate percentile ranges of the same probability content in each marginal distribution. The calculation of such rectangular shapes is based on the assumption that the variables follow a bivariate normal distribution.

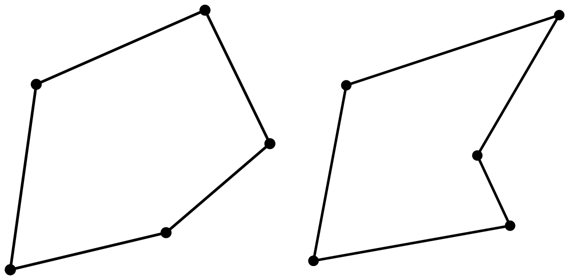
In 2015, Willemsen et al [158] focuses on longitudinal growth measurements within a Bayesian approach and applies the directional quantile theory as described by Kong et al [159] in combination with results from the Multivariate Superimposition by Translation and Rotation (MSITAR) model [160]. For a direction indicated by a unit vector s, the p-th directional quantile in the direction of s corresponds to orthogonal projection of the rv on s. To construct the p-th directional quantile contour, the direction given by the vector s changes to cover all angles. Each contour defines a half space where $(100 - p)\%$ of the observations lie. By taking the intersection of these half spaces, the directional contour is obtained. This application (implemented in `R` via the `modQR` library [161]) is only relevant to longitudinal measurements and goes beyond the remits of this thesis.

I believe that there remains a strong need for a straightforward, data-driven and easy-to-interpret tool to appropriately construct bivariate standards for non-normal data. The rest of this Chapter responds to this need and describes the notions behind my proposal. The next two sections describe the tools necessary for my proposed method.

3.2 Convex hulls

A polygon is a bounded shape linking data points with straight lines. The lines connecting the data can produce interior angles that are either below or above 180°. A convex shape is a polygon for which all interior angles are less or equal to 180°; contrary to a concave shape which has dents in its perimeter (Figure 3-3).

Figure 3-3: Convex (left) and concave (right) drawings of two random variables



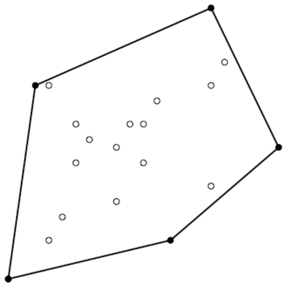
The selection of points that form the perimeter of a convex shape are also referred to as a convex set. The smallest convex set that encapsulates a given selection of points is called the convex hull.

3.3 Sequences of convex hulls

In a scenario where two outcome variables are investigated simultaneously, there is no unique way of jointly ordering these bivariate observations. Convex hulls provide an ordering scheme for multivariate observations and can act as the basis for distribution-free ways of exploring bivariate associations by utilising the geometry of the bivariate scatterplot [135,162]. The density of points inside/outside each convex hull and/or the depth of each point can be seen as a non-parametric feature of the bivariate distribution in question. The depth of each point is the number of convex hulls that contain this point [163], i.e. observations "deeper" inside a clouds of points or closer to the centre of the data will have large depth, contrary to the points near the outskirts of the dataset. Sufficiently deep contours remain robust to outliers.

A depth value can be calculated for each value from a given random sample of bivariate observations $D(Y_{1i}, Y_{2j})$ and the data can be ordered according to their descending or ascending depth value, $D(\cdot)$. For example, points further into the centre of the cloud of points in Figure 3-4 are deeper into the dataset, hence will have higher depth.

Figure 3-4: Notion of depth



The proposed algorithm, described below, follows the principles of convex hull peeling as discussed by Barnett in 1976 [164] and Eddy in 1982 [165]. Convex hull peeling is a procedure that works its way through a dataset by peeling away data that form consecutive convex hulls. The process starts with the convex hull which encloses all sample points. These are then removed, and the next convex hull is constructed. The process is repeated, and a sequence of convex hulls is formed. The process ends when there are no more convex hulls to be created, i.e. less than 3 points.

In 1999, Liu et al [166] wrote a detailed report on multivariate statistics by data depth (including convex hull peeling amongst other methods). Of particular interest here is the mention of the work of two authors, Rousseeuw and Ruts [162,167] and their work in 1996 that preceded Liu's publication. They focused on the calculation of bivariate depth and its representation via the bagplot, i.e. bivariate boxplot and contours surrounding the data/pockets of the data. However, their methods involve an inflation factor chosen to be equal to 3, by which the central bag (as per the authors' terminology) containing 50% of the data is inflated to create a boundary outside which outliers lie. According to the authors the choice of value 3 was the result of experience and simulation studies. The citation used for these simulations does not seem to have been actually published hence has not been obtained.

In 2001, Di Bucchianico et al [135] proposed a new method for constructing non-parametric multivariate tolerance regions. However, it requires pre-specifying the shape of the tolerance region (i.e. ellipsoid, hyperrectangles or convex sets) and resulting regions may not be connected; an unrealistic scenario for many applications such as quality control of items that are produced in continuous patches, i.e. several quality characteristics of items in a production line might be evaluated together to assess their quality, hence separate resulting regions of quality scores would not make practical sense.

In 2007, McDermott et al [168] proposed an alternative method, sequential convex hull (CH) peeling, to the computationally expensive convex hull peeling process, which the authors note is especially relevant to very large datasets. Their proposal depends on an iterative procedure of convex hull peeling on groups of w points at a time from a dataset of size n. It does not peel the entire dataset but instead, this algorithm will only ever peel w points at a time to produce a final 'average' convex percentile. They take w to be between 1000 and 10000, the latter being their recommended value.

They give a message of warning for values higher than 10000 due to the increase in computation time. The comparison of their proposed method, sequential CH peeling against the traditional CH peeling procedure shows superiority of the latter in terms of estimation of the bivariate median, but the tables are turned when it comes to computational power. The average execution time of their proposed method changes linearly with sample size as opposed to quadratically as in other peeling algorithms. The authors did not make the code for their proposed method publicly available.

The most important and relevant publication for this Chapter is Li and Lui's [152] paper from 2008 which builds nonparametric multivariate tolerance regions and discusses the construction of statistically equivalent blocks based on the notion of data depth and spacings. Multivariate spacings are the gaps between two consecutive convex sets. Li and Liu proved that:

- The coverage probabilities of multivariate spacings follow the same distribution as the univariate spacings provided that the depth function is continuous
- The tolerance regions are asymptotically minimal
- The coverage probabilities are distributed Beta $(n-2r+1, 2r)$, where n is the sample size and r is a positive integer such that $r < (n+1)/2$

Applications of the methods cited above have been made available via the implementation of the relevant algorithms in three R libraries; `geometry` [169], `depth` [170] and `DepthProc` [171]. The main disadvantage of the techniques applied in these libraries is that they are not fully driven by the data points themselves and the fact that they do not equip the user with flexibility regarding the calculation of specific percentiles. Findings from Li and Liu's paper are implemented in the `geometry` library which is the `R` adaptation of the more generic `quickhull` algorithm based on

Barber's work [172]. Comparison between the results of the relevant functions from the libraries above and the suggested algorithm from my work will be presented in section 3.6.

My proposed algorithm is a fully non-parametric, exclusively data-driven technique. It uses function `chull` [173] within the `grDevices` R library. It identifies every convex set that surrounds a given data set; from the outer convex hull to the centre of the data, providing an intuitive way of ordering the bivariate sequence of points. Exclusion of the points that form the outermost convex hull, allows the user to move further/deeper into the dataset and create a new convex hull of the subset of data remaining after the exclusion; and so on.

Each convex hull splits the sample of observations in two; those cases outside the convex hull (including those on the boundary of the convex polygon itself) and those contained within the convex hull. Though bivariate centiles cannot be uniquely defined, each convex hull can be thought as an analogue to a centile in the bivariate scale, i.e. bivariate analogue of centiles, BAC, in the sense that it approximately corresponds to a region limiting a certain percentage of cases outside (including those on the hull) and inside this centile.

The algorithm to obtain BACs performs a repetitive procedure of the following steps:

i. Take the outermost convex hull of a cloud of points; BAC_p: the p-th BAC contour, where $0 < p \leq 1$ (or $0 < p_{\%} \leq 100$).
ii. Count/store the number of points that form the convex hull. The number of points on and outside this BAC_p p-th convex hull will be approximately equal to $n \times p$; might be slightly under or over the exact $n \times p$ value, as $n \times p$ might not be an integer value. Subsequently, the reported BAC level might be slightly

under or over p. For instance, the 61st BAC contour of 410 points should be made up by 250.1 points (on and outside the 61st bivariate percentile contour), which translates to approximately 250 points. In fact, the true number of points forming $BAC_{61\%}$ might be slightly under or over 250, e.g. 249, depending on the dependence structure/shape between the two rvs. Therefore, the final reported BAC closest to 61% will be $249/410 = 60.7\%$.

iii. Reduce the cloud of points by deleting the points that form the perimeter of the previously defined BAC, so that the remaining dataset consists of approximately $(n - n \times p) = n(1 - p)$ values and establish the next outermost convex hull.

iv. Re-count the number of points that form the new BAC and append to the previous set of points (step ii).

… repeat until the most central point(s) of the data set has been reached.

The bivariate centre is a central region determined by the shape of the underlying distribution. It might either be a single point (the most central/deep point amongst the cloud of points), or the average between the 2 most central points, or the centroid of the innermost polygon/convex hull obtained by successively peeling away the outermost convex hull layers/defined by the 3 or more points forming it.

This proposed algorithm makes no approximations around data points as all convex sets cross an observed point. It makes no distributional assumptions and is independent of numerical/estimation/convergence issues. Most importantly, it allows for individual, unique centile lines to be picked out and used for further scientific inference. Additionally, it provides the means for a complete exploration of extreme observations (e.g. outside the 5th BAC) and how they compare to univariate results. There is no other publicly available R algorithm that gives a researcher the tools to compare univariate and bivariate extremes.

Constructing a convex hull involves sorting of data, leading to at least $O(n \ln n)$ operation time. In two dimensions, it is reported that $n \times \ln n$ is a lower bound on the complexity of any algorithm that finds a hull [174]. As the data dimension increases, it can be expected that the computation complexity will grow faster than $n \ln n$.

3.4 Simulations Application

A bivariate sample of 1500 values was simulated from a Gumbel copula with $\theta = 1.5$ and Y_1~Normal $(\mu = 1, \sigma = 0.4)$ and Y_2~Weibull $(\mu = 5, \sigma = 25)$ marginals. Figure 3-5 shows the simulated data along with the corresponding contours of the bivariate distribution function. Seventy one convex hulls surround the entire dataset starting from the outer convex hull moving towards the centre of the data, i.e. 71 BACs ranging from 0% to 100% coverage of the data, as shown in Figure 3-6. Each BAC is uniquely defined as there could not be another convex polygon connecting a given selection of observed points.

Figure 3-5: Simulated data from Gumbel bivariate copula

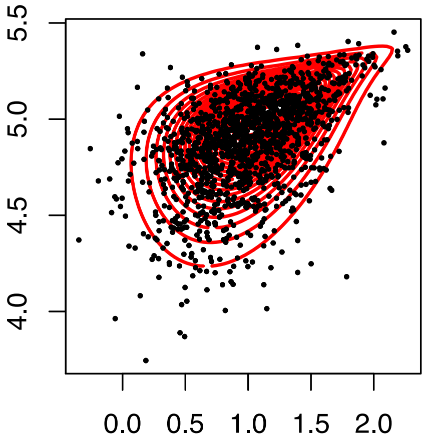
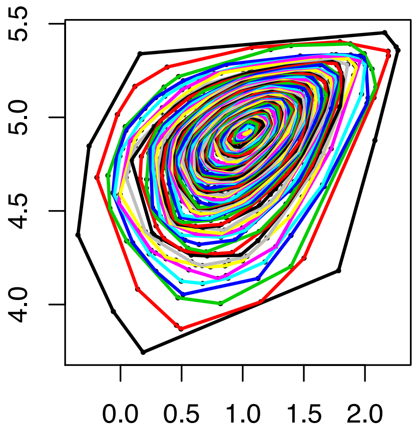


Figure 3-6: All convex sets of Gumbel simulated bivariate copula



The lines shown on Figure 3-7, from the outer to the inner part of the scatterplot, represent the proposed 2.5th, 5th, 50th, 95th and 97.5th BACs. Table 3-2 shows the exact bivariate coverage of these percentiles. The middle line is a bivariate analogue to the median and in this case is very well approximated; a 50.7-49.3% middle split amongst the observed sample. A total of 81 points would be considered as bivariate extremes in this example (i.e. an extreme observation based on both Y_1 and Y_2 variables), falling on or outside the 5% BAC, red convex set.

Figure 3-7: Specific convex sets simulated from a Gumbel bivariate copula

BAC levels: 2.5th, 5th, 50th, 95th, 97.5th

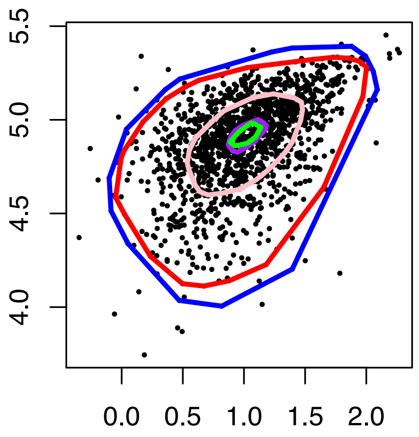


Table 3-2: Exact level of percentage coverage of BACs for simulated Gumbel copula

Desired percentile level	Exact bivariate coverage
2.5%	2.6%
5%	5.4%
50%	50.7%
95%	94.7%
97.5%	97.3%

Table 3-3 shows how many of the 81 bivariate extremes would have been considered as within-the-normal-range observations based on the univariate centile ranges of the Y_1 and Y_2 variables separately. Figure 3-8 and Figure 3-9 graphically display the results of Table 3-3; these have been colour coordinated to facilitate visual representation of the observed differences between univariate centiles and BACs. For the purpose of this example, values outside the 5% centile in the bivariate scale are labelled as ‘extreme’, and similarly values lower than the 2.5% and higher than the 97.5th levels for the univariate scale (but these can in fact be changed to any desired level). The 5% univariate centiles, in lines 1 and 2 of the table, have, as expected, 5% of the data for each rv accounted for as ‘extreme’. The 5th BAC is best approximated by level 5.4 (Table 3-2).

The biggest drop comes in the next two lines, where the number of extremes is almost halved (3.1-2.7%) by simultaneously considering the centile structure of the two rvs. A further reduction (down to 1.5 and 1.2%) is seen on the next two lines of the table which proves the point that a lot less cases would be thought of as ‘extreme’ when the univariate and bivariate analogues of centiles are concurrently reviewed, hence potentially reducing the number of false extremes.

Additionally, based on this algorithm, users are able to identify potential false positives and hidden extremes. False positives (shown in pink in the last subgraph of Figure 3-9) are cases identified as unusual based on the assessment of multiple univariate criteria but are not flagged up as extreme when the numerous criteria are treated multivariately. In this sample, 29 (1.93%) and 34 (2.27%) cases are extreme based on univariate criteria for each of the two outcome variables (below the 2.5th and above the 97.5th centiles), with 5 shared cases, resulting to 58 unique potential false positives (3.87%).

Hidden extremes are cases that are not thought of as extreme based on univariate criteria but are highlighted as extreme based on the bivariate convex hull criteria. In this sample, there were 12 (0.8%) such cases (shown in blue in the middle subgraph of Figure 3-9).

Table 3-3: Number (%) of cases classified as non-normal based on univariate and bivariate criteria

Univariate – below the 2.5th and above the 97.5th centiles		BAC (Y_1, Y_2) – outside 5th convex hull centile	Number (%) of cases
Y_1	Y_2		
☑			75 (5.0%)
	☑		75 (5.0%)
		☑	81 (5.4%)
☑		☑	46 (3.1%)
	☑	☑	41 (2.7%)
☑	☑		23 (1.5%)
☑	☑	☑	**18 (1.2%)**
CONSISTENT EXTREMES			
☒	☒	☑	**12 (0.8%)**
HIDDEN EXTREMES			

Figure 3-8: Outliers according to univariate and bivariate criteria

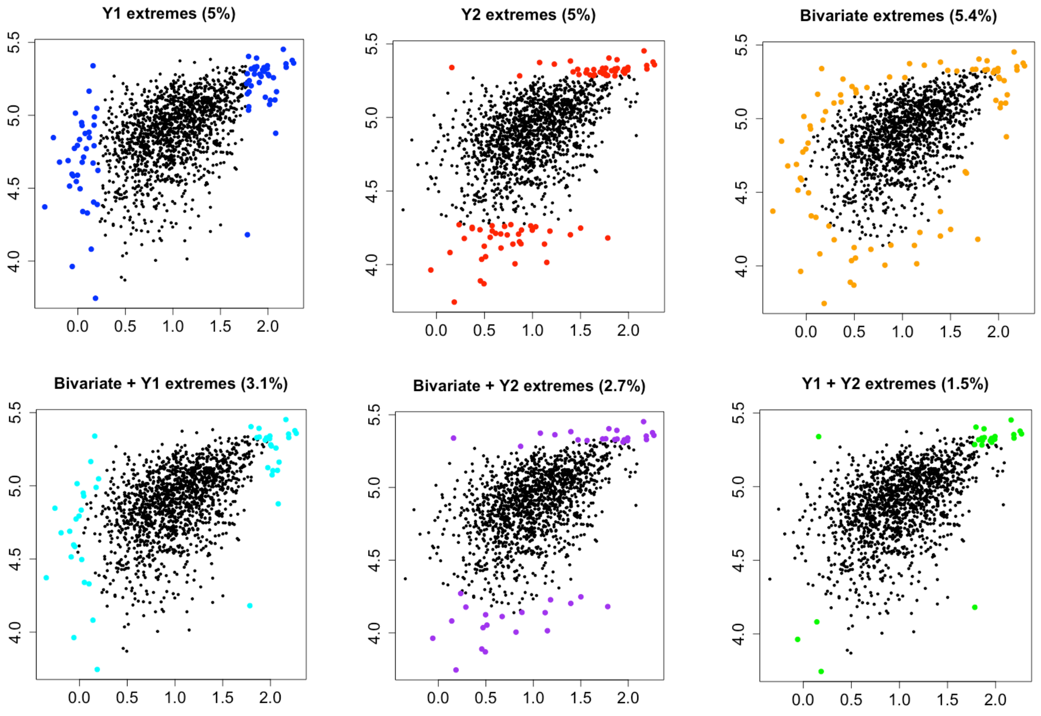


Figure 3-9: Consistent, hidden and false positive extremes

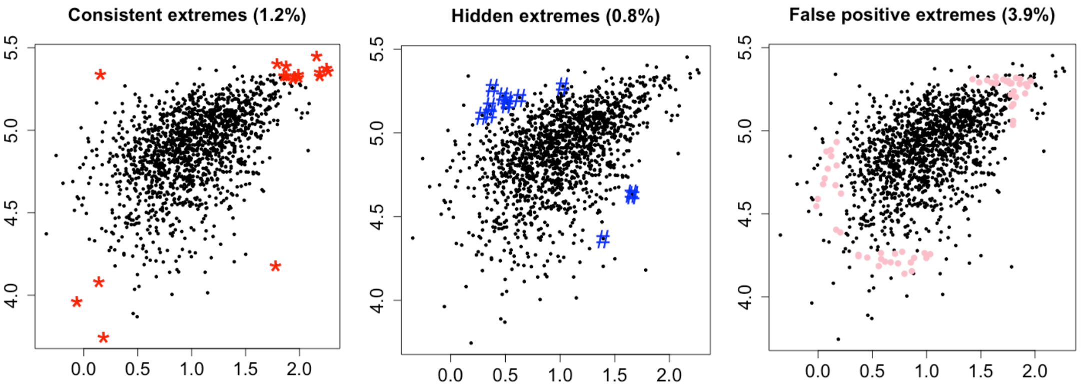
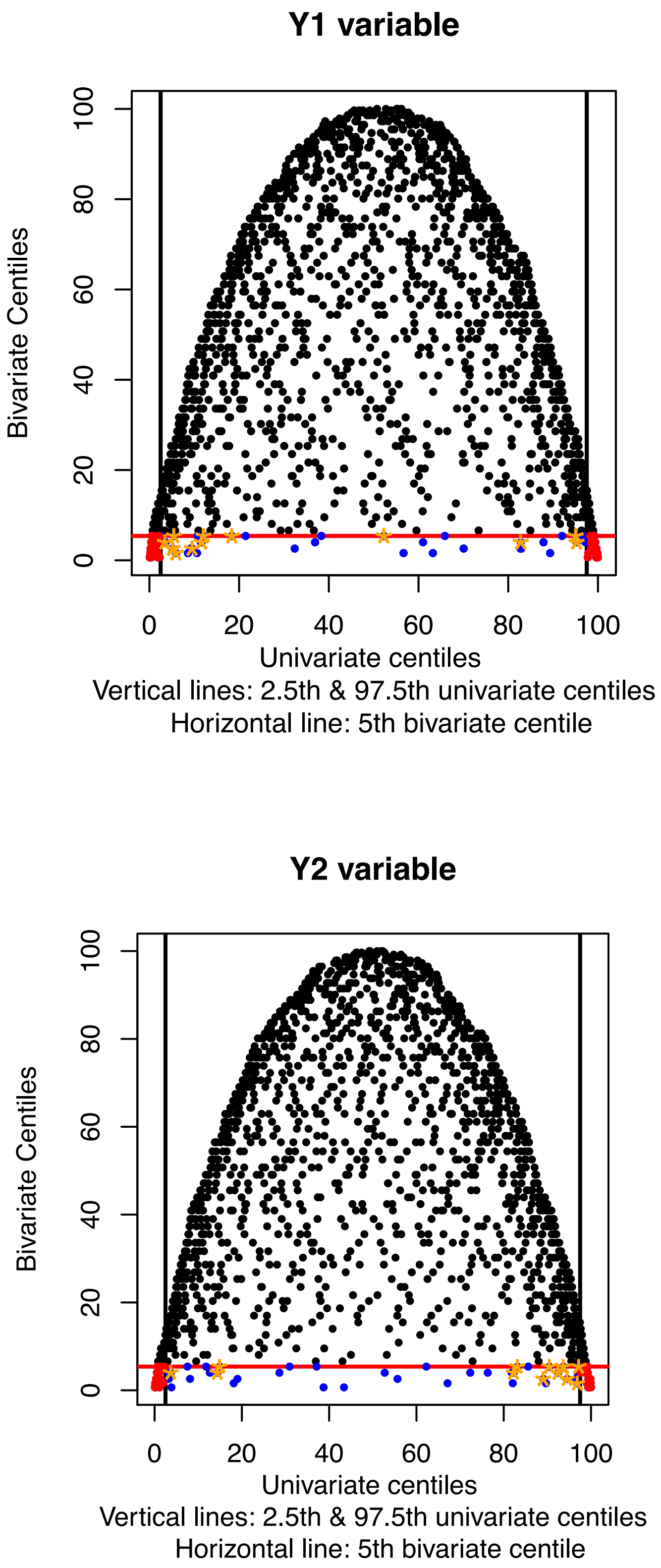


Figure 3-10 displays the relationship between the univariate centiles and BACs with the two vertical and one horizontal line representing the 2.5th and 97.5th univariate centiles and the 5th BACs, respectively. This graph illustrates the direct correspondence of the 3 sets of centile values (two univariate and the bivariate analogue), which in fact is approximated very well by an inverted U-shape (iU) function. The boundary of the iU shows:

- the cases towards the top of the iU shape correspond to the bivariate central cases, i.e. the cases deeper into the dataset, closer to the bivariate and univariate medians
- the degree of differences between univariate and BACs can only vary within the boundary of the iU shape, i.e. the differences are capped so can only vary by a certain degree, defined by the boundary of the iU. For example, there cannot be a univariate centile at 2.5% level which corresponds to a value above, approximately, the 25th BAC as estimated by reviewing the iU plot, i.e. this is the highest y axis value that corresponds to the vertical line drawn at 2.5%

- the small squares on either side at the bottom of the graph contain the consistent extreme values as defined by matching the results of the univariate centile and BAC analysis

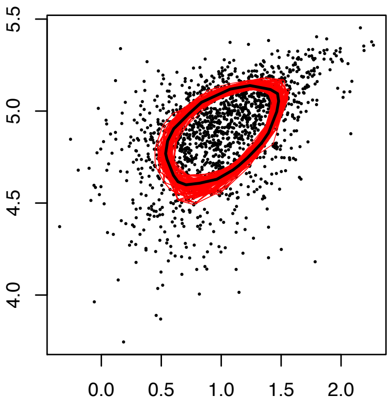
Figure 3-10: Univariate centiles versus BACs



The blue points at the bottom sections of the two graphs in Figure 3-10 represent hidden extremes based on a comparison against univariate centiles alone (35 and 40 respectively). But only 12 of those are bivariate extremes, i.e. fall outside the 5th BAC but within the 2.5th and 97.5th for both response variables (shown as orange asterisks).

To quantify the precision of the BAC obtained, a convex hull envelope of the 50th BAC is presented in Figure 3-11. The black line is based on the first simulated data as introduced in Figure 3-5 Gumbel copula and the remaining red lines are a result of 500 bootstrap samples.

Figure 3-11: Bootstrap envelope of convex hulls



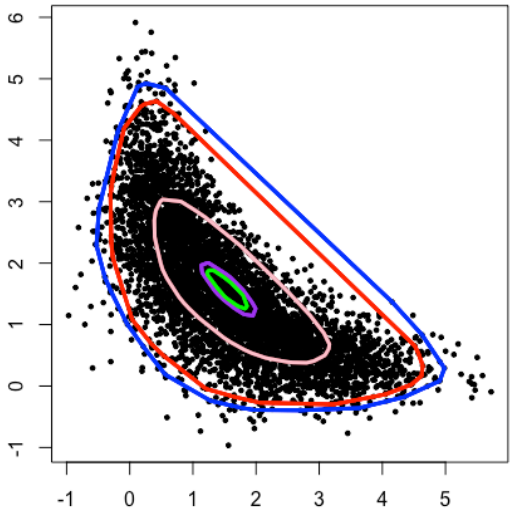
The convex hull envelope can act as a measure of precision of a given BAC, in this case the bivariate median. As it is not possible to order each of the centiles/convex hulls in relation to one another, i.e. from the ‘smallest’ to the ‘largest’, it is not possible to chop off the lowest and highest 2.5% to produce a 95% confidence region of centiles. However, the envelope alone can visually assist researchers when evaluating the fit of the proposed method on a given dataset.

3.5 Further Application

Wei [85] demonstrated how her proposed approach worked for a banana-shaped dataset following a reviewer's suggestion on her manuscript. Wei's method is based on the idea of connecting pairs of opposite points and defining a central interval according to the bivariate distribution of the outcomes conditional on the line connecting the two points (it has not been possible to replicate these results as Wei's algorithm is not publicly available). In Figure 3-12 the results from my proposed algorithm are presented for a banana-shaped scatterplot based on 5000 simulations. This is an interesting example from a family of distributions whose marginals and bivariate density are not normal, though their conditional probabilities in both directions are univariate normal [86].

Figure 3-12: BACs (5th, 50th, 95th) for a banana-shaped data cloud

BAC levels: 2.5th, 5th, 50th, 95th, 97.5th



The algorithm produced the BACs seen in Figure 3-12 and Table 3-4 shows the exact bivariate coverage for 5 selected percentile levels (162 BAC levels in total). As expected, the percentages given in Table 3-4 correspond quite closely to those desired, but Figure 3-12 shows that the fit is not good.

Table 3-4: Exact level of percentage coverage of BACs for banana shaped dataset

Desired percentile level	Exact bivariate coverage
2.5%	2.7%
5%	5.1%
50%	49.8%
95%	94.9%
97.5%	97.5%

Table 3-5 shows the frequency of outliers as identified according to univariate and bivariate criteria. The true 5% coverage for this sample, i.e. the closest percentage to the goal of 5%, is 5.06%. There are 257 (5.14%) false positives (shown in red and green in Figure 3-13 for the Y_1 and Y_2 rvs, respectively) and 26 (0.52%) hidden extremes (shown in blue). If the univariate centiles were the only classification tool (the red horizontal and vertical lines represent the univariate 2.5th and 97.5th centile levels), 5.1% and 0.5% of cases would have been potentially falsely identified and missed as extremes, respectively. The respective iU shape and the convex hull envelope of the 50th BAC are shown in Figure 3-14 and Figure 3-15, respectively.

Table 3-5: Number (%) of cases classified as non-normal based on univariate and bivariate criteria

Univariate – outside 2.5th and 97.5th centile		BAC (Y_1, Y_2) – outside 5th convex hull centile	Number (%) of cases
Y_1	Y_2		
☑			251 (5.0%)
	☑		251 (5.0%)
		☑	253 (5.1%)
☑		☑	118 (2.7%)
	☑	☑	127 (2.5%)

☑	☑		23 (0.5%)
☑	☑	☑	22 (0.4%)
Consistent EXTREMES			
☒ (inside)	☒ (inside)	☑	**30 (0.6%)**
Hidden EXTREMES			

Figure 3-13: Potential false positives for a banana-shaped distribution

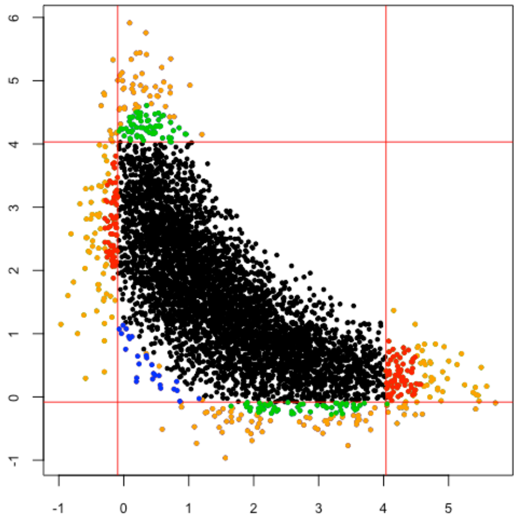
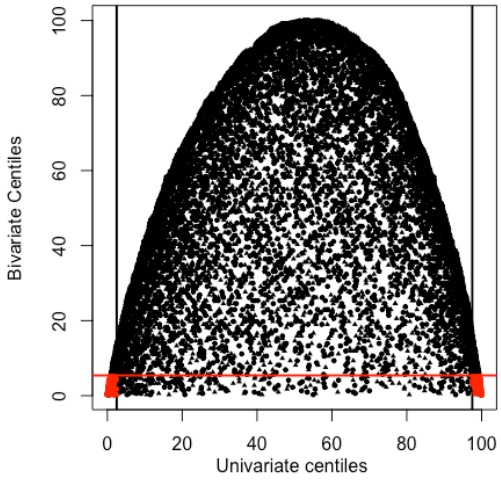
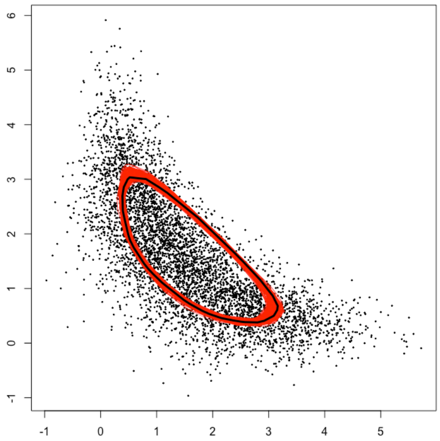


Figure 3-14: Banana-shaped distribution, univariate versus BACs



Vertical lines: 2.5th & 97.5th univariate centiles
Horizontal line: 5th bivariate centile

Figure 3-15: Bootstrap envelope of the 50th BAC for the banana shape



The BACs seen above do not provide a good fit for the banana density due to the non-convexity of its shape. An alternative algorithm that would incorporate concave elements may be more appropriate in this example.

To the best of my knowledge an appropriate convexity index does not already exist and creation of one goes beyond the remit of this thesis. Some theoretical work has been in this area by Porzio and Ragozini [175]. They propose a split of the data between *inner* and *outer* observations depending on how close or far they are from the bulk or the boundary of the data respectively. To identify which are the closest observations, they firstly consider for each convex hull vertex (starting from the outer one) the distances to the remaining points along a radial projection. Then, along each of these directions, they look at the univariate ordering of the points from the closest to the furthest. The presence of gaps, if any, in these univariate orderings will highlight any empty space in the data structure, splitting the outer observations from the inner ones.

3.6 Results from existing R libraries

I have identified four libraries in R containing functions to obtain and display convex hulls.

i. library: `geometry`, function: `convhulln` (also applicable to higher than 2 dimensions)
ii. library: `cxhull`, function: `cxhull`

They are both based on the quickhull algorithm (qhull.org) [172]. The indices of the outermost convex hull are given – these are alternatives to the results of the `chull` function used throughout this Chapter and produce identical results. However, the latter function, `chull`, is more straight forward to use due to the nature of the resulting index (vector), which corresponds directly to data indices, unlike a rather involved list object with mixed order of indices, resulting from either of the above functions.

- library: `depth`, function: `isodepth`

The `isodepth` function is based on Ruts and Rousseeuw's 1996 paper [162]. It draws convex shapes around data points that satisfy certain conditions regarding the depth of the data included in the convex shape (which do not relate to the percentage of points included/excluded from each contour) and are dependent on an 'arbitrary' inflation factor, as discussed on page 112. This is contrary to my method which draws those convex shapes exactly on data points.

In Figure 3-16, all the depth contours of the 1500 values of the simulated Gumbel copula as introduced in Figure 3-6 are presented as per the algorithm of `isodepth` function. The two graphs look very similar but a closer look highlights that the `isodepth` function does not draw the CHs directly on observed data points, unlike the BAC method. The `isodepth` function took more than 10 minutes to run for this 1500-case long dataset, followed by warning messages, in contrast to my proposed BAC procedure which took less than 5 seconds. The warnings imply that it was not possible to draw exact contours for all depth values.

To draw contours of specific depth, the user is asked to define a positive integer, e.g. 75, that corresponds to a contour of specific depth, i.e. 75/1500 = 0.05. The right-hand side graph of Figure 3-17 displays contours of depth $20, 50, 200, 500, 700$ via the `isodepth` function. Even though there were analytical results for depth level $700/1500$, i.e. x and y coordinates of the points forming this level of depth, there was no graphical output for this level, hence only four depth contours are displayed on the graph on the right of Figure 3-17. The graph on the left-hand side of Figure 3-17 shows all 5 contours of the equivalent BAC levels from my proposed algorithm as listed below, where the number of points are 'translated' into percentile coverage:

$20/1500 = 0.013$, exact % BAC coverage 1.6%

$50/1500 = 0.03$, exact % BAC coverage 4%

$200/1500 = 0.13$, exact % BAC coverage 14%

$500/1500 = 0.33$, exact % BAC coverage 33.6%

$700/1500 = 0.46$, exact % BAC coverage 47.1%

Figure 3-16: Depth contours based on the `depth` R library

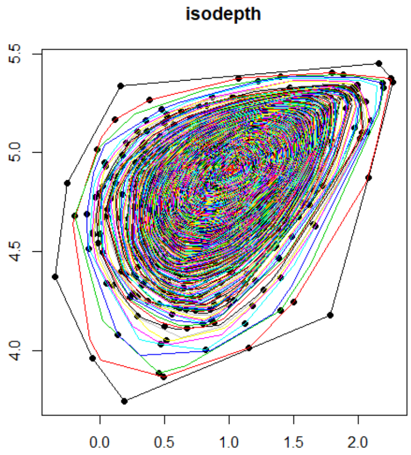
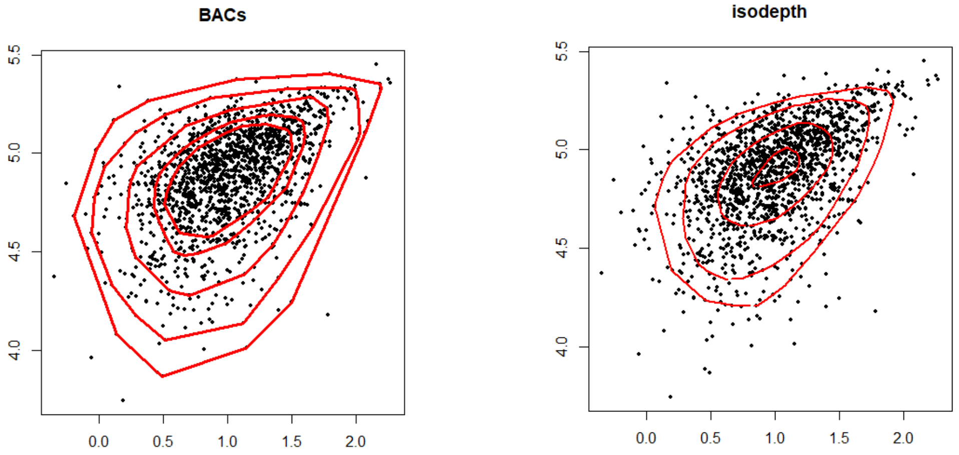


Figure 3-17: Comparison of the BAC algorithm and `isodepth` function results



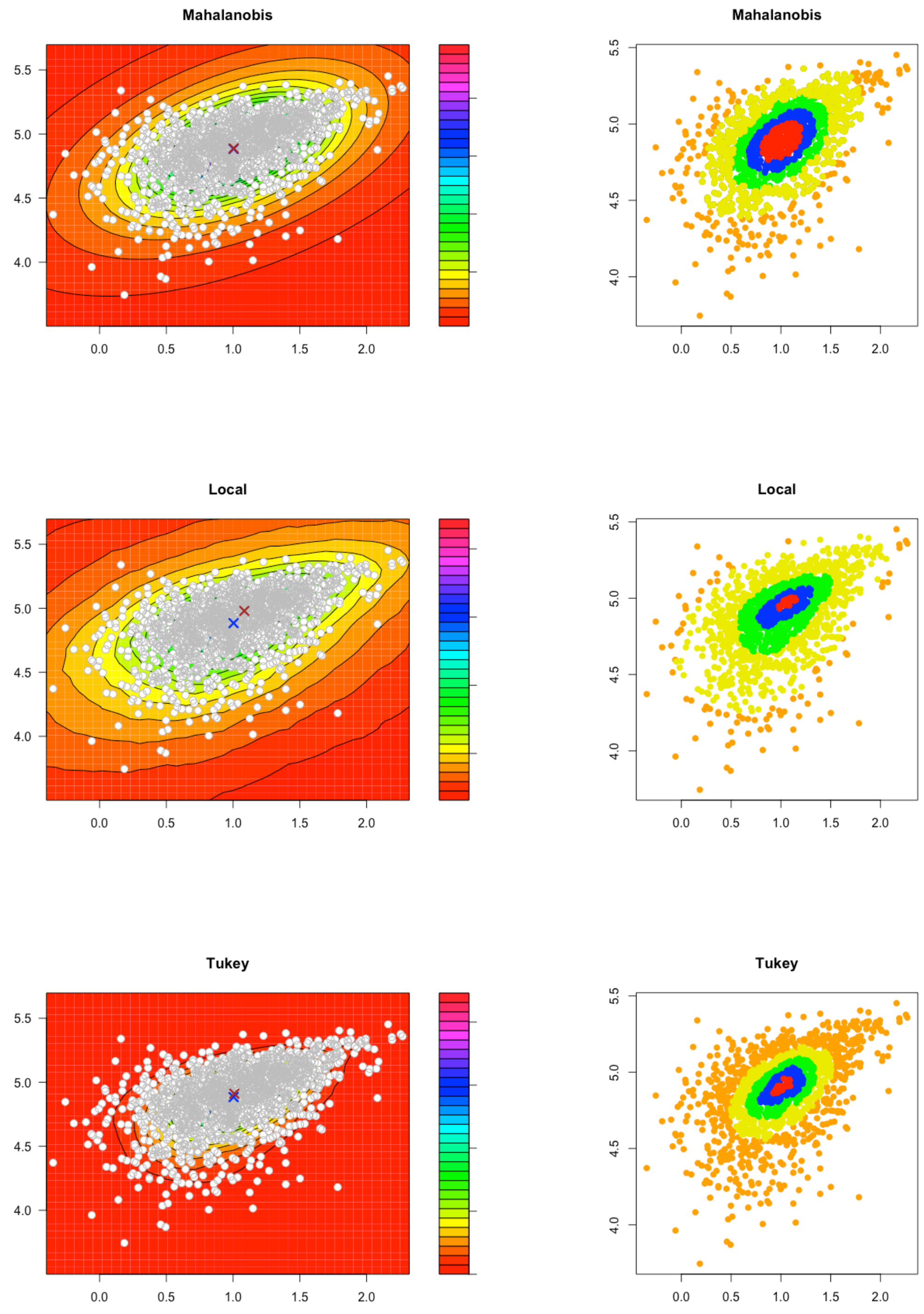
- library: `DepthProc` (depth procedure), function: `depthContour`

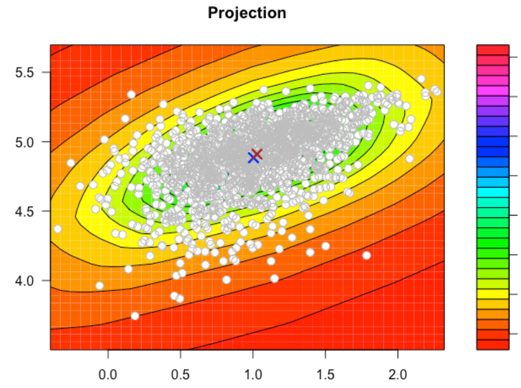
The `DepthProc` library calculates the depth of bivariate data and produces the relevant contour plots based on several different methods: Tukey's, Projection [176], Mahalanobis, Local [177], L^p and Euclidean [166,178–180]. The first two are based on approximate calculations and the rest use exact calculations. There are six rows of graphs in Figure 3-18. Each of the six rows represent the results of the `depthContour` function according to the six different definitions of depth mentioned above. The graphs shown on the left-hand side column of Figure 3-18 are based on the default graphical options of the `depthContour` function which does not offer much flexibility in terms of the size and colour of the data points and their transparency in relation to the drawn contour lines. The colour gradient shown next to each graph serves as a visual representation of depth, from the outer (fewer deep points) to the inner (deeper) ones towards the centre of the scatterplot. To facilitate the graphs' interpretation, I subsequently, redrew these plots and these are presented on the right-hand side column. The data points were assigned different colours according to their depth, matching the depth ranges and colour gradient shown on the graphs on the left, which were assigned by the function by default.

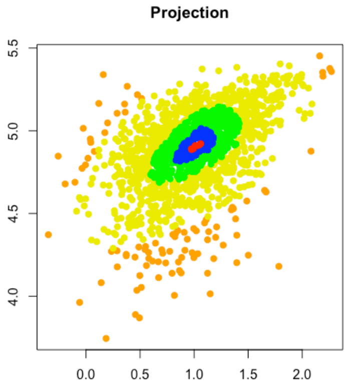
Additionally, the `depthContour` function requires the definition of the number of Y_1 and Y_2 points to be used for each depth contour as well as the number of depth levels to be drawn. The first one dictates the shape of the contours, from square-ish shapes for smaller number of points to polygons for larger number of points. The larger the number of levels, the more separate sections of depth the data set is split into. But neither of these two factors relate to the percentage of points included/excluded from each contour. For this example (Figure 3-18), the number of levels were set to 30 and the number of points was set to 40 (empirically chosen as values below 40 led to very rigid contours and values over 40 to extremely smooth contours).

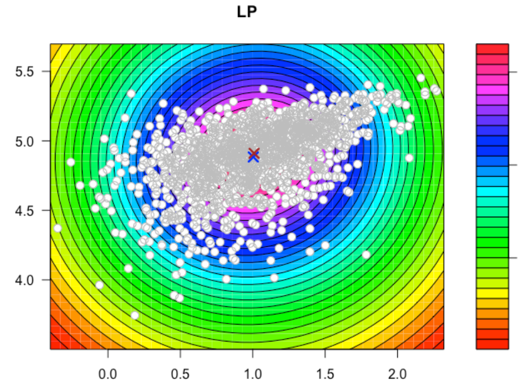
Figure 3-18: Contours from `depthContour` function (library: `DepthProc`)

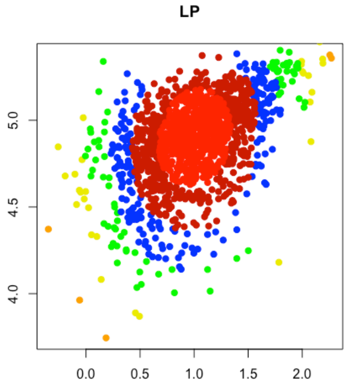
(levels$= 30, n = 40$)

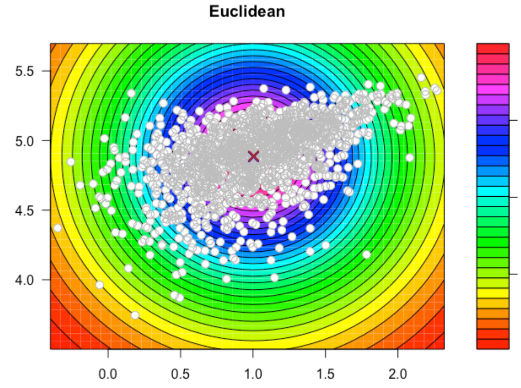



Projection
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0


Projection
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0


LP
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0


LP
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0


Euclidean
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0

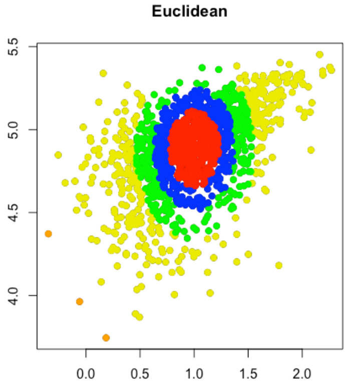

Euclidean
5.5
5.0
4.5
4.0
0.0
0.5
1.0
1.5
2.0

Table 3-6 shows that the resulting bivariate median (50th BAC) from my proposed BAC method approximates very well the results from the two other existing functions. It is exclusively data-driven and does not depend on axillary factors and/or a variety of definitions of the depth function.

Table 3-6: Bivariate median results based on different algorithms/libraries

50th BAC	1.02 , 4.85
`depthMedian` (library: `DepthProc`)	1.02 , 4.91
`ctrmean` (library: `depth`)	1.01 , 4.89

Whilst the previously existing libraries provide ways of calculating the convex hull and/or specific depth contours, they do not allow the user to request specific level BACs in terms of percentile coverage and plot these against their equivalent univariate centiles for each rv.

3.7 Conclusions

Convex hulls utilise the shape of a bivariate association to form analogues to distribution-free centiles in bivariate dimensions. The latter provide an estimate of the bivariate median [181] as well as the means to identify extreme cases according to concurrent measurements on two variables.

My research was motivated by the work of others [158,182,183] and the aim was to produce a robust, transparent, flexible, distribution free and exclusively data-driven tool that clinicians/researchers can use with direct interpretation to their observed sample data.

The number of cases in a sample identified as unusual might be considerably more than 5% when several tests are evaluated independently, depending on the strength of the association between these multiple tests, and values that are multivariately abnormal might be missed. By exploring the bivariate behaviour between rvs, truly unusual cases may be more easily identified.

Despite the numerous papers published in the field of bivariate depth and appropriate contours, the results of the literature review presented at the beginning of this Chapter, show that the use of the term ‘bivariate centile’ is virtually non-existent. A readily available, freely accessible and easy to use algorithm that produces a straightforward comparison between univariate centiles (with which many researchers are already very familiar) and bivariate analogue to centiles (BACs), may facilitate better use of bivariate data within the data analytics field. The use of a term that researchers can see as an extension of something already familiar to them, i.e. extending univariate to bivariate, feels like the appropriate way forward in order to increase the term’s visibility in papers, subsequently its understanding from the readers.

The inclusion of covariates will further enhance the results presented in this Chapter, together with the incorporation of copula models to further characterise the relationships highlighted.

4. Copula Models and Bivariate Analogues of Centiles

This Chapter applies copula models introduced in Chapter 1 to the BAC methodology of Chapter 3. Copula models are used in this Chapter to enhance the applicability of the BACs for convex shapes with and without covariates. The results presented here are based on a simulated example dataset and will lay the necessary foundations (together with previous Chapters) for Chapter 5, where the theories demonstrated in this thesis will be applied to a real dataset.

4.1 Copula bivariate analogues of centiles (CBAC)

As detailed in Chapter 1, copula models are very flexible in quantifying multivariate distribution functions, in particular bivariate relationships, which has been the focus of the thesis from the start.

Having found the best fitted copula for a bivariate scenario, bivariate analogues of centiles (BACs) can be produced on simulated data from this copula. These BACs should match closely with the BACs fitted to the raw data (from which the copula would have resulted) and/or might be smoother versions of the former depending on the goodness of fit of the copula model. This latter point will be explored further in Chapter 5 when a real dataset will be available, and the copula modelling procedure will yield simulated data that can be compared to the raw data. This Chapter will focus on simulated data alone.

Figure 4-1 presents 3000 simulated data points from a Clayton copula with parameter $\theta = 2.2$ and two identical Sinh-Arcsinh (SHASH) [184] marginal distributions; SHASH$(\mu = 20, \sigma = 1.1, \nu = 1.5, \tau = 3)$ for the Y_1 (x-axis) and Y_2 (y-axis) variables, respectively. The blue lines forming the central cross represent the median of each variable plus and minus one quartile on either side. The purple lines in Figure 4-2 represent the contours of the Clayton copula from which the dataset is simulated.

Figure 4-1: Simulated bivariate Clayton copula with SHASH identical marginals

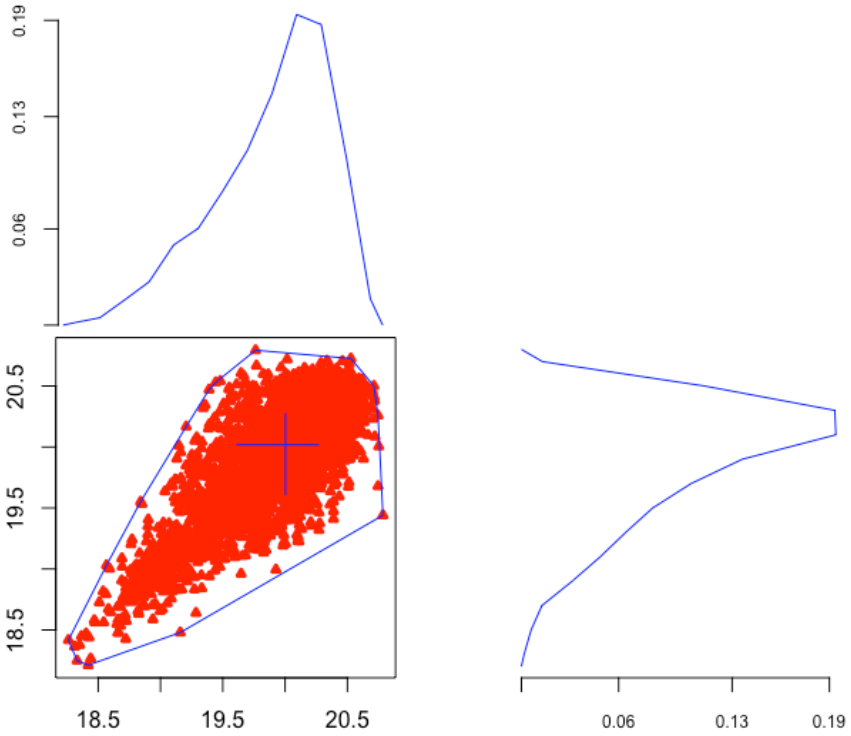


Figure 4-2: Contours of Simulated bivariate Clayton copula

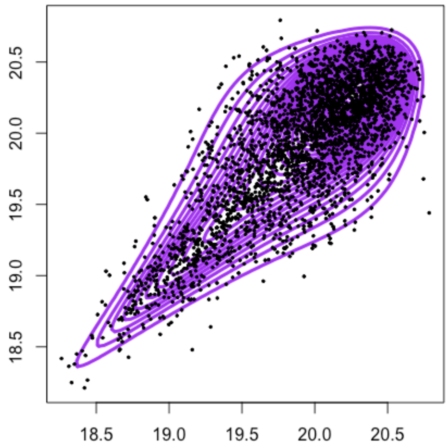
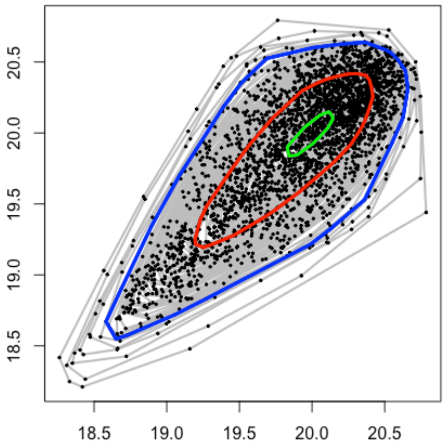


Figure 4-3 shows all 55 BACs produced via my proposed algorithm in grey with the 5th, 50th and 95th BACs represented by the blue, red and green convex hulls respectively. The use of the copula models at this stage provides a basis for the next section which will involve the inclusion of covariates in all the copula's parameters, hence yielding differing BACs for given values of the covariate(s).

Figure 4-3: BACs of simulated copula



4.2 Copula regression models

The investigation of the association between two or more outcome variables will often require conditional modelling both in the marginals and in the dependence parameter. For example, the relationship between the noise and CO_2 emissions of cars may be better understood/explained if the size of the car is also taken into account and perhaps other related features. If the model residuals are not Normally distributed with equal variance for all predictor variables, then transformations may be considered or alternative distributional forms. The class of Generalized Additive Models for Location Scale and Shape (GAMLSS) [185] incorporates many different distributional forms and allows spread, skewness and other distributional parameters to vary according to one or more predictors. For example, if the outcome is log-Normally distributed regression models for the mean and SD of the logged values may be used; if an outcome variable has a Beta distribution, regression models may similarly be fitted, one for each shape parameter. Hence a comprehensive approach is taken to understanding the way predictors influence an outcome whilst ensuring the fitted model has Normally and homoscedastically distributed residuals.

Copulas can extend the GAMLSS approach even further by allowing not only the parameters of the marginals but also the dependence parameter θ of the copula to vary according to one or more predictors.

For example, in the case of a copula with a single θ, Normal $N(\mu_N, \sigma_N)$ and Gamma $GA\ (\mu_G, \sigma_G)$ marginals, all or some of $\theta, \mu_N, \sigma_N, \mu_G, \sigma_G$ could be regressed on predictors. More specifically, each of the parameters can be equal to $g(X_i \beta_i)$ where g is a link function, X some predictor(s) (design matrix) and β the regression coefficients.

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \text{Copula}\left(\begin{matrix} \text{Normal } N(\mu_N, \sigma_N) \\ \text{Gamma } GA\ (\mu_G, \sigma_G) \end{matrix} ; \theta \right)$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \text{Copula}\left(\begin{matrix} \text{Normal } N\big(g_{\mu_N}(X_i\beta_i), g_{\sigma_N}(X_k\beta_k)\big) \\ \text{Gamma } GA\ \big(g_{\mu_G}(X_j\beta_j), g_{\sigma_G}(X_l\beta_l)\big) \end{matrix} ; g_\theta(X_u\beta_u) \right)$$

Depending on which parameters need to be adjusted, the shape, location and dependence structure of the resulting predicted copula distribution can change making copula results very flexible with the ability to represent a wide variety of real-life scenarios.

The choice of which one of the parameters (including θ) will be allowed to vary by specific predictors should reflect expected or evident association patterns and/or knowledge about the context of the research question. If, for example, there is no evidence or reason to believe that the variability of the Normal marginal varies according to a predictor, then the framework allows for formal investigation of this notion and/or to allow σ_N to remain constant.

The optimum model may be selected via comparison of maximum likelihoods or BIC and AIC statistics.

4.3 Conditional Copula Bivariate Analogues of Centiles (CCBAC)

Conditioning statistical results on changing levels/values of confounding variables provides researchers with adjusted estimates that realistically reflect variability in a population.

Given the predicted copula distribution for a specific value of a confounding variable, it will be beneficial for the Conditional Copula Bivariate Analogues of Centiles

(CCBAC) to be computed. For example, in investigating the relationship between weight and height adjusted for age, the conditional copula BAC of a given centile level (i.e. 95, 90, 10, etc.) will inform researchers of the expected distribution for a subject of a given age. Using this adjustment, clinicians will be able to consult the CCBAC of height and weight for a child of a given age and hence compare their values to the general population used for the creation of the BACs.

In order for the above to be possible, simulated data will need to be drawn from a copula conditioned on a specific value of the confounder variable, followed by the application of the BAC algorithm introduced in Chapter 3. The list of steps below details the necessary procedure that would make the above possible:

- Exploration of a variety of copulas for a given dataset
- Choice of the best fitted copula based on numerical criteria such as the BIC
- Adjustment of the appropriate parameters of the fitted copula for covariates
- Prediction of the copula distribution for a set covariate value
- Simulation of bivariate data from the predicted copula distribution
- Implementation of BAC algorithm for a given centile level along with its resulting confidence envelope
- Draw conclusions regarding the predicted bivariate behaviour of a subject with regards to the given percentile values and the confidence region around them

For the remainder of this Chapter, the simulated data presented in Figure 4-1 from a Clayton copula with identical SHASH marginals will be extended to allow the incorporation of covariates, as per Jun Yan's 2006 paper [186].

Each of the SHASH parameters and the copula parameter, θ, will be investigated in turn. Each parameter will be modified by the addition of a parameter shift to reflect a subgroup of the data (such as a binary covariate). For the SHASH parameters, the shift will vary for Y_1 and Y_2. Hence, two additional model coefficients for changes in each of the SHASH parameters and one for the copula parameter will be estimated in turn in the fitting process.

The results of this fitting process are presented in a series of tables (Table 4-1 to Table 4-5) and graphs (Figure 4-4 to Figure 4-8) where the marginal and copula parameters are regressed in turn upon the levels of a simulated binary covariate. Estimates of the copula model parameters are presented with 95% confidence intervals (based on standard error calculations from the estimated Hessian matrix) along with contour plots for each fitted copula. Three BAC levels are presented in all graphs, 5th, 50th and 95th in blue, brown and purple, respectively.

The first section of each set of results in the following pages contains a table that explicitly describes the starting bivariate copula distribution which the presented data are simulated from. For example, in Table 4-1, the mean of each of the two marginals, changes by 0.994 and 1.04, respectively. The following equation represents an ordinary, non-adjusted copula function, whereas the second one represents the adjusted version of the former on μ with a dummy covariate, X: $\{0,1\}$.

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \text{Clayton}\begin{pmatrix} Y_1 \sim SHASH(\mu_1 = 20, \sigma_1 = 1.1, \nu_1 = 1.5, \tau_1 = 3) \\ Y_2 \sim SHASH(\mu_2 = 20, \sigma_2 = 1.1, \nu_2 = 1.5, \tau_2 = 3) \end{pmatrix}, \theta = 2.2\Bigg)$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \text{Clayton}\begin{pmatrix} Y_1 \sim SHASH(\mu_1 = 20 + \beta_{\mu_1} * X_{\{0,1\}}, \sigma_1 = 1.1, \nu_1 = 1.5, \tau_1 = 3) \\ Y_2 \sim SHASH(\mu_2 = 20 + \beta_{\mu_2} * X_{\{0,1\}}, \sigma_2 = 1.1, \nu_2 = 1.5, \tau_2 = 3) \end{pmatrix}, \theta = 2.2\Bigg)$$

The next part of each table lists the estimates of the copula distribution. The graphs presented at the end of each example in the following pages present:

- A scatterplot of the originally simulated values i.e. simulated values from a bivariate copula distribution whose parameters differ between the levels of a binary covariate by the selected factors
- The BAC contours as calculated for the originally simulated values split amongst the levels of the dummy variable representing the binary covariate in red and green points
- The fitted copula contours
- The CCBACs for each example. These are expected to match very closely with the BACs based on the originally simulated data but might highlight some adjustments/changes as a result of the fitting algorithm which updates all the parameters of the conditional copula via an optimisation procedure. The use of the CCBAC will be more informative in Chapter 5 as it will be the result of simulated data from the best fitted copula of an observed dataset.

Table 4-1 and Figure 4-4 display the results of the analysis described above for changes in the μ parameter which mirror a shift in the overall location of the data cloud amongst the levels of the binary covariate.

Table 4-2 and Figure 4-5Table 4-2: Conditional copula on $\boldsymbol{\sigma}$ – parameter estimates show changes in the σ parameter reflecting a narrower distribution for one of the two subgroups of data.

Table 4-3 and Figure 4-6 show an example of changes in the ν parameter which represent a shift in the shape of the lower tail of the bivariate distribution amongst the levels of the binary covariate.

Changes in the upper tail of the bivariate distribution are reflected via a shift in the τ parameter and the results of this analysis are presented in Table 4-4 and Figure 4-7.

Finally, Table 4-5 and Figure 4-8 show the results of possibly one of the most interesting aspects of the analysis described in this Chapter, as the copula dependence parameter, θ, is allowed to differ between the levels of the binary covariate. The correlation between the two response variables was defined to be a lot stronger in one of the subgroups compared to the other (green vs red points); change in θ by 11 units, $\tau = 0.52$ for the red points and $\tau = 0.87$ for the red points (as per the formula connecting the θ and τ for the Clayton copula, seen in Table 1-1).

In all examples, it is evident via the fitted copula contours and the proximity of the BAC and CCBAC contours that the conditional copula modelling procedure was able to capture the tailored shifts in each of the parameters successfully.

Table 4-1: Conditional copula on μ – parameter estimates

$$\binom{Y_1}{Y_2} \sim \text{Clayton}\left(\begin{matrix} Y_1 \sim SHASH\left(\mu_1 = 20 + \beta_{\mu_1} * X_{\{0,1\}}, \sigma_1 = 1.1, \nu_1 = 1.5, \tau_1 = 3\right) \\ Y_2 \sim SHASH\left(\mu_2 = 20 + \beta_{\mu_2} * X_{\{0,1\}}, \sigma_2 = 1.1, \nu_2 = 1.5, \tau_2 = 3\right) \end{matrix}, \theta = 2.2\right)$$

	Y_1 variable (change in μ_1 by 0.994)	Y_2 variable (change in μ_2 by 1.04)
μ	$20.02\ (20.00\ ,20.03)$	$20.02\ (20.01\ ,20.04)$
β_μ	$-0.09\ (-0.07\ ,-0.10)$	$1.01\ (0.99\ ,1.03)$
σ	$1.23\ (1.01\ ,1.50)$	$1.14\ (0.98\ ,1.32)$
ν	$1.62\ (1.39\ ,1.89)$	$1.51\ (1.35\ ,1.70)$
τ	$3.46\ (2.85\ ,4.19)$	$3.22\ (2.79\ ,3.72)$
θ	$2.21\ (2.11\ ,2.32)$	

Figure 4-4: Conditional copula on μ – BAC and copula contours

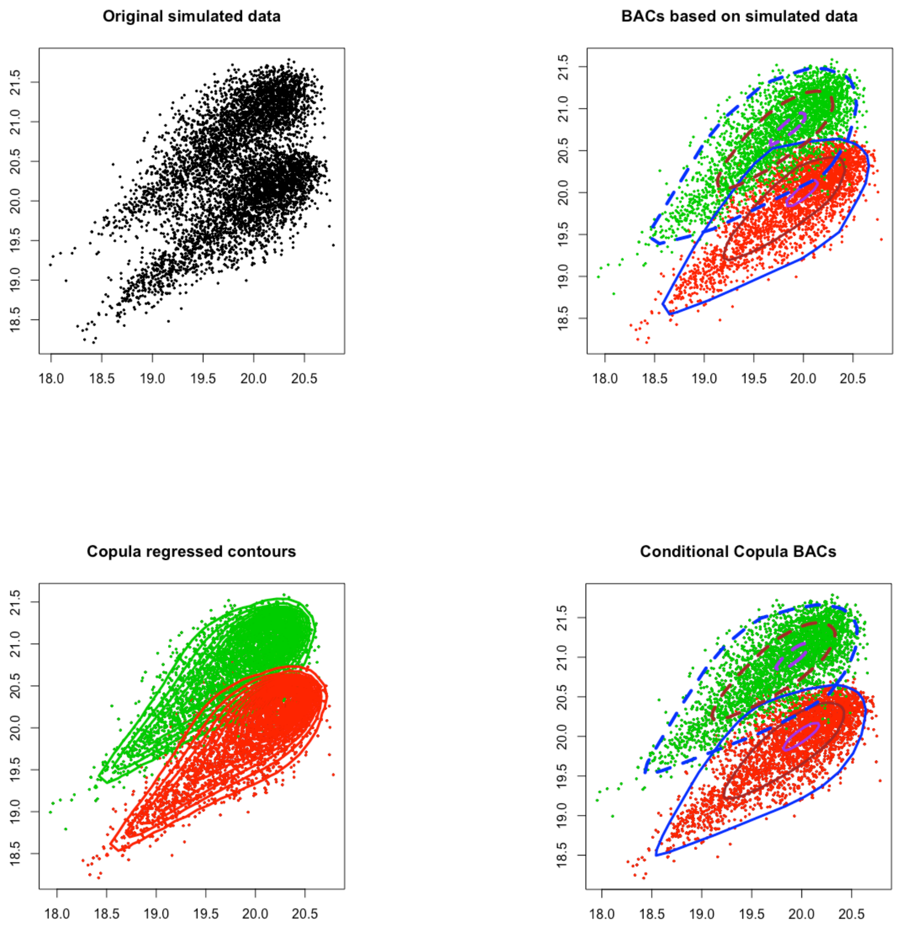


Table 4-2: Conditional copula on σ – parameter estimates

$\begin{pmatrix}Y_1\\Y_2\end{pmatrix} \sim \text{Clayton}\left(\begin{matrix}Y_1 \sim SHASH(\mu_1 = 20, \sigma_1 = 1.1 + \beta_{\sigma_1} * X_{\{0,1\}}, \nu_1 = 1.5, \tau_1 = 3)\\ Y_2 \sim SHASH(\mu_2 = 20, \sigma_2 = 1.1 + \beta_{\sigma_2} * X_{\{0,1\}}, \nu_2 = 1.5, \tau_2 = 3)\end{matrix}, \theta = 2.2\right)$		
	Y_1 variable (change in σ_1 by 1.2)	Y_2 variable (change in σ_2 by 1.3)
μ	20.01 (19.99 , 20.02)	20.01 (20.00 , 20.03)
σ	1.06 (0.92 , 1.22)	1.17 (1.01 , 1.36)
β_σ	0.22 (0.18 , 0.26)	0.36 (0.31 , 0.43)
ν	1.45 (1.30 , 1.61)	1.56 (1.39 , 1.75)
τ	2.93 (2.56 , 3.35)	3.26 (2.83 , 3.76)
θ	2.21 (2.10 , 2.31)	

Figure 4-5: Conditional copula on σ – BAC and copula contours

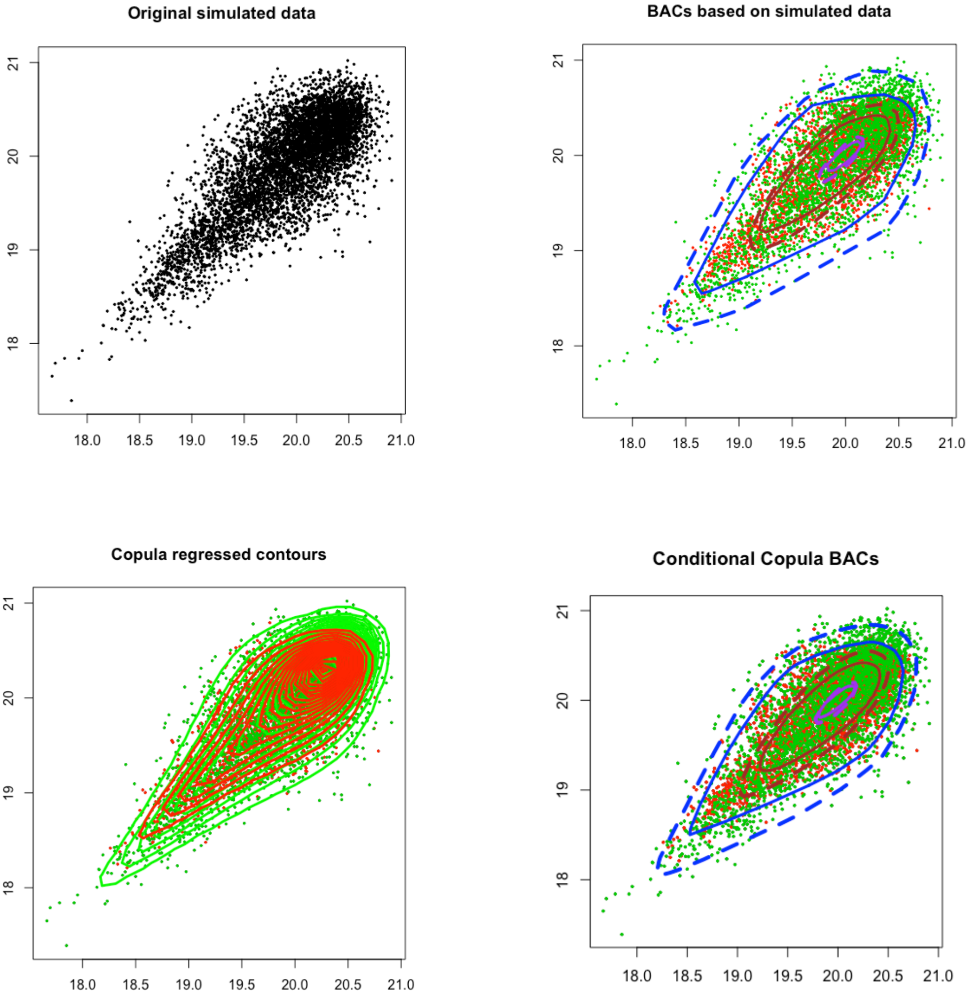
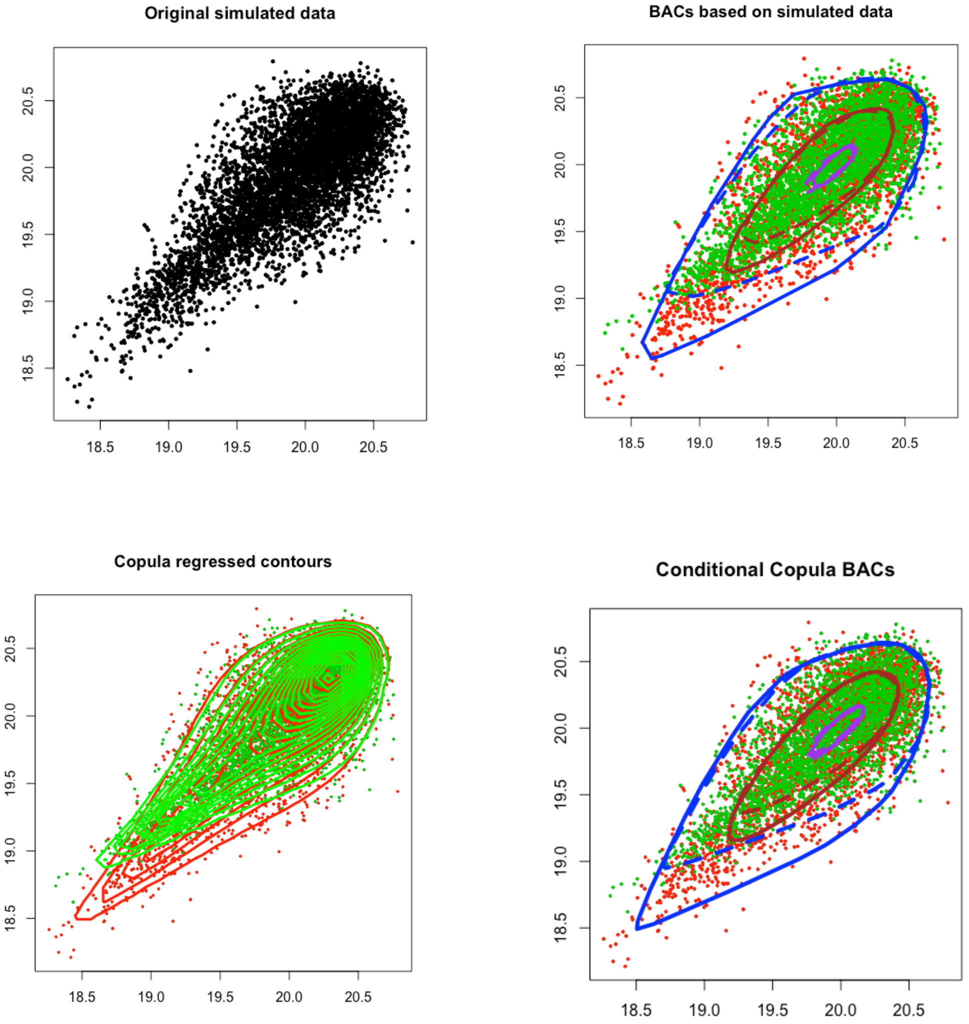


Table 4-3: Conditional copula on ν – parameter estimates

$\begin{pmatrix}Y_1\\Y_2\end{pmatrix} \sim \text{Clayton}\left(\begin{matrix}Y_1 \sim SHASH(\mu_1 = 20, \sigma_1 = 1.1, \nu_1 = 1.5 + \beta_{\nu_1} * X_{\{0,1\}}, \tau_1 = 3)\\ Y_2 \sim SHASH(\mu_2 = 20, \sigma_2 = 1.1, \nu_2 = 1.5 + \beta_{\nu_2} * X_{\{0,1\}}, \tau_2 = 3)\end{matrix}, \theta = 2.2\right)$		
	Y_1 variable (change in v_1 by 1.1)	Y_2 variable (change in v_2 by 1.3)
μ	20.00 (19.99 , 20.01)	20.01 (19.99 , 20.02)
σ	1.34 (0.97 , 1.86)	1.82 (1.03 , 3.22)
ν	1.72 (1.33 , 2.22)	2.21 (1.34 , 3.62)
β_ν	0.22(0.16 , 0.33)	0.82 (0.44 , 1.51)
τ	3.62 (2.67 , 4.93)	4.94 (2.83 , 8.60)
θ	2.31 (2.18 , 2.44)	

Figure 4-6: Conditional copula on ν – BAC and copula contours



Original simulated data

BACs based on simulated data

Copula regressed contours

Conditional Copula BACs

Table 4-4: Conditional copula on τ – parameter estimates

$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \text{Clayton}\left(\begin{matrix} Y_1 \sim SHASH(\mu_1 = 20, \sigma_1 = 1.1, \nu_1 = 1.5, \tau_1 = 3 + \beta_{\tau_1} * X_{\{0,1\}}) \\ Y_2 \sim SHASH(\mu_2 = 20, \sigma_2 = 1.1, \nu_2 = 1.5, \tau_2 = 3 + \beta_{\tau_2} * X_{\{0,1\}}) \end{matrix}, \theta = 2.2 \right)$		
	Y_1 variable (change in τ_1 by 1.5)	Y_2 variable (change in τ_2 by 1.5)
μ	19.95 (19.94 , 19.96)	19.97 (19.96 , 19.98)
σ	0.99 (0.82 , 1.18)	1.51 (1.14 , 2.01)
ν	1.39 (1.22 , 1.58)	1.92 (1.52 , 2.43)
τ	2.63 (2.25 , 3.09)	3.89 (2.97 , 5.11)
β_τ	1.49 (1.17 , 1.91)	1.71 (1.27 , 2.31)
θ	2.42 (2.30 , 2.54)	

Figure 4-7: Conditional copula on τ – BAC and copula contours

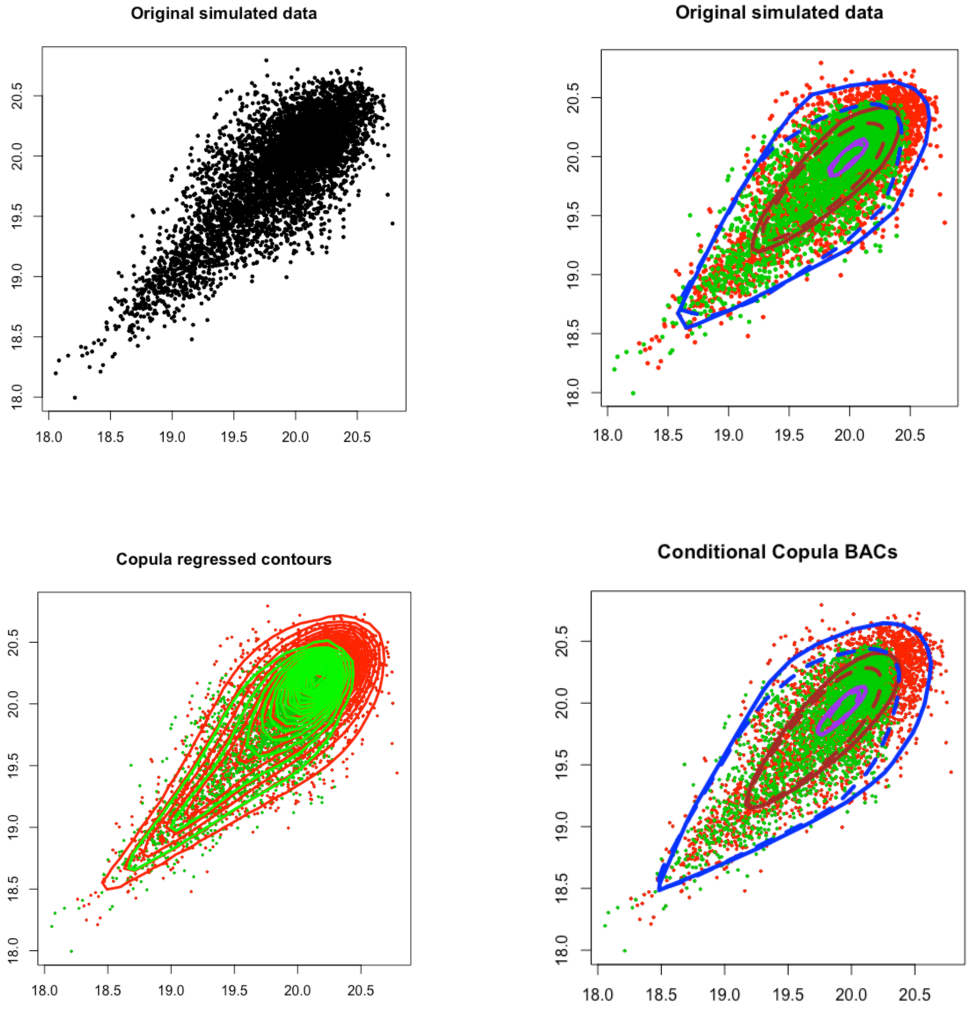
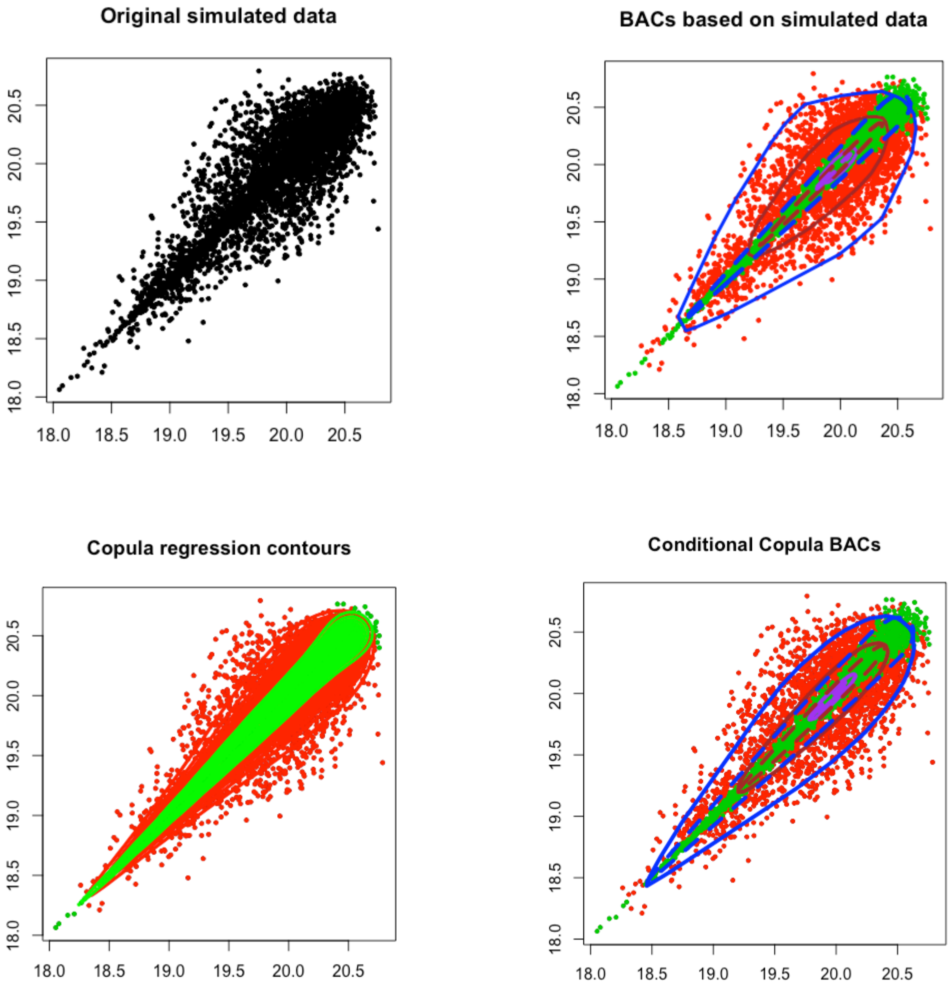


Table 4-5: Conditional copula on θ – parameter estimates

$\binom{Y_1}{Y_2}\sim\text{Clayton}\left(\begin{matrix}Y_1\sim SHASH(\mu_1=20,\sigma_1=1.1,\nu_1=1.5,\tau_1=3)\\Y_2\sim SHASH(\mu_1=20,\sigma_1=1.1,\nu_1=1.5,\tau_1=3)\end{matrix},\theta=2.2+\beta_\theta*X_{\{0,1\}}\right)$		
	Change in θ by 11 units	
	Y_1 variable	Y_2 variable
μ	20.02 (20.01, 20.35)	20.03 (20.01 , 20.04)
σ	1.08 (0.94 , 1.24)	1.12 (0.97 , 1.29)
ν	1.41 (1.28 , 1.57)	1.44 (1.30 , 1.61)
τ	3.10 (2.72 , 3.53)	3.22 (2.82 , 3.68)
θ	4.42 (1.23 , 4.60)	
β_θ	10.85 (5.40 , 17.32)	

Figure 4-8: Conditional copula on θ – BAC and copula contours



4.4 Conclusions

The univariate marginals can be incorporated into the calculation of BACs via the use of copula models and these can be further extended to account for numeric (example shown in Chapter 5) or categorical predictors (as shown earlier in this Chapter). The proposed algorithm equips researchers with the means of performing a straightforward comparison between outliers based on univariate vs bivariate criteria.

A search of the literature regarding 'bivariate centiles' highlighted a gap that the proposed algorithm for bivariate analogues of centiles, goes some way to addressing. This is via a copula which incorporates covariate-adjustment extensions via linear predictors in the marginals' and the copula parameter.

The performance of the proposed algorithm has been satisfactory in terms of run-time (produces results quicker than other existing algorithms, e.g. `isodepth` as seen in section 3.6). A follow-on step from here is to evaluate the identified outliers in terms of their clinical and scientific relevance.

The results of the fitting algorithm are such that they can reflect changes in all copula parameters (marginal and dependent). Small changes are likely to be indicative of clinical scenarios, for example, the right and left eyes might not be expected to be distributed much differently between healthy boys and girls. Relatively small changes were selected for the simulated examples and these were identified successfully via the proposed conditional copula optimisation algorithm.

5. Application

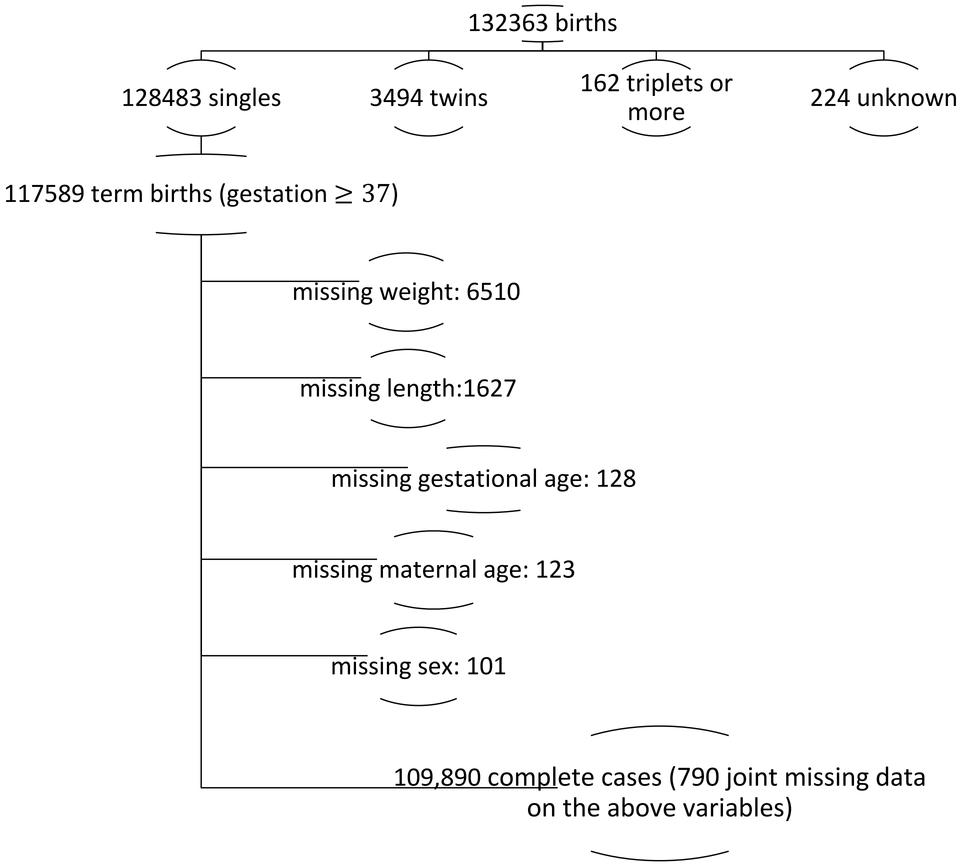
5.1 Introduction

This Chapter applies all the methods described in the earlier Chapters to a real dataset. All live births registered in Mexico City in 2017 (132,363) were obtained via the country's official national health open data resource. The dataset is freely available on the following website (accessed in June 2019):

https://datos.gob.mx/busca/dataset/nacimientos-ocurridos/resource/92e42070-e148-44c7-9ed7-917f4c3bb04f

My focus will be on the relationship between baby's birthweight (in kilograms) and length (in metres) and how this relationship varies according to maternal age, gestational age and baby's gender. The flowchart in Figure 5-1 summarises the dataset and prevalence of missing data. Complete data was available from 109,890 single and term births.

Figure 5-1: Flowchart with sample size changes for live birth data


132363 births
128483 singles
3494 twins
162 triplets or more
224 unknown
117589 term births (gestation ≥ 37)
missing weight: 6510
missing length:1627
missing gestational age: 128
missing maternal age: 123
missing sex: 101
109,890 complete cases (790 joint missing data on the above variables)

Birthweight was measured to the nearest gram and length to the nearest centimetre, hence there was some bunching of the data displayed despite underlying continuums (Figure 5-2). To counteract this, and any effect such bunching may have on the fitting algorithm, a very small amount of noise (Normal(0,0.03)) was added to both outcome variables (birthweight and length), with adequate variance to eliminate the effect of rounding but also not to cause to much distortion in the observed data.

Figure 5-2: Distributions of original birthweight and length variables

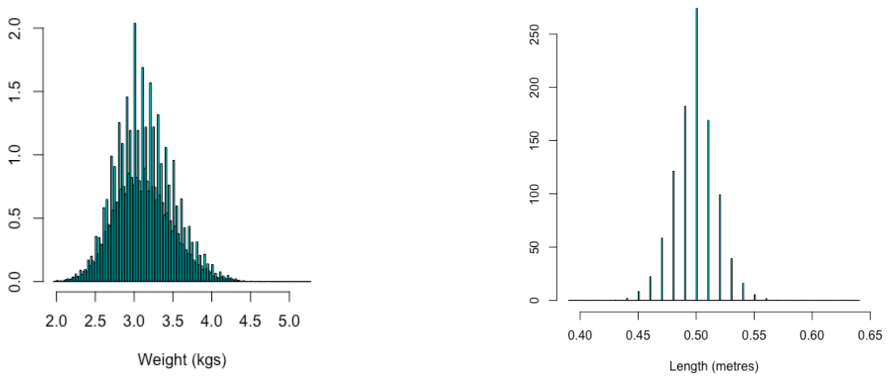


Figure 5-3 shows the scatterplot of the data with the empirical density of the marginal distributions of each response variable. The central lines forming a cross represent the median of each variable and they extend to one quartile on either side. Summary statistics are presented in Table 5.1. There is positive correlation between the two measures as indicated by all 3 types of correlation coefficients. The weight variable is slightly skewed to the right (skewness coefficient just above 0) and a bit more skewed than length, whose value is much closer to 0. Length has thinner tails compared to weight (since kurtotic coefficient is higher than 3). Mardia's statistics (based on a random sample of 5000 cases, as dictated by the capacity of the `R` function `mvn` [187], from the original dataset) imply incompatibility of the observed pattern with the bivariate Normal distribution, i.e. bivariate skewness and kurtosis significantly different to the 'target' Normal values. These significant results are dominated by the large sample size, but the fitting of the marginals presented in section 5.4 also follows the same direction, i.e. non-Normal marginals. More specifically, based purely on mathematical criteria (BIC) the best fitted marginals for weight and length respectively are GIG and BCPEo.

Figure 5-3: Bivariate scatterplot and initial marginal shapes

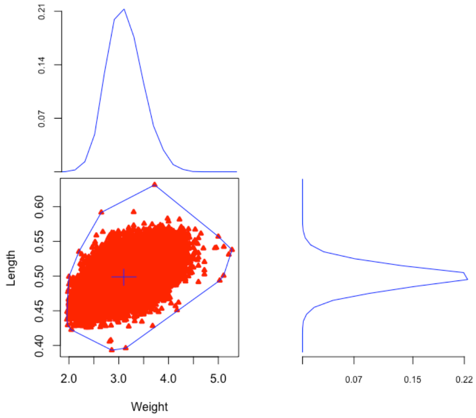
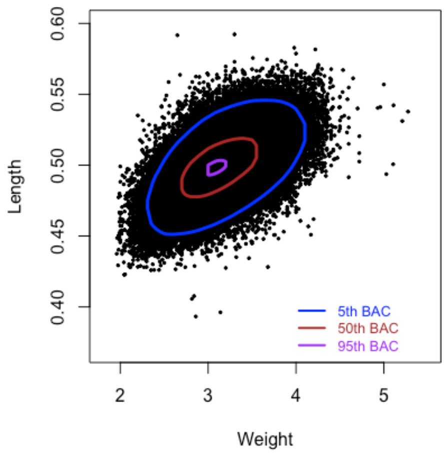


Table 5-1: Descriptive statistics for birthweight and length

	Weight	**Length**	**Correlation:**
Min, Max	1.96, 5.27	0.39, 0.63	Pearson: 0.564, 95% CI (0.560, 0.568) Spearman: 0.550, 95% CI (0.545, 0.554) Kendall: 0.39, 95% CI (0.37, 0.41)
Mean	3.12	0.50	
Median	3.10	0.50	
IQ Range	2.86, 3.35	0.49, 0.51	**Tail dependence:** Lower: 0.14, Upper: 0.10
SD	0.37	0.02	**Mardia's bivariate Normal statistics:**
Skewness	0.30	−0.01	127.9 ($p < 0.001$))
Kurtosis	3.09	3.44	9.8 ($p < 0.001$)

According to the algorithm introduced in Chapter 3, there are 1536 BACs within the original dataset. The 5th, 50th and 95th BACs are shown in Figure 5-4 in blue, brown and purple, respectively. Exact percentage coverage for these levels is 4.99%, 49.98% and 95.00%, respectively.

Figure 5-4: 5th, 50th, 95th BACs for birthweight and length



The graphs in Figure 5-5 show all bivariate extremes (on or outside the 5th BAC) in orange on the first graph and these are subsequently superimposed by different extreme classifications as noted in the subtitles of each graph in brown, grey, green and red, respectively. The 'False positives' subgraph in Figure 5-5 shows in green and blue the univariate extremes for weight and length respectively (below the 2.5th or above the 97.5th univariate centiles) that are not classed as extreme bivariately according to the BAC algorithm (potential false positives). There are 2638 cases for weight, 2599 cases for length and 237 shared cases (shown in pink), adding up to 5000 (4.55%) unique false positive extreme results. There are 920 cases (0.84%) that would have been considered as 'normal' according to their birthweight and length (Table 5-2), but in fact when the bivariate association of these two variables is taken into account, they are flagged up as unusual (hidden extremes, shown as red in the 'Hidden extremes' subgraph). The last graph of Figure 5-5 reaffirms the inverted-U shape formed between univariate centiles and their bivariate analogues as discussed in Chapter 3.

Figure 5-5: Bivariate outliers, hidden and consistent extremes for weight and length

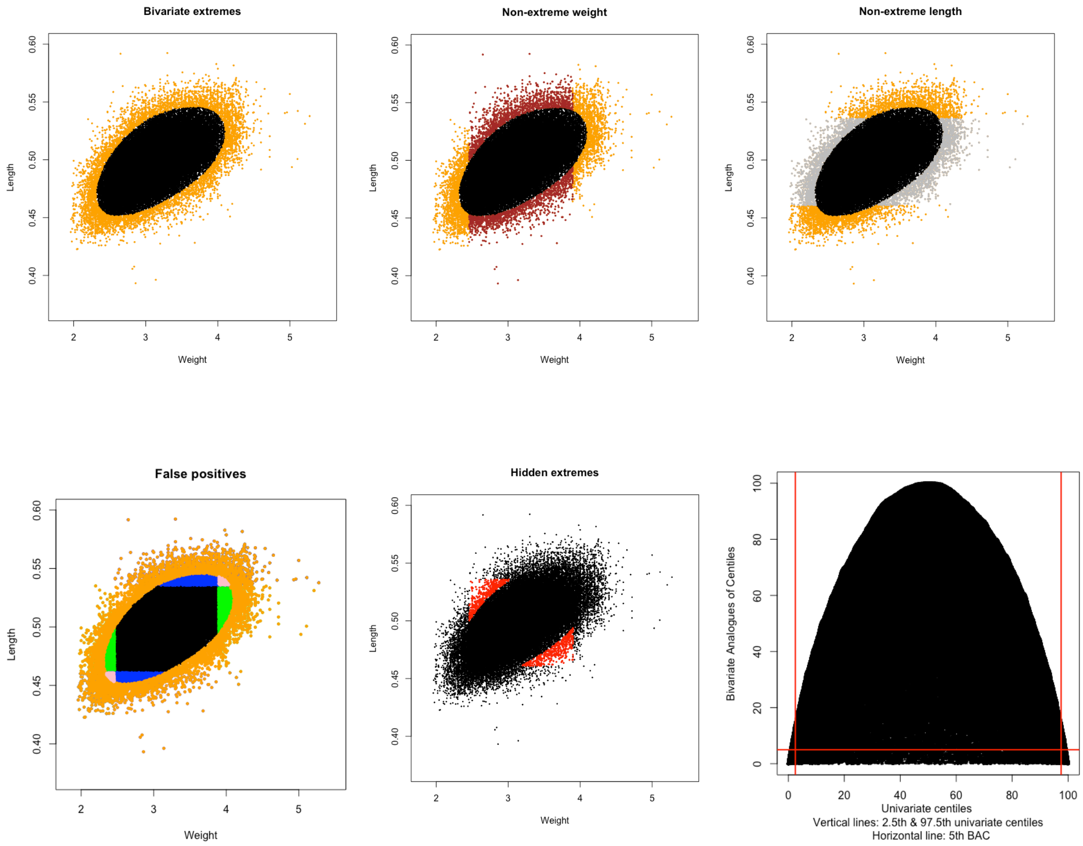


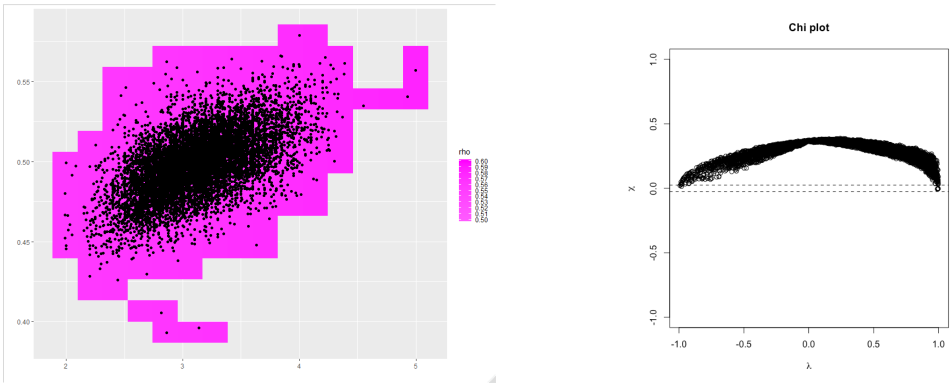
Table 5-2: Number (%) of cases classified as non-normal based on univariate and bivariate criteria

Univariate – below the 2.5th and above the 97.5th centiles		BAC (Weight, Length) – outside 5th convex hull centile	Number (%) of cases
Weight	Length		
☑			5495 (5.00%)
	☑		5495 (5.00%)
		☑	5488 (4.99%)
☑		☑	2857 (2.60%)
	☑	☑	2896 (2.63%)
☑	☑		1422 (1.29%)
☑	☑	☑	**1185 (1.08%)**
CONSISTENT EXTREMES			
☒	☒	☑	**920 (0.84%)**
HIDDEN EXTREMES			

5.2 Local dependence

This section includes the local dependence map and chi plot as described in Chapter 2. Due to the large size of the original dataset, neither of the two `R` functions (`localgauss`[68] and `chi.plot`[188]) were able to produce results for the whole dataset. The graphs in Figure 5-6 present the resulting plots of a random subsample of $10{,}000$ cases from the original dataset. There are no reasons to believe that the equivalent plots for the entire dataset would not follow a similar pattern; this is of positive correlation throughout the entire range of the association between weight and length as indicated by both plots. The linear correlation according to the local dependence map spans values from 0.5 to 0.6 and the chi-plot shows in more detail the changes in the correlation strength. It varies from no correlation for few pairs of points (close to/on the horizontal line) to weak (stronger) correlation on the left (right) above the horizontal line.

Figure 5-6: Local dependence plots



Local dependence map

5.3 Covariate-related BACs

This section explores the descriptive characteristics of the 3 covariates mentioned at the start of this Chapter (sex – binary, gestational age – discrete numeric and maternal age – continuous) and demonstrates the flexibility of the proposed BAC algorithm for a dataset over 100,000 cases long.

Sex

Figure 5-7 shows the bivariate distribution between weight and length whilst accounting for the baby's sex (coded as 0 for boys and 1 for girls). There were 57,031 (51.90%) boys and 52,859 (48.10%) girls. The overall pattern of the association of weight and length appears to be similar for boys and girls and this is evident from the BACs in Figure 5-8 too. There is minimal change in the length values between the two sexes. There were 939 and 902 BACs obtained from the data for boys and girls with 0.84% and 0.79% hidden extremes, respectively, as seen in Table 5-3.

Table 5-3: BAC results by sex

Sex	No. of BAC levels	Exact BAC coverage			Hidden extremes
		5%	50%	95%	
Girls	939	5.03	50.07	94.97	480 (0.84%)
Boys	902	5.05	50.05	95.02	418 (0.79%)

Figure 5-7: Relationship between birthweight and length for boys and girls

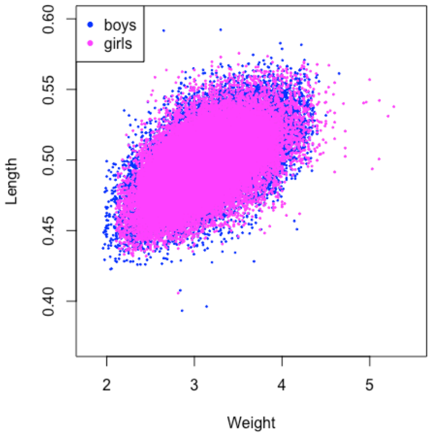
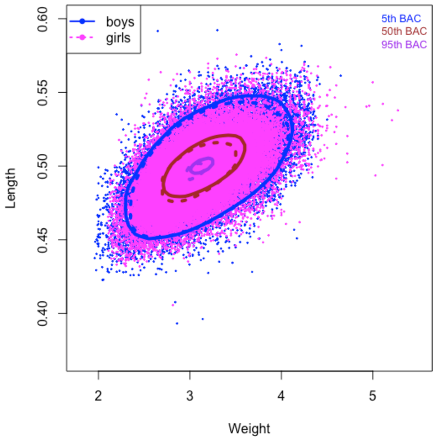


Figure 5-8: Sex-related BACs between birthweight and length



Gestational age

The two graphs in Figure 5-9 show the individual relationship between each of the response variables and gestational age (Table 5-4). There is an increase in both weight and length for higher gestational ages with minimal change in the variation within each gestation group, with the exceptions of weeks 40 and 42 for weight and week 42 for length.

Table 5-4: Distribution of gestational ages

Gestation	37	38	39	40	41	42	43
Frequency	11849 (11%)	27245 (25%)	30548 (28%)	31350 (29%)	7310 (7%)	1568 (1%)	20 (0.002%)

Figure 5-9: Gestational age vs birthweight and length

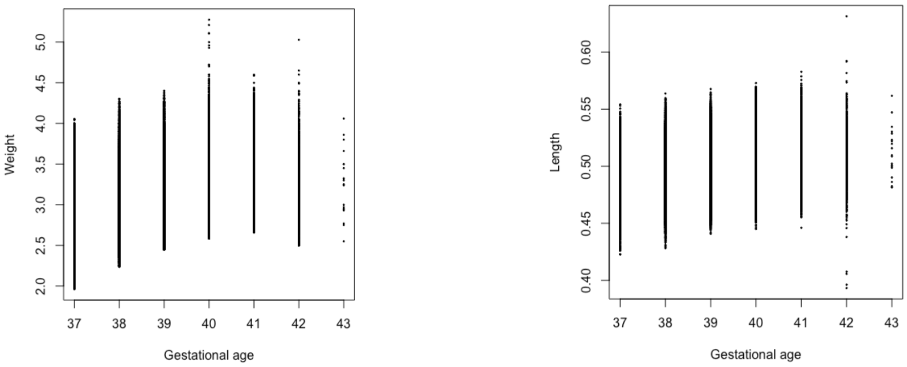


Figure 5-10 introduces the gestational age into the bivariate distribution, where different colours represent different gestational ages, from completed week 37 to completed week 43. The concurrent increase in weight and length is apparent with the increase in gestational age.

The vertical lining visible on the lower bound for weight within each gestational age group is the result of the original rounded values of weight despite the addition of random noise. It was never intended to remove the original bunching entirely, but instead to dilute it to eliminate potential issues with optimisation algorithms used in the later sections of this Chapter.

Figure 5-10: Birthweight and length by gestational age

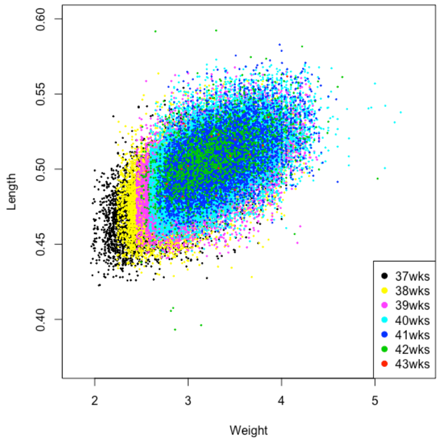


Figure 5-11 displays the fitted BACs according to gestational age. The same colour has been used for all 3 levels (5th, 50th and 95th from the outer to the inner, respectively). There are no BAC convex hulls drawn for the oldest gestational group of 43 weeks as it contained just 20 data points and yielded only 3 BACs, capturing 45%, 75% and all of the bivariate data, which are not comparable to the 5th, 50th and 95th levels of the other gestation groups, hence not plotted. The next oldest gestational age of 42 weeks is the most variable group, where its 5th BAC reaches all the way down to the 5th BAC of the 39 weeks' gestation group.

Figure 5-11: Gestational age-related BACs between birthweight and length

Weeks	**37**	**38**	**39**	**40**	**41**	**42**	**43**
BACs	297	556	595	614	213	74	3

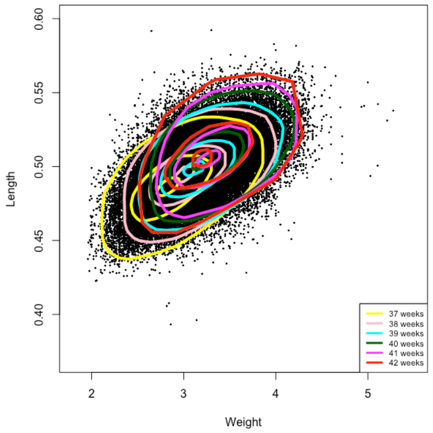


Table 5-5 presents the results of the BAC algorithm in terms of total number of percentile levels identified in this dataset, exact coverage for the commonly used levels of 5, 50 and 95% and the frequency of potential false and hidden extremes within each gestational group. The algorithm estimates the presented percentile levels closely and, as expected, it does better with larger sample size.

Table 5-5: BAC results by gestational age

Gestational age	Sample size	No. of BAC levels	Exact BAC coverage			Hidden extremes
			5%	50%	95%	
37 weeks	11849	297	4.87	49.91	94.98	85 (0.72%)
38 weeks	27245	556	5.05	49.92	94.95	179 (0.62%)
39 weeks	30548	595	4.98	50.02	94.98	213 (0.70%)
40 weeks	31350	614	4.94	50.00	94.96	233 (0.70%)
41 weeks	7310	213	4.95	49.90	94.95	45 (0.61%)
42 weeks	1568	74	5.16	49.55	95.41	9 (0.57%)

Maternal age

Maternal age was reported to the closest year, but the dataset also contained date of birth of each child and date of mother's birth, so these were used to give more accurate values. The IQR spanned ages 22 to 32 with minimum age of 9 and maximum of 55 years. Figure 5-12 shows the relationship between maternal age and birthweight in purple and length in grey; there is no pattern emerging between weight or length and maternal age. This is evident in the BACs shown in Figure 5-14 too; these are calculated for each quartile boundary of maternal age, i.e. ages below 9.01, between 21.96 and 26.74, between 26.74 and 32.09 and above 32.09. The results from the 4 different groups are very close for each BAC level, indicating minimal differences in the joint behaviour of weight and length by maternal age. This message is reinforced by the results in Table 5-6, which change very slightly between each quartile range of maternal ages.

Figure 5-12: Birthweight (in purple) and length (in grey) by maternal age

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Maternal age	9.01	21.96	26.74	27.25	32.09	55.19

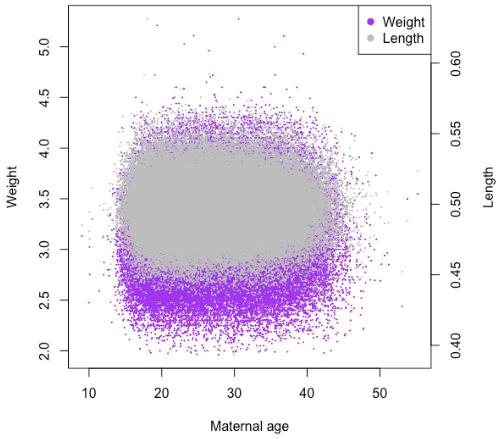


Figure 5-13: Maternal age-related BACs for birthweight and length

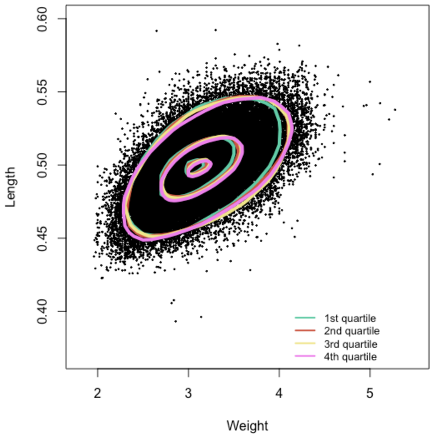


Table 5-6: BAC results by maternal age

Maternal age (years)	Sample size	No. of BAC levels	Exact BAC coverage			Hidden extremes
			5%	50%	95%	
$\leq \mathbf{21.96}$	27480	569	5.08	49.89	95.04	216
$> \mathbf{21.96}$ & $\leq \mathbf{26.74}$	27472	564	4.97	49.93	94.94	242
$> \mathbf{26.74}$ & $\leq \mathbf{32.09}$	27469	551	4.95	49.99	94.96	207
$> \mathbf{32.09}$	27469	538	5.07	49.93	95.07	248

Summary

The exploratory analysis conducted in this section has been explicitly data driven and applied to the entirety of the original dataset; except a small number of results which were based on a random subsample from the original dataset; the multivariate Normal characteristics, the local dependence map and the chi plot. The results have been informative in terms of emerging patterns between the bivariate outcome of weight and length and the 3 covariates and has allowed me to identify hidden extremes for each covariate level in a non-parametric way. The covariates were selected to cover a range of data types: binary, discrete and continuous numeric to illustrate the fitting of each.

The next sections focus on fitting copulas after deciding on the best fitted marginal distributions for weight and length. The copula will then be extended to incorporate the 3 covariates and copula-based BACs for each covariate will be produced. The ultimate aim is to be able to generate an adjusted BAC (CCBAC) for children with specific characteristics, for example, for boys born at 38 weeks whose mothers were 25 years old.

Despite the fact that the BAC algorithm has performed well so far for the entirety of the live-births dataset, i.e. produced results fast (0-10 seconds in a computer with processor characteristics: 2.5 GHz Intel Core i7 and memory: 16 GB 1600 MHz DDR3), this is not the case for the optimisation algorithm used for the conditional copula models. Therefore, to facilitate computation for the remainder of this Chapter, a random subsample of 20,000 cases from the original dataset will be analysed. All the results shown in the coming sections relate to this smaller unbiased selection from the original sample.

5.4 Marginal distributions

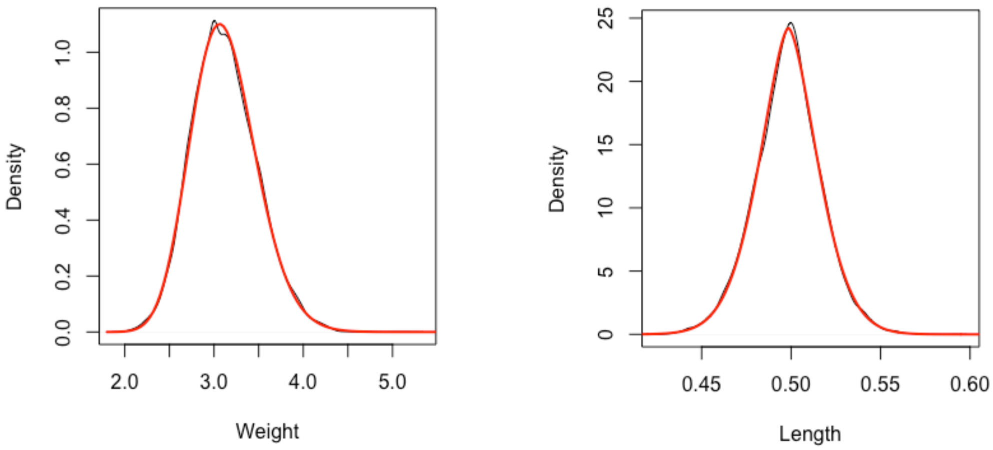
The `gamlss` library [185] in `R` was utilised for the investigation of the best fitted distribution for each of the two variables. Table 5-7 shows the BIC results of several univariate marginals. Weight was found to be best described with a Generalised Inverse Gamma distribution, GIG $(\mu = \exp(1.14) = 3.13, \sigma = \exp(-2.12) = 0.12, \nu = 21.86)$; In link function for μ and σ (identity for ν). Length was best approximated by a Box-Cox Power Exponential-original distribution, BCPEo $(\mu = \exp(-0.70) = 0.50\,, \sigma = \exp(-3.30) = 0.04, \nu = 1.20, \tau = \exp(0.5) = 1.65)$; In link functions for μ, σ and τ (identity for ν). The fitted marginals are shown in Figure 5-14, where the red lines represent the density of the best fitting marginals superimposed over the true density of each rv.

Table 5-7: Marginal distribution comparison [ranks]

Distribution	BIC – Weight	BIC – Length
Generalised Inverse Gamma	16306 [1]	−102791 [3]
SHASH	16309 [2]	−102446 [4]
Box Cox Power Exponential original	16318 [3]	−102951 [1]
Normal	16628 [4]	−102794 [2]
Gumbel	22039 [5]	−98972[5]

There is minimal change in the fit of each of these marginals on the two response variables. The choice of the best fitted marginal is purely based on the BIC. The GIG has three parameters and is able to capture the slight skewness of weight to the right satisfactorily. The BCPEo has four parameters and is able to capture the kurtotic pattern in length, i.e. thinner tails on both sides.

Figure 5-14: Fitted marginal distributions for birthweight and length



5.5 Copulas

Using the marginal distributions found to best fit the data in Section 5.4, copulas of varying shapes were considered. The investigations in sections 5.1 and 5.2 did not support the bivariate Normal distributional assumption, whose marginals are univariate Normal, but the overall pattern of the bivariate association has thus far not deviated from a symmetric shape. The Normal copula with GIG and BCPEo marginals for weight and length respectively is found to be the best fitted copula according to the comparison of BIC values, Table 5-8 yielding an elliptical copula.

Table 5-8: Comparison of copula functions

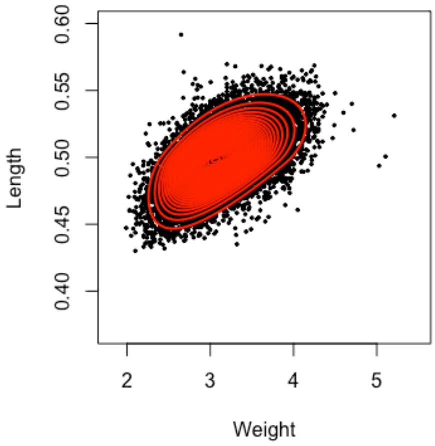
Copula:	**BIC:**
Normal	-94380
Frank	-93831
t-EV	-93741
Gumbel	-93733
AMH	-93349
Clayton	-93020

The updated complete copula function can now be written as follows and the fitted contours are shown in Figure 5-15:

Equation 5-1:

$$\begin{pmatrix}\text{Weight}\\ \text{Length}\end{pmatrix} \sim \text{Normal}\begin{pmatrix}\text{GIG}(\mu = 3.12, \sigma = 0.12, \nu = -1.57)\\ \text{BCPEo}(\mu = 0.50, \sigma = 0.04, \nu = 1.00, \tau = 1.63)\end{pmatrix}; \theta = 0.6\Bigg)$$

Figure 5-15: Contours of best fitted copula for birthweight and length

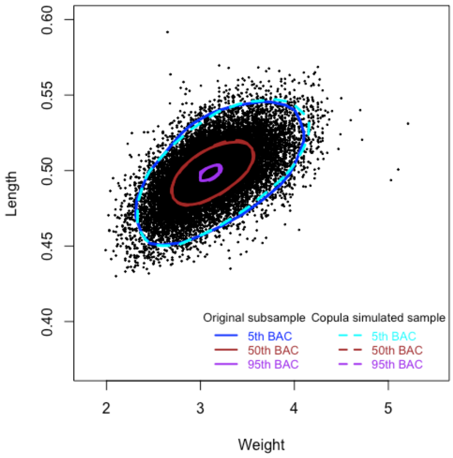


The Normal copula is able to capture the symmetric ellipsoid nature of the bivariate pattern.

5.6 Copula BACs (CBAC - unadjusted)

Simulated data are drawn from a Normal copula with GIG and BCPEo marginal distributions with the estimated parameters as shown in Equation 5-1. Figure 5-16 shows the correspondence between the BACs of the original subsample (dashed lines) and the simulated data. The difference in the outer BAC is most noticeable, not surprisingly as the edges of any distribution are the most difficult to estimate; nevertheless, the 50th and 95th BACs virtually coincide.

Figure 5-16: Copula BACs for birthweight and length



5.7 Conditional Copula BACs

The covariate-related BACs seen earlier are now enhanced via copula regression modelling of the location parameter μ of both marginals.

The location parameter for the GIG and BCPEo marginals is conditioned on sex, gestational age, maternal age and their combination. Hence, four conditional copula models are fitted. Each set of results shown in the following pages contains tables of the estimated copula parameters and the frequency of false and hidden extremes based on the BAC results as well as graphical displays of the conditional copula BACs.

The starting point for each of the conditional copulas is the copula model detailed in Equation 5-1.

Table 5-9 and Figure 5-17 show minimal effect of sex on the bivariate distribution of weight and length. The drawn BACs virtually coincide and so does the prevalence of potential false and hidden extremes, i.e. the sex coefficient for both marginal parameters is very small. The parameter estimates for the sex variable are both significant, although confidence intervals are narrow with both limits close to zero.

Table 5-10 and Figure 5-18 show small changes in the location parameter for each gestational age group and a similar pattern like that seen in Figure 5-11.

Table 5-11 and Figure 5-19 show the change in the location parameter by changes in maternal age. The estimates and confidence intervals of the coefficient for the location parameter both for birthweight and length is very close to 0.

Finally, Table 5-12 shows the results of the conditional copula model with multiple covariates. The two sets of convex hulls in Figure 5-20 display the predicted 5th, 50th and 95th BACs for a boy and girl, respectively, born at 37 weeks with maternal age at 25 years old.

Figure 5-21 and Figure 5-22 show the 5th BACs for girls born at 37 and 41 weeks, respectively, with the same maternal age (32 years old) along with a bootstrap envelope based on 500 simulations. New born girls, whose mothers' age is equal to 32 years, with observed weight and length laying outside the relevant 5th BAC, depending on their gestational age, would be classed as unusual and might require targeted interventions or follow up.

The conditional copula BACs with a single covariate have been found to be very close to their equivalent covariate-related BACs, illustrating the appropriate fit of this model form. The conditional copula BACs with multiple covariates has provided a direct extension of univariate centiles for varying levels/values of several predictors.

Table 5-9: Conditional Copula BAC results for sex

$$\binom{\text{Weight}}{\text{Length}} \sim \text{Normal}\left(\begin{matrix}\text{GIG}(\boldsymbol{\mu = 3.12 + \beta_{\mu_1} * \textbf{Sex}_{\{0,1:girls\}}, \sigma = 0.12, \nu = -1.6})\\ \text{BCPEo}(\boldsymbol{\mu = 0.50 + \beta_{\mu_2} * \textbf{Sex}_{\{0,1:girls\}}, \sigma = 0.04, \nu = 1.00, \tau = 1.63})\end{matrix}; \boldsymbol{\theta = 0.6}\right)$$

	Birthweight	Length
$\boldsymbol{\mu}$	$3.12\ (3.11, 3.13)$	$0.499\ (0.498, 0.500)$
$\boldsymbol{\beta_{\mu}}$	$0.2 \times 10^{-4}\ (0.8 \times 10^{-6}, 0.01)$	$0.2\ \times 10^{-5}(0.2\ x\ 10^{-6}, 0.5\ \times 10^{-3})$
$\boldsymbol{\sigma}$	$0.118\ (0.117, 0.119)$	$0.037\ (0.036, 0.038)$
$\boldsymbol{\nu}$	$1.88\ (1.21, 2.13)$	$0.98\ (0.68, 1.28)$
$\boldsymbol{\tau}$		$1.64\ (1.59, 1.68)$
$\boldsymbol{\theta}$	$0.57\ (0.56, 0.58)$	
BIC	-94406.61	

Figure 5-17: Conditional Copula BAC graph for sex

Sex	Girls	Boys
Hidden extremes	168 (0.84%)	179 (0.89%)
False extremes	971 (4.58%)	963 (4.81%)

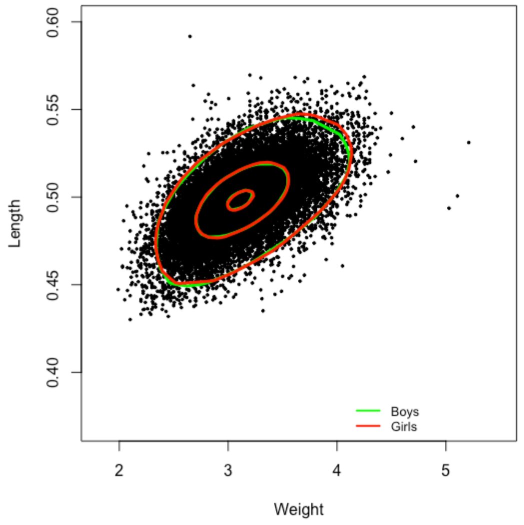


Table 5-10: Conditional Copula BAC results for gestational age

$$\binom{\text{Weight}}{\text{Length}} \sim \text{Normal}\left(\begin{matrix} \text{GIG}(\boldsymbol{\mu} = \mathbf{3.12} + \boldsymbol{\beta}_{\boldsymbol{\mu}_1} * \textbf{GestAge}, \boldsymbol{\sigma} = \mathbf{0.12}, \boldsymbol{\nu} = \mathbf{-1.6}) \\ \text{BCPEo}(\boldsymbol{\mu} = \mathbf{0.50} + \boldsymbol{\beta}_{\boldsymbol{\mu}_2} * \textbf{GestAge}, \boldsymbol{\sigma} = \mathbf{0.04}, \boldsymbol{\nu} = \mathbf{1.00}, \boldsymbol{\tau} = \mathbf{1.63}) \end{matrix}; \boldsymbol{\theta} = \mathbf{0.6}\right)$$

	Birthweight	**Length**
$\boldsymbol{\mu}$	$0.68\ (0.60, 0.76)$	$0.36\ (0.35, 0.37)$
$\boldsymbol{\beta_\mu}$	$0.063\ (0.061, 0.065)$	$0.003\ (0.003, 0.004)$
$\boldsymbol{\sigma}$	$0.112\ (0.111, 0.113)$	$0.036\ (0.035, 0.036)$
$\boldsymbol{\nu}$	$0.31(0.20, 0.64)$	$0.89\ (0.57, 1.20)$
$\boldsymbol{\tau}$		$1.65\ (1.61, 1.70)$
$\boldsymbol{\theta}$	$0.53\ (0.52,0.54)$	
BIC	-96875.48	

Figure 5-18: Conditional Copula BAC graph for gestational age

Weeks	**37**	**38**	**39**	**40**	**41**	**42**
Hidden extremes	154 (0.77%)	160 (0.80%)	152 (0.76%)	151 (0.75%)	174 (0.87%)	174 (0.87%)
False extremes	1004 (5.02%)	983 (4.91%)	1005 (5.02%)	985 (4.92%)	979 (4.89%)	984 (4.92%)

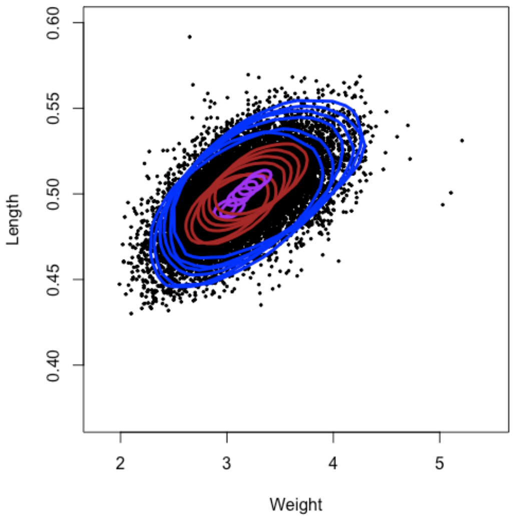


Table 5-11: Conditional Copula BAC results for maternal age

$\begin{pmatrix}\text{Weight}\\ \text{Length}\end{pmatrix} \sim\text{Normal}\left(\begin{matrix}\text{GIG}(\boldsymbol{\mu} = \mathbf{3.12} + \boldsymbol{\beta}_{\mu_1} * \textbf{MaternAge}, \boldsymbol{\sigma} = \mathbf{0.12}, \boldsymbol{\nu} = -\mathbf{1.6})\\ \text{BCPEo}(\boldsymbol{\mu} = \mathbf{0.50} + \boldsymbol{\beta}_{\mu_2} * \textbf{MaternAge}, \boldsymbol{\sigma} = \mathbf{0.04}, \boldsymbol{\nu} = \mathbf{1.00}, \boldsymbol{\tau} = \mathbf{1.63})\end{matrix}; \boldsymbol{\theta} = \mathbf{0.6}\right)$		
	Birthweight	**Length**
$\boldsymbol{\mu}$	$3.04\ (3.02, 3.05)$	$0.498\ (0.497, 0.499)$
$\boldsymbol{\beta_\mu}$	$0.003\ (0.003, 0.004)$	$0.3 \times 10^{-5}(0.2 \times 10^{-6}, 0.5\ \times 10^{-4})$
$\boldsymbol{\sigma}$	$0.118\ (0.116, 0.119)$	$0.037\ (0.036, 0.038)$
$\boldsymbol{\nu}$	$0.345\ (0.27, 3.42)$	$0.95\ (0.65, 1.25)$
$\boldsymbol{\tau}$		$1.63\ (1.59,1.68)$
$\boldsymbol{\theta}$	$0.57\ (0.56, 0.58)$	
BIC	-94500.95	

Figure 5-19: Conditional Copula BAC graph for maternal age

Age centiles	**5th**	**25th**	**50th**	**75th**	**95th**
Hidden extremes	179 (0.89%)	162 (0.81%)	168 (0.84%)	156 (0.78%)	186 (0.93%)
False extremes	964 (4.82%)	975 (4.87%)	962 (4.81%)	979 (4.89%)	957 (4.78%)

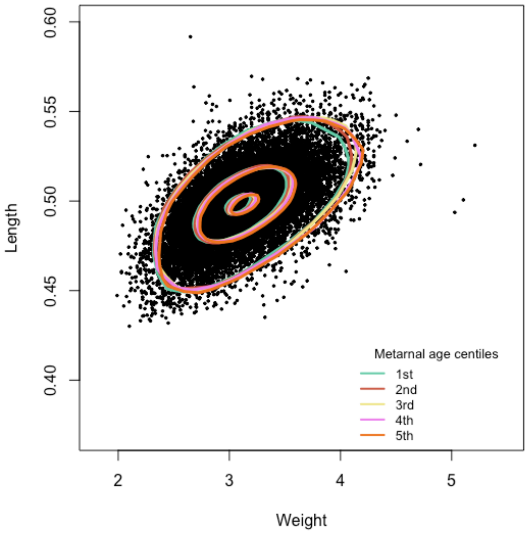


Table 5-12: Conditional Copula BAC results for sex, gestational age (GA) and maternal age (MA)

$$\binom{\text{Weight}}{\text{Length}} \sim \text{Normal}\left(\begin{matrix}\text{GIG}\left(\boldsymbol{\mu = 3.12 + \beta^{S}_{\mu_1} * \text{Sex} + \beta^{GA}_{\mu_1} * \text{GA} + \beta^{MA}_{\mu_1} * \text{MA}, \sigma = 0.12, \nu = -1.6}\right) \\ \text{BCPEo}\left(\boldsymbol{\mu = 0.50 + \beta^{S}_{\mu_2} * \text{Sex} + \beta^{GA}_{\mu_2} * \text{GA} + \beta^{MA}_{\mu_2} * \text{MA}, \sigma = 0.04, \nu = 1.00, \tau = 1.63}\right)\end{matrix}; \boldsymbol{\theta = 0.6}\right)$$

	Birthweight	Length
$\boldsymbol{\mu}$	$0.44\ (0.35, 0.55)$	$0.409\ (0.402, 0.415)$
$\boldsymbol{\beta^{S}_{\mu}}$	$0.10\ (0.09, 0.11)$	$0.0046\ (0.0039, 0.0053)$
$\boldsymbol{\beta^{GA}_{\mu}}$	$0.061\ 0.059\ 0.064$	$0.0020\ (0.0018, 0.0022)$
$\boldsymbol{\beta^{MA}_{\mu}}$	$0.009\ (0.008, 0.011)$	$0.0004\ (0.0003, 0.0005)$
$\boldsymbol{\sigma}$	$0.114\ (0.112, 0.116)$	$0.037\ (0.036, 0.038)$
$\boldsymbol{\nu}$	$-0.437\ (-0.8, -0.2 \times 10^{-5})$	$4.31\ (3.59, 5.02)$
$\boldsymbol{\tau}$		$1.57\ (1.48, 1.68)$
$\boldsymbol{\theta}$	$0.52\ (0.50, 0.55)$	
BIC	-23555.17	

Figure 5-20: Conditional Copula BAC graph for sex, gestational age and maternal age

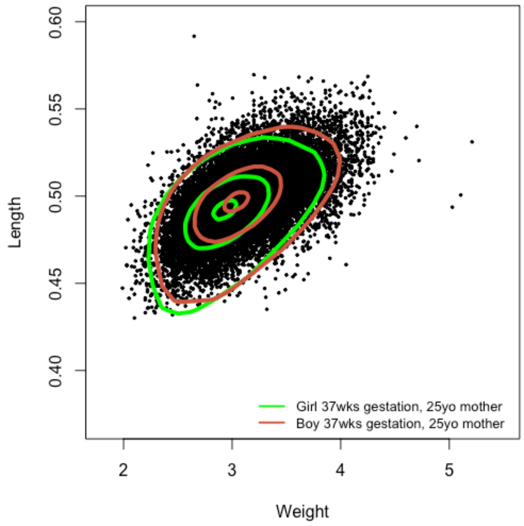


Figure 5-21: Conditional Copula BAC graph with bootstrap envelope, 38 weeks gestation

Girls born at 38 weeks and maternal age equal to 31 years old

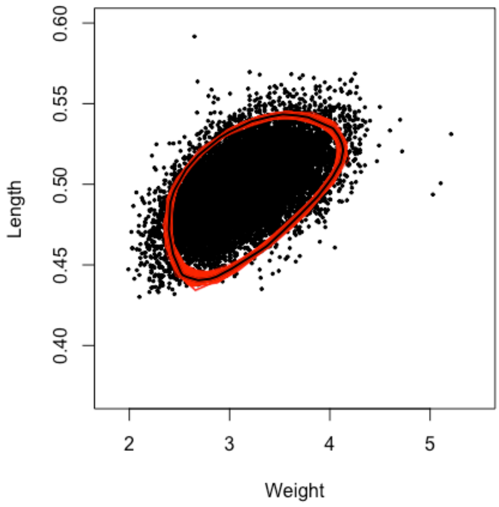
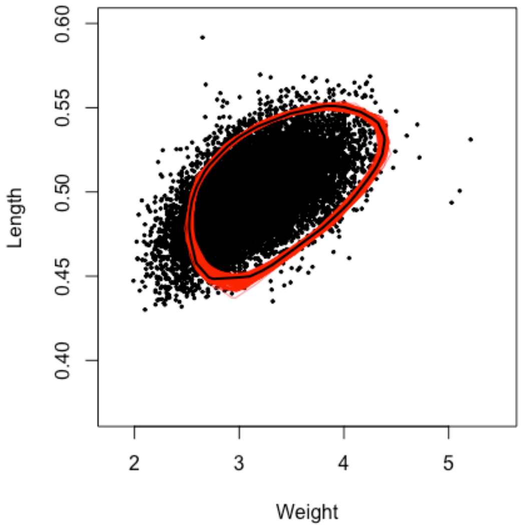


Figure 5-22: Conditional Copula BAC graph with bootstrap envelope, 41 weeks gestation

Girls born at 41 weeks and maternal age equal to 31 years old



5.8 Conclusions

The techniques described in Chapters 2 to 4 of this thesis have allowed me to analyse birth data from Mexico City in a comprehensive way. The relationship between birthweight and length was of interest and it was enhanced over and above the conventional use of correlation measures.

A fitted copula model represented the bivariate distribution between weight and length based on their best fitted marginals and an estimate of the dependence parameter.

These results were then adjusted for three covariates (sex, gestational age and maternal age). The relationship between birthweight and length was symmetric across all values of these covariates. In fact, sex and gestational age do not contribute statistically in the model, however clinically they are often regarded as important confounders. Therefore, they have been retained in the final model as an example of the flexibility of the copula model and the proposed BAC algorithm.

Covariate-related and conditional copula adjusted bivariate analogues of centiles were calculated along with hidden and false extremes at each stage. There was minimal change in the bivariate relationship between weight and length for boys and girls and changing values of maternal age, whilst there was a small shift in the location of the bivariate distribution for later gestational ages.

Generating adjusted BACs for multiple covariates has been possible via the novel combination of copulas and the proposed BAC algorithm. Graphs such as those shown in Figure 5-21 and Figure 5-22 have the potential to become a tool for everyday use for clinicians and may subsequently improve the course of a patient’s care and ultimately their quality of life.

The weight and length values of a new born baby can be mapped against the bivariate distribution of weight and length from a given population and the BAC level to which it lays closer to, will be informative of the ranking of this baby within the population that the BACs were based on. This ranking may vary between children of the similar weight and length depending on their other relevant characteristics, such as gestational and maternal ages.

BAC contours have the potential to change the way bivariate associations are explored and understood in an epidemiological setting (and beyond). The resulting percentile coverage within a bivariate distribution, which can shift depending on covariate values, may lead to significant changes in a clinician’s decision-making process and understanding of joint associations of related clinical outcomes.

6. Discussion

In this thesis, I have shown how copula models can be usefully applied to the analysis of health data. The exploration of bivariate associations has been the focus of my work; this started by reviewing some of the principles of scalar correlation measures and copula functions. The former are not always adequate measures of complicated bivariate relationships that can change both in strength and direction (as discussed at the first half of Chapter 1) whereas the latter (copulas) provide great flexibility in modelling bivariate distribution functions.

There is a wide selection of copula functions in the literature and some of the most commonly used models were described in Chapter 1. They can accommodate a variety of shapes for bivariate associations depending on the choice of marginal distributions and the value of their dependence parameter.

Moving from scalar coefficient measures to localised equivalent measures such as the local dependence function shown in Chapter 2, granted benefits in the analysis of the relationship between two rvs. This extended to graphical displays too with the local dependence map and the chi-plot complimenting the conventional scatterplot. The chi-plot is a scatterplot of transformed values from any given dataset and focuses on the strength and direction of the observed association, which might be often missed or overlooked within a scatterplot. Similarly, the local dependence map addresses the gaps in the assessment of bivariate associations that cannot be adequately summarised via single scalar values; and does this by assessing correlation in smaller sections of the data. A specific application of the local dependence function for three copulas with Beta marginals was presented and I believe that the resulting LDF results presented in this thesis go some way in enabling further applicability of local dependence.

In Chapter 3, I turned my focus to ways of calculating centiles for bivariate data enabling the creation of bivariate analogues of centiles (BACs). Several other ways exist in the literature and claim to also produce such centiles, however my proposed BAC method is unique in that it is completely data driven, with no kernel, inflation or smoothing parameters involved. A natural progression is the novel application of the identification of hidden extremes and potentially falsely identified extremes, which can result to better targeted treatments. The proposed algorithm produced results fast when applied to a dataset just under 110,000 cases long, hence providing a concrete start for future work on large datasets.

In Chapter 4, I explored how the combination of copulas and bivariate analogues of centiles can produce informative results for researchers for the classification of unusual cases in a bivariate setting. In most medical examples, the use of covariates is imperative in the investigation of relationships between several outcomes. The BAC algorithm can produce covariate-related bivariate analogues of centiles (equivalent to the univariate covariate-related centiles), but when the number of covariates increases, this process becomes cumbersome. Introducing copula models in the analysis allowed the proposed BACs to be adjusted for several covariates simultaneously.

In Chapter 5, all previous findings of this thesis were applied to observed data from term, live births during 2017 in Mexico City. Initial exploration of the bivariate association between birthweight and length based on correlation coefficients was limited, whereas the local dependence map and chi-plot were more informative. These two plots indicated small deviation from positive correlation, which at certain areas was a lot weaker than others. The BAC algorithm produced covariate-related centiles for sex, maternal age and gestational age. The latter covariate was the only one associated with small changes in the bivariate association of birthweight and length. These results were further explored via conditional copulas models whose fitted values matched the already observed patterns.

A further advancement of the covariate-related BACs came via multiple-covariate adjustment of the copula distribution. Being able to extend copula models to adjust for covariates is very relevant to epidemiological studies. Copula models allow for the adjustment of all marginal parameters as well as the copula dependence parameter which controls the strength and/or direction of the bivariate association. I was able to produce an extension of univariate centiles to the bivariate scale, which has the potential of becoming a very useful tool for clinicians (and other researchers) in the identification of unusual observations.

It has been possible for each BAC level to be presented with a bootstrap envelope of BACs of the same level and such results can certainly improve a study's statistical inference properties.

Collectively, I believe that the methods and results presented in this thesis provide an important step towards the exploration of a significant clinical task; flexible, straight forward and data-driven exploration of bivariate associations. Several of the results described in this thesis were presented at various conferences and events in the UK during the course of this PhD, as shown in Appendix 2, and attracted attention for their applicability within research areas where clinicians (and others) are very familiar with their univariate equivalents, i.e. univariate centiles. In 2007, Schweizer [28] pointed out that the research community had not yet reached a point of acceptance and appropriate recognition of joint outcomes. I hope that the results presented in this thesis will assist the future progression of this recognition process.

6.1 Limitations

This thesis has evolved around bivariate relationships and has not explored the extension of the results to higher than two dimensions. This is a very interesting field for further research as the `R` copula library can accommodate copulas for higher than 2 dimensions. There are examples in epidemiological research where it is desirable to model concurrently more than 2 outcomes. For example, lung function is commonly assessed via a battery of measures (FEV, FVC, LCI etc). Joint behaviour of a suite of outcomes can be treated as a multivariate distribution function.

I have also only dealt with numerical continuous outcome measurements in this thesis and it would be of interest to explore the compatibility of these results for discrete and/or categorical variables. An extensive description of copulas for different types of data are discussed in great detail by Genest and Nelsehova [189,190] and include binary, ordinal categorical and count data.

Moreover, the BAC results presented here are only applicable to convex shapes. A natural next step forward, is to enable the algorithm to work both for convex and concave shapes. Such work would start with the creation of a convexity index which could potentially inform the algorithm when to create a concave shape, instead of convex. However, shapes that look concave at their boundaries might actually become convex in more central areas, i.e. the deeper into the bivariate scatterplot we move. Nevertheless, the identification of extremes will be affected by those boundaries, either to a greater or smaller extent depending on the actual data pattern.

Finally, the necessary theory is not yet in place for some of the theoretical properties of convex hulls, hence there are no standard errors available for the estimated BACs. More work needs to be done to develop inferential methods for these estimators; methods such as the bootstrapping, as applied on the proposed BACs, might go some way in addressing this issue.

6.2 Future work

Areas that I would wish to concentrate on for future work are the extensions to higher than 2 dimensions, the application of the proposed methods to other types of copulas; such as vine copulas (which facilitate the construction of models for higher than 2 dimensions) and rotated copulas [191] (which allow modelling negative dependences in copulas, such as the Clayton and Gumbel) as well as to investigate extensions to concave bivariate shapes.

I would also like to see the implementation of the proposed algorithm in more clinical settings and evaluate its usefulness in the eyes of a clinician in terms of assessing how good it is in identifying clinically-interesting extreme cases. The results presented here can be viewed as a stepping stone for a straightforward and data-driven tool that clinicians and other researchers can easily relate to and view as an extension of something already familiar to them (univariate centiles).

As there is not a single unique way of ordering bivariate data, with the proposed BAC method being one of the possible answers, it is important to allow for comparisons between newly proposed techniques. An example of this is the future implementation of the methodologies therein this thesis to a Bayesian framework. Stander et al [192] recently published findings regarding identification of unusual visual acuity measurements in children based on the bivariate posterior predictive distribution of two rvs which forms an interesting clinical example for comparison. Normally distributed priors and a variety of copulas were used, and outliers were identified according to the order of their posterior predictive density values.

6.3 Conclusions

The work presented in this thesis paves the way for better understanding of bivariate associations in epidemiological research and practice. This has been achieved by acknowledging the limitations of conventional correlation coefficient measures then addressing those via local dependence measures and copula models. Identifying extreme observations between two joint clinical outcome variables and how these might differ according to covariates is a natural consequence of improved exploration of bivariate relationships. Their comparison with extreme cases identified via univariate analysis will ensure improved and better targeted treatments for patients.

Bibliography

1. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.r-project.org/. 2019.

2. Fréchet M. Sur les tableaux de corrélation dont les marges sont données. *Ann Univ Lyon Sect A*. 1951;9:53-77.

3. Hoeffding W, Fisher NI, Sen PK. *The Collected Works of Wassily Hoeffding*. New York: Springer; 1994.

4. Pearson K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philos Trans R Soc A Math Phys Eng Sci*. 1896;187:253-318. doi:10.1098/rsta.1896.0007

5. Balakrishnan N, Lai C. Continuous Bivariate Distributions. *Springer-Verlag, New York*. 2009.

6. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika*. 1965;52(3/4):591-611.

7. Mardia K V. Measures of multivariate skewness and kurtosis with applications. *Biometrika*. 1970;57(3):519-530.

8. Mardia K V. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā Indian J Stat Ser B*. 1974:115-128.

9. Student. The probable error of a mean. *Biometrica*. 1908.

10. Fisher RA. Applications of “Student’s” distribution. *Metron*. 1925;5(3):90-104.

11. Kotz S, Nadarajah S. *Multivariate T-Distributions and Their Applications*.; 2004.

12. Hofert M. On sampling from the multivariate t distribution. *R J*. 2013;5(2):129-136.

13. Mari DD, Kotz S, Samuel M, Mari DD, Kotz S. *Correlation and Dependence*. World Scientific Publishing Company; 2001.

14. Joe H. *Multivariate Models and Multivariate Dependence Concepts*. Vol 73. Chapman & Hall/CRC; 1997.

15. Gibbons JD, Chakraborti S. *Nonparametric Statistical Inference*. Vol 168. CRC press; 2003.

16. Barbour AD, Costi MB. Correlation tests for non-linear alternatives. *J R Stat Soc Ser B*. 1993;55(2):541-548.

17. Spearman C. The proof and measurement of association between two things. *Am J Psychol*. 1904;15(1):72-101.

18. Kendall MG. *Rank Correlation Methods*. London: Griffin; 1948.

19. Daniels HE. The relation between measures of correlation in the universe of sample permutations. *Biometrika*. 1944;33(2):129-135.

20. Frahm G, Junker M, Schmidt R. Estimating the tail-dependence coefficient: Properties and pitfalls. *Insur Math Econ*. 2005;37(1):80-100. doi:10.1016/j.matheco.2005.05.008

21. Sibuya M. Bivariate extreme statistics. *Ann Inst Stat Math*. 1959;11(2):195-210.

22. Collins DE. Collins English Dictionary. *Collins Enlgish Dict - Complet Unabridged 10th Ed*. 2012.

23. Hoeffding W. Scale-invariant correlation measures for discontinuous distributions. *Collect Work Wassily Hoeffding New York Springer-Verlag*. 1941:109-133.

24. Sklar A. Fonctions de répartition à n dimensions et leurs marges. *Publ Inst Stat*

Univ Paris. 1959;8(1):11.

25. Dall'Aglio G, Kotz S, Salinetti G. *Advances in Probability Distributions with given Marginals*. Kluwer Dordrecht; 1991.

26. Nelsen RB. *An Introduction to Copulas*. Springer-Verlag; 1999.

27. Nelsen RB. *An Introduction to Copulas*. Springer; 2006.

28. Schweizer B. Introduction to copulas. *J Hydrol Eng*. 2007;12(4):346.

29. Schweizer B, Sklar A. *Probabilistic Metric Spaces*. Dover Publications; 1983.

30. Angus JE. The probability integral transform and related results. *SIAM Rev*. 1994;36(4):652-654.

31. Schweizer B, Wolff EF. On nonparametric measures of dependence for random variables. *Ann Stat*. 1981:879-885.

32. Hoeffding W. A non-parametric test of independence. *Ann Math Stat*. 1948;19(4):546-557.

33. Lehmann EL. Some concepts of dependence. *Ann Math Stat*. 1966:1137-1153.

34. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.

35. Akaike H. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*. 1979;66(2):237-242.

36. Morgenstern D. Einfache beispiele zweidimensionaler verteilungen. *Mitt Math Stat*. 1956;8(1):234-235.

37. Gumbel EJ. Distributions des valeurs extrêmes en plusieurs dimensions. *Publ Inst Stat Univ Paris*. 1960;9:171-173.

38. Farlie DJG. The performance of some correlation coefficients for a general bivariate distribution. *Biometrika*. 1960:307-323.

39. Genest C, Mackay RJ. Archimedean copulas and families of bidimensional laws for which the marginals are given. *Can J Stat Can Stat*. 1986;14(2):145-159. doi:10.2307/3314660

40. Quinn C. *Using Copulas to Measure Association between Ordinal Measures of Health and Income*. Health, Econometrics and Data Group, HEDG working paper, Department of Economics, University of York; 2007.

41. Gumbel EJ. *Statistics of Extremes*. Courier Dover Publications; 2004.

42. Genest C, Rivest L-P. A characterization of Gumbel's family of extreme value distributions. *Stat Probab Lett*. 1989;8(3):207-211.

43. Frank MJ. On the simultaneous associativity ofF (x, y) andx+y− F (x, y). *Aequationes Math*. 1979;19(1):194-226.

44. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*. 1978;65(1):141-151.

45. Cook RD, Johnson ME. A family of distributions for modelling non-elliptically symmetric multivariate data. *J R Stat Soc Ser B*. 1981:210-218.

46. Kimaldorf G, Sampson A. One-parameter families of bivariate distributions with fixed marginals. *Commun Stat Methods*. 4(3):293-301.

47. Kimeldori G, Sampson A. Uniform representations of bivariate distributions. *Commun Stat Methods*. 4(7):617-627.

48. Kelker D. Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā Indian J Stat Ser A*. 1970:419-430.

49. Fang K-T, Kotz S, Ng KW. *Symmetric Multivariate and Related Distributions, Monographs on Statistics and Applied Probability, 36*. London: Chapman and Hall Ltd. MR1071174; 1990.

50. Landsman ZM, Valdez EA. Tail conditional expectations for elliptical

distributions. *North Am Actuar J*. 2003;7(4):55-71.

51. Lindskog F, Mcneil A, Schmock U. Kendall’s tau for elliptical distributions. *Credit risk Meas Eval Manag*. 2003:149-156.

52. Frahm G, Junker M, Szimayer A. Elliptical copulas: applicability and limitations. *Stat Probab Lett*. 2003;63(3):275-286.

53. Fang H-B, Fang K-T, Kotz S. The meta-elliptical distributions with given marginals. *J Multivar Anal*. 2002;82(1):1-16.

54. Lee L-F. Generalized econometric models with selectivity. *Econom J Econom Soc*. 1983:507-512.

55. Embrechts P, McNeil A, Straumann D. Correlation and dependence in risk management: properties and pitfalls. *Risk Manag value risk beyond*. 2002:176-223.

56. Gudendorf G, Segers J. Extreme-value copulas. *Copula Theory Its Appl*. 2010:127-145.

57. Deheuvels P. Probabilistic aspects of multivariate extremes. In: *Statistical Extremes and Applications*. Dordrecht: Springer; 1984:117-130.

58. Galambos J. *The Asymptotic Theory of Extreme Order Statistics*. Vol 352. Wiley New York; 1978.

59. Pickands J. Multivariate extreme value distributions. In: *Proceedings 43rd Session International Statistical Institute*. Vol 2. ; 1981:859-878.

60. Hüsler J, Reiss R-D. Maxima of normal random vectors: between independence and complete dependence. *Stat Probab Lett*. 1989;7(4):283-286.

61. Abberger K. *Exploring Local Dependence*. Universität Konstanz, Fachbereich für Wirtschaftswissenschaften; 2002.

62. Fisher NI, Switzer P. Chi-plots for assessing dependence. *Biometrika*. 1985;72(2):253-265.

63. Fisher NI, Switzer P. Graphical assessment of dependence: Is a picture worth 100 tests? *Am Stat*. 2001;55(3):233-239. doi:10.1198/000313001317098248

64. Holland PW, Wang YJ. Dependence function for continuous bivariate densities. *Commun Stat Methods*. 1987;16(3):863-876.

65. Jones MC. The local dependence function. *Biometrika*. 1996;83(4):899-904.

66. Jones MC. Constant local dependence. *J Multivar Anal*. 1998;64(2):148-155.

67. Jones MC, Koch I. Dependence maps: Local dependence in practice. *Stat Comput*. 2003;13(3):241-255.

68. localgauss: Estimating local Gaussian parameters. *R Packag*. 2013;(0.32).

69. Olkin I, Liu R. A bivariate beta distribution. *Stat Probab Lett*. 2003;62(4):407-412.

70. Gupta AK, Orozco-Castañeda JM, Nagar DK. Non-central bivariate beta distribution. *Stat Pap*. 2011;52(1):139-152. doi:10.1007/s00362-009-0215-y

71. Jones MC. Multivariate t and beta distributions associated with the multivariate F distribution. *Metrika*. 2002;54(3):215-231.

72. El-Bassiouny AH, Jones MC. A bivariate F distribution with marginals on arbitrary numerator and denominator degrees of freedom, and related bivariate beta and t distributions. *Stat Methods Appl*. 2009;18(4):465-481. doi:10.1007/s10260-008-0103-y

73. Gupta R, Kirmani S, HM Srivastava. Local dependence functions for some families of bivariate distributions and total positivity. *Appl Math Comput*. 2010;216(4):1267-1279.

74. Koutoumanou E, Wade A, Cortina-Borja M. Local dependence in bivariate

copulae with beta marginals. *Rev Colomb Estad*. 2017;40(2). doi:10.15446/rce.v40n2.59404

75. Wolfram Research I. Mathematica. 2018.

76. Rose C, Smith MD. mathStatica: Mathematical Statistics with Mathematica. In: New York: Springer; 2002. doi:10.1007/978-3-642-57489-4_66

77. d'Este GM. A Morgenstern-type bivariate gamma distribution. *Biometrika*. 1981;68(1):339-340.

78. Gupta AK, Wong CF. On three and five parameter bivariate Beta distributions. *Metrika*. 1985;32(1):85-91.

79. Schucany WR, Parr WC, Boyer JE. Correlation structure in Farlie-Gumbel-Morgenstern distributions. *Biometrika*. 1978;65(3):650-653.

80. Azzalini A, Dalla Valle A. *The Multivariate Skew-Normal Distribution*. Vol 83.; 1996.

81. Petersen JH. Two bivariate geometrically defined reference regions with applications to male reproductive hormones and human growth. *Stat Med*. 2003;22(16):2603-2618. doi:10.1002/sim.1480

82. Peckham CS, Dezateux C. Issues underlying the evaluation of screening programmes. *Br Med Bull*. 1998;54(4):767-778. doi:10.1093/oxfordjournals.bmb.a011728

83. Boyd JC. Reference regions of two or more dimensions. *Clin Chem Lab Med*. 2004;42(7):739-746. doi:10.1515/CCLM.2004.125

84. Boyd JC, Lacher DA. The multivariate reference range: an alternative interpretation of multi-test profiles. *Clin Chem*. 1982;28(2).

85. Wei Y. An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *J Am Stat Assoc*. 2008;103(481):397-409. doi:10.1198/016214507000001472

86. Gelman A, Meng X-L. A Note on Bivariate Distributions That are Conditionally Normal. *Am Stat*. 1991;45(2):125-126. doi:10.1080/00031305.1991.10475784

87. Rossi PE. *Bayesian Non- and Semi-Parametric Methods and Applications*. Princeton University Press; 2014.

88. Web of Science. http://wok.mimas.ac.uk/.

89. B Rossi LP and GBAP. A new method for monitoring body fluid variation by bioimpedance analysis the RXc graph. *Elsevier*. 1994.

90. Guttmann I. Statistical tolerance regions: classical and Bayesian. *Griffin's Stat Monogr Courses*. 1970;26.

91. Thompson W. Biological applications of normal range and associated significance tests in ignorance of original distribution forms. *Ann Math Stat*. 1938;9(4):281-287.

92. Wilks S. Determination of sample sizes for setting tolerance limits. *Ann Math Stat*. 1941;12:91-96.

93. Wilks S. Statistical prediction with special reference to the problem of tolerance limits. *Ann Math Stat*. 1942;18:400-409.

94. Wald A. An extension of Wilks' method for setting tolerance limits. *Ann Math Stat*. 1943;14:45-55.

95. Tukey JW. Non-Parametric Estimation II. Statistically Equivalent Blocks and Multivariate Tolerance Regions--The Continuous Case. *Ann Math Stat*. 1947:529--539.

96. Scheffe H, Tukey JW. Non-parametric estimation. 1. Validation of order statistics. *Ann Math Stat*. 1945;16(2):187-192. doi:10.1214/aoms/1177731119

97. Tukey JW. Nonparametric estimation. 3. Statistically equivalent blocks and multivariate tolerance regions - the discontinuous case. *Ann Math Stat*. 1948;19(1):30-39. doi:10.1214/aoms/1177730287

98. Tukey JW. Nonparametric Estimation, III. Statistically Equivalent Blocks and Multivariate Tolerance Regions--The Discontinuous Case. *Ann Math Stat*. 1948;19(1):30-39. doi:10.1214/aoms/1177730287

99. Murphy RB. Non-Parametric Tolerance Limits. *Ann Mathmatical Stat*. 1948;19(4):581-589. doi:doi:10.1214/aoms/1177730154

100. Fraser D. Sequentially determined statistically equivalent blocks. *Ann Math Stat*. 1951;22:372-381.

101. Fraser D. Nonparametric Tolerance Regions. *Ann Math Stat*. 1953;24(1):44-55.

102. JHB Kemperman. Generalized tolerance limits. *Ann Math Stat*. 1956;27(1):180-186.

103. Somerville PN. Tables for obtaining non-parametric tolerance limits. *Ann Math Stat*. 1958;29(2):599-601.

104. Buret J, Monfort F. Multivariate reference values-theory and application to ionogram. *Ann Biol Clin (Paris)*. 1975;33(3):248-248.

105. Abt K. Skalenunabhiingige multivariate nichtparametrische Toleranzbereiche. *Math Methoden der Medizin*. 1977.

106. Kagedal B, Sandstrom A, Tibbling G. Determination of a trivariate reference region for free thyroxine index, free triiodothyronine index, and thyrotropin from results obtained in a health survey of middle-aged women. *Clin Chem*. 1978;24(10):1744-1750.

107. Larsson L, Anjou C, Kagedal B, Norr A. Trivariate reference region for serum-calcium, phosphate and alkaline-phosphatase in serum. *Miner Electrolyte Metab*. 1979;2(4-5):244-245.

108. Mahalanobis P. On the generalized distance in statistics. *Proc Natl Inst Sci*. 1936;(2):49-55.

109. Gelsema ES, Leijnse B, Wulkan RW. A multi-dimensional analysis of three chemical quantities in the blood. *Med Informatics*. 1991;16(1):43-54. doi:10.3109/14639239109025294

110. Regland, B., Abrahamsson, L., Blennow, K., Gottfries, C. G., & Wallin A. Vitamin-b12 in csf - reduced csf serum-b12 ratio in demented men. *ACTA Neurol Scand*. 1992;85(4):276-281.

111. Klee GG. Clinical interpretation of reference intervals and reference limits. A plea for assay harmonization. *Clin Chem Lab Med*. 2004;42(7). doi:10.1515/CCLM.2004.127

112. Lekadir K, Keenan N, Pennell D, Yang G-Z. Shape-based myocardial contractility analysis using multivariate outlier detection. In: Ayache, N and Ourdelin, S and Maeder A, ed. *Medical Image Computing And Computer-Assisted Intervention- Miccai 2007, Pt 2, Proceedings*. Vol 4792. Lecture Notes in Computer Science. ; 2007:834+.

113. Mattsson A, Svensson D, Schuett B, Osterziel KJ, Ranke MB. Multidimensional reference regions for IGF-I, IGFBP-2 and IGFBP-3 concentrations in serum of healthy adults. *GROWTH Horm IGF Res*. 2008;18(6):506-516. doi:10.1016/j.ghir.2008.04.005

114. Zengli S, Jiu W, Yong G. Establishment of Multivariate Reference Range for P-wave Amplitude. In: Zhu, K and Zhang H, ed. *Recent Advance In Statistics Application And Related Areas, Vols I And II*. ; 2009:1892-1895.

115. Ross HA, den Heijer M, Hermus ARMM, Sweep FCGJ. Composite Reference Interval for Thyroid-Stimulating Hormone and Free Thyroxine, Comparison with Common Cutoff Values, and Reconsideration of Subclinical Thyroid Disease. *Clin Chem*. 2009;55(11):2019-2025. doi:10.1373/clinchem.2009.124560

116. Ward L, Winall A, Isenring E, et al. Assessment of Bilateral Limb Lymphedema by Bioelectrical Impedance Spectroscopy. *Int J Gynecol CANCER*. 2011;21(2):409-418. doi:10.1097/IGC.0b013e31820866e1

117. Willemsen SP, Eilers PHC, Steegers-Theunissen RPM, et al. Impact of vitamin D-related serum PTH reference values on the diagnosis of mild primary hyperparathyroidism, using bivariate calcium/PTH reference regions. *PLoS One*. 2012;76(3):179-194. doi:10.1371/journal.pone.0033990

118. Fillée C, Keller T, Mourad M, Brinkmann T, Ketelslegers JM. Impact of vitamin D-related serum PTH reference values on the diagnosis of mild primary hyperparathyroidism, using bivariate calcium/PTH reference regions. *Clin Endocrinol (Oxf)*. 2012;76(6):785-789. doi:10.1111/j.1365-2265.2011.04285.x

119. Hoermann R, Larisch R, Dietrich JW, Midgley JEM. Derivation of a multivariate reference range for pituitary thyrotropin and thyroid hormones: diagnostic efficiency compared with conventional single-reference method. *Eur J Endocrinol*. 2016;174(6):735-743. doi:10.1530/EJE-16-0031

120. Selmeryd J, Henriksen E, Dalen H, Hedberg P. Derivation and Evaluation of Age-Specific Multivariate Reference Regions to Aid in Identification of Abnormal Filling Patterns The HUNT and VaMIS Studies. *JACC-CARDIOVASCULAR IMAGING*. 2018;11(3):400-408. doi:10.1016/j.jcmg.2017.04.019

121. Linnet K. Influence of sampling variation and analytical errors on the performance of the multivariate reference region. *Methods Inf Med*. 1988;27(01):37-42. doi:10.2307/2348765

122. Fuchs C, Kenett RS. Multivariate Tolerance Regions and F-Tests. *J Qual Technol*. 1987;19(3):122-131. doi:10.1080/00224065.1987.11979053

123. Fuchs C, Kenett RS. Appraisal of Ceramic Substrates by Multivariate Tolerance Regions. *Stat*. 1988;37(4/5):401. doi:10.2307/2348765

124. Rode RA, Chinchilli VM. The Use of Box-Cox Transformations in the Development of Multivariate Tolerance Regions with Applications to Clinical Chemistry. *Am Stat*. 1988;42(1):23. doi:10.2307/2685257

125. Boente G, Farall AA. Robust Multivariate Tolerance Regions: Influence

Function and Monte Carlo Study. *Technometrics*. 2008;50(4):487-500. doi:10.1198/004017008000000398

126. Kauerz U, Bernhardt W, Weisner B, Rehpenning W. The concentration of glucose in ventricular, cisternal and lumbar cerebrospinal-fluid - bivariate reference ranges csf-serum - a clinical-assessment. *J Clin Chem Clin Biochem*. 1980;18(10):745-746.

127. Boyd JC, Lacher D., Savory J, Bruns D, Renoe B, Wills MK. The multivariate reference range in the clinical-chemistry laboratory. In: *Joint Meeting of the American Association for Clinical Chemistry and the Canadian Society of Clinical Chemists*. Boston; 1980.

128. Abt K, Ackermann H. Univariate and multivariate normal values in medicine. *Med Welt*. 1981;32(13):409-413.

129. Abt K. Scale-independent non-parametric multivariate tolerance regions and their application in medicine. *Biometrical J*. 1982;24(1):27-48. doi:10.1002/bimj.4710240104

130. Makosch G, Ackermann H, Hövels O. Bivariate tolerance region for weight and length of newborns in view of diagnostic criteria. *Monatsschr Kinderheilkd*. 1982;130(5):276-279.

131. Makosch G, Ackermann H, Hovels O. Multivariate tolerance regions for parameters of bone-picture, kidney-picture, and blood-picture of children in view. *Eur J Pediatr*. 1982;138(1):99-99.

132. Massart DL. Observer: Multi-or uni-variate reference regions? *Trends Anal Chem*. 1983;2(6).

133. Ackermann H. Multivariate non-parametric tolerance regions: a new construction technique. *BIOMETRICAL J*. 1983;25:351-359.

134. Ackermann H, Abt K. Designing the Sample Size for Non-parametric, Multivariate Tolerance Regions. *Biometrical J*. 1984;26(7):723-734.

doi:10.1002/bimj.4710260705

135. Di Bucchianico A, Einmahl JHJ, Mushkudiani NA. Smallest nonparametric tolerance regions. *Ann Stat*. 2001;29(5):1320-1343. doi:10.1214/aos/1013203456

136. Chatterjee SK, Patra NK. Asymptotically Minimal Multivariate Tolerance Sets. *Calcutta Stat Assoc Bull*. 1980;29(1-2):73-94. doi:10.1177/0008068319800106

137. Wold S, Esbensen K, Geladi K. Principal component analysis. *Chemom Intell Lab*. 1987;2(1-3):37-52.

138. Karjalainen EJ, Karjalainen UP. Finding the “natural” vector bases for multidimensional reference values. *Scand J Clin Lab Invest*. 1995;55(sup222):61-67. doi:10.3109/00365519509088451

139. Hekking M, Lindemans J, Gelsema ES. Design and representation of multivariate patient-based reference regions for arterial ph, pco2 and base excess values. *Clin Biochem*. 1995;28(6):581-585. doi:10.1016/0009-9120(95)02008-X

140. Piccoli A, Rossi B, Pillon L, Bucciante G. Body fluid overload and bioelectrical impedance analysis in renal patients. *Miner Electrolyte Metab*. 1996;22(1-3):76-78.

141. Piccoli A, Pillon L, Favaro E. Asymmetry of the total body water prediction bias using the impedance index. *Nutrition*. 1997;13(5):438-441. doi:10.1016/S0899-9007(97)91282-X

142. Piccoli A, Nigrelli S, Caberlotto A, et al. Bivariate normal values of the bioelectrical impedence vector in adult and elderly populations. *Am J Clin Nutr*. 1995;61:269-270.

143. Piccoli A, Fanos V, Peruzzi L, et al. Reference values of the bioelectrical impedance vector in neonates in the first week after birth. *NUTRITION*. 2002;18(5):383-387. doi:10.1016/S0899-9007(02)00795-5

144. Espinosa-Cuevas M, Rivas-Rodriguez L, Cristal Gonzalez-Medina E, Atilano-Carsi X, Miranda-Alatriste P, Correa-Rotter R. Bioimpedance vector analysis for body composition in Mexican population. *Rev Investig Clin Transl Investig*. 2007;59(1):15-24.

145. L'Abée C, Poorts-Borger PH, Gorter EHGMGM, et al. The bioelectrical impedance vector migration in healthy infants. *Clin Nutr*. 2010;29(2):222-226. doi:10.1016/j.clnu.2009.08.007

146. Margutti AVB, Monteiro JP, Camelo JS, Camelo Jr. JS. Reference distribution of the bioelectrical impedance vector in healthy term newborns. *Br J Nutr*. 2010;104(10):1508-1513. doi:10.1017/S000711451000245X

147. Donega Toffano RB, Hillesheim E, Barban Margutti AV, et al. Bioelectrical Impedance Vector Analysis in Healthy Term Infants in the First Three Months of Life in Brazil. *J Am Coll Nutr*. 2018;37(2):93-98. doi:10.1080/07315724.2017.1364678

148. Thompson M LOU, Fatti LP. Construction of multivariate centile charts for longitudinal measurements. *Stat Med*. 1997;16(4):333-345. doi:10.1002/(SICI)1097-0258(19970228)16:4<333::AID-SIM416>3.0.CO;2-0

149. Hekking M, Lindemans J, Gelsema ES. A computer program for constructing multivariate reference models. *Comput Methods Programs Biomed*. 1997;53(3):191-200. doi:10.1016/S0169-2607(97)00018-7

150. Capitani E, Laiacona M, Barbarotto R, Cossa FM. How Can We Evaluate Interference in Attentional Tests? A Study Based on Bi-Variate Non-Parametric Tolerance Limits. *J Clin Exp Neuropsychol*. 1999;21(2):216-228. doi:10.1076/jcen.21.2.216.934

151. David L. Duewer KLR, Reeder DJ. RFLP Band Size Standards: NIST Standard Reference Material® 2390. *J Forensic Sci*. 2000;45(5):1093-1105.

152. Li J, Liu RY. Multivariate spacings based on data depth: I. Construction of nonparametric multivariate tolerance regions. *Ann Stat*. 2008;36(3):1299-1323.

doi:10.1214/07-AOS505

153. Yu K, Jones MC. Local Linear Quantile Regression. *J Am Stat Assoc*. 1998;93(441):228-237. doi:10.1080/01621459.1998.10474104

154. Scheike T, Petersen J. Non-parametric estimation of conditional quantiles: locally weighted regression quantiles. *Tech Rep 11, Dep Biostat Univ Copenhagen, Denmark*. 1997.

155. Amin R, Li K, Bengel O. Tolerance limits based on the multivariate MaxMin chart. *Commun Stat Comput*. 2008;37(5):1020-1037. doi:10.1080/03610910801943628

156. Petersen JH. A Non-parametric Conditional Bivariate Reference Region with an Application to Height/Weight Measurements on Normal Girls. *Biometrical J*. 2009;51(4):697-709. doi:10.1002/bimj.200800146

157. Wellek S. On easily interpretable multivariate reference regions of rectangular shape. *Biometrical J*. 2011;53(3):491-511. doi:10.1002/bimj.201000147

158. Willemsen SP, Eilers PHC, Steegers-Theunissen RPM, Lesaffre E. A multivariate Bayesian model for embryonic growth. *Stat Med*. 2015;34(8):1351-1365.

159. Kong L, Mizera I. Quantile tomography: using quantiles with multivariate data. *Stat Sin*. 2012;22(4):1589-1610. doi:10.5705/ss.2010.224

160. Cole T, Donaldson M, Ben-Shlomo Y. SITAR - a useful instrument for growth curve analysis. *Int J Epidemiol*. 2010;39(6):1558–1566.

161. Boček P, Siman M. Directional Quantile Regression in R. *Kybernetika*. 2017;53(3):480-492. doi:10.14736/kyb-2017-3-0480

162. Ruts I, Rousseeuw PJ. Computing depth contours of bivariate point clouds. *Comput Stat Data Anal*. 1996;23(1):153-168.

163. Liu RY. On a notion of simplicial depth. *Statistics (Ber)*. 1988;85:1732-1734.

164. Barnett V. The Ordering of Multivariate Data. *J R Stat Soc Ser A*. 1976;139(3):318-355. doi:10.2307/2344839

165. Eddy WF. Convex Hull Peeling. In: *COMPSTAT 1982 5th Symposium Held at Toulouse 1982*. Heidelberg: Physica-Verlag HD; 1982:42-47. doi:10.1007/978-3-642-51461-6_4

166. Liu R, Parelius J, Singh K. Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by Liu and Singh). *Ann Stat*. 1999;27(3):783-858.

167. Rousseeuw PJ, Ruts I. Algorithm AS 307: Bivariate location depth. *J R Stat Soc Ser C (Applied Stat*. 1996;45(4):516-526.

168. McDermott JP, Lin DKJ. Quantile contours and multivariate density estimation for massive datasets via sequential convex hull peeling. *IIE Trans*. 2007;39(6):581-591. doi:10.1080/07408170600899599

169. Habel K, Grasman R, Gramacy RB, Stahel A, Sterratt D. geometry: Mesh Generation and Surface Tesselation. 2015.

170. Genest M, Masse J-C, Plante J-F. depth: Nonparametric Depth Functions for Multivariate Analysis. 2017.

171. Kosiorowski D, Zawadzki Z. DepthProc: Statistical Depth Functions for Multivariate Analysis. 2018.

172. Barber C, Dobkin D, Huhdanpaa H. The quickhull algorithm for convex hull. *ACM Trans Math Softw*. 1996;22(4):469-483.

173. Eddy WF. A New Convex Hull Algorithm for Planar Sets. *ACM Trans Math Softw*. 1977;3(4):398-403. doi:10.1145/355759.355766

174. de Berg M, van Kreveld M, Overmars M, Schwarzkopf O. Computational Geometry in C. In: *Computational Geometry in C*. Cambridge: Cambridge University Press; 1998. doi:10.1007/978-3-662-03427-9_1

175. Porzio G, Stat GR-Q, 2000 undefined. Peeling multivariate data sets: a new approach. *Quad Stat*. 2000;2.

176. Liu X, Zuo Y. Computing projection depth and its associated estimators. *Stat Comput* . 2014;24(1):51-63.

177. Paindaveine D, Van bever G. From Depth to Local Depth: A Focus on Centrality. *J Am Stat Assoc*. 2013;108(503):1105-1119. doi:10.1080/01621459.2013.813390

178. Becker C, Fried R, Kuhnt S, Gather U. *Robustness and Complex Data Structures : Festschrift in Honour of Ursula Gather*.

179. Rousseeuw PJ, Struyf A. Computing location depth and regression depth in higher dimensions. *Stat Comput*. 1998;8(3):193-203. doi:10.1023/A:1008945009397

180. Zuo Y, Serfling R. General notions of statistical depth function. *Ann Stat*. 2000;28(2):461-482.

181. Goldberg KM, Iglewicz B. Bivariate extensions of the boxplot. *Technometrics*. 1992;34(3):307-320.

182. Kong L, Mizera I. Quantile tomography: using quantiles with multivariate data. *Stat Sin*. 2012:1589-1610.

183. Hyndman RJ. Computing and graphing highest density regions. *Am Stat*. 1996;50(2):120-126.

184. Jones MC, Pewsey A. Sinh-arcsinh distributions. *Biometrika*. 2009;96(4):761-780.

185. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *J R Stat Soc Ser C (Applied Stat*. 2005;54(3):507-554.

186. Yan J. Enjoy the joy of copulas: With a package copula. *J Stat Softw*. 2007;21(4):1-21.

187. Korkmaz S, Goksuluk D, Zararsiz G. MVN: An R Package for Assessing Multivariate Normality. *R J*. 2014;6(2):151-162.

188. Aho K. asbio: A Collection of Statistical Tools for Biologists, R package. 2019.

189. Genest C, Nešlehová J, Neslehova J. A primer on copulas for count data. *ASTIN Bull*. 2007;37(2):475-515. doi:10.2143/ast.37.2.2024077

190. Nikoloulopoulos AK, Karlis D. Modeling multivariate count data using copulas. *Commun Stat Comput*. 2009;39(1):172-187.

191. Nikoloulopoulos AK, Joe H, Haijun L. Vine copulas with asymmetric tail dependence and applications to financial return data. *Comput Stat Data Anal*. 2012;56(11):3659-3673.

192. Stander J, Dalla Valle L, Taglioni C, Liseo B, Wade A, Cortina-Borja M. Analysis of paediatric visual acuity using Bayesian copula models with sinh-arcsinh marginal densities. *Stat Med*. 2019:3421-3443. doi:10.1002/sim.8176

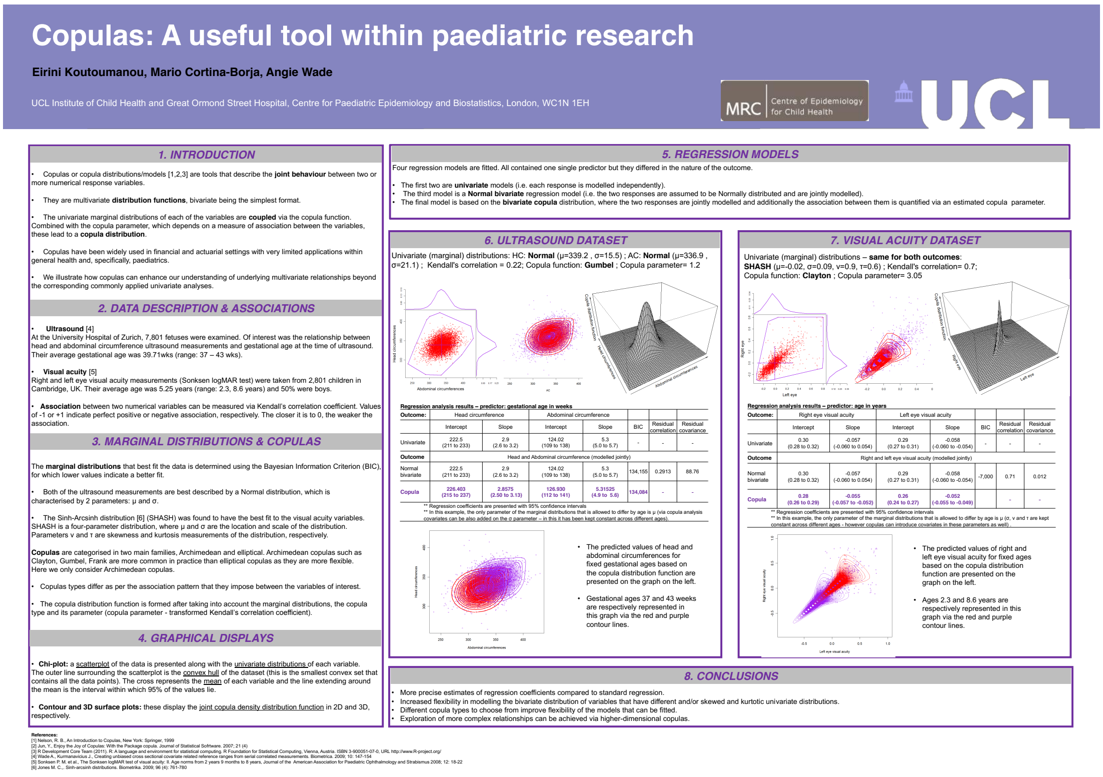
Appendix 1: R code for the BAC algorithm

(omitted)

Appendix 2: Presentations and Posters

A2.1 Copulas: A useful tool with paediatric research

- 2011 Open Day of the UCL Great Ormond Street Institute of Child Health – Poster Presentation
- 2012 Poster Competition, UCL Graduate School, London, UK



Copulas: A useful tool within paediatric research

Eirini Koutoumanou, Mario Cortina-Borja, Angie Wade

UCL Institute of Child Health and Great Ormond Street Hospital, Centre for Paediatric Epidemiology and Biostatistics, London, WC1N 1EH

1. INTRODUCTION

- Copulas or copula distributions/models [1,2,3] are tools that describe the **joint behaviour** between two or more numerical response variables.

- They are multivariate **distribution functions**, bivariate being the simplest format.

- The univariate marginal distributions of each of the variables are **coupled** via the copula function. Combined with the copula parameter, which depends on a measure of association between the variables, these lead to a **copula distribution**.

- Copulas have been widely used in financial and actuarial settings with very limited applications within general health and, specifically, paediatrics.

- We illustrate how copulas can enhance our understanding of underlying multivariate relationships beyond the corresponding commonly applied univariate analyses.

2. DATA DESCRIPTION & ASSOCIATIONS

- **Ultrasound** [4]

At the University Hospital of Zurich, 7,801 fetuses were examined. Of interest was the relationship between head and abdominal circumference ultrasound measurements and gestational age at the time of ultrasound. Their average gestational age was 39.71wks (range: 37 – 43 wks).

- **Visual acuity** [5]

Right and left eye visual acuity measurements (Sonksen logMAR test) were taken from 2,801 children in Cambridge, UK. Their average age was 5.25 years (range: 2.3, 8.6 years) and 50% were boys.

- **Association** between two numerical variables can be measured via Kendall's correlation coefficient. Values of -1 or +1 indicate perfect positive or negative association, respectively. The closer it is to 0, the weaker the association.

3. MARGINAL DISTRIBUTIONS & COPULAS

The **marginal distributions** that best fit the data is determined using the Bayesian Information Criterion (BIC), for which lower values indicate a better fit.

- Both of the ultrasound measurements are best described by a Normal distribution, which is characterised by 2 parameters: μ and σ.

- The Sinh-Arcsinh distribution [6] (SHASH) was found to have the best fit to the visual acuity variables. SHASH is a four-parameter distribution, where μ and σ are the location and scale of the distribution. Parameters v and τ are skewness and kurtosis measurements of the distribution, respectively.

Copulas are categorised in two main families, Archimedean and elliptical. Archimedean copulas such as Clayton, Gumbel, Frank are more common in practice than elliptical copulas as they are more flexible. Here we only consider Archimedean copulas.

- Copulas types differ as per the association pattern that they impose between the variables of interest.

- The copula distribution function is formed after taking into account the marginal distributions, the copula type and its parameter (copula parameter - transformed Kendall's correlation coefficient).

4. GRAPHICAL DISPLAYS

- **Chi-plot:** a scatterplot of the data is presented along with the univariate distributions of each variable. The outer line surrounding the scatterplot is the convex hull of the dataset (this is the smallest convex set that contains all the data points). The cross represents the mean of each variable and the line extending around the mean is the interval within which 95% of the values lie.

- **Contour and 3D surface plots:** these display the joint copula density distribution function in 2D and 3D, respectively.

References:
[1] Nelson, R. B., An Introduction to Copulas, New York: Springer, 1999
[2] Jun, Y., Enjoy the Joy of Copulas: With the Package copula. Journal of Statistical Sofrtware. 2007; 21 (4)
[3] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/
[4] Wade A., Kurmanavicius J., Creating unbiased cross sectional covariate related reference ranges from serial correlated measurements. Biometrica. 2009; 10: 147-154
[5] Sonksen P. M. et al., The Sonksen logMAR test of visual acuity: II. Age norms from 2 years 9 months to 8 years, Journal of the American Association for Paediatric Ophthalmology and Strabismus 2008; 12: 18-22
[6] Jones M. C.,. Sinh-arcsinh distributions. Biometrika. 2009; 96 (4): 761-780

5. REGRESSION MODELS

Four regression models are fitted. All contained one single predictor but they differed in the nature of the outcome.

- The first two are **univariate** models (i.e. each response is modelled independently).
- The third model is a **Normal bivariate** regression model (i.e. the two responses are assumed to be Normally distributed and are jointly modelled).
- The final model is based on the **bivariate copula** distribution, where the two responses are jointly modelled and additionally the association between them is quantified via an estimated copula parameter.

6. ULTRASOUND DATASET

Univariate (marginal) distributions: HC: **Normal** (μ=339.2 , σ=15.5) ; AC: **Normal** (μ=336.9 , σ=21.1) ; Kendall's correlation = 0.22; Copula function: **Gumbel** ; Copula parameter= 1.2

Regression analysis results – predictor: gestational age in weeks

Outcome:	Head circumference		Abdominal circumference				
	Intercept	Slope	Intercept	Slope	BIC	Residual correlation	Residual covariance
Univariate	222.5 (211 to 233)	2.9 (2.6 to 3.2)	124.02 (109 to 138)	5.3 (5.0 to 5.7)	-	-	-
Outcome	Head and Abdominal circumference (modelled jointly)						
Normal bivariate	222.5 (211 to 233)	2.9 (2.6 to 3.2)	124.02 (109 to 138)	5.3 (5.0 to 5.7)	134,155	0.2913	88.76
Copula	**226.403** **(215 to 237)**	**2.8575** **(2.50 to 3.13)**	**126.930** **(112 to 141)**	**5.31525** **(4.9 to 5.6)**	**134,084**	-	-

** Regression coefficients are presented with 95% confidence intervals
** In this example, the only parameter of the marginal distributions that is allowed to differ by age is μ (via copula analysis covariates can be also added on the σ parameter – in this it has been kept constant across different ages).

- The predicted values of head and abdominal circumferences for fixed gestational ages based on the copula distribution function are presented on the graph on the left.

- Gestational ages 37 and 43 weeks are respectively represented in this graph via the red and purple contour lines.

7. VISUAL ACUITY DATASET

Univariate (marginal) distributions – **same for both outcomes**:
SHASH (μ=-0.02, σ=0.09, v=0.9, τ=0.6) ; Kendall's correlation= 0.7;
Copula function: **Clayton** ; Copula parameter= 3.05

Regression analysis results – predictor: age in years

Outcome:	Right eye visual acuity		Left eye visual acuity				
	Intercept	Slope	Intercept	Slope	BIC	Residual correlation	Residual covariance
Univariate	0.30 (0.28 to 0.32)	-0.057 (-0.060 to 0.054)	0.29 (0.27 to 0.31)	-0.058 (-0.060 to -0.054)	-	-	-
Outcome	Right and left eye visual acuity (modelled jointly)						
Normal bivariate	0.30 (0.28 to 0.32)	-0.057 (-0.060 to 0.054)	0.29 (0.27 to 0.31)	-0.058 (-0.060 to -0.054)	-7,000	0.71	0.012
Copula	**0.28** **(0.26 to 0.29)**	**-0.055** **(-0.057 to -0.052)**	**0.26** **(0.24 to 0.27)**	**-0.052** **(-0.055 to -0.049)**		-	-

** Regression coefficients are presented with 95% confidence intervals
** In this example, the only parameter of the marginal distributions that is allowed to differ by age is μ (σ, v and τ are kept constant across different ages - however copulas can introduce covariates in these parameters as well) .

- The predicted values of right and left eye visual acuity for fixed ages based on the copula distribution function are presented on the graph on the left.

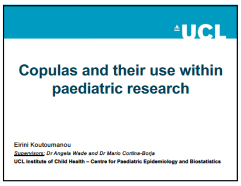
- Ages 2.3 and 8.6 years are respectively represented in this graph via the red and purple contour lines.

8. CONCLUSIONS

- More precise estimates of regression coefficients compared to standard regression.
- Increased flexibility in modelling the bivariate distribution of variables that have different and/or skewed and kurtotic univariate distributions.
- Different copula types to choose from improve flexibility of the models that can be fitted.
- Exploration of more complex relationships can be achieved via higher-dimensional copulas.

A2.2 Copulas and their use within paediatric research

2012 Research Students Conference, Southampton, UK – Talk (a sample of slides shown below)

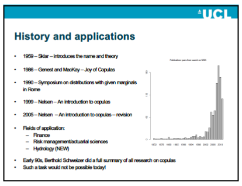


Copulas and their use within paediatric research

Eirini Koutoumanou

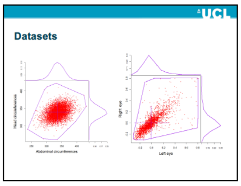
Supervisors: Dr Angela Wade and Dr Mario Cortina-Borja

UCL Institute of Child Health – Centre for Paediatric Epidemiology and Biostatistics

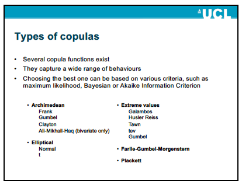


History and applications

- 1959 – Sklar – introduces the name and theory
- 1986 – Genest and MacKay – Joy of Copulas
- 1990 – Symposium on distributions with given marginals in Rome
- 1999 – Nelsen – An introduction to copulas
- 2005 – Nelsen – An introduction to copulas – revision
- Fields of application:
  - Finance
  - Risk management/actuarial sciences
  - Hydrology (NEW)
- Early 90s, Berthold Schweizer did a full summary of all research on copulas
- Such a task would not be possible today!



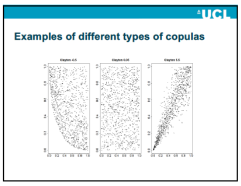
Datasets



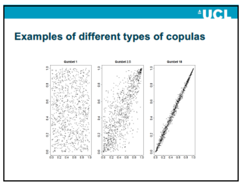
Types of copulas

- Several copula functions exist
- They capture a wide range of behaviours
- Choosing the best one can be based on various criteria, such as maximum likelihood, Bayesian or Akaike Information Criterion

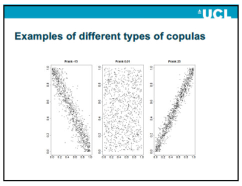
- **Archimedean**
  Frank
  Gumbel
  Clayton
  Ali-Mikhail-Haq (bivariate only)
- **Elliptical**
  Normal
  t
- **Extreme values**
  Galambos
  Husler Reiss
  Tawn
  tev
  Gumbel
- **Farlie-Gumbel-Morgenstern**
- **Plackett**



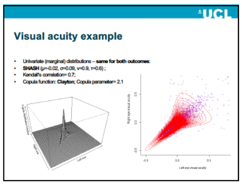
Examples of different types of copulas



Examples of different types of copulas

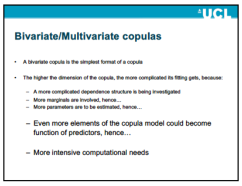


Examples of different types of copulas



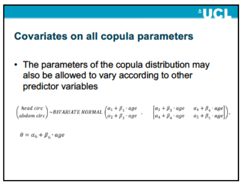
Visual acuity example

- Univariate (marginal) distributions – **same for both outcomes:**
- **SHASH** (μ=-0.02, σ=0.09, ν=0.9, τ=0.6) ;
- Kendall's correlation= 0.7;
- Copula function: **Clayton**; Copula parameter= 2.1



Bivariate/Multivariate copulas

- A bivariate copula is the simplest format of a copula
- The higher the dimension of the copula, the more complicated its fitting gets, because:
  - A more complicated dependence structure is being investigated
  - More marginals are involved, hence…
  - More parameters are to be estimated, hence…
  - Even more elements of the copula model could become function of predictors, hence…
  - More intensive computational needs

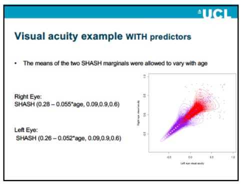


Covariates on all copula parameters

- The parameters of the copula distribution may also be allowed to vary according to other predictor variables

$$\begin{pmatrix} head\ circ \\ abdom\ circ \end{pmatrix} \sim BIVARIATE\ NORMAL \left(\begin{pmatrix} \alpha_1 + \beta_1 \cdot age \\ \alpha_2 + \beta_2 \cdot age \end{pmatrix}, \begin{bmatrix} \alpha_3 + \beta_3 \cdot age & \alpha_4 + \beta_4 \cdot age \\ \alpha_4 + \beta_4 \cdot age & \alpha_5 + \beta_5 \cdot age \end{bmatrix} \right),$$

$$\theta = \alpha_6 + \beta_6 \cdot age$$

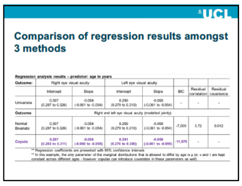


Visual acuity example WITH predictors

- The means of the two SHASH marginals were allowed to vary with age

Right Eye:
SHASH (0.28 – 0.055*age, 0.09,0.9,0.6)

Left Eye:
SHASH (0.26 – 0.052*age, 0.09,0.9,0.6)



Comparison of regression results amongst 3 methods

Regression analysis results – predictor: age in years

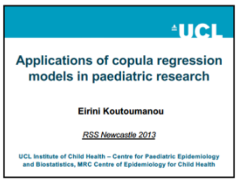
Outcome:	Right eye visual acuity		Left eye visual acuity				
	Intercept	Slope	Intercept	Slope	BIC	Residual correlation	Residual covariance
Univariate	0.307 (0.287 to 0.326)	-0.058 (-0.061 to -0.054)	0.299 (0.279 to 0.318)	-0.058 (-0.061 to -0.054)	-	-	-
Outcome	Right and left eye visual acuity (modelled jointly)						
Normal Bivariate	0.307 (0.287 to 0.326)	-0.058 (-0.061 to -0.054)	0.299 (0.279 to 0.318)	-0.058 (-0.061 to -0.054)	-7,000	0.72	0.012
Copula	**0.297** **(0.283 to 0.311)**	**-0.058** **(-0.060 to -0.055)**	**0.291** **(0.276 to 0.306)**	**-0.058** **(-0.061 to -0.055)**	-11,079	-	-

** Regression coefficients are presented with 95% confidence intervals

** In this example, the only parameter of the marginal distributions that is allowed to differ by age is μ (σ, ν and τ are kept constant across different ages - however copulas can introduce covariates in these parameters as well) .

A2.3 Applications of copula regression models in paediatric research

2013 Royal Statistical Society, Newcastle, UK – Talk (a sample of slides shown below)



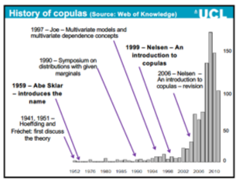
UCL

Applications of copula regression models in paediatric research

Eirini Koutoumanou

RSS Newcastle 2013

UCL Institute of Child Health – Centre for Paediatric Epidemiology and Biostatistics, MRC Centre of Epidemiology for Child Health



History of copulas (Source: Web of Knowledge) UCL

1997 – Joe – Multivariate models and multivariate dependence concepts

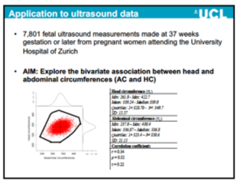
1999 – Nelsen – An introduction to copulas

1990 – Symposium on distributions with given marginals

2006 – Nelsen – An introduction to copulas – revision

1959 – Abe Sklar – introduces the name

1941, 1951 – Hoeffding and Fréchet: first discuss the theory



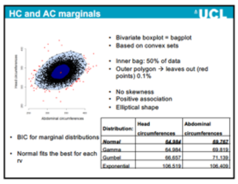
Application to ultrasound data UCL

- 7,801 fetal ultrasound measurements made at 37 weeks gestation or later from pregnant women attending the University Hospital of Zurich
- **AIM: Explore the bivariate association between head and abdominal circumferences (AC and HC)**

Head circumference (Y_1)
Min: 261.0 - Max: 422.7
Mean: 339.24 - Median: 339.0
Quartiles: 1st: 328.70 - 3rd: 349.7
SD: 15.57

Abdominal circumference (Y_2)
Min: 237.0 - Max: 430.4
Mean: 336.87 - Median: 336.8
Quartiles: 1st: 323.4 - 3rd: 350.6
SD: 21.15

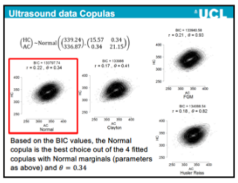
Correlation coefficients:
$r = 0.34$
$\rho = 0.32$
$\tau = 0.22$



HC and AC marginals UCL

- Bivariate boxplot = bagplot
- Based on convex sets
- Inner bag: 50% of data
- Outer polygon → leaves out (red points) 0.1%
- No skewness
- Positive association
- Elliptical shape
- BIC for marginal distributions
- Normal fits the best for each rv

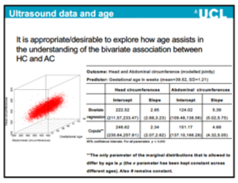
Distribution:	Head circumferences	Abdominal circumferences
Normal	***64,984***	***69,767***
Gamma	64,984	69,819
Gumbel	66,657	71,139
Exponential	106,519	106,409



Ultrasound data Copulas UCL

$$\binom{HC}{AC} \sim \text{Normal}\left(\binom{339.24}{336.87}, \begin{pmatrix}15.57 & 0.34\\ 0.34 & 21.15\end{pmatrix}\right)$$

Based on the BIC values, the Normal copula is the best choice out of the 4 fitted copulas with Normal marginals (parameters as above) and $\theta = 0.34$



Ultrasound data and age UCL

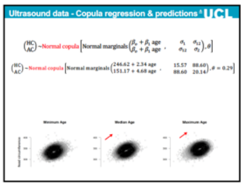
It is appropriate/desirable to explore how age assists in the understanding of the bivariate association between HC and AC

Outcome: Head and Abdominal circumference (modelled jointly)
Predictor: Gestational age in weeks (mean=39.52, SD=1.21)

	Head circumferences		Abdominal circumferences	
	Intercept	Slope	Intercept	Slope
Bivariate regression	222.52 (211.57,233.47)	2.95 (2.68,3.23)	124.02 (109.48,138.56)	5.39 (5.02,5.75)
Copula**	246.62 (235.64,257.61)	2.34 (2.07,2.62)	151.17 (137.10,166.28)	4.68 (4.32,5.05)

95% confidence intervals. For all parameters $p < 0.005$

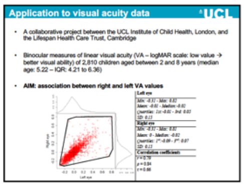
**The only parameter of the marginal distributions that is allowed to differ by age is μ (the σ parameter has been kept constant across different ages). Also θ remains constant.



Ultrasound data - Copula regression & predictions UCL

$$\binom{HC}{AC} \sim \text{Normal copula}\left[\text{Normal marginals}\left(\begin{matrix}\beta_o+\beta_1\,\text{age}\\ \beta_o+\beta_1\,\text{age}\end{matrix}, \begin{matrix}\sigma_1 & \sigma_{12}\\ \sigma_{12} & \sigma_2\end{matrix}\right), \theta\right]$$

$$\binom{HC}{AC} \sim \text{Normal copula}\left[\text{Normal marginals}\left(\begin{matrix}246.62+2.34\,\text{age}\\ 151.17+4.68\,\text{age}\end{matrix}, \begin{matrix}15.57 & 88.60\\ 88.60 & 20.14\end{matrix}\right), \theta = 0.29\right]$$



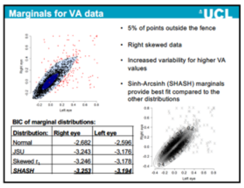
Application to visual acuity data UCL

- A collaborative project between the UCL Institute of Child Health, London, and the Lifespan Health Care Trust, Cambridge
- Binocular measures of linear visual acuity (VA – logMAR scale: low value → better visual ability) of 2,810 children aged between 2 and 8 years (median age: 5.22 – IQR: 4.21 to 6.36)
- **AIM: association between right and left VA values**

Left eye
Min: -0.31 - Max: 0.82
Mean: -0.01 - Median: -0.02
Quartiles: 1st: -0.01 - 3rd: 0.03
SD: 0.13

Right eye
Min: -0.31 - Max: 0.81
Mean: 0 - Median: -0.02
Quartiles: 1st: -0.09 - 3rd: 0.07
SD: 0.13

Correlation coefficients
$r = 0.79$
$\rho = 0.84$
$\tau = 0.66$

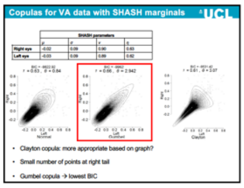


Marginals for VA data UCL

- 5% of points outside the fence
- Right skewed data
- Increased variability for higher VA values
- Sinh-Arcsinh (SHASH) marginals provide best fit compared to the other distributions

BIC of marginal distributions:

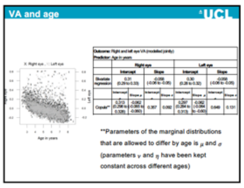
Distribution:	Right eye	Left eye
Normal	-2,682	-2,596
JSU	-3,243	-3,176
Skewed t_1	-3,246	-3,178
SHASH	***-3,253***	***-3,194***



Copulas for VA data with SHASH marginals UCL

	SHASH parameters			
	μ	σ	ν	η
Right eye	-0.02	0.09	0.90	0.63
Left eye	-0.03	0.09	0.89	0.62

- Clayton copula: more appropriate based on graph?
- Small number of points at right tail
- Gumbel copula → lowest BIC

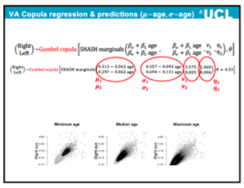


VA and age UCL

Outcome: Right and left eye VA (modelled jointly)
Predictor: Age in years

	Right eye				Left eye			
	Intercept		Slope		Intercept		Slope	
Bivariate regression	0.31 (0.29 to 0.33)		-0.058 (-0.06 to -0.05)		0.30 (0.28 to 0.32)		-0.058 (-0.06 to -0.05)	
	Intercept μ	Slope μ	Intercept σ	Slope σ	Intercept μ	Slope μ	Intercept σ	Slope σ
Copula**	0.313 (0.298 to 0.328)	-0.062 (-0.065 to -0.060)	0.357	0.092	0.297 (0.284 to 0.313)	-0.062 (-0.064 to -0.060)	0.649	0.131

**Parameters of the marginal distributions that are allowed to differ by age is μ and σ (parameters ν and η have been kept constant across different ages)



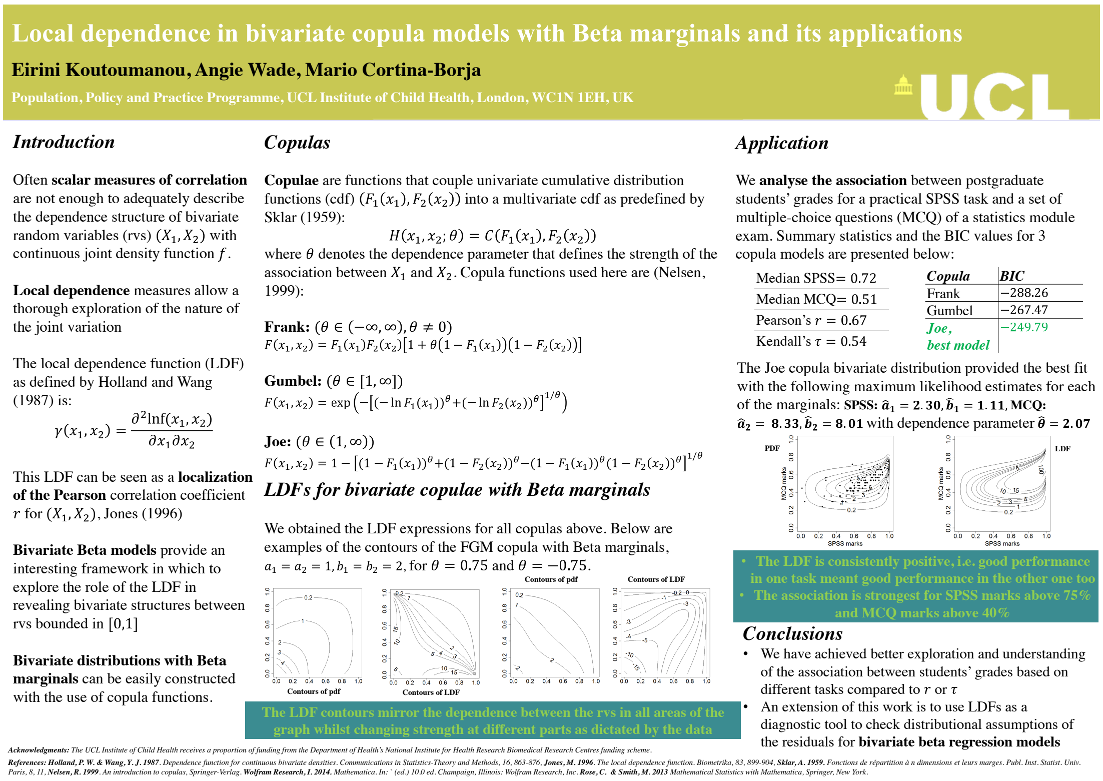
VA Copula regression & predictions ($\mu\sim$age, $\sigma\sim$age) UCL

$$\binom{\text{Right}}{\text{Left}} \sim \text{Gumbel copula}\left[\text{SHASH marginals}\left(\begin{matrix}\beta_o+\beta_1\,\text{age}\\ \beta_o+\beta_1\,\text{age}\end{matrix}, \begin{matrix}\beta_o+\beta_1\,\text{age}\\ \beta_o+\beta_1\,\text{age}\end{matrix}, \begin{matrix}\nu_1\\ \nu_2\end{matrix}, \begin{matrix}\eta_1\\ \eta_2\end{matrix}\right), \theta\right]$$

$$\binom{\text{Right}}{\text{Left}} \sim \text{Gumbel copula}\left[\text{SHASH marginals}\left(\underbrace{\begin{matrix}0.313-0.062\,\text{age}\\ 0.297-0.062\,\text{age}\end{matrix}}_{\mu_1,\ \mu_2}, \underbrace{\begin{matrix}0.357-0.092\,\text{age}\\ 0.694-0.131\,\text{age}\end{matrix}}_{\sigma_1,\ \sigma_2}, \underbrace{\begin{matrix}5.575\\ 0.855\end{matrix}}_{\nu_1,\ \nu_2}, \underbrace{\begin{matrix}0.060\\ 0.006\end{matrix}}_{\eta_1,\ \eta_2}\right), \theta = 4.33\right]$$

A2.4 Local dependence in bivariate copula models with Beta marginals and its applications

2014 Joint Statistical Meetings, Boston, USA – Contributed Poster Presentation



Local dependence in bivariate copula models with Beta marginals and its applications

Eirini Koutoumanou, Angie Wade, Mario Cortina-Borja

Population, Policy and Practice Programme, UCL Institute of Child Health, London, WC1N 1EH, UK

Introduction

Often **scalar measures of correlation** are not enough to adequately describe the dependence structure of bivariate random variables (rvs) (X_1, X_2) with continuous joint density function f.

Local dependence measures allow a thorough exploration of the nature of the joint variation

The local dependence function (LDF) as defined by Holland and Wang (1987) is:

$$\gamma(x_1, x_2) = \frac{\partial^2 \ln f(x_1, x_2)}{\partial x_1 \partial x_2}$$

This LDF can be seen as a **localization of the Pearson** correlation coefficient r for (X_1, X_2), Jones (1996)

Bivariate Beta models provide an interesting framework in which to explore the role of the LDF in revealing bivariate structures between rvs bounded in [0,1]

Bivariate distributions with Beta marginals can be easily constructed with the use of copula functions.

Copulas

Copulae are functions that couple univariate cumulative distribution functions (cdf) $(F_1(x_1), F_2(x_2))$ into a multivariate cdf as predefined by Sklar (1959):

$$H(x_1, x_2; \theta) = C(F_1(x_1), F_2(x_2))$$

where θ denotes the dependence parameter that defines the strength of the association between X_1 and X_2. Copula functions used here are (Nelsen, 1999):

Frank: $(\theta \in (-\infty, \infty), \theta \neq 0)$

$$F(x_1, x_2) = F_1(x_1)F_2(x_2)[1 + \theta(1 - F_1(x_1))(1 - F_2(x_2))]$$

Gumbel: $(\theta \in [1, \infty])$

$$F(x_1, x_2) = \exp\left(-\left[(-\ln F_1(x_1))^\theta + (-\ln F_2(x_2))^\theta\right]^{1/\theta}\right)$$

Joe: $(\theta \in (1, \infty))$

$$F(x_1, x_2) = 1 - \left[(1 - F_1(x_1))^\theta + (1 - F_2(x_2))^\theta - (1 - F_1(x_1))^\theta (1 - F_2(x_2))^\theta\right]^{1/\theta}$$

LDFs for bivariate copulae with Beta marginals

We obtained the LDF expressions for all copulas above. Below are examples of the contours of the FGM copula with Beta marginals, $a_1 = a_2 = 1, b_1 = b_2 = 2$, for $\theta = 0.75$ and $\theta = -0.75$.

Contours of pdf

Contours of LDF

Contours of pdf

Contours of LDF

The LDF contours mirror the dependence between the rvs in all areas of the graph whilst changing strength at different parts as dictated by the data

Application

We **analyse the association** between postgraduate students' grades for a practical SPSS task and a set of multiple-choice questions (MCQ) of a statistics module exam. Summary statistics and the BIC values for 3 copula models are presented below:

Median SPSS= 0.72
Median MCQ= 0.51
Pearson's r = 0.67
Kendall's τ = 0.54

Copula	*BIC*
Frank	−288.26
Gumbel	−267.47
Joe, best model	−249.79

The Joe copula bivariate distribution provided the best fit with the following maximum likelihood estimates for each of the marginals: **SPSS:** $\boldsymbol{\hat{a}_1 = 2.30, \hat{b}_1 = 1.11}$, **MCQ:** $\boldsymbol{\hat{a}_2 = 8.33, \hat{b}_2 = 8.01}$ with dependence parameter $\boldsymbol{\hat{\theta} = 2.07}$

PDF

LDF

- **The LDF is consistently positive, i.e. good performance in one task meant good performance in the other one too**
- **The association is strongest for SPSS marks above 75% and MCQ marks above 40%**

Conclusions

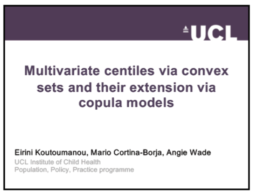
- We have achieved better exploration and understanding of the association between students' grades based on different tasks compared to r or τ
- An extension of this work is to use LDFs as a diagnostic tool to check distributional assumptions of the residuals for **bivariate beta regression models**

Acknowledgments: *The UCL Institute of Child Health receives a proportion of funding from the Department of Health's National Institute for Health Research Biomedical Research Centres funding scheme.*

References: ***Holland, P. W. & Wang, Y. J. 1987****. Dependence function for continuous bivariate densities. Communications in Statistics-Theory and Methods, 16, 863-876,* ***Jones, M. 1996****. The local dependence function. Biometrika, 83, 899-904,* ***Sklar, A. 1959.*** *Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris, 8, 11,* ***Nelsen, R. 1999****. An introduction to copulas, Springer-Verlag.* **Wolfram Research, I. 2014.** *Mathematica. In: ` (ed.) 10.0 ed. Champaign, Illinois: Wolfram Research, Inc.* ***Rose, C. & Smith, M. 2013*** *Mathematical Statistics with Mathematica, Springer, New York.*

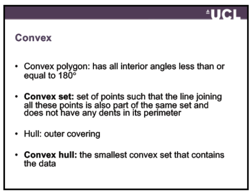
A2.5 Multivariate centiles via convex sets and their extension via copula models

2015 Research Students Conference, Leeds, UK (a sample of slides shown below)



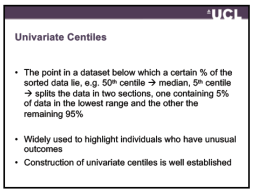
Multivariate centiles via convex sets and their extension via copula models

Eirini Koutoumanou, Mario Cortina-Borja, Angie Wade
UCL Institute of Child Health
Population, Policy, Practice programme



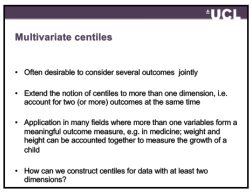
Convex

- Convex polygon: has all interior angles less than or equal to 180°
- **Convex set:** set of points such that the line joining all these points is also part of the same set and does not have any dents in its perimeter
- Hull: outer covering
- **Convex hull:** the smallest convex set that contains the data



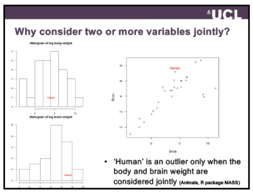
Univariate Centiles

- The point in a dataset below which a certain % of the sorted data lie, e.g. 50^{th} centile → median, 5^{th} centile → splits the data in two sections, one containing 5% of data in the lowest range and the other the remaining 95%
- Widely used to highlight individuals who have unusual outcomes
- Construction of univariate centiles is well established



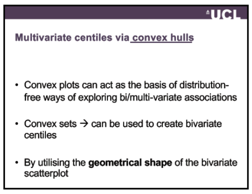
Multivariate centiles

- Often desirable to consider several outcomes jointly
- Extend the notion of centiles to more than one dimension, i.e. account for two (or more) outcomes at the same time
- Application in many fields where more than one variables form a meaningful outcome measure, e.g. in medicine; weight and height can be accounted together to measure the growth of a child
- How can we construct centiles for data with at least two dimensions?



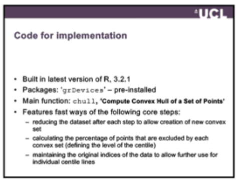
Why consider two or more variables jointly?

- 'Human' is an outlier only when the body and brain weight are considered jointly (Animals, R package MASS)



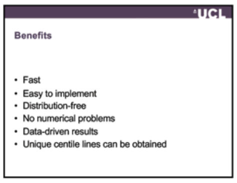
Multivariate centiles via convex hulls

- Convex plots can act as the basis of distribution-free ways of exploring bi/multi-variate associations
- Convex sets → can be used to create bivariate centiles
- By utilising the **geometrical shape** of the bivariate scatterplot



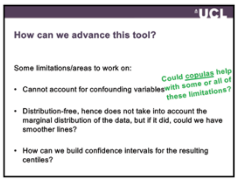
Code for implementation

- Built in latest version of R, 3.2.1
- Packages: '`grDevices`' – pre-installed
- Main function: `chull`, **'Compute Convex Hull of a Set of Points'**
- Features fast ways of the following core steps:
  - reducing the dataset after each step to allow creation of new convex set
  - calculating the percentage of points that are excluded by each convex set (defining the level of the centile)
  - maintaining the original indices of the data to allow further use for individual centile lines



Benefits

- Fast
- Easy to implement
- Distribution-free
- No numerical problems
- Data-driven results
- Unique centile lines can be obtained

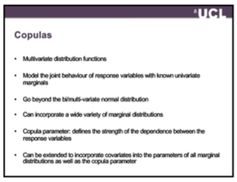


How can we advance this tool?

Some limitations/areas to work on:

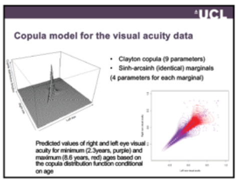
- Cannot account for confounding variables
- Distribution-free, hence does not take into account the marginal distribution of the data, but if it did, could we have smoother lines?
- How can we build confidence intervals for the resulting centiles?

Could copulas help with some or all of these limitations?



Copulas

- Multivariate distribution functions
- Model the joint behaviour of response variables with known univariate marginals
- Go beyond the bi/multi-variate normal distribution
- Can incorporate a wide variety of marginal distributions
- Copula parameter: defines the strength of the dependence between the response variables
- Can be extended to incorporate covariates into the parameters of all marginal distributions as well as the copula parameter

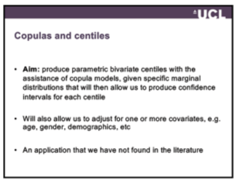


Copula model for the visual acuity data

- Clayton copula (9 parameters)
- Sinh-arcsinh (identical) marginals

(4 parameters for each marginal)

Predicted values of right and left eye visual acuity for minimum (2.3years, purple) and maximum (8.6 years, red) ages based on the copula distribution function conditional on age

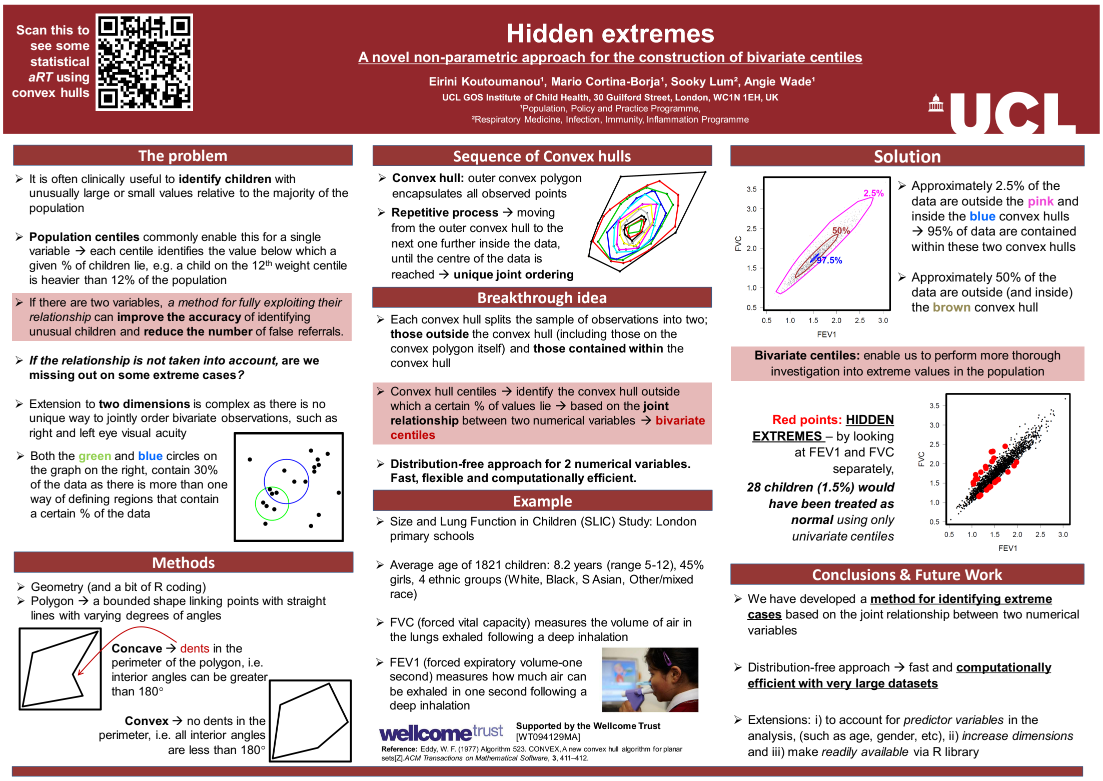


Copulas and centiles

- **Aim:** produce parametric bivariate centiles with the assistance of copula models, given specific marginal distributions that will then allow us to produce confidence intervals for each centile
- Will also allow us to adjust for one or more covariates, e.g. age, gender, demographics, etc
- An application that we have not found in the literature

A2.6 Hidden extremes. A novel non-parametric approach for the construction of bivariate centiles

- 2017 Open day of the UCL Great Ormond Street Institute of Child Health – Poster Presentation
- 2018 Young Statisticians Meeting, Oxford, UK



Scan this to see some statistical *aRT* using convex hulls

Hidden extremes

A novel non-parametric approach for the construction of bivariate centiles

Eirini Koutoumanou[1], Mario Cortina-Borja[1], Sooky Lum[2], Angie Wade[1]

UCL GOS Institute of Child Health, 30 Guilford Street, London, WC1N 1EH, UK

[1]Population, Policy and Practice Programme,

[2]Respiratory Medicine, Infection, Immunity, Inflammation Programme

The problem

- It is often clinically useful to **identify children** with unusually large or small values relative to the majority of the population
- **Population centiles** commonly enable this for a single variable → each centile identifies the value below which a given % of children lie, e.g. a child on the 12th weight centile is heavier than 12% of the population
- If there are two variables, *a method for fully exploiting their relationship* can **improve the accuracy** of identifying unusual children and **reduce the number** of false referrals.
- ***If the relationship is not taken into account,* are we missing out on some extreme cases*?***
- Extension to **two dimensions** is complex as there is no unique way to jointly order bivariate observations, such as right and left eye visual acuity
- Both the green and blue circles on the graph on the right, contain 30% of the data as there is more than one way of defining regions that contain a certain % of the data

Methods

- Geometry (and a bit of R coding)
- Polygon → a bounded shape linking points with straight lines with varying degrees of angles

Concave → dents in the perimeter of the polygon, i.e. interior angles can be greater than 180°

Convex → no dents in the perimeter, i.e. all interior angles are less than 180°

Sequence of Convex hulls

- **Convex hull:** outer convex polygon encapsulates all observed points
- **Repetitive process** → moving from the outer convex hull to the next one further inside the data, until the centre of the data is reached → **unique joint ordering**

Breakthrough idea

- Each convex hull splits the sample of observations into two; **those outside** the convex hull (including those on the convex polygon itself) and **those contained within** the convex hull
- Convex hull centiles → identify the convex hull outside which a certain % of values lie → based on the **joint relationship** between two numerical variables → **bivariate centiles**
- **Distribution-free approach for 2 numerical variables. Fast, flexible and computationally efficient.**

Example

- Size and Lung Function in Children (SLIC) Study: London primary schools
- Average age of 1821 children: 8.2 years (range 5-12), 45% girls, 4 ethnic groups (White, Black, S Asian, Other/mixed race)
- FVC (forced vital capacity) measures the volume of air in the lungs exhaled following a deep inhalation
- FEV1 (forced expiratory volume-one second) measures how much air can be exhaled in one second following a deep inhalation

Supported by the Wellcome Trust [WT094129MA]

Reference: Eddy, W. F. (1977) Algorithm 523. CONVEX, A new convex hull algorithm for planar sets[Z]. *ACM Transactions on Mathematical Software*, **3**, 411–412.

Solution

- Approximately 2.5% of the data are outside the pink and inside the blue convex hulls → 95% of data are contained within these two convex hulls
- Approximately 50% of the data are outside (and inside) the brown convex hull

Bivariate centiles: enable us to perform more thorough investigation into extreme values in the population

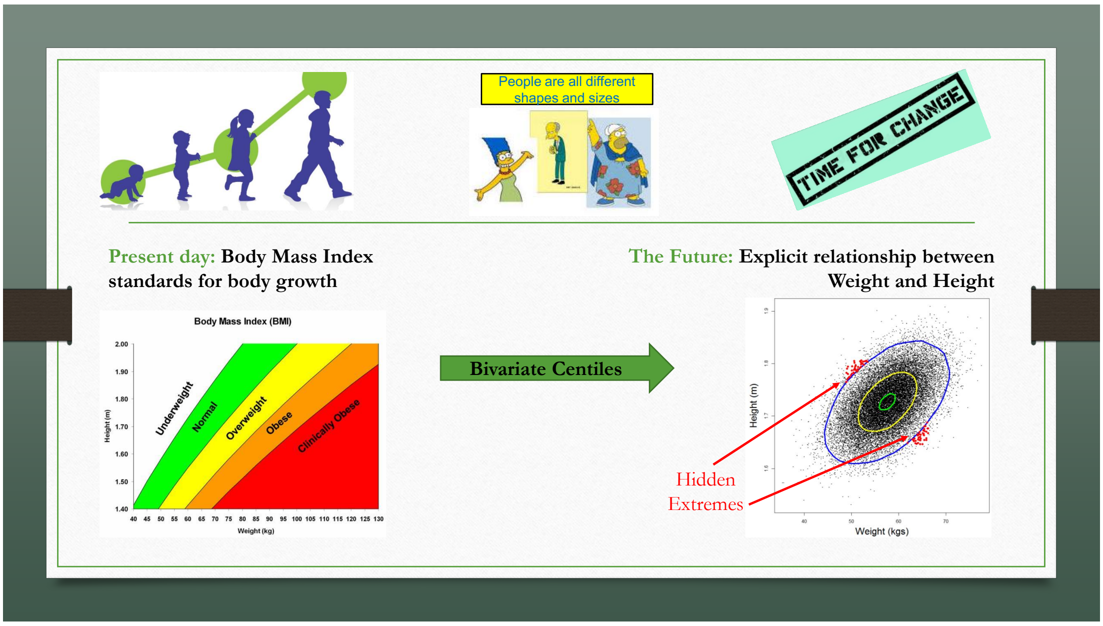
Red points: HIDDEN EXTREMES – by looking at FEV1 and FVC separately, ***28 children (1.5%) would have been treated as normal*** *using only univariate centiles*

Conclusions & Future Work

- We have developed a **method for identifying extreme cases** based on the joint relationship between two numerical variables
- Distribution-free approach → fast and **computationally efficient with very large datasets**
- Extensions: i) to account for *predictor variables* in the analysis, (such as age, gender, etc), ii) *increase dimensions* and iii) make *readily available* via R library

A2.7 Three Minute Thesis (3MT) Competition

Runner up at the 2018 ICH 3MT competition (single, static slide shown below)



People are all different shapes and sizes

TIME FOR CHANGE

Present day: Body Mass Index standards for body growth

Bivariate Centiles

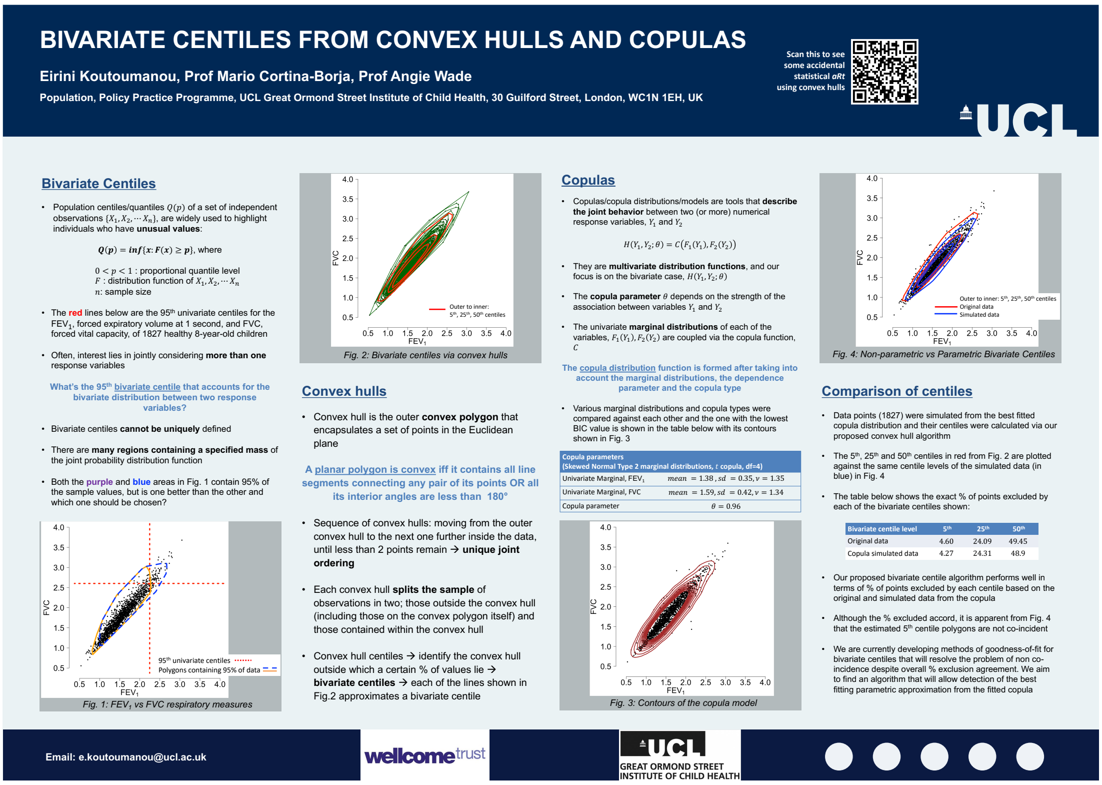
The Future: Explicit relationship between Weight and Height

Hidden Extremes

Body Mass Index (BMI)
Height (m)
2.00
1.90
1.80
1.70
1.60
1.50
1.40
Underweight
Normal
Overweight
Obese
Clinically Obese
40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130
Weight (kg)
Height (m)
1.9
1.8
1.7
1.6
40
50
60
70
Weight (kgs)

A2.8 Bivariate centiles from convex hulls and copulas

2018 Royal Statistical Society, Cardiff, UK – Poster



BIVARIATE CENTILES FROM CONVEX HULLS AND COPULAS

Eirini Koutoumanou, Prof Mario Cortina-Borja, Prof Angie Wade

Population, Policy Practice Programme, UCL Great Ormond Street Institute of Child Health, 30 Guilford Street, London, WC1N 1EH, UK

Scan this to see some accidental statistical *aRt* using convex hulls

Bivariate Centiles

- Population centiles/quantiles $Q(p)$ of a set of independent observations $\{X_1, X_2, \cdots X_n\}$, are widely used to highlight individuals who have **unusual values**:

$$\boldsymbol{Q(p) = inf\{x: F(x) \geq p\}}, \text{ where}$$

$0 < p < 1$: proportional quantile level
F : distribution function of $X_1, X_2, \cdots X_n$
n: sample size

- The **red** lines below are the 95th univariate centiles for the FEV$_1$, forced expiratory volume at 1 second, and FVC, forced vital capacity, of 1827 healthy 8-year-old children
- Often, interest lies in jointly considering **more than one** response variables

What's the 95th bivariate centile that accounts for the bivariate distribution between two response variables?

- Bivariate centiles **cannot be uniquely** defined
- There are **many regions containing a specified mass** of the joint probability distribution function
- Both the **purple** and **blue** areas in Fig. 1 contain 95% of the sample values, but is one better than the other and which one should be chosen?

Fig. 1: FEV$_1$ vs FVC respiratory measures

Fig. 2: Bivariate centiles via convex hulls

Convex hulls

- Convex hull is the outer **convex polygon** that encapsulates a set of points in the Euclidean plane

A planar polygon is convex iff it contains all line segments connecting any pair of its points OR all its interior angles are less than 180°

- Sequence of convex hulls: moving from the outer convex hull to the next one further inside the data, until less than 2 points remain → **unique joint ordering**
- Each convex hull **splits the sample** of observations in two; those outside the convex hull (including those on the convex polygon itself) and those contained within the convex hull
- Convex hull centiles → identify the convex hull outside which a certain % of values lie → **bivariate centiles** → each of the lines shown in Fig.2 approximates a bivariate centile

Copulas

- Copulas/copula distributions/models are tools that **describe the joint behavior** between two (or more) numerical response variables, Y_1 and Y_2

$$H(Y_1, Y_2; \theta) = C\big(F_1(Y_1), F_2(Y_2)\big)$$

- They are **multivariate distribution functions**, and our focus is on the bivariate case, $H(Y_1, Y_2; \theta)$
- The **copula parameter** θ depends on the strength of the association between variables Y_1 and Y_2
- The univariate **marginal distributions** of each of the variables, $F_1(Y_1), F_2(Y_2)$ are coupled via the copula function, C

The copula distribution function is formed after taking into account the marginal distributions, the dependence parameter and the copula type

- Various marginal distributions and copula types were compared against each other and the one with the lowest BIC value is shown in the table below with its contours shown in Fig. 3

Copula parameters (Skewed Normal Type 2 marginal distributions, t copula, df=4)	
Univariate Marginal, FEV$_1$	$mean = 1.38, sd = 0.35, \nu = 1.35$
Univariate Marginal, FVC	$mean = 1.59, sd = 0.42, \nu = 1.34$
Copula parameter	$\theta = 0.96$

Fig. 3: Contours of the copula model

Fig. 4: Non-parametric vs Parametric Bivariate Centiles

Comparison of centiles

- Data points (1827) were simulated from the best fitted copula distribution and their centiles were calculated via our proposed convex hull algorithm
- The 5th, 25th and 50th centiles in red from Fig. 2 are plotted against the same centile levels of the simulated data (in blue) in Fig. 4
- The table below shows the exact % of points excluded by each of the bivariate centiles shown:

Bivariate centile level	5th	25th	50th
Original data	4.60	24.09	49.45
Copula simulated data	4.27	24.31	48.9

- Our proposed bivariate centile algorithm performs well in terms of % of points excluded by each centile based on the original and simulated data from the copula
- Although the % excluded accord, it is apparent from Fig. 4 that the estimated 5th centile polygons are not co-incident
- We are currently developing methods of goodness-of-fit for bivariate centiles that will resolve the problem of non co-incidence despite overall % exclusion agreement. We aim to find an algorithm that will allow detection of the best fitting parametric approximation from the fitted copula

Email: e.koutoumanou@ucl.ac.uk

GREAT ORMOND STREET
INSTITUTE OF CHILD HEALTH

Appendix 3: Publications

Local dependence in Bivariate Copulae with Beta Marginals

Journal: Revista Colombiana De Estadística, 2017

(omitted)