Predictive Inference with Copulas for Bivariate Data

Noryanti Muhammad

A Thesis presented for the degree of
Doctor of Philosophy

Statistics and Probability Research Group
Department of Mathematical Sciences
University of Durham
England

February 2016

Dedicated to

My beloved husband; Imran, children; Aliah and Amirul,

parents,

my brothers and sisters.

Predictive Inference with Copulas for Bivariate Data

Noryanti Muhammad

Submitted for the degree of Doctor of Philosophy
February 2016

Abstract

Nonparametric predictive inference (NPI) is a statistical approach with strong frequentist properties, with inferences explicitly in terms of one or more future observations. NPI is based on relatively few modelling assumptions, enabled by the use of lower and upper probabilities to quantify uncertainty. While NPI has been developed for a range of data types, and for a variety of applications, thus far it has not been developed for multivariate data. This thesis presents the first study in this direction. Restricting attention to bivariate data, a novel approach is presented which combines NPI for the marginals with copulas for representing the dependence between the two variables. It turns out that, by using a discretization of the copula, this combined method leads to relatively easy computations. The new method is introduced with use of an assumed parametric copula. The main idea is that NPI on the marginals provides a level of robustness which, for small to medium-sized data sets, allows some level of misspecification of the copula.

As parametric copulas have restrictions with regard to the kind of dependency they can model, we also consider the use of nonparametric copulas in combination with NPI for the marginals. As an example application of our new method, we consider accuracy of diagnostic tests with bivariate outcomes, where the weighted combination of both variables can lead to better diagnostic results than the use of either of the variables alone. The results of simulation studies are presented to provide initial insights into the performance of the new methods presented in this thesis, and examples using data from the literature are used to illustrate applications

of the methods. As this is the first research into developing NPI-based methods for multivariate data, there are many related research opportunities and challenges, which we briefly discuss.

Declaration

The work in this thesis is based on research carried out at the Statistics and Probability Research Group, the Department of Mathematical Sciences, University of Durham, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Copyright © 2016 by Noryanti Muhammad.

"The copyright of this thesis rests with the author. No quotations from it should be published without the author's prior written consent and information derived from it should be acknowledged".

Acknowledgements

Alhamdullilah, Allah my God, I am truly grateful for your countless blessings you have bestowed on me generally, and especially in accomplishing this thesis.

I would like to express my main deepest appreciation to my supervisors, Prof. Frank Coolen and Dr. Tahani Coolen-Maturi for their unlimited support, expert advice and guidance. Their patience, kindness, enthusiasm, and untiring support and calm advice have been invaluable to me.

I am immensely grateful to my husband, Muhamad Imran for his unwavering belief in me and unlimited support. His patience and kindness in helping to manage the family, especially housework and our children, are really appreciated and respected. To my beloved children, Aliah and Amirul, thank you very much for understanding that your mom's work kept her very busy.

Special thanks to my mother and mother-in-law for their frequent loving prayers for me, the substantial amount of their unconditional love surrounds me. To my great sadness, my father and father-in-law passed away on May 2014, they always encourage me to do the right things. To my brothers and sisters, thank you for your support and encouragement.

To all my friends who support and help me either directly or indirectly for this research, specifically by giving motivation and encouragement to finish this research, thank you so much. May God bless and ease whatever you do.

My final thanks to the Ministry of Higher Education of Malaysia (MOHE) and Universiti Malaysia PAHANG (UMP) for giving me the opportunity to pursue my studies at Durham University, supported by a full scholarship. Thanks also to the Department of Mathematical Sciences for offering such an enjoyable academic atmosphere and for the facilities that have enabled me to study smoothly.

Contents

Chapter 1

Introduction

1.1 Overview

Identifying and modelling dependencies between two or more related random quantities is a main challenge in statistics and is important in many application areas. Taking dependence into account is important to model, estimate and predict weather, risk and aspects of other applications more efficiently. Analyses of dependencies are of considerable importance in many sectors as an aid to better understanding the interaction of variables in a certain field of study and also as an input in every aspect of our life including engineering, health, finance, insurance and agriculture.

Statistical dependence is a relationship between any two or more characteristics of units under study or review. These units may, for example, be individuals, objects, or various aspects of environment. The dependence structure is important in order to know whether a particular model or inference might be suitable for a given application or data set. Several types of dependence can occur, for example positive and negative dependence, exchangeable or flexible dependence and dependence decreasing with lag (for data with a time index) [55]. A popular method for modelling dependencies is the use of a copula [14, 80]. Generally, a copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform [55, 73]. Many researchers have addressed and studied dependence using copulas including Genest et al. [42], Embrechts et al. [36], Scaillet and Fermanian [82] and Tsukahara [94]. Often, in their studies they estimate dependence

parameter(s). The dependence is also important in prediction where it plays a key role in decision making processes, classifying and other aspects that involve the dependence. For example, in risk of failure trajectory (e.g. effect of random external actions like wind, or unexpected reactions of the drivers), the dependence structure between vehicle criteria and safety acceptance of the models is considered to reduce road accidents rate [60].

This thesis presents a new method for predictive inference taking into account the dependence structure. It uses Nonparametric Predictive Inference (NPI) for the marginals combined with a copula. We restricted attention to bivariate data. The important general idea in this thesis is to look at the prediction of the two random quantities. We consider the dependence structure between these two random quantities using copula, as copula gives an interesting tool for describing the dependence structures. The idea that we considering the dependence structure between the two random quantities using parametric copula for small data sets and nonparametric copula specifically kernel-based method for large data sets. The NPI on the marginals with the estimated copulas, presenting in this thesis is somewhat different to the usual statistical approaches based on imprecise probabilities [2]. Our method uses a discretized version of the copula which fits perfectly with the NPI method for the marginals and leads to relatively straight forward computations because there is no need to estimate the marginals and the copula simultaneously. By using the NPI for the marginals, the information shortage is most likely to be about the dependence structure.

NPI has been developed over the last two decades, with many applications in statistics, reliability, risk and operations research (see www.npi-statistics.com). It has excellent frequentist properties, but relies on the natural ordering of the observed data or of a reasonable underlying latent variable representation with a natural ordering (e.g. used for Bernoulli and categorical observations [19]). So far, NPI has only been introduced for one-dimensional (univariate) data, this is the first thesis introducing a method which attempts to generalize NPI to bivariate data.

In Section 1.2 we present the main idea of NPI and a detailed outline of this thesis is given in Section 1.3, with details of related publications.

1.2 Nonparametric predictive inference

Nonparametric Predictive Inference (NPI) is a frequentist statistical framework for inference on a future observation based on past data observations [19]. NPI uses lower and upper probabilities, also known as imprecise probability [2], to quantify uncertainty and is based on only few assumptions.

NPI is based on the assumption $A_{(n)}$, proposed by Hill [50], which gives direct conditional probabilities for a future real-valued random quantity, conditional on observed values of n related random quantities [1, 18]. Effectively, it assumes that the rank of the future observation among the observed values is equally likely to have each possible value $1, \ldots, n+1$. Hence, this assumption is that the next observation has probability $1/(n+1)$ to be in each interval of the partition of the real line as created by the n observations. Suppose that $X_1, X_2, ..., X_n, X_{n+1}$ are continuous and exchangeable real-valued random quantities. Let the ordered observed values of $X_1, X_2, ..., X_n$ be denoted by $x_{(1)} < x_{(2)} < ... < x_{(n)}$, let $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$. For a future observation X_{n+1}, the assumption $A_{(n)}$ is

$$P(X_{n+1} \in (x_{(i-1)}, x_{(i)})) = \frac{1}{n+1}$$

for all $i = 1, 2, ..., n+1$. We assume here, for ease of presentation, that there are no tied observations. These can be dealt with by assuming that such observations differ by a very small amount, a common method to break ties in statistics [51].

Inferences based on $A_{(n)}$ are predictive and nonparametric, and can be considered suitable if there is hardly any knowledge about the random quantity of interest, other than the n observations, or if one does not want to use any such further information in order to derive inferences that are strongly based on the data. The assumption $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but it provides bounds for probabilities via the 'fundamental theorem of probability' [30], which are lower and upper probabilities in imprecise probability theory [1, 2]. The lower and upper probabilities for event A are denoted by $\underline{P}(A)$ and $\overline{P}(A)$, respectively, and can be interpreted in several ways [18]. For example, $\underline{P}(A)$ ($\overline{P}(A)$) can be interpreted as the supremum buying (infimum selling) price for the gamble on event A, which pays 1 if A occur and 0 if not. Alternatively, $\underline{P}(A)$ ($\overline{P}(A)$) can just be

regarded as the maximum lower (minimum upper) bound for a precise probability for A that follows from the assumptions made, we use this interpretation in this thesis. Generally, in imprecise probability theory [2], $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$ and $\underline{P}(A) = 1 - \overline{P}(A^c)$ where A^c is the complement any event to A. These properties hold for all methods in this thesis.

NPI typically leads to lower and upper probabilities for events of interest, which are based on Hill's assumption $A_{(n)}$ and have strong properties from frequentist statistics perspective. As events of interest are explicitly about a future observation, or a function of such an observation, NPI is indeed explicitly about prediction. The NPI lower and upper probabilities have a frequentist interpretation that could be regarded as 'confidence statements' related to repeated application of the same procedure. From this perspective, corresponding lower and upper probabilities can be interpreted as bounds for the confidence level for the event of interest. However, this method does provide neither predictions nor prediction intervals in the classical sense, as e.g. appear in frequentist regression methods. Prediction intervals tend to relate to confidence intervals for model parameter estimates combined with variability included in the model, in NPI no variability is explicitly included in a model and there are clearly no parameters to be estimated.

Augustin and Coolen [1] proved that NPI has attractive inferential properties, it is also exactly calibrated from frequentist statistics perspective [62], which allows interpretation of the NPI lower and upper probabilities as bounds on the long-term ratio with which the event of interest occurs upon repeated application of statistical procedure. One attractive aspect of the NPI approach is that the amount of information available in the data is directly related to the differences between corresponding upper and lower probability, providing a new dimension to uncertainty quantification when compared to statistical methods which use only precise probabilities, such as standard Bayesian and frequentist methods including most commonly used nonparametric methods [23].

As mentioned in Section 1.1, NPI has been developed for a wide range of applications as NPI methods are available for Bernoulli data [17], real-valued data [1], data including right-censored observations [24], ordinal data [34] and multinomial

data [3, 20, 21].

1.3 Outline of the thesis

In Chapter 2 we introduce the main contribution of this thesis, novelty a new method for predictive inference which combines NPI for the marginals with an estimated parametric copula. We investigate the performance of this method via simulations, with particular attention to robustness with regard to the assumed copula in case of small data sets. A paper based on Chapter 2 has been accepted for publication in Journal of Statistical Theory and Practice [25]. In Chapter 3 we combine NPI with nonparametric copulas specifically using a kernel-based method, and we investigate the performance of this method via simulations. This chapter has been presented at the 23^{rd} National Symposium on Mathematical Sciences (*Simposium Kebangsaan Sains Matematik Ke*-23) at Malaysia and a short paper based on it was published in the conference proceedings [72]. We present and illustrate the application of the method proposed in Chapter 2 to a real world scenario in Chapter 4, concerning accuracy of diagnostic tests using Receiver Operating Characteristic (ROC) curves. In this chapter, we introduce a weighted average of bivariate diagnostic test results and we consider the dependence structure in order to maximise the accuracy of the tests involved on combined measurements. We study the performance of the method by simulations. This method raises interesting questions for future research, some brief comments and general conclusions are included in Chapter 5. In Chapters 2 - 4, illustrative examples are presented using data from the literature. In addition to the presentation of results for Chapter 3 mentioned above, this chapter have been regularly presented at several seminars and conferences, including at Northern Postgraduate Mini-Conference in Statistics (NPMCS) 2014, Newcastle University (Oral presentation), Royal Statistical Society (RSS) 2014 International Conference at Sheffield (Poster presentation), Durham Risk Day conference 2014 at Durham (Poster presentation), 4th Annual Survival Analysis for Junior Researchers Conference 2015 at Keele University (Poster presentation), NPMCS 2015 at Durham University (Oral presentation) and European Meeting of Statisticians (EMS) 2015 at

Vrije University, Amsterdam (Poster presentation). For Chapter 4, the results have been presented at Statistics seminar (Oral presentation) and recently at Stat4Grads seminar (Oral presentation) in Durham University.

Chapter 2

NPI with parametric copula

2.1 Introduction

In this chapter, we present main contribution of this thesis, a new novel method for predictive inference which combines NPI for the marginals with an estimated parametric copula. We propose a new semi-parametric method for predictive inference for a future bivariate observation. The proposed method combines NPI in the marginals with an estimated copula to take dependence into account. The proposed method can be used with any parametric copula. Of course, if one has specific knowledge in favour of a particular family of copulas for the application considered, then using this family is most sensible and should lead to best results, if indeed this knowledge is correct. Any of the available methods to estimate the copula parameter can be used, where advantages and disadvantages of specific estimation methods are carried over. In our numerical studies, to investigate the performance of the proposed method and to illustrate its use, we will mention the specific estimation method applied.

Semi-parametric methods using copulas for statistical inference have been presented before, see e.g. [13, 56, 94]. The main approach presented herein involves combining the empirical estimators for the marginals with a parametric copula, in nature this is close to the method presented in this chapter. Even more, Chen et al. [13] use a rescaled empirical estimator which, effectively, deals with the marginals in the same manner as the method used in this chapter. However, these presented

methods in the literature all consider estimation, while our approach in this thesis is explicitly developed for predictive inference.

In Section 2.2 we briefly give an introduction on copulas and specifically parametric copulas. In Section 2.3 we introduce how NPI can be combined with an estimated parametric copula to provide a semi-parametric predictive method. Section 2.4 demonstrates how the proposed semi-parametric predictive method can be used for inference about different events of interest. In Section 2.5 we investigate the performance of this method via simulations, with particular attention to robustness with regard to the assumed copula in case of small data sets. Two examples are presented in Section 2.6 to illustrate application of the method to real world scenarios, these examples use data from the literature. This method raises interesting questions for future research, some brief comments on this are included in Section 2.7.

2.2 Copula

Copula is a statistical concept for modelling dependence of random variables. The copula was invented and first used [73] by Sklar in 1959 [89]. Nelsen [73] presents a detailed introduction and overview. The word copula has been derived from the Latin word "copulare" which means to link, or to connect [36, 73], which is appropriate as the copula models the way in which random quantities are linked or connected.

By the well-known theorem by Sklar [89], every joint cumulative distribution function F of continuous random quantities (X, Y) can be written as $F(x, y) = C(F_X(x), F_Y(y))$, for all $(x, y) \in \mathbb{R}^2$, where F_X and F_Y are the continuous marginal distributions and $C : [0, 1] \times [0, 1] \to [0, 1]$ is a unique copula corresponding to this joint distribution. So, a copula is a joint cumulative distribution function whose marginals are uniformly distributed on $[0, 1]$ [14, 73].

Copulas have become popular tools for modelling dependence between random quantities in many application areas, including finance [14, 80], actuarial science [39, 79], risk management [36], hydrology [41] and reliability analysis [92]. Copulas

are attractive due to their ability to model dependence between random quantities separately from their marginal distributions [14, 73]. Throughout this thesis, attention is restricted to bivariate data, the proposed methods can straightforwardly be generalized to more dimensional data but its performance would need to be studied in detail, this is left as a topic for future research. In this thesis we use some parametric copula models and some nonparametric copula methods. Parametric copulas are used in this chapter and introduced below. Nonparametric copulas are introduced in Section 3.2.

Many parametric families of copulas have been presented in the literature, see e.g. [14, 55, 73]. In this research, we use four common bivariate one-parameter copulas, namely the Normal (or Gaussian), Clayton [15], Frank [38] and Gumbel [45] copulas, these are briefly reviewed below.

The Normal copula, with parameter θ_n, has cumulative distribution function (cdf)

$$C_n(u, v|\theta_n) = \Phi_B(\Phi^{-1}(u), \Phi^{-1}(v)|\theta_n)$$

where Φ is the cdf of the standard normal distribution, and Φ_B is the cdf of the standard bivariate normal distribution with correlation parameter $\theta_n \in (-1, 1)$. The Normal copula is easy to compute and to extend to more dimensions [66]. Moreover, the Normal copula is uniquely defined by the correlation of marginal distributions, thus it is easy to calibrate as this only requires calculating the pairwise correlation. However, Normal copula does not allow tail dependence to see modelled and it is symmetric, therefore it cannot capture interdependence among extreme events and does not allow asymmetric dependence among variables [66].

The Clayton copula [15] has cdf

$$C_c(u, v|\theta_c) = \max[(u^{-\theta_c} + v^{-\theta_c} - 1)^{-1/\theta_c}, 0]$$

with dependence parameter $\theta_c \in [-1, 0) \cup (0, +\infty)$. It is an asymmetric copula, exhibiting greater dependence in the negative tail than in the positive.

The Frank copula [38] has cdf

$$C_f(u, v|\theta_f) = -\theta_f^{-1} \ln \left\{ 1 + \frac{(e^{-\theta_f u} - 1)(e^{-\theta_f v} - 1)}{e^{-\theta_f} - 1} \right\}$$

with dependence parameter $\theta_f \in (-\infty, 0) \cup (0, +\infty)$. It is a symmetric copula.

The Gumbel copula [45] has cdf

$$C_g(u, v|\theta_g) = \exp(-[(-\ln u)^{\theta_g} + (-\ln v)^{\theta_g}]^{1/\theta_g})$$

with dependence parameter $\theta_g \in [1, +\infty)$. The Gumbel copula (also known as Gumbel-Hougaard copula [73]) is an asymmetric copula. The Gumbel copula models strong right-tail dependence and relatively weak left-tail dependence [93].

These four commonly used copulas all have their own characteristics as mentioned above. There is a one-to-one relationship between the dependence parameters of these four copulas and the concordance measure Kendall's tau, τ, as given in Table 2.1 [14], note that the Gumbel copula cannot be used to model negative dependence, so it can only correspond to $\tau \geq 0$, and Frank copula does not allow $\tau = 0$.

Family	Parameter range	τ
Normal	$\theta_n \in (-1, 1)$	$\frac{2}{\pi} \arcsin \theta_n$
Clayton	$\theta_c \in [-1, 0) \cup (0, +\infty)$	$\theta_c/(\theta_c + 2)$
Frank	$\theta_f \in (-\infty, 0) \cup (0, +\infty)$	$1 - 4/\theta_f[1 - D_1(\theta_f)]$
Gumbel	$\theta_g \in [1, +\infty)$	$1 - 1/\theta_g$

Note: $D_1(\theta) = \int_0^\theta (x/\theta)/(e^x - 1)dx$ is the first Debye function [14].

Table 2.1: Relationship between dependence parameters and Kendall's tau, τ

Many methods to estimate the parameter of a copula have been presented in the literature, see e.g. in [14, 80, 93]. There are several well known methods for estimating the parameter of a parametric copula, such as maximum likelihood estimator (MLE), inference functions of margins (IFM) [55], pseudo maximum likelihood estimation or canonical maximum likelihood [14] and method-of-moment [61]. The IFM estimation method is a two-stage estimation method which is based on MLE and is also known as multi-stage maximum likelihood (MSML) estimation [55]. This method allows us to estimate the parameters separately for the marginals and the copula. The method-of-moment approaches are based on the inversion of a consistent estimator of a moment of the copula, such as Spearmans rho, these are discussed

in detail in [61]. In the presentation of our method, we will denote a parameter estimate by $\hat{\theta}$ without the need to specify a particular estimation method.

There are advantages and disadvantages of the estimation methods, for example, MLE can be computationally intensive in the case of high dimensional distributions, because the number of parameters to be estimated simultaneously can be large. The problem might also occur when we have a very large sample size. The estimation of the estimator covariance matrices of the IFM is difficult both analytically and numerically due to the need to compute many derivatives in higher dimension [55], which should be considered when to generalize the method proposed to more than two dimensions. In addition, these two parametric methods are not robust against misspecification of the marginal distributions [58]. This problem has been argued by many researchers who advocate that the estimation of θ should not be affected by the choice of marginal distribution functions. The pseudo maximum likelihood estimation method solves this problem, it is discussed in details in Genest et al. [42] and in Shih and Louis [87].

2.3 Combining NPI with a parametric copula

In this section we present NPI with a parametric copula to provide a semi-parametric predictive method. The proposed semi-parametric predictive method consists of two steps. The first step is to use NPI for the marginals, the second step is to use a bivariate parametric copula and estimate the parameter value, to take the dependence structure in the data into account.

The first step is to use NPI for the marginals. Suppose that we have n bivariate real-valued observations (x_i, y_i), $i = 1, \ldots, n$, which are the observed values of n exchangeable bivariate random quantities. Henceforth, to simplify notation, we will actually use x_i and y_j to denote the ordered observations when considering the marginals, so $x_1 < \ldots < x_i < \ldots < x_n$ and $y_1 < \ldots < y_j < \ldots < y_n$. So it is important that, with the plain indices now related to the separately ordered data related to the marginals, the values x_i and y_i do not form an observed pair. It should be emphasized that the information about the actual observation pairs is only used

in the second step, where the parameter value of the assumed copula is estimated, the first step considers the marginals and hence only uses the information consisting of either the n observations x_i or the n observations y_j.

We are interested in prediction of one future bivariate observation, denoted by (X_{n+1}, Y_{n+1}). Using the assumption $A_{(n)}$ we derive a partially specified predictive probability distribution for X_{n+1}, given the observations $x_1, \ldots, x_n$, and similarly a partially specified predictive probability distribution for Y_{n+1}, given the observations $y_1, \ldots, y_n$. These are as follows:

$$P(X_{n+1} \in (x_{i-1}, x_i)) = \frac{1}{n+1} \quad \text{and} \quad P(Y_{n+1} \in (y_{j-1}, y_j)) = \frac{1}{n+1}$$

for $i, j = 1, 2, \ldots, n+1$, where $x_0 = -\infty$, $x_{n+1} = \infty$, $y_0 = -\infty$ and $y_{n+1} = \infty$ are introduced for simplicity of notation.

To link this first step to the second step, where the dependence structure in the observed data is taken into account in order to provide a partially specified predictive distribution for the bivariate (X_{n+1}, Y_{n+1}), we introduce a natural transformation of these two random quantities individually. Let $\widetilde{X}_{n+1}$ and $\widetilde{Y}_{n+1}$ denote transformed versions of the random quantities X_{n+1} and Y_{n+1}, respectively, following from the natural transformations related to the marginal $A_{(n)}$ assumptions,

$$\begin{gathered}(X_{n+1} \in (x_{i-1}, x_i), Y_{n+1} \in (y_{j-1}, y_j)) \Longleftrightarrow \\ \left(\widetilde{X}_{n+1} \in \left(\frac{i-1}{n+1}, \frac{i}{n+1}\right), \widetilde{Y}_{n+1} \in \left(\frac{j-1}{n+1}, \frac{j}{n+1}\right)\right)\end{gathered}$$

for $i, j = 1, 2, \ldots, n+1$. This is a transformation from the real plane $\mathbb{R}^2$ into $[0,1]^2$ where, based on n bivariate data, $[0,1]^2$ is divided into $(n+1)^2$ equal-sized squares. The $A_{(n)}$ assumptions for the marginals lead to

$$\begin{aligned} P(\widetilde{X}_{n+1} \in \left(\frac{i-1}{n+1}, \frac{i}{n+1}\right)) &= P(X_{n+1} \in (x_{i-1}, x_i)) = \frac{1}{n+1} \\ P(\widetilde{Y}_{n+1} \in \left(\frac{j-1}{n+1}, \frac{j}{n+1}\right)) &= P(Y_{n+1} \in (y_{j-1}, y_j)) = \frac{1}{n+1} \end{aligned}$$

for $i, j = 1, 2, \ldots, n+1$. Note that, following these transformations of the marginals, we have discretized uniform marginal distributions on $[0,1]$, which therefore fully correspond to copulas, as any copula will provide exactly the same discretized uniform marginal distributions. Hence, this basic transformation shows that the NPI

approach for the marginals can be easily combined with any copula model to reflect the dependence structure, leading naturally to the second step of our method.

The second step is to assume a bivariate parametric copula and estimate the parameter value. In this second step, the proposed method deals with the information, in the observed data, with regard to dependence of the two random quantities X_{n+1} and Y_{n+1}. A bivariate parametric copula is assumed, with parameter θ. Using the data, the parameter can be estimated by any statistical method, e.g. maximum likelihood estimation or a convenient (for computation) variation to it, resulting in a point estimate $\hat{\theta}$. In order to correspond to the transformation method for the marginals, and to avoid having to consider the marginals whilst estimating the copula parameter, at this stage we use also transformed data, where each observed pair (x_i, y_i), $i = 1, \ldots, n$, is replaced by $(r_i^x/(n+1), r_i^y/(n+1))$, with r_i^x the rank of the observation x_i among the n x-observations (where the smallest value has rank 1), and similarly r_i^y the rank of y_i among the n y-observations. It should be noticed that, as this estimation process does not involve any estimation of the marginals, it can be performed in a computationally efficient manner, as it is often the simultaneous estimation of the copula and related marginals that may cause computational difficulties in other statistical methods using copulas.

NPI on the marginals can now be combined with the estimated copula by defining the following probability for the event that the transformed pair $(\widetilde{X}_{n+1}, \widetilde{Y}_{n+1})$ belongs to a specific square from the $(n+1)^2$ squares into which the space $[0,1]^2$ has been partitioned,

$$h_{ij}(\hat{\theta}) = P_C(\widetilde{X}_{n+1} \in \left(\frac{i-1}{n+1}, \frac{i}{n+1}\right), \widetilde{Y}_{n+1} \in \left(\frac{j-1}{n+1}, \frac{j}{n+1}\right) | \hat{\theta}) \tag{2.1}$$

for $i, j = 1, 2, \ldots, n+1$, with $P_C(\cdot|\hat{\theta})$ representing the copula-based probability with estimated parameter value $\hat{\theta}$, and the corresponding cumulative distribution function,

$$H_{ij}(\hat{\theta}) = P_C(\widetilde{X}_{n+1} \leq \frac{i}{n+1}, \widetilde{Y}_{n+1} \leq \frac{j}{n+1} | \hat{\theta}) = \sum_{k=1}^{i} \sum_{l=1}^{j} h_{kl}(\hat{\theta}) \tag{2.2}$$

Equations (2.1) and (2.2) can be represented by Figure 2.1. These $(n+1)^2$ values $h_{ij}(\hat{\theta})$, which sum up to 1, provide the complete discretized probability distribution

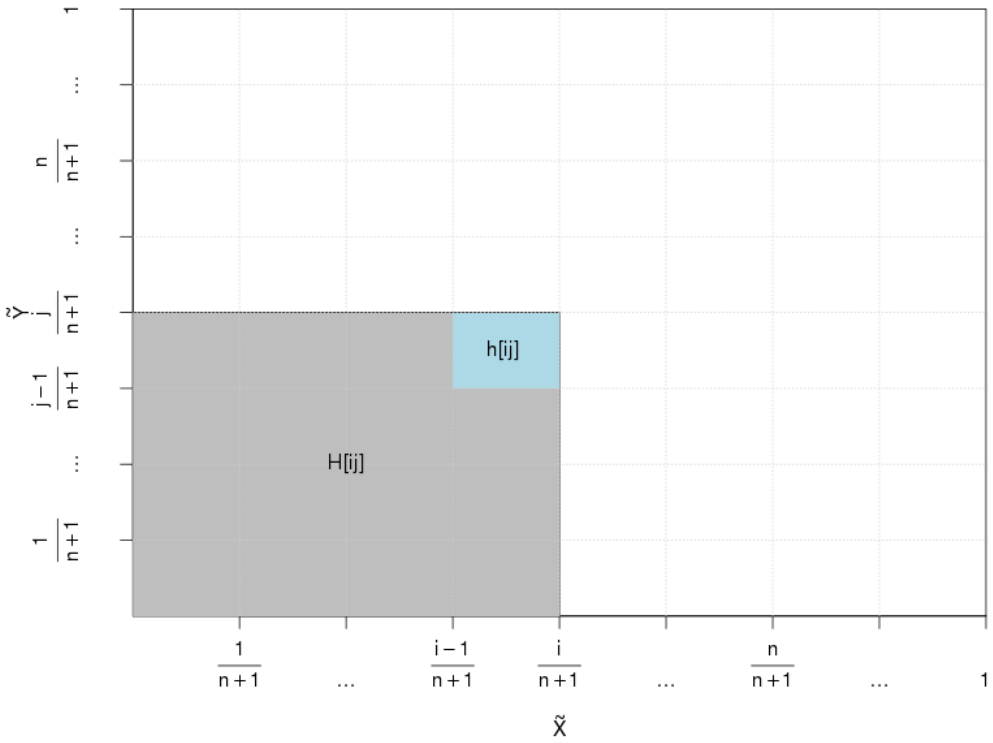


Figure 2.1: Presentation of probabilities h_{ij} and H_{ij} with an estimated copula

for the transformed future observations, which can be used for statistical inference on the actual future observation (X_{n+1}, Y_{n+1}) or an event of interest involving this bivariate random quantity, as explained in the next section. The probabilities h_{ij} must satisfy the following conditions;

1. $\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij} = 1$
2. $\sum_{j=1}^{n} h_{ij} = \frac{1}{n+1}, \forall i \in (1, ..., n+1)$, and $\sum_{i=1}^{n} h_{ij} = \frac{1}{n+1}, \forall j \in (1, ..., n+1)$
3. $h_{ij} \geq 0, \forall i, j = 1, ..., n+1$.

These conditions will hold by the choice of a proper parametric copula. Note that, although a completely specified copula is used initially, for our inferences we only use the discretized version on the $(n+1)^2$ equal-sized squares with probabilities $h_{ij}(\hat{\theta})$. In this discretized setting, $h_{ij}(\hat{\theta}) = \frac{1}{(n+1)^2}$ for all $i, j = 1, \ldots, n+1$ would indicate complete independence of $\widetilde{X}_{n+1}$ and $\widetilde{Y}_{n+1}$, and hence of X_{n+1} and Y_{n+1}.

Furthermore, $h_{ij}(\hat{\theta}) = \frac{1}{(n+1)}$ for all $i = j = 1, \ldots, n+1$ would correspond to correlation 1 between these random quantities (both for the transformed and the actual future observations), while correlation -1 would correspond to $h_{ij}(\hat{\theta}) = \frac{1}{(n+1)}$ for all $j = (n+2) - i$ with $i = 1, \ldots, n+1$. For example, consider $n = 4$ and the corresponding h_{ij} for -1.00 correlation, 1.00 correlation and no correlation are given in Tables 2.2, 2.3 and 2.4.

$j = 5$	0.2000	0.0000	0.0000	0.0000	0.0000
$j = 4$	0.0000	0.2000	0.0000	0.0000	0.0000
$j = 3$	0.0000	0.0000	0.2000	0.0000	0.0000
$j = 2$	0.0000	0.0000	0.0000	0.2000	0.0000
$j = 1$	0.0000	0.0000	0.0000	0.0000	0.2000
h_{ij}	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$

Table 2.2: The probability of h_{ij}

$j = 5$	0.0000	0.0000	0.0000	0.0000	0.2000
$j = 4$	0.0000	0.0000	0.0000	0.2000	0.0000
$j = 3$	0.0000	0.0000	0.2000	0.0000	0.0000
$j = 2$	0.0000	0.2000	0.0000	0.0000	0.0000
$j = 1$	0.2000	0.0000	0.0000	0.0000	0.0000
h_{ij}	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$

Table 2.3: The probability of h_{ij}

$j = 5$	0.0400	0.0400	0.0400	0.0400	0.0400
$j = 4$	0.0400	0.0400	0.0400	0.0400	0.0400
$j = 3$	0.0400	0.0400	0.0400	0.0400	0.0400
$j = 2$	0.0400	0.0400	0.0400	0.0400	0.0400
$j = 1$	0.0400	0.0400	0.0400	0.0400	0.0400
h_{ij}	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$

Table 2.4: The probability of h_{ij}

2.4 Semi-parametric predictive inference

In this section, the semi-parametric predictive method presented in Section 2.3 is used for inference about an event which involves the next bivariate observation (X_{n+1}, Y_{n+1}). Let $E(X_{n+1}, Y_{n+1})$ denote the event of interest. Let $\underline{P}(E(X_{n+1}, Y_{n+1}))$ and $\overline{P}(E(X_{n+1}, Y_{n+1}))$ be the lower and upper probabilities, based on our semi-parametric method, for this event to be true. As explained in the previous section, the observed data (x_i, y_i), $i = 1, \ldots, n$, divide $\mathbb{R}^2$ into $(n+1)^2$ blocks $B_{ij} = (x_{i-1}, x_i) \odot (y_{j-1}, y_j)$, for $i, j = 1, \ldots, n+1$ (with, as before, $x_0 = -\infty, x_{n+1} = \infty, y_0 = -\infty, y_{n+1} = \infty$ defined for ease of notation). Figure 2.2 shows the area of blocks $B_{ij} = (x_{i-1}, x_i) \odot (y_{j-1}, y_j)$, for $i, j = 1, \ldots, n+1$.

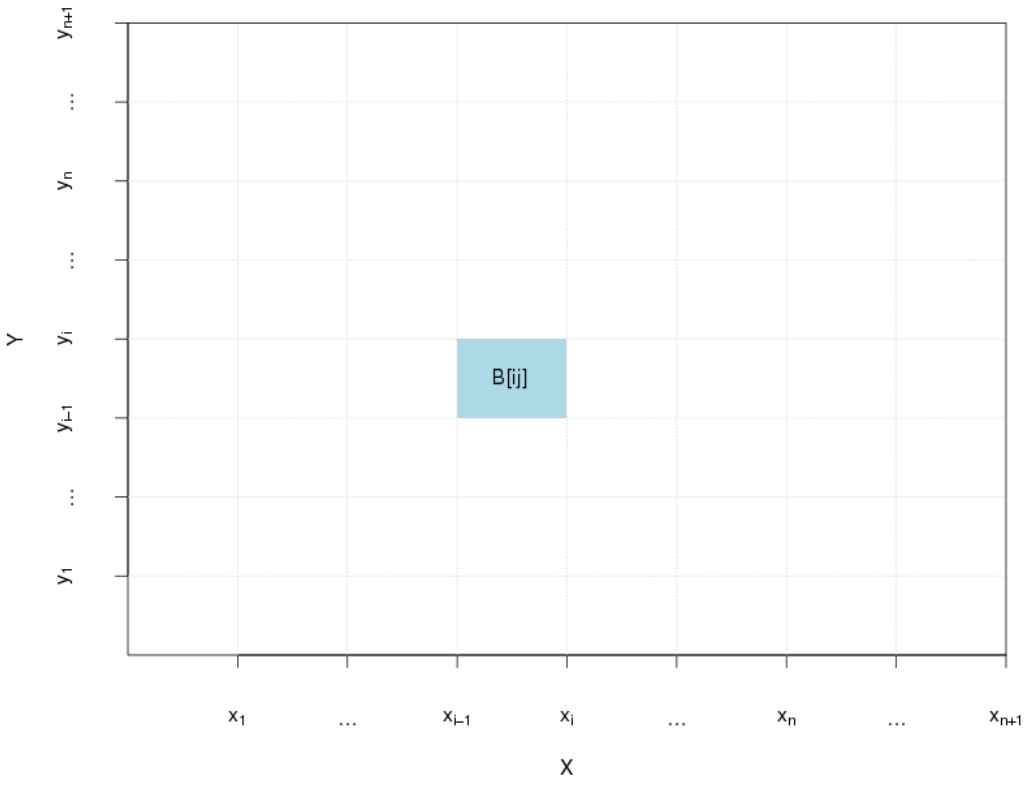


Figure 2.2: Presentation of area of blocks $B_{ij} = (x_{i-1}, x_i) \odot (y_{j-1}, y_j)$

We further define

$$E(x, y) = \begin{cases} 1 & \text{if } E(X_{n+1}, Y_{n+1}) \text{ is true for } X_{n+1} = x \text{ and } Y_{n+1} = y \\ 0 & \text{else.} \end{cases}$$

The fact that we work with a discretized probability distribution leads to imprecise probabilities as follows [2]. We define $\overline{E}_{ij} = \max_{(x,y)\in B_{ij}} E(x,y)$, so $\overline{E}_{ij} = 1$ if there is at least one $(x,y) \in B_{ij}$ for which $E(x,y) = 1$, else $\overline{E}_{ij} = 0$. Furthermore, we define $\underline{E}_{ij} = \min_{(x,y)\in B_{ij}} E(x,y)$, so $\underline{E}_{ij} = 1$ if $E(x,y) = 1$ for all $(x,y) \in B_{ij}$, else $\underline{E}_{ij} = 0$. The semi-parametric method presented in the previous section leads to the following lower and upper probabilities for the event $E(X_{n+1}, Y_{n+1})$,

$$\underline{P}(E(X_{n+1}, Y_{n+1})) = \sum_{i,j} \underline{E}_{ij}\ h_{ij}(\hat{\theta}) \tag{2.3}$$

$$\overline{P}(E(X_{n+1}, Y_{n+1})) = \sum_{i,j} \overline{E}_{ij}\ h_{ij}(\hat{\theta}) \tag{2.4}$$

Many events of interest can be considered with the summations over all $i, j = 1, ..., n+1$. Suppose, for example, that we are interested in the sum of the next observations X_{n+1} and Y_{n+1}, say $T_{n+1} = X_{n+1} + Y_{n+1}$. Then the lower probability for the event that the sum of the next observations will exceed a particular value t is

$$\underline{P}(T_{n+1} > t) = \sum_{(i,j)\in L_t} h_{ij}(\hat{\theta}) \tag{2.5}$$

with $L_t = \{(i,j) : x_{i-1} + y_{j-1} > t, 1 \leq i \leq n+1, 1 \leq j \leq n+1\}$, and the corresponding upper probability is

$$\overline{P}(T_{n+1} > t) = \sum_{(i,j)\in U_t} h_{ij}(\hat{\theta}) \tag{2.6}$$

with $U_t = \{(i,j) : x_i + y_j > t, 1 \leq i \leq n+1, 1 \leq j \leq n+1\}$. Equations (2.5) and (2.6) represent the lower and upper survival functions for the future observation T_{n+1}, based on our newly presented semi-parametric method, we denote these by $\underline{S}(t) = \underline{P}(T_{n+1} > t)$ and $\overline{S}(t) = \overline{P}(T_{n+1} > t)$ and will use them in our analysis of the predictive performance of our method in Section 2.5.

Before analysing the performance of this new semi-parametric method, it is useful to explain the idea behind it. As mentioned in Section 1.2, NPI has been developed over the last two decades for many applications and it has excellent frequentist properties, but it relies on the natural ordering of the observed data (or on an assumed underlying latent variable representation with a natural ordering [19]). Moving to multivariate observations, however, causes problems due to the absence of a natural

ordering. At the same time, copulas have proved to be powerful tools to model dependence, and, as shown in Section 2.3, they can be linked in an attractive manner to NPI on the marginals, via discretization after a straightforward transformation. The resulting semi-parametric method is, however, a heuristic approach, in that it lacks the theoretical properties which make NPI for real-valued (one-dimensional) observations an attractive frequentist statistics method.

In Section 2.5 we show how the predictive performance of this method can be analysed, focussing on a case where interest is in the sum of X_{n+1} and Y_{n+1}. This will also illustrate aspects of the imprecision in relation to the number of data observations and the dependence structure in the data.

2.5 Predictive performance

To investigate the predictive performance of the semi-parametric method presented in Sections 2.3 and 2.4, we conduct a simulation study. In each run of the simulation $N = 10,000$ bivariate samples are generated, each of size $n+1$, where we have used $n = 10, 50, 100$. For each simulated sample, the first n pairs are used as the data for the proposed semi-parametric predictive model, with the additional simulated pair used to test the predictive performance of this method.

In this analysis, we focus on the sum of of the next observations, so $T_{n+1} = X_{n+1} + Y_{n+1}$, as presented in Section 2.4. Let (x_i^j, y_i^j) be the jth simulated sample, consisting of n pairs, so with subscript $i = 1, 2, \ldots, n$ indicating the pair within one sample, and superscript $j = 1, 2, \ldots, N$ indicating the specific simulated sample. Let (x_f^j, y_f^j) be the additional simulated ('future') pair for sample j, and let the corresponding sum be denoted by $t_f^j = x_f^j + y_f^j$, for $j = 1, 2, \ldots, N$. For $q \in (0, 1)$, the inverse values of the lower and upper survival functions of T_{n+1} in equations (2.5) and (2.6), can be defined as

$$\underline{t}_q = \underline{S}^{-1}(q) = \inf_{t \in \mathbb{R}} \{\underline{S}(t) \leqslant q\} \tag{2.7}$$

$$\overline{t}_q = \overline{S}^{-1}(q) = \inf_{t \in \mathbb{R}} \{\overline{S}(t) \leqslant q\} \tag{2.8}$$

where $\underline{t}_q \leq \overline{t}_q$ obviously holds. It is reasonable to claim that the proposed semi-

parametric predictive method performs well if the two following inequalities hold,

$$p_1 = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}(t_f^j \geqslant \bar{t}_q^j) \leq q \tag{2.9}$$

$$p_2 = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}(t_f^j \geqslant \underline{t}_q^j) \geq q \tag{2.10}$$

We will investigate the performance in this manner for $q = 0.25, 0.50, 0.75$. One could of course investigate different quantiles but these values will provide a good picture of the performance of the method, together with some particular aspects which are important to illustrate. To perform the simulation, we consider different values of Kendall's τ in order to study the method under different levels of dependence. For each, we simulate from an assumed parametric copula with the parameter set at the value which corresponds to τ as presented in Table 2.1.

We consider two main scenarios. First, we actually assume in our semi-parametric method a copula from the same parametric family as used for simulation. Secondly, we use an assumed parametric copula in our method which differs from the copula used for the simulation. For the first case, we expect the method to perform well. Of course, this scenario is highly unlikely in practice, but it is important to study the performance of the method in this case, and the simulations will also enable study of the level of imprecision in the predictive inferences. The second scenario is more important, as it represents a more likely practical situation, namely where a parametric copula is assumed but this is actually not fully in line with the data generating mechanism. This can be considered as misspecification, and it is in such scenarios that we hope that our method will provide sufficient robustness to still provide relatively good quality predictive inference.

Given the simulated data in a single run, we estimate the parameter of the assumed parametric copula using the pseudo maximum likelihood method, which was briefly reviewed in Section 2.2 and is included in the R package `VineCopula` [83]. We used this estimation method because we need to have a fast algorithm in order to use the copula parameter estimation as part of the method proposed in Sections 2.3 and 2.4. In addition, this method was considered the best estimation method

by [42]. However, any alternative estimation methods can be used; of course these may lead to slightly different results, but the overall performance of the method is unlikely to be affected much by minor differences in the estimation method. With the estimate $\hat{\theta}$ for the copula parameter, we obtain the probabilities $h_{ij}(\hat{\theta})$ as given in equation (2.1), and these form the basis for all possible inferences of interest.

We have run $N = 10,000$ simulations with sample sizes $n = 10, 50, 100$, and with $q = 0.25, 0.50, 0.75$ and $\tau = -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75$. We restricted attention to the four parametric copulas discussed in Section 2.2, noting that the Frank copula does not allow $\tau = 0$ and the Gumbel copula cannot be used to model negative dependence.

First, we applied our semi-parametric method with the assumed copula actually belonging to the same parametric family used for the data generation. Tables 2.5, 2.6, 2.7 and 2.8 present the results for the Clayton, Frank, Normal and Gumbel copula, respectively. These tables report the values p_1 and p_2 for the different values of τ and n, as described in equations (2.9) and (2.10), for $q = 0.25, 0.50, 0.75$. For good performance of our method, we require $p_1 \leq q \leq p_2$. Furthermore, these tables also present a value $\hat{\theta}$, this is the average of the 10,000 estimates of the parameter, so for these tables this value is expected to be close to the value for θ which corresponds directly to the τ used, and which is given in the second column of each table. However, we will not focus on these estimated values as it is really the predictive performance that is important to consider, due to the predictive nature of our approach. It is clear though that the parameter estimates tend to be closer to the real value for larger values of n, which is of course fully as expected. It may be of interest to implement other estimation methods for the copula parameter, which may provide a slightly better performance, detailed study of this is left as a topic for future research.

Most cases in Tables 2.5, 2.6, 2.7 and 2.8 have $q \in [p_1, p_2]$, which shows an overall good performance of our semi-parametric predictive method, which is fully in line with expectations due to the use of the same parametric copula family in our method as the one that was actually used to simulate the data.

These tables illustrate two important aspects of the imprecision in our method.

First, for corresponding cases with increasing n, the imprecision, reflected through the difference $p_2 - p_1$, decreases. This is logical from the perspective that more data allow more precise inferences, which is common in statistical methods using imprecise probabilities [2]. Indeed, if one increases the value of n further, imprecision will decrease to 0 in the limit, where, informally, limit arguments are based on NPI for the marginals converging to the empirical marginal distributions, which in turn will converge to the underlying distributions, and with the assumed copula actually belonging to the same family as the one used to generate the data, this also will ensure an increasingly good performance of the method for increasing n.

A perhaps somewhat less expected feature of our method is seen by comparing corresponding cases with the same absolute value of τ, but negative τ compared to positive τ. For such cases, the imprecision $p_2 - p_1$ is always greater with the negative correlation than with the positive correlation, and this effect is stronger the larger the absolute value of the correlation. This feature occurs due to the fact that we are considering events $T_{n+1} = X_{n+1} + Y_{n+1} > t$, and can be explained by considering the probabilities $h_{ij}(\hat{\theta})$ which are the key ingredients of our method for inference. In case of positive correlation, the $h_{ij}(\hat{\theta})$ tend to be largest for values of i and j close to each other, while for negative correlation this is the case for values of i and j with sum near to $n + 2$, and this effect is stronger the larger the absolute value of the correlation. Calculating the lower and upper probabilities, equations (2.5) and (2.6) tends to include several more $h_{ij}(\hat{\theta})$ values in the latter than in the former, and for events $T_{n+1} > t$ these extra $h_{ij}(\hat{\theta})$ included in the upper probability tend to have the sum of their subscripts i and j about constant. Hence, for positive correlation these extra $h_{ij}(\hat{\theta})$ tend to include a few larger values for most values of t. For negative correlation the effect is quite different, as then these extra $h_{ij}(\hat{\theta})$ tend to include relatively small values for small and for large values of t, in relation to the observed data, but when t is closer to the center of the empirical distribution of the values $x_i + y_i$, corresponding to the n data pairs (x_i, y_i), then many of the extra $h_{ij}(\hat{\theta})$ are quite large, resulting in large imprecision. This effect can also be seen from plots of the lower and upper survival functions for T_{n+1} shown in Figure 2.3 and Figure 2.4 for the Clayton and Frank copulas, respectively. The plots of the lower and upper

			$n = 10$			$n = 50$			$n = 100$		
τ	θ	q	$\hat{\theta}$	p_1	p_2	$\hat{\theta}$	p_1	p_2	$\hat{\theta}$	p_1	p_2
-0.75	-6.0000	0.25	-9.7782	0.0806	0.5130	-5.8932	0.1859	0.2904	-5.8691	0.2233	0.2741
		0.5		0.2171	0.7735		0.3963	0.5992		0.4428	0.5653
		0.75		0.4770	0.9193		0.7009	0.7992		0.7376	0.7841
-0.5	-2.0000	0.25	-3.1214	0.1581	0.4114	-2.1369	0.2234	0.2732	-2.0693	0.2383	0.2653
		0.5		0.3350	0.6711		0.4526	0.5545		0.4712	0.5286
		0.75		0.5935	0.8427		0.7207	0.7710		0.7377	0.7640
-0.25	-0.6667	0.25	-1.3182	0.1995	0.3742	-0.7863	0.2436	0.2820	-0.7358	0.2405	0.2584
		0.5		0.3919	0.6188		0.4743	0.5312		0.4840	0.5186
		0.75		0.6381	0.8095		0.7235	0.7579		0.7354	0.7528
0.25	0.6667	0.25	1.3232	0.1737	0.2939	0.7934	0.2342	0.2587	0.7349	0.2380	0.2518
		0.5		0.4289	0.5627		0.4784	0.5081		0.4876	0.5018
		0.75		0.7143	0.8119		0.7451	0.7658		0.7457	0.7561
0.5	2.0000	0.25	3.0532	0.1836	0.2953	2.1431	0.2455	0.2711	2.0681	0.2380	0.2516
		0.5		0.4487	0.5522		0.4962	0.5200		0.4916	0.5028
		0.75		0.7091	0.7931		0.7460	0.7651		0.7479	0.7563
0.75	6.0000	0.25	10.1198	0.1970	0.2979	5.8992	0.2342	0.2596	5.8700	0.2458	0.2569
		0.5		0.4587	0.5526		0.4922	0.5098		0.4933	0.5028
		0.75		0.7132	0.8039		0.7337	0.7535		0.7427	0.7529

Table 2.5: Predictive performance, Clayton copula

survival functions for T_{n+1} for the Normal and Gumbel copulas was very similar. For all these copulas, positive correlation leads to imprecision for the events considered here being fairly similar over the whole range, while for negative correlation there is little imprecision in the tails but much imprecision near the center of the empirical distribution of the T_{n+1}.

			$n = 10$			$n = 50$			$n = 100$		
τ	θ_f	q	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2
-0.75	-14.1385	0.25	-15.5793	0.0675	0.4846	-13.9428	0.1927	0.2960	-14.0058	0.2084	0.2677
		0.50		0.2364	0.7453		0.4232	0.5663		0.4467	0.5270
		0.75		0.4924	0.9249		0.6934	0.8006		0.7204	0.7784
-0.50	-5.7363	0.25	-6.9835	0.1578	0.4040	-5.8859	0.2263	0.2817	-5.7992	0.2320	0.2624
		0.50		0.3494	0.6661		0.4635	0.5480		0.4725	0.5144
		0.75		0.6092	0.8569		0.7282	0.7838		0.7259	0.7552
-0.25	-2.3719	0.25	-3.0634	0.1769	0.3533	-2.4751	0.2340	0.2727	-2.4138	0.2377	0.2572
		0.50		0.3941	0.6099		0.4797	0.5323		0.4787	0.5088
		0.75		0.6482	0.8207		0.7349	0.7688		0.7375	0.7580
0.25	2.3719	0.25	3.0129	0.2045	0.3026	2.4784	0.2364	0.2604	2.4088	0.2452	0.2549
		0.50		0.4376	0.5583		0.4854	0.5135		0.4889	0.5048
		0.75		0.6980	0.8052		0.7345	0.7583		0.7447	0.7580
0.50	5.7363	0.25	6.9335	0.1962	0.2989	5.8935	0.2382	0.2578	5.7972	0.2401	0.2526
		0.50		0.4498	0.5517		0.4843	0.5075		0.4922	0.5025
		0.75		0.7065	0.8052		0.7370	0.7568		0.7432	0.7554
0.75	14.1385	0.25	15.6739	0.1960	0.2898	13.8912	0.2429	0.2643	14.0050	0.2443	0.2551
		0.50		0.4541	0.5487		0.4927	0.5127		0.4943	0.5053
		0.75		0.7135	0.7998		0.7398	0.7607		0.7481	0.7557

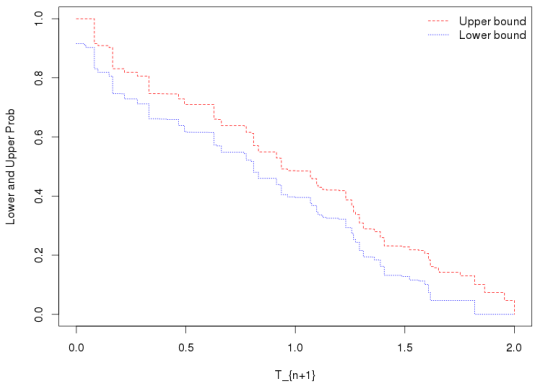
Table 2.6: Predictive performance, Frank copula

			$n = 10$			$n = 50$			$n = 100$		
τ	θ_n	q	$\hat{\theta}_n$	p_1	p_2	$\hat{\theta}_n$	p_1	p_2	$\hat{\theta}_n$	p_1	p_2
-0.75	-0.9239	0.25	-0.9181	0.0854	0.5099	-0.9212	0.2002	0.3015	-0.9228	0.2202	0.2761
		0.50		0.2477	0.7533		0.4187	0.5871		0.4566	0.5544
		0.75		0.4911	0.9153		0.7045	0.8026		0.7311	0.7810
-0.50	-0.7071	0.25	-0.7462	0.1534	0.4002	-0.7235	0.2355	0.2919	-0.7169	0.2465	0.2691
		0.50		0.3342	0.6466		0.4641	0.5529		0.4848	0.5292
		0.75		0.5798	0.8355		0.7252	0.7797		0.7344	0.7604
-0.25	-0.3827	0.25	-0.4473	0.1942	0.3672	-0.4128	0.2406	0.2767	-0.3997	0.2408	0.2597
		0.50		0.3943	0.6121		0.4728	0.5296		0.4894	0.5156
		0.75		0.6386	0.8084		0.7303	0.7639		0.7412	0.7570
0.00	0	0.25	-0.0010	0.1877	0.3139	-0.0008	0.2362	0.2635	0.0000	0.2431	0.2566
		0.50		0.4102	0.5723		0.4711	0.5105		0.4933	0.5141
		0.75		0.6665	0.7971		0.7323	0.7626		0.7466	0.7598
0.25	0.3827	0.25	0.4478	0.1847	0.2956	0.4113	0.2279	0.2505	0.4004	0.2454	0.2556
		0.50		0.4286	0.5538		0.4766	0.5074		0.4908	0.5026
		0.75		0.6968	0.8057		0.7369	0.7580		0.7437	0.7540
0.50	0.7071	0.25	0.7469	0.2011	0.2931	0.7224	0.2394	0.2595	0.7164	0.2440	0.2525
		0.50		0.4500	0.5554		0.4788	0.5033		0.4898	0.5026
		0.75		0.7021	0.7978		0.7326	0.7537		0.7489	0.7602
0.75	0.9239	0.25	0.9174	0.2009	0.2865	0.9211	0.2430	0.2629	0.9224	0.2417	0.2524
		0.50		0.4465	0.5441		0.4980	0.5168		0.4933	0.5039
		0.75		0.6986	0.7961		0.7411	0.7607		0.7430	0.7527

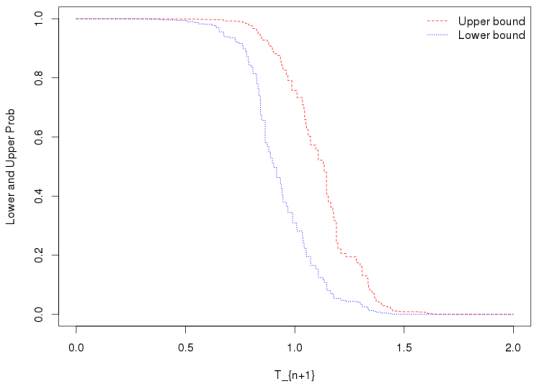
Table 2.7: Predictive performance, Normal copula

			$n = 10$			$n = 30$			$n = 50$			$n = 100$		
τ	θ_g	q	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2
0	1.0000	0.25	1.2216	0.1735	0.2955	1.0699	0.2195	0.2642	1.0467	0.2311	0.2568	1.0266	0.2367	0.2515
		0.5		0.4251	0.5871		0.4722	0.5331		0.4837	0.5199		0.4931	0.5113
		0.75		0.7063	0.8231		0.7372	0.7827		0.7469	0.7753		0.7491	0.7636
0.25	1.3333	0.25	1.6911	0.1865	0.2861	1.4397	0.2237	0.2573	1.3973	0.2330	0.2548	1.3680	0.2451	0.2546
		0.5		0.4288	0.5623		0.4695	0.5212		0.4776	0.5053		**0.5008**	0.5156
		0.75		0.7032	0.8151		0.7355	0.7757		0.7396	0.7642		**0.7551**	0.7693
0.5	2.0000	0.25	2.6425	0.1961	0.2912	2.1723	0.2342	0.2684	2.1015	0.2371	0.2584	2.0514	0.2479	0.2582
		0.5		0.4387	0.5488		0.4865	0.5257		0.4877	0.5128		**0.5013**	0.5134
		0.75		0.7011	0.8072		0.7346	0.7710		0.7452	0.7673		**0.7556**	0.7679
0.75	4.0000	0.25	5.9120	0.2005	0.2870	4.1538	0.2335	0.2639	4.0598	0.2384	0.2575	4.0221	0.2502	0.2601
		0.5		0.4557	0.5481		0.4835	0.5152		0.4881	0.5058		0.4997	0.5099
		0.75		0.7012	0.7994		0.7287	0.7608		0.7384	0.7609		0.7445	0.7562

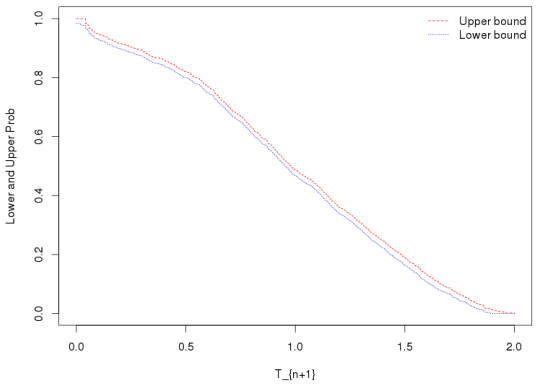
Table 2.8: Predictive performance, Gumbel copula



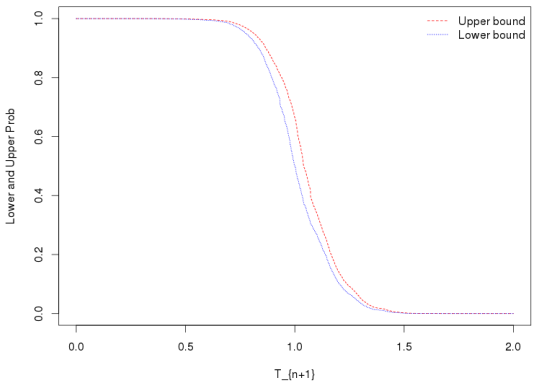
(a) $\tau = 0.75$; $n = 10$



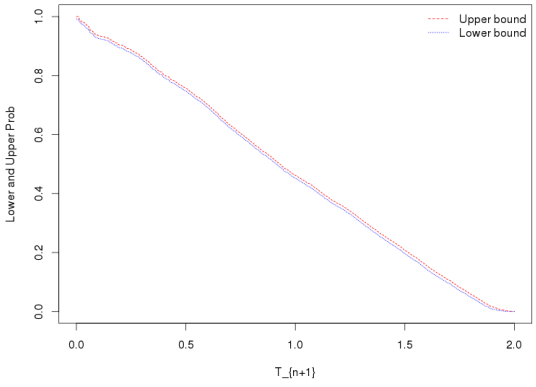
(b) $\tau = -0.75$; $n = 10$



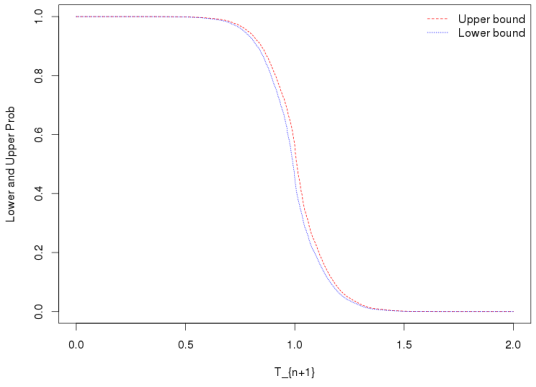
(c) $\tau = 0.75$; $n = 50$



(d) $\tau = -0.75$; $n = 50$

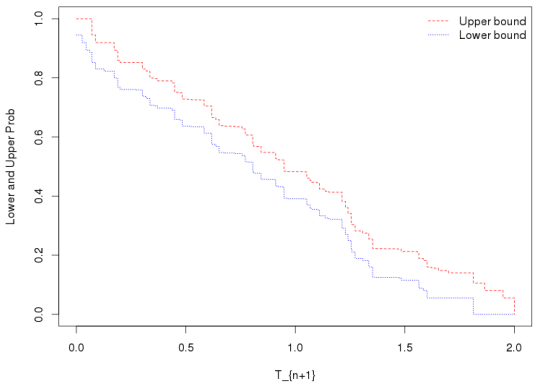


(e) $\tau = 0.75$; $n = 100$

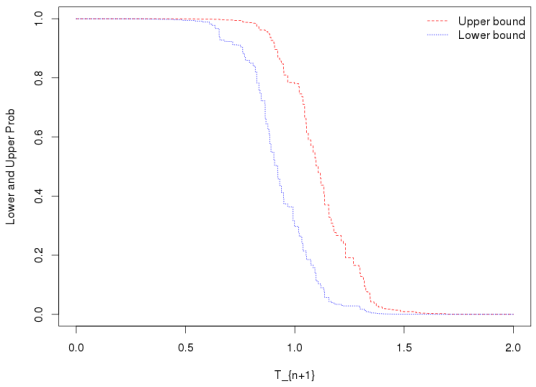


(f) $\tau = -0.75$; $n = 100$

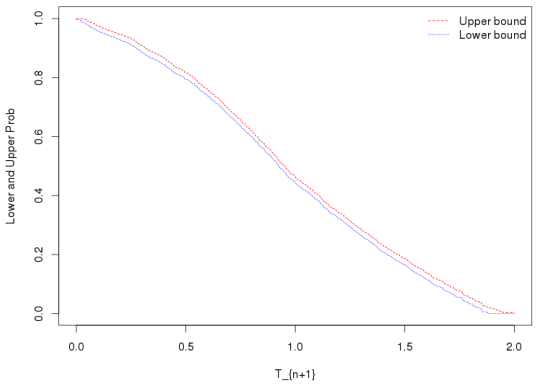
Figure 2.3: Lower and upper NPI probabilities for T_{n+1}, Clayton copula



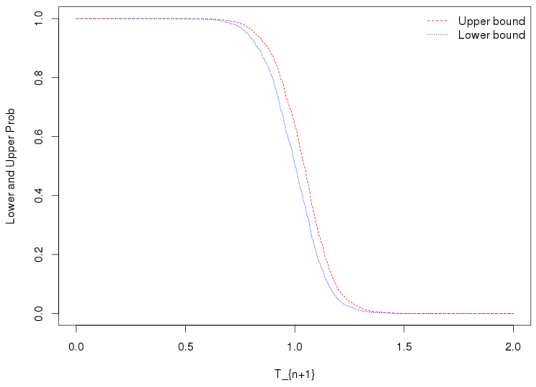
(a) $\tau = 0.75$; $n = 10$



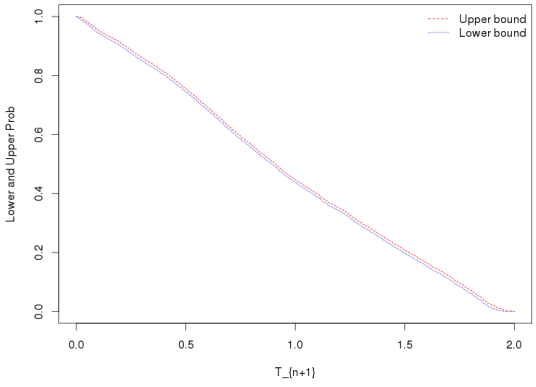
(b) $\tau = -0.75$; $n = 10$



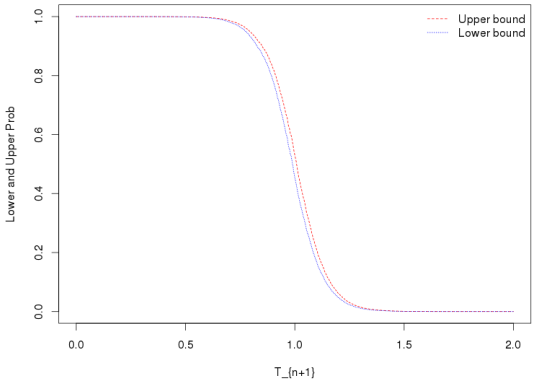
(c) $\tau = 0.75$; $n = 50$



(d) $\tau = -0.75$; $n = 50$



(e) $\tau = 0.75$; $n = 100$



(f) $\tau = -0.75$; $n = 100$

Figure 2.4: Lower and upper NPI probabilities for T_{n+1}, Frank copula

As mentioned before, the main idea of the new method presented in this chapter is to provide a quite straightforward method for prediction of a bivariate random quantity, where imprecision in the marginals provides robustness with regard to the assumed copula. This is attractive in practice, because one often has less knowledge about the dependence structure than about the marginals, in particular if one has a relatively small data set available. The practical usefulness of the method is therefore dependent on its ability to provide reasonable quality predictive inference in case one does not assume to know the parametric family of copulas, which generated the data, exactly. To study the performance of our semi-parametric predictive inference method, we perform simulations as before, but now we generate the data from one of the four mentioned copula families, while we assume a different parametric copula for the second step of our method. The simulations are further performed in the same manner as those above, with attention again on prediction of $T_{n+1} = X_{n+1} + Y_{n+1}$.

We report again first simulation results for just a few scenarios, the other combinations of real and assumed copulas, out of the four parametric copula families discussed before, provided very similar results, as did repeated simulations of the same scenarios. Table 2.9 presents the results with data generated from the Frank copula whilst assuming the Normal copula in our method. While we mostly focus on the predictive performance, it is important to briefly consider the parameter estimate $\hat{\theta}_n$, where we have added subscript n to indicate this is the parameter of the Normal copula. Of course, this is not an estimate of the parameter θ_f as used in the Frank copula for generating the data, the values θ_n corresponding to the respective values for τ is shown in the same table. These estimated values for θ_n are now a bit further from the values given, which results from the fact that the data are not generated from the Normal copula but from the Frank copula.

It is more important to consider the predictive performance of our method. The values of p_1 and p_2 in Table 2.9 are mostly pretty similar to those in Table 2.6 and Table 2.7, although there are now a few cases for which q is not contained in the interval $[p_1, p_2]$. These are highlighted by bold font numbers in the table. For $n = 10$ there are no such cases, indeed the imprecision in the method provides sufficient robustness to still have $q \in [p_1, p_2]$. For $n = 50$ this is also mostly the

case, although there is one case here, for $\tau = 0.5$ and $q = 0.75$, where $p_2 < q$, albeit only just. For $n = 100$ there are substantially more cases where the interval $[p_1, p_2]$ does not contain the corresponding q, although in these cases q tends to be only just outside the interval. This is in line with expectation, because for larger n the method has only small imprecision and assuming the wrong parametric copula starts to have a stronger effect. Table 2.10 presents the results of a similar simulation with the data generated from the Normal copula and the Frank copula assumed in our method. The results for this case are very similar to those just described.

				$n = 10$			$n = 50$			$n = 100$		
τ	θ_f	θ_n	q	$\hat{\theta}_n$	p_1	p_2	$\hat{\theta}_n$	p_1	p_2	$\hat{\theta}_n$	p_1	p_2
-0.75	-14.1385	-0.9239	0.25	-0.9137	0.0737	0.4991	-0.9020	0.1757	0.2774	-0.8967	0.1967	0.2506
			0.50		0.2391	0.7566		0.4242	0.5738		0.4639	0.5449
			0.75		0.4932	0.9228		0.7203	0.8272		**0.7514**	0.8018
-0.50	-5.7363	-0.7071	0.25	-0.7424	0.1580	0.4120	-0.6964	0.2203	0.2726	-0.6840	0.2237	0.2525
			0.50		0.3447	0.6599		0.4603	0.5429		0.4794	0.5221
			0.75		0.5899	0.8458		0.7326	0.7851		**0.7517**	0.7803
-0.25	-2.3719	-0.3827	0.25	-0.4323	0.1847	0.3525	-0.3900	0.2383	0.2756	-0.3756	0.2272	**0.2450**
			0.50		0.3845	0.6100		0.4798	0.5365		0.4853	0.5145
			0.75		0.6380	0.8085		0.7424	0.7800		0.7394	0.7574
0.25	2.3719	0.3827	0.25	0.4307	0.1906	0.3024	0.3901	0.2403	0.2644	0.3762	**0.2508**	0.2633
			0.50		0.4340	0.5569		0.4886	0.5158		0.4918	0.5066
			0.75		0.6939	0.8047		0.7355	0.7594		0.7367	**0.7489**
0.50	5.7363	0.7071	0.25	0.7432	0.2035	0.2987	0.6966	0.2416	0.2643	0.6837	**0.2585**	0.2703
			0.50		0.4452	0.5407		0.4815	0.5010		0.4950	0.5052
			0.75		0.6949	0.7965		0.7269	**0.7490**		0.7346	**0.7442**
0.75	14.1385	0.9239	0.25	0.9142	0.2048	0.2974	0.9019	0.2478	0.2668	0.8969	**0.2602**	0.2725
			0.50		0.4511	0.5450		0.4938	0.5141		**0.5034**	0.5119
			0.75		0.7016	0.7936		0.7320	0.7501		0.7368	**0.7458**

Table 2.9: Simulations from Frank copula; Normal copula assumed for inference

				$n = 10$			$n = 50$			$n = 100$		
τ	θ_n	θ_f	q	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2
-0.75	-0.9239	-14.1385	0.25	-15.7767	0.0739	0.4897	-13.6590	0.1907	0.2933	-13.6472	0.2201	0.2690
			0.50		0.2331	0.7605		0.4177	0.5873		0.4552	0.5457
			0.75		0.5088	0.9203		0.7176	0.8110		0.7330	0.7856
-0.50	-0.7071	-5.7363	0.25	-6.9087	0.1566	0.3969	-5.8457	0.2382	0.2894	-5.7489	0.2332	0.2599
			0.50		0.3451	0.6580		0.4607	0.5449		0.4673	0.5162
			0.75		0.6087	0.8464		0.7200	0.7732		0.7270	0.7534
-0.25	-0.3827	-2.3719	0.25	-3.0572	0.1902	0.3622	-2.4593	0.2393	0.2746	-2.4218	**0.2530**	0.2715
			0.50		0.3971	0.6135		0.4677	0.5198		0.4951	0.5256
			0.75		0.6523	0.8201		0.7235	0.7620		0.7484	0.7662
0	0	-	0.25	-0.0383	0.1924	0.3195	-0.0032	0.2399	0.2662	-0.0031	0.2456	0.2595
			0.50		0.4199	0.5844		0.4803	0.5200		0.4933	0.5136
			0.75		0.6773	0.8054		0.7422	0.7704		0.7476	0.7607
0.25	0.3827	2.3719	0.25	2.9621	0.2011	0.3089	2.4619	0.2297	0.2516	2.4183	0.2404	0.2523
			0.50		0.4490	0.5743		0.4848	0.5113		0.4967	0.5109
			0.75		0.7050	0.8118		0.7404	0.7640		**0.7504**	0.7612
0.50	0.7071	5.7363	0.25	7.0106	0.1993	0.2933	5.8423	0.2298	0.2522	5.7466	0.2299	**0.2396**
			0.50		0.4478	0.5535		0.4922	0.5132		0.4868	**0.4990**
			0.75		0.7080	0.8095		**0.7514**	0.7716		0.7490	0.7596
0.75	0.9239	14.1385	0.25	15.7494	0.1991	0.2951	13.6822	0.2430	0.2615	13.6889	0.2357	**0.2460**
			0.50		0.4640	0.5504		0.4898	0.5101		0.4951	0.5070
			0.75		0.7150	0.8034		0.7493	0.7689		**0.7538**	0.7634

Table 2.10: Simulations from Normal copula; Frank copula assumed for inference

Tables 2.11 and 2.12 present the results of similar simulation studies with data generated from the Clayton and Gumbel copulas, respectively. For both these cases the Frank copula was assumed for our method; in further simulations, with the Normal copula assumed instead, the results were very similar. For $n = 10$ the robustness is again sufficient to always get $q \in [p_1, p_2]$, indeed we have not encountered any simulation, for any combination of these four copulas, where this was not the case. For $n = 50$ and $n = 100$ the results are now slightly worse than before, but where q is outside the interval $[p_1, p_2]$ it is always close to it. This reflects that the Clayton and Gumbel copulas differ more from the Frank copula than the Normal copula does. We also included the case $n = 30$ here, for which the results were all fine.

				$n=10$			$n=30$			$n=50$			$n=100$		
τ	θ_g	θ_f	q	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2
0	1	-	0.25	0.0116	0.1937	0.3130	-0.0031	0.2283	0.2730	-0.0019	0.2369	0.2659	0.0079	0.2370	0.2501
			0.50		0.4143	0.5813		0.4652	0.5247		0.4824	0.5195		0.4885	0.5076
			0.75		0.6793	0.8088		0.7253	0.7699		0.7367	0.7656		0.7349	**0.7484**
0.25	1.3333	2.3719	0.25	3.0423	0.1974	0.2958	2.5644	0.2165	0.2507	2.5089	0.2225	**0.2419**	2.4531	0.2372	**0.2478**
			0.50		0.4270	0.5586		0.4610	0.5092		0.4703	**0.4993**		0.4817	**0.4957**
			0.75		0.7030	0.8074		0.7336	0.7770		0.7441	0.7698		**0.7516**	0.7645
0.50	2.0000	5.7363	0.25	7.0647	0.1976	0.2858	6.0249	0.2274	0.2572	5.8939	0.2245	**0.2444**	5.8077	0.2308	**0.2410**
			0.50		0.4275	0.5379		0.4733	0.5141		0.4734	**0.4941**		0.4689	**0.4814**
			0.75		0.7085	0.8177		0.7477	0.7835		0.7446	0.7686		**0.7525**	0.7626
0.75	4.0000	14.1385	0.25	16.2068	0.2035	0.2946	13.8853	0.2286	0.2580	13.8537	0.2290	**0.2460**	13.7948	**0.2502**	0.2594
			0.50		0.4480	0.5417		0.4732	0.5070		0.4860	0.5062		**0.5023**	0.5118
			0.75		0.7119	0.8092		0.7348	0.7688		0.7460	0.7678		**0.7630**	0.7738

Table 2.12: Simulations from Gumbel copula; Frank copula assumed for inference

				$n=10$			$n=30$			$n=50$			$n=100$		
τ	θ_c	θ_f	q	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2	$\hat{\theta}_f$	p_1	p_2
-0.75	-6.0000	-14.1385	0.25	-14.8626	0.1122	0.4171	-14.0165	0.1556	0.3041	-13.7984	0.1828	0.2741	-13.7111	0.2004	0.2503
			0.5		0.2834	0.7234		0.3554	0.6472		0.4025	0.5960		0.4497	0.5648
			0.75		0.5907	0.8923		0.6947	0.8444		0.7244	0.8177		**0.7561**	0.7996
-0.50	-2.0000	-5.7363	0.25	-3.1599	0.1509	0.4119	-6.0220	0.2033	0.2879	-5.8851	0.2157	0.2650	-5.7672	0.2295	0.2541
			0.50		0.3325	0.6651		0.4278	0.5773		0.4453	0.5445		0.4709	0.5268
			0.75		0.5896	0.8395		0.7132	0.7952		0.7262	0.7793		0.7446	0.7669
-0.25	-0.6667	-2.3719	0.25	-3.0626	0.1870	0.3534	-2.5641	0.2264	0.2805	-2.5090	0.2353	0.2711	-2.4492	0.2422	0.2630
			0.50		0.3929	0.6195		0.4522	0.5491		0.4726	0.5278		0.4841	0.5132
			0.75		0.6596	0.8248		0.7165	0.7746		0.7277	0.7612		0.7364	0.7546
0.25	0.6667	2.3719	0.25	3.0639	0.1809	0.2959	2.5553	0.2214	0.2637	2.5017	0.2313	0.2567	2.4415	0.2375	**0.2493**
			0.50		0.4424	0.5745		0.4970	0.5457		**0.5058**	0.5338		**0.5181**	0.5329
			0.75		0.7001	0.7985		0.7401	0.7762		0.7498	0.7733		0.7545	0.7645
0.50	2.0000	5.7363	0.25	7.1205	0.1866	0.2968	6.0366	0.2177	0.2572	5.8780	0.2254	0.2505	5.7896	0.2284	**0.2416**
			0.50		0.4630	0.5732		0.4958	0.5354		**0.5081**	0.5321		**0.5144**	0.5259
			0.75		0.7095	0.7975		0.7305	0.7612		0.7433	0.7618		**0.7534**	0.7636
0.75	6.0000	14.1385	0.25	16.3807	0.1904	0.2908	13.9919	0.2298	0.2642	13.8441	0.2355	0.2580	13.7415	0.2458	0.2575
			0.50		0.4670	0.5626		0.4915	0.5248		0.4962	0.5149		**0.5031**	0.5134
			0.75		0.7107	0.7953		0.7387	0.7686		0.7412	0.7583		**0.7531**	0.7619

Table 2.11: Simulations from Clayton copula; Frank copula assumed for inference

This simulation study has illustrated our new semi-parametric method and revealed some interesting aspects, as discussed above. The main conclusion we draw from it, is that for small values of n the imprecision provides sufficient robustness for the predictive inferences to have good frequentist properties. This depends on the copulas used, the random quantity considered, and also the percentiles considered. Differences would show more strongly if one considers quite extreme percentiles. If

data were generated with a very different dependence structure than can be modelled through the assumed parametric copula, then the method would also perform worse. However, we would hope that in such cases, either there is background knowledge about the dependence structure, which can be used to select a more suitable copula, or that the data already show a certain pattern to make us aware of the unlikely success of the proposed method with a basic copula.

The main idea of the larger research project to which this chapter presents the first step is as follows. To take dependence into account, and ideally based only on the observed data, would require a substantial amount of data in the bivariate setting (and this is of course far worse in higher dimensional scenarios). If one has much data available, it may be possible to use nonparametric copula methods in combination with NPI for the marginals, in order to arrive at good predictive inference. For smaller data sets, however, it is unlikely that the data reveal much information about the dependence between the random quantities X_{n+1} and Y_{n+1}. The method proposed in this chapter aims at being robust in light of such absence of detailed information, by using the imprecision in NPI on the marginals, together with the discretization of the estimated copula, with the hope that for many scenarios of interest the resulting heuristic method will have a good performance. Of course, if even small or medium sized data sets already reveal a particular (likely) dependence structure, then this should be taken into account in the selection of the copula in our method. But if the data do not strongly indicate a specific dependence structure, then we propose to use a family of parametric copulas which is quite flexible and convenient for computation. In addition, the method used for estimation of the parameter will normally not be that relevant due to the robustness that is implicit in our approach, although of course there are situations where care will be needed (e.g. if the likelihood function has multiple modes one may wish to find an alternative to maximum likelihood estimation; these are well-known general considerations that do not require detailed attention in this chapter but which provide interesting topics for future research).

Interestingly, one could consider the way in which imprecision is used in this chapter as being somewhat different to the usual statistical approaches based on

imprecise probabilities [2]. Typically, it is advocated to add imprecision to parts of a problem where one has less information, indeed to reflect the absence of detailed information. Yet in our presented method, the imprecision is mainly a result from using NPI for the marginals, while the information shortage is most likely to be about the dependence structure. Of course, the discretization of the copula also provides some imprecision, but the main idea is that the imprecise predictive method used for the marginals, which is straightforward, provides robustness with regard to taking the dependence structure into account, which is normally the harder part of such inferences. Furthermore, it turns out that, with NPI used for the marginals, the resulting second step involving the copula estimation can be kept conveniently simple. This is an important advantage of this method, in particular if one would consider implementing it in (more or less) automated inference situations which require fast computation.

The performance of the proposed method is measured by verifying whether the future observation falls in between the quantiles chosen earlier or not. This predictive performance does not evaluate or measure every aspect of the performance (in this sense, it is not an ideal performance measure). The method used in this section is useful to investigate the frequentist performance of our method with regard to the quantiles considered, using the imprecision in our method. However, for large n there is very little imprecision, so it will happen more often that the value q is not in the interval defined by p_1 and p_2. For such cases, the method used in this section does not give a good indication of how far the q is from the p_1 and p_2 interval. For further investigation, we can measure further the performance for the misspecification scenario by calculating the minimum distance of q to the interval $[p_1, p_2]$, d_N and the maximum distance to any point within this interval, d_F. These distances are calculates or measures of how much the misspecification scenario has been missed or far away from the bound of p_1 and p_2. This measurement can be taken an average of N = 10000 times. If the q is in the interval of p_1 and p_2, of course we do not have any nearest distance but we can have the furthest distance which indicates how far is the q from p_1 and p_2. This shows us how wide is the interval. If $q < p_1$ or $q > p_2$, we can see how close is the q to p_1 and p_2 using the

nearest distance. While, the furthest distance can show us how far is the q from p_1 and p_2. Hence, we measure how well is the proposed method even if the q is not in the interval of p_1 and p_2.

2.6 Examples

In this section, two examples are presented using a data set from the literature to illustrate the method proposed in Sections 2.3 and 2.4.

2.6.1 Insurance example

Consider the data set in Table 2.13 on casuality insurance [59, p. 403], which records both the loss and the expenses that are directly related to the payment of the loss (the 'allocated loss adjustment expenses', ALAE) for an insurance company on twenty claims. The loss and the ALAE are usually positively correlated [59], there is some suggestion that this is also the case in these data as can be seen from Figure 2.5. The original data consist of 24 bivariate data observations, to illustrate our approach we have removed four 'outliers' and we have adjusted the data to avoid tied observations (namely 2501, 7001, 51 are used instead of 2500, 7000, 50). There are many ways to deal with outliers as discussed in [6, 48]. In this research, the outliers are not our main concern but it does affect the linear dependence structure between these two variables. There is no strong need to exclude outlying data from the analysis when our method is used, but the effect of data which strongly influence the copula estimation requires further study, for example into the use of copulas with multiple parameters that can separate different dependence relations over the ranges of the data considered. This is left as an important topic for future research, in particular to compare when it is better to use more complicated parametric copulas and when it is better to use nonparametric copulas. In addition, it should also emphasize that our method does not only deal with the linear dependence. For this data set, if we include the outliers, the Pearson correlation is reduced from 0.2080 to 0.0838, it still shows a positive correlation between Loss and ALAE but reduces the strength of the dependency.

Loss	ALAE	Loss	ALAE
1,500	301	10,000	1,174
2,000	3,043	11,750	2,530
2,500	415	12,500	165
2,501	4,940	14,000	175
4,500	395	15,000	2,072
5,000	25	17,500	6,328
7,000	50	19,833	212
7,001	10,593	30,000	2,172
7,500	51	33,033	7,845
9,000	406	44,887	2,178

Table 2.13: Losses and corresponding ALAE values, Example 2.6.1

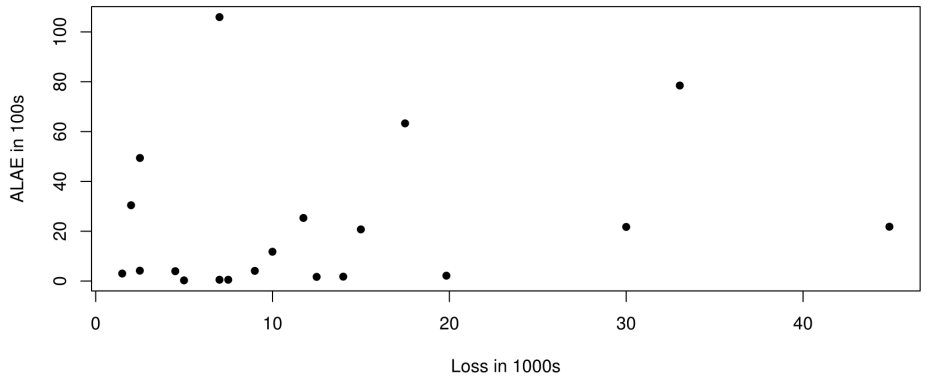


Figure 2.5: Losses and corresponding ALAE values, Example 2.6.1

In line with the earlier presentation in this chapter, Loss will be the X variable and ALEA the Y variable. Suppose that we are interested in the event that the sum of the next Loss and ALAE will exceed t, that is $T_{n+1} = X_{n+1} + Y_{n+1} > t$, based on the available data (x_i, y_i), $i = 1, 2, \ldots, 20$. We apply the new semi-parametric method presented in Section 2.4, where we assume the Normal copula, Clayton copula, Gumbel copula and Frank copula, and we use pseudo maximum likelihood method to estimate the copula parameter, the method is available in the R package `VineCopula` [83], which also used in Section 2.5.

The lower and upper probabilities for the event $T_{n+1} > t$ are presented in Figure 2.6 only for the Normal copula, and Table 2.14 shows the NPI lower and upper probabilities for the event $T_{n+1} > t$ for different parametric copulas, for selected

values of t. These results can be used in a variety of ways, depending on the actual question of interest. From this table, we can see that the value of NPI lower and upper probabilities are different at each t among the parametric copulas. Figure 2.6 shows that the imprecision, which is the difference between corresponding upper and lower probabilities, is pretty similar through the main range of empirical values for $x_i + y_i$. This is due to the effect discussed for the simulations in Section 2.5, namely the positive correlation between Loss and ALEA combined with interest in the sum of these quantities. If the data would have indicated a negative correlation, then imprecision would vary more substantially for the sum of the two quantities. In Figure 2.7, we show the imprecision for different parametric copulas considered. From this figure, we can see that imprecision is quite similar for these parametric copulas.

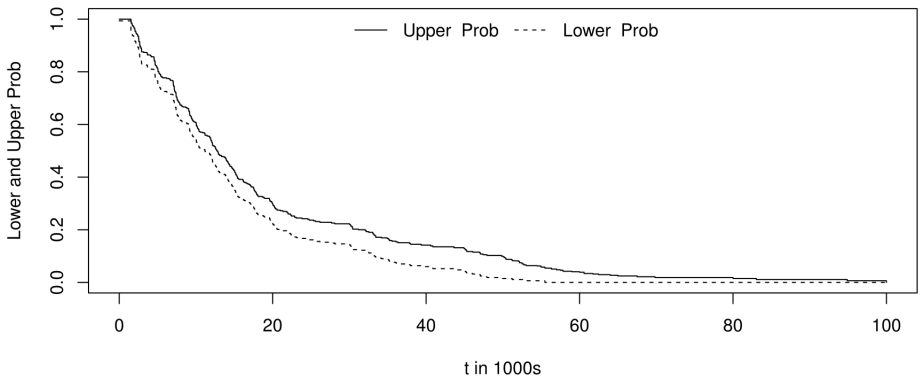


Figure 2.6: Lower and upper probabilities for $T_{n+1} > t$, Example 2.6.1

t in 1000s	Normal		Clayton		Gumbel		Frank	
	$\underline{P}(T_{n+1} > t)$	$\overline{P}(T_{n+1} > t)$	$\underline{P}(T_{n+1} > t)$	$\overline{P}(T_{n+1} > t)$	$\underline{P}(T_{n+1} > t)$	$\overline{P}(T_{n+1} > t)$	$\underline{P}(T_{n+1} > t)$	$\overline{P}(T_{n+1} > t)$
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0.8695	0.9080	0.8779	0.9122	0.8770	0.9159	0.8767	0.9160
10	0.7242	0.7725	0.7432	0.7887	0.7283	0.7783	0.7309	0.7793
15	0.6073	0.6633	0.6291	0.6843	0.6063	0.6646	0.6139	0.6697
20	0.5302	0.5889	0.5493	0.6089	0.5265	0.5872	0.5369	0.5955
25	0.4654	0.5284	0.4821	0.5476	0.4598	0.5244	0.4717	0.5349
30	0.4120	0.4788	0.4277	0.4978	0.4050	0.4732	0.4190	0.4858
35	0.3347	0.4033	0.3424	0.4171	0.3259	0.3953	0.3392	0.4092
40	0.2845	0.3532	0.2902	0.3654	0.2768	0.3451	0.2891	0.3596
45	0.2568	0.3266	0.2618	0.3387	0.2492	0.3187	0.2606	0.3331
50	0.2345	0.3065	0.2328	0.3135	0.2283	0.2984	0.2344	0.3104
55	0.2066	0.2656	0.2036	0.2651	0.2014	0.2590	0.2062	0.2660
60	0.1880	0.2474	0.1843	0.2462	0.1841	0.2413	0.1876	0.2479
66	0.1647	0.2251	0.1602	0.2225	0.1608	0.2192	0.1633	0.2247
70	0.1487	0.2085	0.1437	0.2058	0.1463	0.2034	0.1471	0.2082
75	0.1352	0.1889	0.1295	0.1845	0.1344	0.1863	0.1337	0.1883
80	0.1169	0.1693	0.1114	0.1655	0.1165	0.1681	0.1147	0.1682
85	0.1006	0.1517	0.0938	0.1463	0.1001	0.1502	0.0978	0.1501
90	0.0904	0.1411	0.0822	0.1347	0.0913	0.1404	0.0874	0.1396
96	0.0733	0.1248	0.0666	0.1193	0.0755	0.1245	0.0706	0.1233
106	0.0635	0.1157	0.0586	0.1114	0.0654	0.1158	0.0613	0.1144
110	0.0520	0.1038	0.0452	0.0980	0.0532	0.1030	0.0480	0.1012
116	0.0386	0.0859	0.0291	0.0792	0.0411	0.0868	0.0334	0.0834
121	0.0291	0.0761	0.0191	0.0692	0.0344	0.0782	0.0235	0.0735
125	0.0185	0.0644	0.0110	0.0539	0.0244	0.0659	0.0140	0.0591
130	0.0150	0.0491	0.0083	0.0373	0.0216	0.0530	0.0108	0.0432
135	0.0150	0.0397	0.0083	0.0272	0.0216	0.0464	0.0108	0.0333
140	0.0064	0.0269	0.0028	0.0166	0.0125	0.0346	0.0038	0.0210
150	0.0064	0.0234	0.0028	0.0138	0.0125	0.0317	0.0038	0.0178
156	0.0000	0.0110	0.0000	0.0056	0.0000	0.0179	0.0000	0.0074

Table 2.14: Lower and upper probabilities for $T_{n+1} > t$, Example 2.6.1

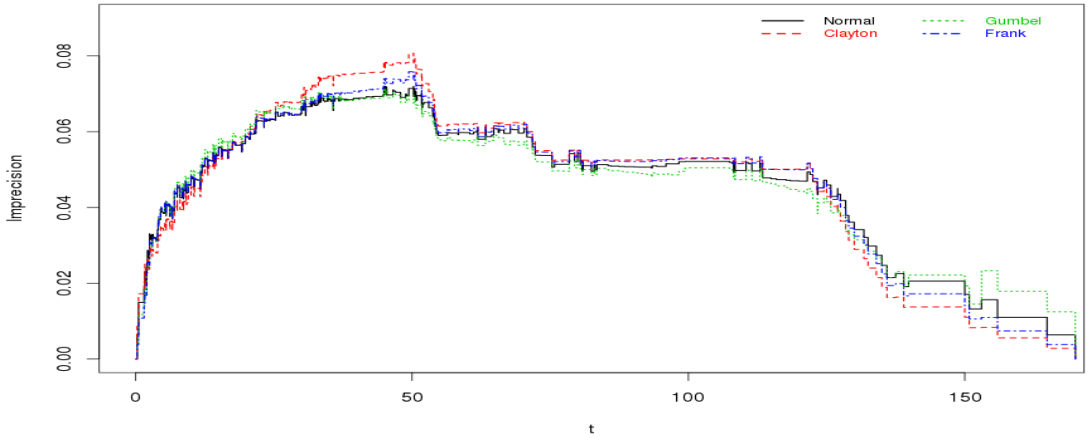
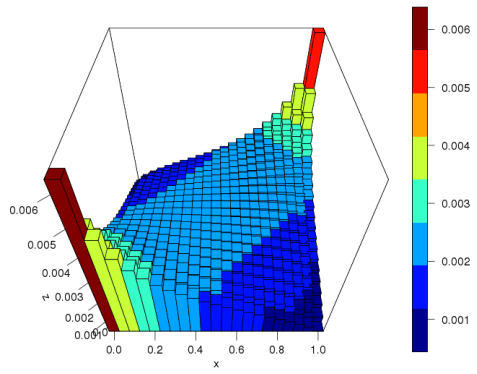
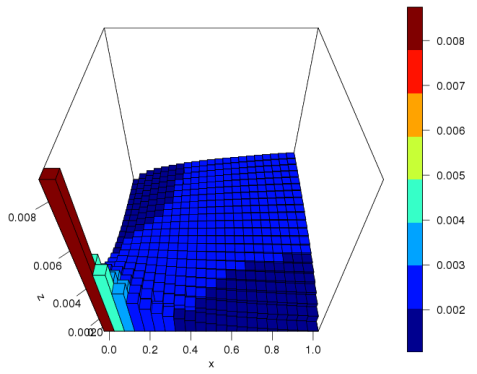


Figure 2.7: Imprecision for different parametric copulas, Example 2.6.1

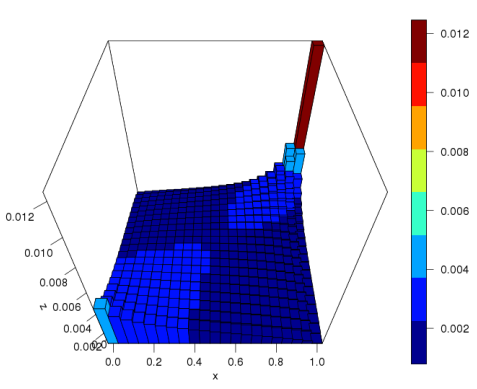
Figure 2.8 shows the 3D-plots of the probabilities $h_{ij}(\hat{\theta})$ for these data for different parametric copulas. The plots of probabilities $h_{ij}(\hat{\theta})$ in this section are given with x, y and z are equal to 0 at the left-front corner, and at the right-back corner x, y and z are equal to 1. We can see that the Normal and Frank copulas give a symmetric shape for the probabilities $h_{ij}(\hat{\theta})$, but different values of probabilities $h_{ij}(\hat{\theta})$ for each cell. For the Clayton copula, it shows that the probabilities $h_{ij}(\hat{\theta})$ are higher at the left-front corner, while for Gumbel copula, the probabilities $h_{ij}(\hat{\theta})$ are higher at the right-back corner. These features occur due to the parametric copula characteristics mentioned in Section 2.2. Consequently, the NPI lower and upper probabilities for the event $T_{n+1} > t$ in Table 2.14 are different. For example, from Figure 2.7, at $t \geq 125$, the imprecision for the Clayton copula is smallest among the four parametric copulas considered due to lower probabilities $h_{ij}(\hat{\theta})$, shown in Figure 2.8, at the right-back corner of the Clayton copula 3D-plot. The Gumbel copula leads to the largest imprecision at the same t value due to large probabilities $h_{ij}(\hat{\theta})$, shown in Figure 2.8, at the right-back corner of the Gumbel copula 3D-plot.



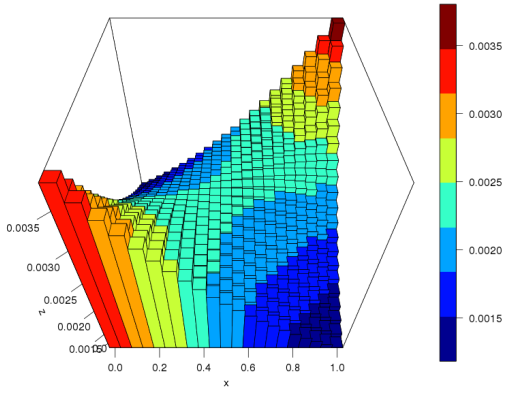
(a) Normal copula



(b) Clayton copula



(c) Gumbel copula



(d) Frank copula

Figure 2.8: 3D-Plots for probabilities $h_{ij}(\hat{\theta})$, Example 2.6.1

2.6.2 Body-Mass Index example

Thus far, we have illustrated our method by considering the sum of the two values in the next bivariate observation, $X_{n+1} + Y_{n+1}$. In order to illustrate application to scenarios where interest is in a different function of (X_{n+1}, Y_{n+1}), consider the data presented in Table 2.15 and Figure 2.9 [46]. These present the heights (m) and weights (kg) of $n = 30$ eleven-year-old girls attending Heaton Middle School in Bradford. Let heights be X and weights be Y random quantities. Suppose that one is interested in the body-mass index (BMI) of a further girl, where one can imagine

there having been 31 girls with one selected randomly to not be included in the data set, and whose BMI one would wish to predict after learning the heights and weights of the other 30 girls. Interest in the BMI may be in order to investigate whether they have healthy weight, are underweight or overweight, or even obese, so we derive the lower and upper probabilities for the thirty-first girl to be in each of these categories, based on our semi-parametric method. The BMI is calculated using the well-known formula,

$$\text{BMI} = \frac{\text{Weight (kg)}}{[\text{Height (m)}]^2}$$

Height (m)	Weight (kg)	BMI	Height (m)	Weight (kg)	BMI
1.35	26	14.27	1.33	31	17.53
1.46	33	15.48	1.49	34	15.31
1.53	55	23.50	1.41	32	16.10
1.54	50	21.08	1.64	47	17.47
1.39	32	16.56	1.46	37	17.36
1.31	25	14.57	1.49	46	20.72
1.49	44	19.82	1.47	36	16.66
1.37	31	16.52	1.52	47	20.34
1.43	36	17.60	1.40	33	16.84
1.46	35	16.42	1.43	42	20.54
1.41	28	14.08	1.48	32	14.61
1.36	28	15.14	1.49	32	14.41
1.54	36	15.18	1.41	29	14.59
1.51	48	21.05	1.37	34	18.11
1.55	36	14.98	1.35	30	16.46

Table 2.15: The heights (m), weights (kg) and BMI of 30 eleven-year-old girls

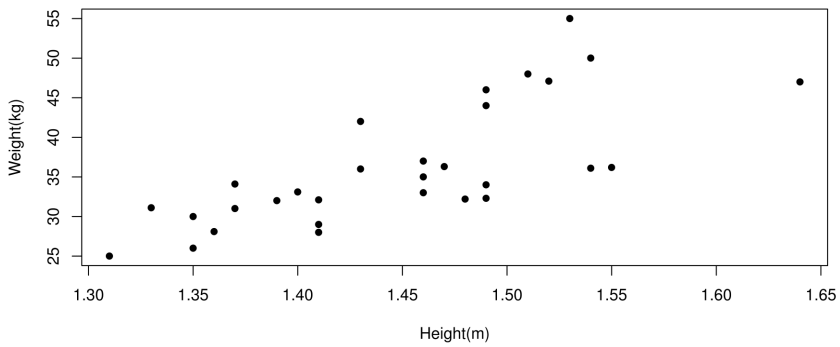


Figure 2.9: Heights (m) and corresponding weights (kg) values, Example 2.6.2

For this illustrative example, we use the classification of BMI values provided by the Center for Disease Control and Prevention (www.cdc.gov), according to which an eleven-year-old girl is considered underweight if her BMI is less than 14.08, has healthy weight if the BMI is between 14.08 and 19.50, is overweight if the BMI is between 19.50 and 24.14, and obese if the BMI is at least 24.14. The lower and upper probabilities for these events of interest are given in Table 2.16. These are calculated using equations (2.3) and (2.4) with the same parametric copulas and estimation method used in Example 2.6.1. To avoid difficulties due to the functional form of the BMI, we restricted the range of possible values for the height and weight quantities by setting finite end-points for the ranges used in NPI for the marginals. We set these values at $x_0 = 1.25$, $x_{31} = 1.70$, $y_0 = 20$ and $y_{31} = 60$, which seem quite realistic and lead to corresponding minimum BMI 6.92 and maximum BMI 38.40, which are included in the ranges in Table 2.16. Choosing different values for x_0, x_{31}, y_0 and y_{31} will have some impact on the lower and upper probabilities resulting from our method, but the effect of minor differences to these values is quite minimal. There are many ways to interpret the results in Table 2.16. For example, from the table, we can see using the Normal copula, our method gives lower and upper probabilities for the event that a future eleven-year-old girl is in healthy weight are 0.6521 and 0.8107, respectively. The results are quite similar for all four parametric copulas considered.

	BMI$\in$	Normal		Clayton		Gumbel		Frank	
		$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
Underweight	[6.92,14.08)	0.0303	0.1010	0.0313	0.1098	0.0520	0.1245	0.0479	0.1080
Healthy weight	[14.08,19.50)	0.6521	0.8107	0.6514	0.7869	0.6078	0.7733	0.6479	0.7862
Overweight	[19.50,24.14)	0.1368	0.2456	0.1331	0.2236	0.1431	0.2636	0.1300	0.2377
Obese	[24.14,38.40)	0.0013	0.0222	0.0152	0.0487	0.0042	0.0217	0.0064	0.0360

Table 2.16: NPI lower and upper probabilities for different parametric copula

2.7 Concluding remarks

This chapter presents a new semi-parametric method for predictive inference about a future bivariate observation, which can be used to consider any function of in-

terest involving the two quantities in such an observation. It combines NPI on the marginals, which is predictive by nature, with the use of a parametric copula to take dependence into account and the parameter of the copula estimated based on available data. This method can be used with a wide variety of estimation methods because only a single point estimator is used. For the semi-parametric predictive method presented in this chapter, any of the available methods to estimate the copula parameter can be used, of course advantages and disadvantages of specific estimation methods are carried over. A possible generalization of the method is by introducing some further robustness, or imprecision, in the copula, either by using a range of parameter values (e.g. related to a confidence interval) or a set of copulas. Implementing these straightforward ideas would require further research, as they would lead to imprecise probabilities instead of the probabilities $h_{ij}(\hat{\theta})$ which are central to our method.

By combining NPI with an estimated copula, the proposed method does not fully adopt the strong frequentist properties of NPI, and hence has a heuristic nature. We have investigated its performance via simulation studies, more detailed research of its performance in a wider range of applications will be of benefit. The main idea of this research is that the robustness provided by our method, with the use of a quite basic copula, will often lead to satisfactory inferences for small to medium sized data sets. Of course this is not an argument for neglecting important information about the dependence structure, but it will enable, for many applications, trustworthy predictive inference with the use of a relatively basic copula. For larger data sets, it is expected that the method may work well using a nonparametric copula instead of a parametric copula, this will be investigated in Chapter 3.

Throughout this work, we restricted attention to a single future observation. In practice, one may be interested in multiple future observations, in NPI the inter-dependence of such multiple future observations is taken into account [19]. It will be of interest to develop the bivariate method presented in this chapter, for multiple future observations.

A major advantage of the presented method is its relatively easy computations, as the use of NPI on the marginals combines naturally with the discretization of the

copula. Hence, the computational complexity is only with regard to the estimation of the copula parameter, which for the copulas considered in this chapter is a routine procedure for which standard software is available. It may be attractive to use copulas with multi-dimensional parameters, which would provide better opportunities to take more information about dependence in the data into account. As long as suitable estimation methods are available, this can be implemented in our method without any difficulties.

The bivariate method presented here can straightforwardly be generalized to multivariate data, where the curse of dimensionality [32, 85] implies that the number of data required to get meaningful inferences grows exponentially with the dimension of the data. We restricted attention to the bivariate case in order to introduce, illustrate and investigate the method, application to higher dimensional situations is an important topic for future research.

It should be emphasized that the method used in Section 2.5 (i.e. predictive performance) is useful to investigate the frequentist performance of our method with regard to the quantiles considered, using the imprecision in our method. However, for large n there is very little imprecision, so it will happen more that q is not in the p_1 and p_2 interval. Further investigation into methods for performance evaluation for our method is an important topic for future research, as such methods for imprecise predictive methods have not yet been studied in detail.

Chapter 3

NPI with nonparametric copula

3.1 Introduction

In this chapter, we introduce our new method of predictive inference for bivariate data by combining NPI for the marginals with an estimated nonparametric copula, where we restrict attention to kernel-based methods. Kernel-based copulas provide more flexibility than the parametric copulas used in Chapter 2. The main interest in this chapter is to introduce our new method with the use of nonparametric copulas. We investigate its performance via simulations, both for small and large data sets.

Section 3.2 is a brief introduction to nonparametric copulas. In Section 3.3 we introduce how NPI for the marginals can be combined with an estimated nonparametric kernel-based copula to provide a nonparametric predictive method, and demonstrates how the proposed method can be used for inference about different events of interest. Section 3.3 is relatively similar to Sections 2.3 and 2.4. In Section 3.4 we investigate the performance of this method via simulations, considering different bandwidths for the kernel copula. Two examples are presented in Section 3.5 to illustrate our method, these are the same data sets used in Chapter 2. Some brief comments, conclusions and suggestions for further research are included in Section 3.6.

3.2 Nonparametric copula

There are many nonparametric methods to estimate the dependence structure between two random quantities, such as Deheuvels' empirical copula [31], polynomial approximation copula [14, 65] and kernel smoothing copulas [12, 14]. In this research we use kernel type estimators because this method offers a flexible alternative compared to parametric copulas and the method is the most commonly used in nonparametric estimation of copulas. In this research, we use the R package `np` [49] to estimate the copula using the kernel method.

Generally, we have two main different kinds of kernel in literature, which can be classified as 'classical statistics' kernel [64, 85, 88] and 'machine learning' kernel [29, 52]. In the classical statistics literature, a kernel is a nonparametric method for estimating the probability density function (pdf) of a continuous random variable. Any probability density can be used for the kernel [85]. The kernel estimate places a probability mass of size $1/n$ (where n is the sample size) in the shape of the kernel, which has been scaled by a smoothing parameter, centered on each data point. These probability masses are then summed up at each point to give the kernel estimate.

In machine learning, kernel methods are a class of algorithms used mainly for pattern analysis, for example in support vector machine (SVM) [52]. SVM are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis [29].

Some other density estimation algorithms, equivalent to classical kernel method, use weights $1/n$ but have an adaptive bandwidth, for example kth nearest neighbour estimator. This type of kernel method is estimating the pdf depending on nearest neighbours of the observations. It is related to distance of any point to its nearest observations and it is centered on that point [64, 85, 90]. Loftsgaarden and Quesenberry [70] used the nearest neighbour density estimator for multivariate data. The distance can be any types of distance but the most popular used is Euclidean distance [64, 85].

The main difference between these two approaches is how the kernel method is used in each area. In machine learning, the kernel method is used to express the machine learning algorithms in terms of dot products instead of feature vectors.

So, the machine learning algorithms can work with highly complex, efficient-to-compute, and high performing kernels without ever having to write down huge and potentially infinite dimensional feature vector [52]. While, in classical statistics, the kernel method is used to put weights to the data (points) when estimating the probability density function. The machine learning kernel method has been also used in the classical statistics literature, especially for multivariate data [49, 64].

Let x_i, $i = 1, ..., n$, be a random sample from a distribution with an unknown probability density function, $f(x)$. A standard kernel density estimator for $f(x)$ is given by [74]

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b}\right) \tag{3.1}$$

where $K(\cdot)$ is a univariate kernel function and $b > 0$ is a bandwidth, where $b \to 0$ as $n \to \infty$. As mentioned earlier, in this research we focus on bivariate data. Suppose that we have a bivariate sample (x_i, y_i), $i = 1, \ldots, n$, then the kernel bivariate density function of empirical data is given by [11]

$$\hat{f}(x, y) = \frac{1}{nb_X b_Y} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b_X}, \frac{y - y_i}{b_Y}\right) \tag{3.2}$$

where $b_X, b_Y > 0$ are bandwidths and $K : \mathbb{R}^2 \to \mathbb{R}$ is a kernel function. In general, one could use any probability density function as the $K(\cdot)$ but it is advisable to choose a bivariate kernel with a simple covariance structure [85]. According to Silverman [88], the appropriate sample size to be used in kernel bivariate density estimation is $n \geq 19$.

There are two ways to interpret the kernel density estimator, from a local and global point of view. From a local view, we see a point estimate as a weighted average of frequencies in a neighbourhood of a point. The weighting is conducted according to the kernel function $K(\cdot)$ and the size of the neighbourhood is controlled by the bandwidth. From a global point of view, an estimate of the density is constructed as follows: Centered upon each observation, a bump in the shape of a scaled kernel, $K(\cdot)$, is placed and all the bumps are averaged to give the whole surface of the density. We should emphasize that $K(\cdot)$ can be from any density function [85]. It is well-known that, for estimation, the choice of $K(\cdot)$ is not as

important as the bandwidth, and it does not strongly affect the density estimation, but the smoothness of the density function depends on it [85, 88].

For estimation, the most important feature in kernel method is to choose the appropriate bandwidth or smoothing parameter, b. One may want to choose b as small as the data allow, however there is always a trade-off between the bias of the estimator and its variance. A large bandwidth leads to an estimate with a small variance but a large bias. In contrast, a small bandwidth induces a small bias and a larger variance. There are two ways to choose b, rule-of-thumb or plug-in method and least squares cross-validation (LSCV) [8, 64, 85, 88, 96]. The LSCV method proposed by Rudemo [81] is a fully automatic data-driven method for selecting b. The LSCV method is based on the principle of selecting the value of b that minimizes the integrated squared error of the resulting estimate, i.e. it provides an optimal bandwidth (not over-smooth or under-smooth) tailored to fitting of all data in estimating the probability density function [63, 88]. The normal reference rule-of-thumb bandwidth is given as [85]

$$b_z = 1.06\sigma_z n^{(-1/4)} \tag{3.3}$$

where z denotes either variable X or Y, σ_z is an adaptive measure of spread of the continuous variable z, defined as min(standard deviation, interquatile range /1.349), n is the sample size.

Another kernel method in classical statistics is the kernel method which estimates the pdf function depending on nearest neighbours of the observations and it is related to the distance of any point to its nearest observations. Let $d_k(x, y)$ be the Euclidean distance from (x, y) to the kth nearest data point in two dimensions, and let $V_k(x, y)$ be the volume of the two dimensional sphere of radius $d_k(x, y)$; thus, $V_k(x, y) = \pi d_k(x, y)^2$ [88]. The nearest neighbours density estimate is then defined by Silverman [88]

$$\hat{f}(x, y) = \frac{k/n}{V_k(x, y)} = \frac{k/n}{\pi d_k(x, y)^2} \tag{3.4}$$

Consider the kernel estimate based on the kernel

$$K(x, y) = \begin{cases} \frac{1}{\pi}, & \text{if } \mid x, y \mid \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{3.5}$$

Then, the generalized kth nearest neighbour (generalized-nn) estimate for two dimensions is defined by

$$\hat{f}(x,y) = \frac{1}{nd_k(x,y)^2} \sum_{i=1}^{n} K\left(\frac{x-x_i}{d_k(x,y)}, \frac{y-y_i}{d_k(x,y)}\right) \tag{3.6}$$

where K is defined in equation (3.5). An estimate of $f(x,y)$ can be obtained by choosing k such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. In this case, k/n plays a similar role to the fixed smoothing parameter b for the kernel estimator. The conditions $k \to \infty$ and $k/n \to 0$ are similar to $n \to \infty$ and $b \to 0$. However, the generalized-nn method is not very satisfactory for overall estimates (global point of view) because they are likely to suffer from local noise (unexplained variation in a sample), to produce an estimate with very heavy tails (peakedness), and the density estimate will have discontinuities because the function $d_k(x,y)$ is not differentiable due to unknown density [88, 96]. In addition, the integral over the estimated density function is not equal to 1 and, in general, diverges.

Adaptive kernel estimation is one of the methods used to overcome the problems of the nearest neighbour method [10, 85, 88]. It combines features of the kernel and the nearest neighbour approaches. Adaptive kernel estimation or adaptive nearest neighbour (adaptive-nn) estimation is an approach that adapts sparseness of data using a wider kernel over observations located in areas of low density. In other words, a large bandwidth is used for area where the data points are far-off from each other and the density is smooth (low density is provided). But when the data points are close to each other a small bandwidth is used, allowing the kernel density function to provide high density estimation in those parts of the distribution. The adaptive kernel density estimate, $\hat{f}(x,y)$ given by Breiman et al. [10] is

$$\hat{f}(x,y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{d_i(x,y)^2} K\left(\frac{x-x_i}{d_i(x,y)}, \frac{y-y_i}{d_i(x,y)}\right) \tag{3.7}$$

where K is a bivariate kernel function and $d_i(x,y)$ is the distance from the point (x_i, y_i) to its kth nearest neighbour.

Basically in practice, a pilot estimate is obtained for the unknown density function at the sample points, whereby an initial density estimate is computed using a pilot estimate (fixed bandwidth) to get an idea of the density at each of the data

points [10]. In `np` package [49], the adaptive nearest neighbour (adaptive-nn) bandwidth method is given as in equation (3.3), but the value 1.06 is replaced by k_z i.e. $b_z = k_z \sigma_z n^{(-1/4)}$ where k_z is an integer value.

There are many books and papers about the choice of the bandwidth, for example see [64, 69, 85]. However, generally, there is no evidence or proof which method is more appropriate and reliable either for estimation or prediction purpose. It is well-known that selecting an appropriate bandwidth is very important as under-smoothing or over-smoothing can substantially reduce the precision of estimation, and it might also reduce accuracy in prediction. As we have discussed above, different types of bandwidths and different bandwidth selections offered different advantages and disadvantages. For example, fixed bandwidth gives the same value of bandwidth to all observation points when estimating the density [64, 85, 88]. While the adaptive-nn and the generalized-nn give different value of bandwidth to each observation point when estimate the density [64, 85, 88]. All the methods discussed in this section will be considered in Chapter 3.

As mentioned earlier in Section 2.2, a copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. Consider (X_i, Y_i) as a random quantity with marginal distributions $F_X(X_i)$ and $F_Y(Y_i)$ where $i = 1, ..., n$, and let $F(x, y)$ be its joint marginal distribution. Let $(U, V) \sim [0, 1]$ be random quantities with joint distribution C and corresponding probability density function, $c : [0, 1]^2 \to \mathbb{R}$. In line with equation (3.2), the kernel smoothing copula can be denoted as

$$\hat{c}(u, v) = \frac{1}{n b_U b_V} \sum_{i=1}^{n} K\left(\frac{u - U_i}{b_U}, \frac{v - V_i}{b_V}\right) \tag{3.8}$$

where for all $u, v \in [0, 1]$, $K : \mathbb{R}^2 \to \mathbb{R}$ is a bivariate kernel function and $b_U, b_V > 0$ is a bandwidth where $b \to 0$ as $n \to \infty$. Then, it well-known that $F_X \sim U[0, 1]$ and $F_Y \sim U[0, 1]$. The corresponding copula, C is then defined as the distribution function of $(F_X(X_i), F_Y(Y_i))$. This explains a copula as the distribution of uniformly distributed random quantities.

Many researchers argue that the kernel estimator is not suitable for the unit-squared copula densities, mainly because it is heavily affected by boundary bias

issues for estimation purpose [40, 97]. In addition, most common copulas admit unbounded densities, and kernel methods are not consistent in that case. Therefore, many researchers study and provide solutions to the boundary bias, including Gijbels and Mielniczuk [44], Charpentier et al. [12], Geenens et al. [40] and recently, Wen and Wu [97]. As discussed in Section 2.3, we use the NPI on the marginals combined with the discretization of the copula, the problem does not occur due to the transformations of variables that are used to estimate the densities, which is free of boundary bias.

As mentioned earlier, we use the R package `np` [49] to estimate the copula pdf using kernel method. In this package, the coding allows us to choose the bandwidth selection methods and type of bandwidths which discussed in this section. Furthermore, the package allows us to choose either the coding to give the value of the bandwidth or we give it manually.

3.3 Combining NPI with kernel-based copula

In this section, we present how NPI for the marginals can be combined with a nonparametric copula. The idea is effectively the same as in Section 2.3. Let (X_{n+1}, Y_{n+1}) be a future bivariate observation and $\widetilde{X}_{n+1}$ and $\widetilde{Y}_{n+1}$ denote transformed versions of the random quantities X_{n+1} and Y_{n+1}, respectively, following from the natural transformations related to the marginal $A_{(n)}$ assumptions as presented in Section 2.3. For an assumed kernel smoothing copula, equation (3.8), an estimate $\hat{c}$ can be defined as

$$\widehat{c}(x,y) = \frac{1}{nb_Xb_Y}\sum_{i=1}^{n} K\left(\frac{x - F_X(\widetilde{X}_i)}{b_X}, \frac{y - F_Y(\widetilde{Y}_i)}{b_Y}\right) \tag{3.9}$$

where $K : \mathbb{R}^2 \to \mathbb{R}$ is a bivariate kernel function satisfying, $b_X, b_Y > 0$ are bandwidths, and $F_X(\widetilde{X}_i) = \frac{r_x^i}{n+1}$ and $F_Y(\widetilde{Y}_i) = \frac{r_y^i}{n+1}$ for $i = 1, ..., n+1$, with r_x^i and r_y^i are rank values of x_i and y_i, where these ranks are only among the x and y observations, respectively. As mentioned in Section 2.3, the discretization using NPI which correspond to copulas shows that the NPI approach for the marginals can be easily combined with this nonparametric kernel-based copula to reflect the dependence structure.

Using the same natural transformations related to the marginal $A_{(n)}$ assumptions as given in Section 2.3 and equation (3.9), NPI on the marginals can be combined with the estimated kernel-based nonparametric copula, $\hat{c}$ as follows,

$$h_{ij}(\hat{c}) = P_C(\widetilde{X}_{n+1} \in \left(\frac{i-1}{n+1}, \frac{i}{n+1}\right), \widetilde{Y}_{n+1} \in \left(\frac{j-1}{n+1}, \frac{j}{n+1}\right) |\hat{c}) \tag{3.10}$$

where $i, j = 1, ..., n+1$ and $P_C(\cdot|\hat{c})$ represents the nonparametric kernel-based copula probability with estimated density function, $\hat{c}$, and the corresponding cumulative distribution function is

$$H_{ij}(\hat{c}) = P_C(\widetilde{X}_{n+1} \leq \frac{i}{n+1}, \widetilde{Y}_{n+1} \leq \frac{j}{n+1}|\hat{c}) = \sum_{k=1}^{i}\sum_{l=1}^{j} h_{kl}(\hat{c}) \tag{3.11}$$

As mentioned in Section 3.2, we use the `np` package in R [49] to estimate the kernel in equation (3.9), $\widehat{c}(x, y)$ resulting probabilities $h_{ij}(\hat{c})$ and $H_{ij}(\hat{c})$ are used for inference about the future observation (X_{n+1}, Y_{n+1}) as in Section 2.4, using lower and upper probabilities.

As in Chapter 2, our method consists of two steps. First we consider NPI for the marginals and the second step is to use the bivariate nonparametric kernel-based copula, where we estimate the copula as in equation (3.9). At this stage, the bandwidths b affect the probabilities $h_{ij}(\hat{c})$. As mentioned in Section 2.3, the probabilities $h_{ij}(\hat{c})$ must satisfy the three conditions discussed in such section.

We present an example using simulated data and study the types of bandwidths and bandwidth selections discussed in Section 3.2, in order to investigate how the probabilities $h_{ij}(\hat{c})$ are dispersed in the $(n+1)^2$ equal-sized squares. In order to investigate the probabilities $h_{ij}(\hat{c})$ and to get an insight how the proposed method works with the nonparametric copula, we present an example using a small simulated data set. We should emphasize that, as mentioned in Section 3.2, there are many nonparametric methods can also be used instead of kernel-based copula methods. However, the performance of the proposed method with other nonparametric copulas should be studied and investigated. We left these as a topic for future research.

3.3.1 Example: Simulated data

Consider a set of bivariate data, (x_i, y_i) where $i = 1, ..., 9$. Using the proposed method in this chapter, we calculate the probabilities $h_{ij}(\hat{c})$, which is equation (3.10)

for different types of bandwidth selections and different types of bandwidths. In this example, we simulated data from the Frank copula with Kendall's $\tau = 0.75$, which indicates a strong positive association between the two random quantities. The data and the scatter plot are shown in Table 3.1 and Figure 3.1.

x	0.0654	0.0988	0.2234	0.2515	0.3010	0.3640	0.5440	0.8986	0.9660
y	0.0692	0.1118	0.1825	0.3419	0.3642	0.3973	0.6839	0.8058	0.8314

Table 3.1: Simulation data from Frank copula, $\tau = 0.75$

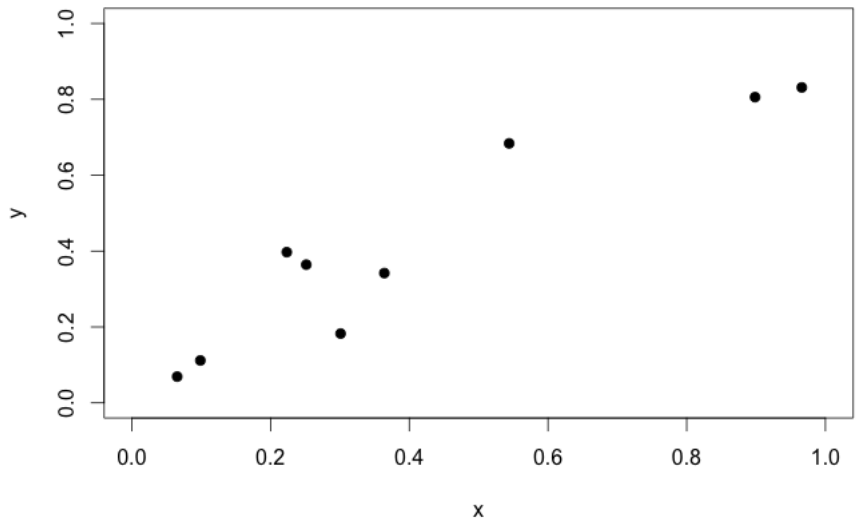


Figure 3.1: Scatter plot of the simulation data

Tables 3.2 - 3.5 show the probabilities $h_{ij}(\hat{c})$ and $H_{ij}(\hat{c})$ based on different types of bandwidth selections and types of bandwidths. These tables are presented this way in order to show the natural corresponding to the bivariate plots of simulated data in Figure 3.1 with the corresponding probabilities $h_{ij}(\hat{c})$ and $H_{ij}(\hat{c})$.

For Table 3.2, the normal reference rule-of-thumb from equation (3.3), discussed in Section 3.2, has been used. We see that the sum of $h_{ij}(\hat{c})$ is equal to 1, each row and column is equal to $\frac{1}{n+1}$, and all $h_{ij}(\hat{c}) \geq 0$. Figure 3.2 shows a 3D-plot of the probabilities $h_{ij}(\hat{c})$. The plots of probabilities $h_{ij}(\hat{c})$ in this section are given with x, y and z are equal to 0 at the left-front corner, and at the right-back corner x, y and

z are equal to 1. This figure shows that the probabilities $h_{ij}(\hat{c})$ are higher at left-front corner and right-back corner compared to other corners due to the simulated data, where we have two points with small x and y values and also two with large x and y values, but no points with one value small and one large. The corresponding bandwidths, b for X and Y are given in Table 3.6.

		i=1	2	3	4	5	6	7	8	9	10
$H_{ij}(\hat{c})$	j=10	0.1001	0.2004	0.3003	0.4012	0.5009	0.6005	0.7013	0.8010	0.9027	1.0000
	9	0.0995	0.1985	0.2965	0.3942	0.4891	0.5814	0.6709	0.7546	0.8341	0.9027
	8	0.0978	0.1935	0.2870	0.3788	0.4663	0.5491	0.6261	0.6944	0.7546	0.8010
	7	0.0949	0.1854	0.2721	0.3558	0.4340	0.5062	0.5712	0.6259	0.6707	0.7013
	6	0.0905	0.1739	0.2515	0.3248	0.3919	0.4525	0.5056	0.5484	0.5811	0.6005
	5	0.0845	0.1589	0.2259	0.2872	0.3421	0.3908	0.4326	0.4652	0.4887	0.5009
	4	0.0762	0.1401	0.1952	0.2437	0.2860	0.3227	0.3536	0.3773	0.3936	0.4012
	3	0.0648	0.1164	0.1586	0.1941	0.2237	0.2488	0.2696	0.2852	0.2958	0.3003
	2	0.0495	0.0868	0.1157	0.1386	0.1567	0.1714	0.1832	0.1921	0.1980	0.2004
	1	0.0288	0.0493	0.0642	0.0752	0.0831	0.0891	0.0937	0.0971	0.0993	0.1001
$h_{ij}(\hat{c})$	10	0.0006	0.0013	0.0020	0.0031	0.0048	0.0074	0.0113	0.0160	0.0222	0.0288
	9	0.0017	0.0033	0.0045	0.0059	0.0074	0.0095	0.0124	0.0155	0.0193	0.0222
	8	0.0029	0.0052	0.0068	0.0082	0.0093	0.0105	0.0121	0.0135	0.0154	0.0159
	7	0.0044	0.0072	0.0090	0.0104	0.0111	0.0115	0.0119	0.0119	0.0121	0.0111
	6	0.0061	0.0089	0.0107	0.0119	0.0122	0.0119	0.0113	0.0102	0.0092	0.0072
	5	0.0083	0.0105	0.0119	0.0128	0.0127	0.0120	0.0108	0.0090	0.0071	0.0046
	4	0.0114	0.0124	0.0128	0.0131	0.0126	0.0116	0.0101	0.0080	0.0058	0.0031
	3	0.0153	0.0142	0.0133	0.0126	0.0116	0.0104	0.0089	0.0068	0.0047	0.0021
	2	0.0207	0.0168	0.0140	0.0119	0.0101	0.0087	0.0073	0.0055	0.0037	0.0015
	1	0.0288	0.0205	0.0150	0.0110	0.0079	0.0060	0.0046	0.0033	0.0022	0.0009

Table 3.2: $H_{ij}(\hat{c})$ and $h_{ij}(\hat{c})$ with Normal reference rule-of-thumb and fixed bandwidth

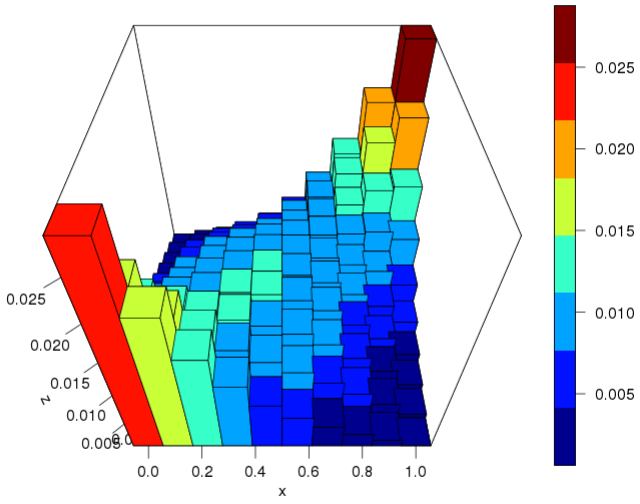


Figure 3.2: 3D-plot of probabilities $h_{ij}(\hat{c})$ for normal reference rule-of-thumb

In Table 3.3, the probabilities $h_{ij}(\hat{c})$ are obtained using the LSCV bandwidth selection and fixed bandwidth (discussed in Section 3.2). The logical conditions for the $h_{ij}(\hat{c})$ are always satisfied except when explicitly mentioned. The corresponding 3D-plot in Figure 3.3 shows that the probabilities $h_{ij}(\hat{c})$ are large close to the observation points, this happens because the method for estimating the density is based on the minimum distance measure between $\hat{f}(x, y)$ and $f(x, y)$ (i.e. the integrated squared error mentioned in Section 3.2). This also reflects the bandwidths obtained from this method is smaller than normal reference rule-of-thumb which shown in Table 3.6. The same 3D-plot also shows that the probabilities $h_{ij}(\hat{c})$ are higher in three main areas, namely left-front corner, the right-back corner and in the middle of the 3D-plot, which reflects the simulated data.

		i=1	2	3	4	5	6	7	8	9	10
	j=10	0.1004	0.2006	0.3014	0.4011	0.5009	0.6006	0.7004	0.8012	0.9010	1.0000
	9	0.1004	0.2006	0.3014	0.4011	0.5007	0.5999	0.6960	0.7836	0.8529	0.9013
	8	0.1004	0.2005	0.3008	0.3996	0.4981	0.5944	0.6804	0.7456	0.7840	0.8015
	7	0.1003	0.1995	0.2963	0.3884	0.4807	0.5702	0.6403	0.6803	0.6961	0.7004
	6	0.0997	0.1949	0.2797	0.3532	0.4302	0.5105	0.5688	0.5938	0.5998	0.6006
$H_{ij}(\hat{c})$	5	0.0983	0.1850	0.2493	0.2968	0.3541	0.4240	0.4764	0.4969	0.5006	0.5009
	4	0.0967	0.1752	0.2220	0.2483	0.2844	0.3370	0.3802	0.3978	0.4010	0.4011
	3	0.0939	0.1661	0.2039	0.2182	0.2344	0.2622	0.2881	0.2992	0.3013	0.3014
	2	0.0822	0.1395	0.1667	0.1744	0.1789	0.1873	0.1957	0.1996	0.2003	0.2003
	1	0.0517	0.0820	0.0941	0.0968	0.0975	0.0985	0.0996	0.1001	0.1002	0.1002
	10	0.0000	0.0000	0.0000	0.0000	0.0001	0.0006	0.0036	0.0132	0.0306	0.0506
	9	0.0000	0.0001	0.0005	0.0009	0.0011	0.0029	0.0101	0.0224	0.0308	0.0310
	8	0.0001	0.0009	0.0035	0.0066	0.0062	0.0068	0.0159	0.0252	0.0226	0.0132
	7	0.0006	0.0040	0.0120	0.0187	0.0152	0.0091	0.0119	0.0149	0.0098	0.0034
	6	0.0014	0.0085	0.0205	0.0259	0.0197	0.0103	0.0059	0.0045	0.0023	0.0006
$h_{ij}(\hat{c})$	5	0.0016	0.0082	0.0175	0.0212	0.0212	0.0174	0.0091	0.0029	0.0006	0.0001
	4	0.0028	0.0063	0.0091	0.0120	0.0199	0.0247	0.0174	0.0064	0.0011	0.0001
	3	0.0118	0.0148	0.0106	0.0066	0.0116	0.0195	0.0174	0.0073	0.0013	0.0001
	2	0.0305	0.0270	0.0151	0.0049	0.0039	0.0073	0.0074	0.0033	0.0006	0.0000
	1	0.0517	0.0303	0.0121	0.0027	0.0007	0.0010	0.0011	0.0005	0.0001	0.0000

Table 3.3: $H_{ij}(\hat{c})$ and $h_{ij}(\hat{c})$ with LSCV and fixed bandwidth

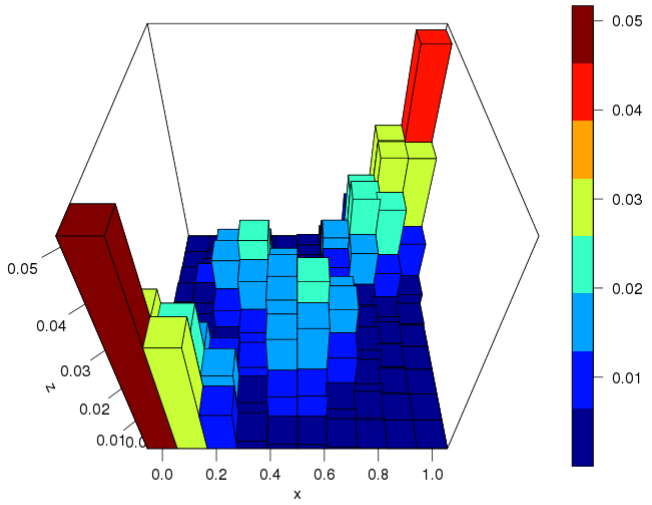


Figure 3.3: 3D-plot of probabilities $h_{ij}(\hat{c})$ for LSCV

For an adaptive-nn type of bandwidth discussed in Section 3.2, the probabilities $h_{ij}(\hat{c})$ are shown in Table 3.4 and the corresponding 3D-plot of probabilities $h_{ij}(\hat{c})$ given in Figure 3.4. From the 3D-plot, we see that the shape of the figure is quite similar with Figure 3.3 where the probabilities $h_{ij}(\hat{c})$ are higher in three main areas, but the probabilities $h_{ij}(\hat{c})$ for each cell are different. Providing the type of bandwidth used, the probabilities $h_{ij}(\hat{c})$ are distributes based on the observed data. Basically, the adaptive-nn type of bandwidth uses a large bandwidth to the data that sparse, providing a low density, and a small bandwidth to the data that close to each other, providing a high density. The corresponding bandwidths for X and Y this method are shown in Table 3.6.

		i=1	2	3	4	5	6	7	8	9	10
	j=10	0.1006	0.2001	0.3008	0.4003	0.5001	0.6012	0.7010	0.8017	0.9009	1.0000
	9	0.1005	0.2000	0.3006	0.3998	0.4985	0.5949	0.6838	0.7677	0.8434	0.9001
	8	0.1005	0.1997	0.2995	0.3973	0.4937	0.5837	0.6575	0.7200	0.7705	0.8006
	7	0.1003	0.1979	0.2940	0.3857	0.4746	0.5544	0.6106	0.6523	0.6831	0.7003
	6	0.0993	0.1918	0.2777	0.3551	0.4282	0.4951	0.5382	0.5669	0.5877	0.6006
$H_{ij}(\hat{c})$	5	0.0970	0.1799	0.2495	0.3069	0.3603	0.4143	0.4498	0.4726	0.4894	0.5009
	4	0.0936	0.1657	0.2193	0.2584	0.2941	0.3334	0.3607	0.3784	0.3917	0.4011
	3	0.0879	0.1493	0.1906	0.2169	0.2381	0.2610	0.2770	0.2875	0.2956	0.3015
	2	0.0740	0.1196	0.1484	0.1651	0.1760	0.1855	0.1917	0.1957	0.1988	0.2011
	1	0.0456	0.0687	0.0828	0.0908	0.0955	0.0983	0.0996	0.1003	0.1009	0.1013
	10	0.0000	0.0000	0.0001	0.0004	0.0011	0.0047	0.0109	0.0168	0.0235	0.0424
	9	0.0000	0.0003	0.0008	0.0014	0.0023	0.0064	0.0152	0.0215	0.0251	0.0266
	8	0.0002	0.0016	0.0037	0.0060	0.0076	0.0103	0.0175	0.0208	0.0198	0.0128
	7	0.0010	0.0051	0.0102	0.0144	0.0158	0.0128	0.0131	0.0130	0.0101	0.0043
	6	0.0023	0.0096	0.0163	0.0200	0.0197	0.0130	0.0076	0.0058	0.0039	0.0015
$h_{ij}(\hat{c})$	5	0.0034	0.0108	0.0161	0.0181	0.0178	0.0146	0.0083	0.0051	0.0035	0.0021
	4	0.0057	0.0107	0.0123	0.0128	0.0144	0.0165	0.0113	0.0072	0.0052	0.0036
	3	0.0139	0.0159	0.0124	0.0097	0.0102	0.0133	0.0098	0.0065	0.0050	0.0036
	2	0.0284	0.0225	0.0147	0.0086	0.0063	0.0067	0.0048	0.0033	0.0025	0.0019
	1	0.0456	0.0231	0.0141	0.0080	0.0047	0.0028	0.0013	0.0008	0.0006	0.0004

Table 3.4: $H_{ij}(\hat{c})$ and $h_{ij}(\hat{c})$ with LSCV and adaptive-nn bandwidth

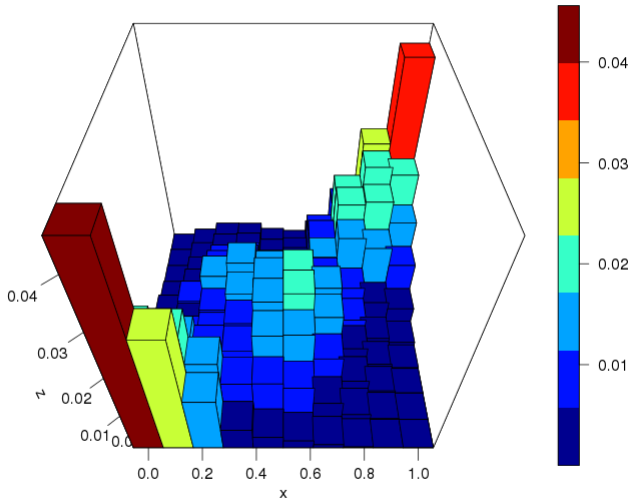


Figure 3.4: 3D-plot of probabilities $h_{ij}(\hat{c})$ for adaptive-nn

Finally, in Table 3.5, we show the probabilities $h_{ij}(\hat{c})$ for generalized-nn type of bandwidth. From this table, we can see that the sum of the probabilities $h_{ij}(\hat{c})$ are not 1, each row and column is not equal with $\frac{1}{n+1}$ and there are probabilities $h_{ij}(\hat{c})$ values less than 0. The probabilities $h_{ij}(\hat{c})$ obtained are clearly seen not as they should be, hence, this method cannot be used further. These features reflect the generalized-nn used and it problem mentioned in Section 3.2 where the integral over the estimated density function does not equal 1 and tends to produce an estimate with very heavy tails.

		i=1	2	3	4	5	6	7	8	9	10
	j=10	0.0971	0.2015	0.3004	0.3987	0.4973	0.5956	0.6919	0.7825	0.8648	0.8879
	9	0.0961	0.2017	0.3014	0.4011	0.5009	0.6004	0.6987	0.7912	0.8633	0.8648
	8	0.0913	0.2017	0.3012	0.4007	0.4997	0.5970	0.6870	0.7562	0.7912	0.7825
	7	0.0848	0.2013	0.2978	0.3902	0.4830	0.5733	0.6472	0.6870	0.6987	0.6919
	6	0.0755	0.1985	0.2811	0.3520	0.4295	0.5099	0.5730	0.5970	0.6004	0.5957
$H_{ij}(\hat{c})$	5	0.0646	0.1922	0.2528	0.2975	0.3561	0.4244	0.4809	0.4996	0.5009	0.4971
	4	0.0538	0.1854	0.2258	0.2484	0.2864	0.3376	0.3846	0.4004	0.4011	0.3980
	3	0.0441	0.1785	0.2094	0.2192	0.2358	0.2621	0.2905	0.3009	0.3014	0.2995
	2	0.0335	0.1560	0.1784	0.1830	0.1864	0.1920	0.1989	0.2015	0.2017	0.2012
	1	0.0120	0.0338	0.0449	0.0541	0.0640	0.0746	0.0845	0.0913	0.0961	0.0970
	10	0.0009	-0.0011	-0.0008	-0.0014	-0.0012	-0.0012	-0.0021	-0.0018	0.0102	0.0216
	9	0.0048	-0.0048	0.0001	0.0003	0.0007	0.0022	0.0083	0.0233	0.0372	0.0102
	8	0.0065	-0.0062	0.0032	0.0070	0.0062	0.0070	0.0161	0.0293	0.0233	-0.0019
	7	0.0093	-0.0064	0.0137	0.0216	0.0153	0.0098	0.0108	0.0159	0.0083	-0.0021
	6	0.0109	-0.0047	0.0221	0.0261	0.0188	0.0121	0.0067	0.0053	0.0021	-0.0009
$h_{ij}(\hat{c})$	5	0.0108	-0.0041	0.0201	0.0222	0.0206	0.0171	0.0094	0.0029	0.0005	-0.0007
	4	0.0097	-0.0029	0.0096	0.0126	0.0215	0.0249	0.0186	0.0054	0.0003	-0.0012
	3	0.0105	0.0120	0.0084	0.0052	0.0132	0.0207	0.0215	0.0077	0.0004	-0.0014
	2	0.0215	0.1006	0.0113	-0.0045	-0.0066	-0.0050	-0.0030	-0.0042	-0.0047	-0.0014
	1	0.0120	0.0218	0.0111	0.0092	0.0099	0.0106	0.0099	0.0068	0.0048	0.0009

Table 3.5: $H_{ij}(\hat{c})$ and $h_{ij}(\hat{c})$ with LSCV and generalized-nn bandwidth

Table 3.6 shows the corresponding bandwidths, b of the two random quantities for the two bandwidth selections and the four type of bandwidths discussed above in Section 3.2.

Bandwidth selection	Type of bandwidth	b_X	b_Y	Table
Normal Reference rule-of-thumb	fixed	0.2089	0.2089	Table 3.2
Least Square Cross-Validation	fixed	0.0868	0.0926	Table 3.3
	average adaptive-nn	0.1649	0.4019	Table 3.4
	average generalized-nn	0.1648	0.2009	Table 3.5

Table 3.6: Bandwidth selections and type of bandwidths

As discussed in Section 2.4, equations (3.10) and (3.11) can be considered to infer about an event E that involves the next observation (X_{n+1}, Y_{n+1}). Given the same definition in Section 2.4, the nonparametric method presented in Section 3.3 leads to the lower and upper probabilities for the event $E(X_{n+1}, Y_{n+1})$ as in equations (2.3) and (2.4). Consider the similar event as in Section 2.4, where we are interested in the sum of the next observations X_{n+1} and Y_{n+1}, say $T_{n+1} = X_{n+1} + Y_{n+1}$. Then

the lower and upper probability for the event that the sum of the next observations will exceed a particular value t are following equations (2.5) and (2.6), where in this chapter, we use nonparametric copula instead of parametric copula.

3.3.2 Example: Insurance data

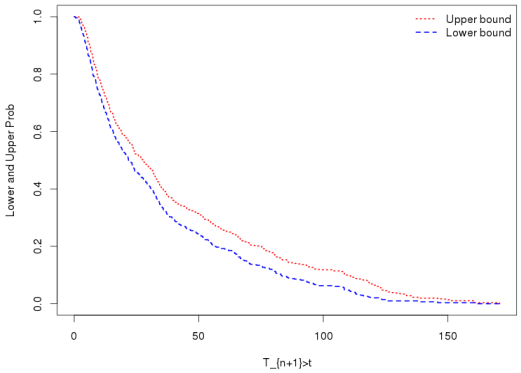
We consider the data from casualty insurances, given in Example 2.6.1 with the X value representing Loss and the Y value ALAE (allocated loss adjustment expenses). We are interested in the event that the sum of the two values for the next observation is greater than a certain value t, so, $T_{n+1} = X_{n+1} + Y_{n+1} > t$. In this example we used only the bandwidth selections and the types of bandwidths discussed in Section 3.3.1, which did not lead to problem with the h_{ij} values, i.e. normal reference rule-of-thumb and LSCV bandwidth selections, with fixed and adaptive-nn bandwidths. The results are presented in Table 3.7.

Bandwidth selection	Type of bandwidth	b_X	b_Y
Normal Reference rule-of-thumb	fixed	0.1647	0.1647
Least Square Cross Validation	fixed	0.1673	0.2577
	average adaptive-nn	0.5329	0.9325

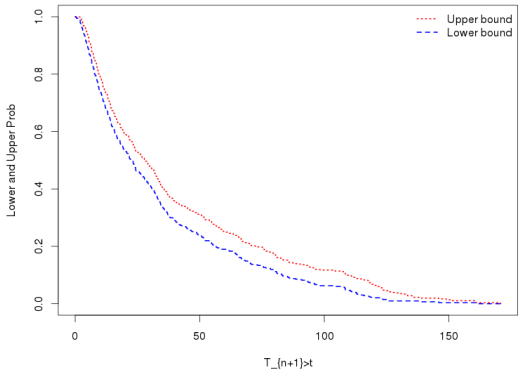
Table 3.7: Bandwidth selections and type of bandwidths

The results in Table 3.7 show that the bandwidths of the Loss and ALAE variables are 0.1647 when using the normal reference rule-of-thumb. The value is the same for both variables because its a fixed bandwidth type. For the LSCV method, the fixed bandwidth type gives bandwidth for Loss 0.1673 and for ALAE 0.2577. The bandwidths for these two random quantities are different because the LSCV method chooses the bandwidth based on minimizing the integrated squared error as discussed in Section 3.2. For Loss and ALAE variables, the algorithm produced 4-th and 7-th adaptive-nn, respectively. The corresponding bandwidth values are $b = 0.5329$ and $b = 0.9325$, respectively, using the formula given in Section 3.2 i.e. $b_z = k_z \sigma_z n^{(-1/4)}$. Figure 3.5 shows lower and upper probabilities according to our method for the event $T_{n+1} > t$ corresponding to these bandwidth selections and types of bandwidths. This figure can be interpreted and implemented in many ways

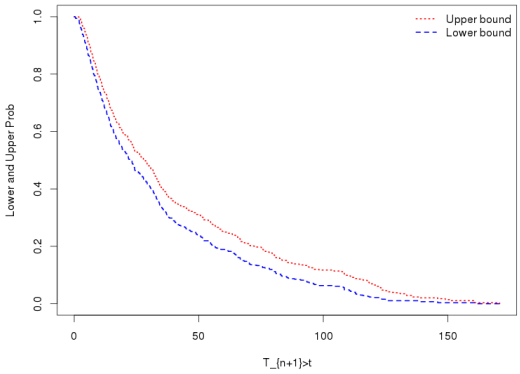
depending on applications or events of interest. From this figure we cannot see much differences, it seem all bandwidth selections and types of bandwidth used give an identical figure of lower and upper probabilities for the event $T_{n+1} > t$. However, the bandwidths obtained in Table 3.7 shows that the LSCV method gives higher bandwidth for ALAE compared to normal reference rule-of thumb which shows the LSCV method is over-smoothing the probabilities h_{ij}. As the probabilities h_{ij} are the most important part in this study, we show the 3D-plot of this data set for all bandwidth selections and types of bandwidths in Figure 3.6. From this figure, the probabilities h_{ij} are quite higher at left-front corner and right-back corner for each subfigure (i.e. Figures 3.5(a), 3.5(b) and 3.5(c)), and the probabilities h_{ij} are scattered at most of the cells for all bandwidth selections and the types of bandwidths. However, the probabilities h_{ij} are different in small amount among the cells. These features are the reason that the lower and upper probabilities for the event $T_{n+1} > t$ is quite identical in Figure 3.5. Another noticeable feature when the proposed method applied for this data set, the probabilities h_{ij} are more scattered in Figure 3.6 compared to Figure 2.8 in Section 2.6.1.



(a) Normal reference rule-of-thumb; fixed bandwidth

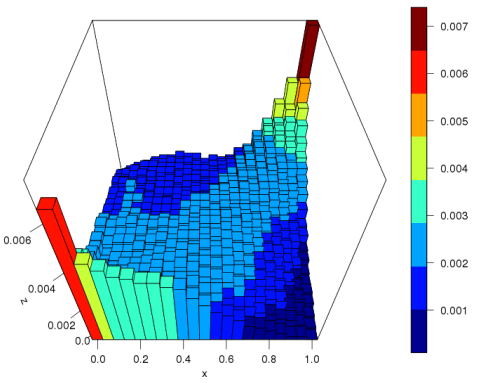


(b) LSCV; fixed bandwidth

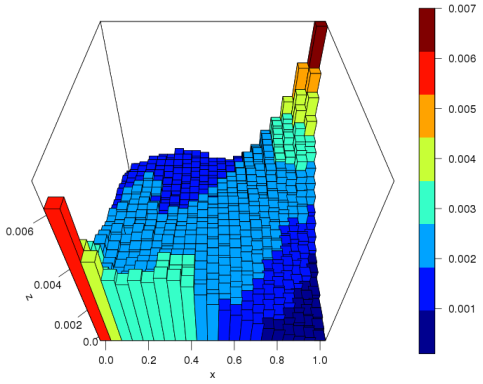


(c) LSCV; Adaptive-nn bandwidth $k_x = 4$ and $k_y = 7$

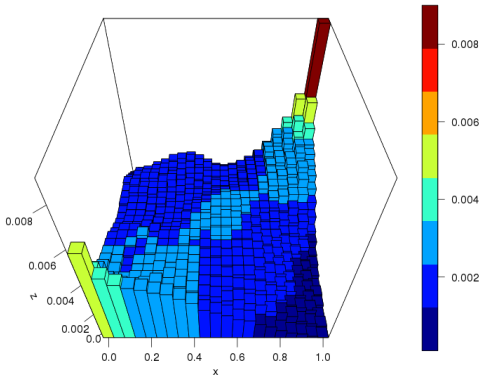
Figure 3.5: Lower and upper probabilities for the event $T_{n+1} > t$



(a) Normal reference rule-of-thumb; fixed bandwidth



(b) LSCV; fixed bandwidth

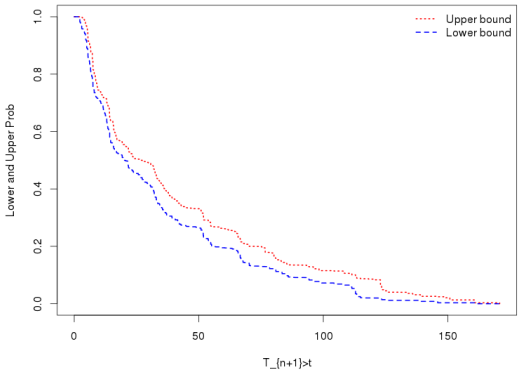


(c) LSCV; Adaptive-nn bandwidth $k_x = 4$ and $k_y = 7$

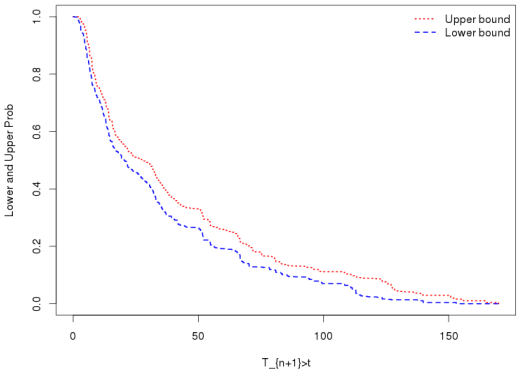
Figure 3.6: 3D-plot of probabilities h_{ij} for bandwidth selections and types of bandwidths

Due to quite identical lower and upper probabilities for the event $T_{n+1} > t$ in Figure 3.5, we use different values of k_z for the adaptive-nn bandwidth. k_z is the kth nearest neighbours of the observations and it is related to the distance of any point to its nearest observations as discussed in Section 3.2. We only use this type of bandwidth because this method allows us to determine the nearest neighbour to be used, while the other types of bandwidth do not offer this possibility. Another reason why we investigate this method further for different values of k_z is, as mentioned in

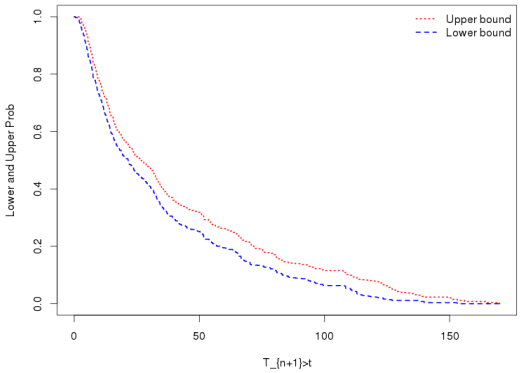
Section 3.2, estimation is influenced by the values of k_z, so, the values of k_z might also affect prediction. Figure 3.7 shows the lower and upper probabilities for the event $T_{n+1} > t$ for our method, for different values of k_z. From this figure, we can see that as k_z increases, the lower and upper survival functions become smoother. This is due to the adaptive-nn bandwidth used in our method and the specific event of interest. For example, when we consider $k_z = 2$ in adaptive-nn bandwidth, this type of bandwidth uses two nearest points from the point that need to be estimated, and this gives a few peaks. If we consider $k_z = 5$, the adaptive-nn bandwidth uses the five nearest points from the point that need to be estimated, and this gives fewer peaks. In other words, the 5-th nearest neighbour uses a broader distance for estimating the points. In addition, as we consider the sum event of the bivariate random quantities, the possibility probabilities h_{ij} to be included or not is depending on how the probabilities h_{ij} are scattered. We show the 3D-plots of probabilities h_{ij} for $k_z = 2$ and $k_z = 5$ in Figure 3.8. Figure 3.8 shows that the probabilities h_{ij} are scattered differently between $k_z = 2$ and $k_z = 5$, whereby the probabilities h_{ij} are higher at $k_z = 2$ compared to $k_z = 5$. So, these 3D-plots suggest that the k_z for adaptive-nn bandwidth does affect the prediction. This is due to the nearest point used and the way of adaptive-nn bandwidth work. As the value of k_z increases, the adaptive-nn method over-smooth the probabilities h_{ij}. Consequently, the three conditions for the probabilities h_{ij} discussed in Section 2.3 are not satisfied. In the following section, we will investigate the bandwidth selection related to the predictive performance of our method.



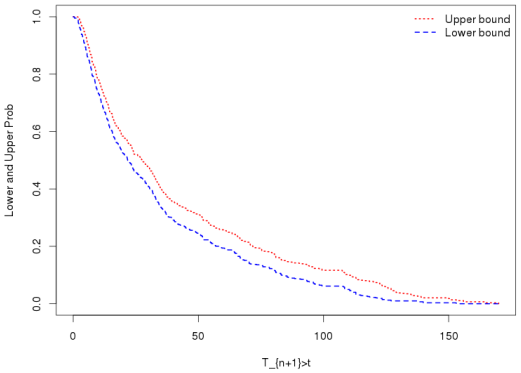
(a) $k_x = k_y = 1$



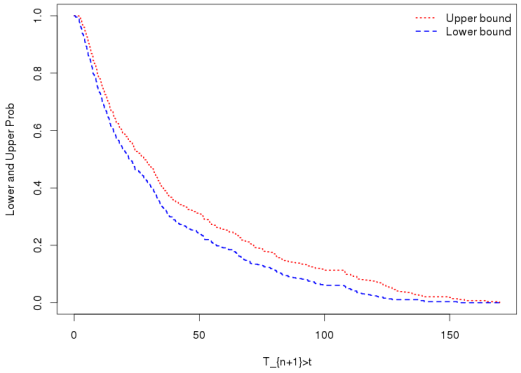
(b) $k_x = k_y = 2$



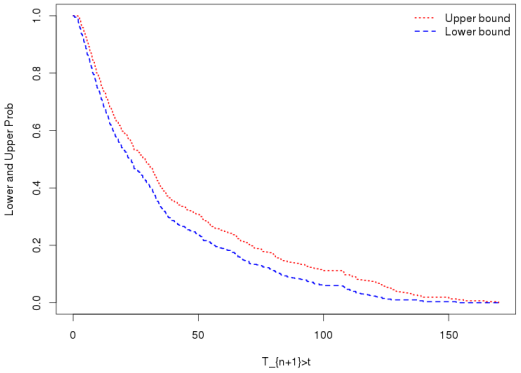
(c) $k_x = k_y = 3$



(d) $k_x = k_y = 4$

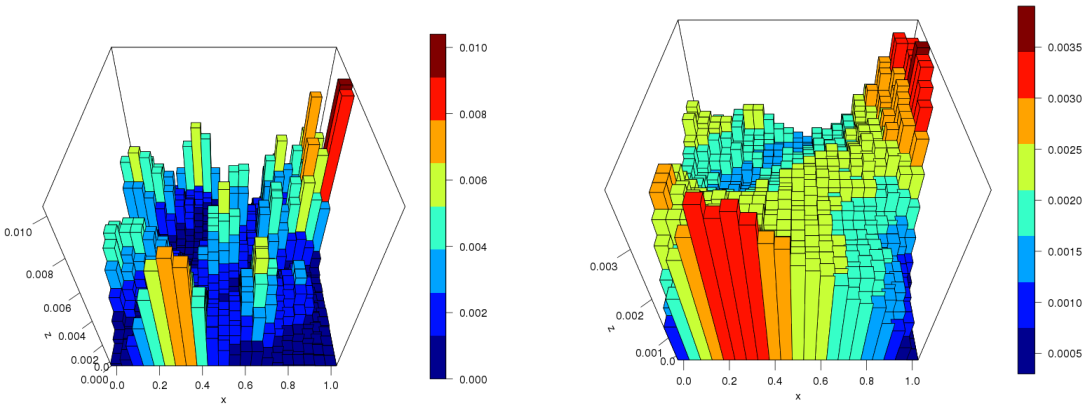


(e) $k_x = k_y = 5$



(f) $k_x = k_y = 6$

Figure 3.7: Lower and upper probabilities for the event $T_{n+1} > t$, adaptive-nn bandwidth for different k_z, in each case $k_x = k_y$



(a) $k_z = 2$ (b) $k_z = 5$

Figure 3.8: 3D-plot of probabilities h_{ij} for $k_x = k_y = 2$ and $k_x = k_y = 5$ adaptive-nn bandwidth.

3.4 Predictive performance

We conducted a simulation study to obtain an indication of the predictive performance of this method. We used a similar method as discussed in Section 2.5 to indicate the predictive performance of our method, but now using a nonparametric copula. The results are based on $N = 10,000$ bivariate simulated samples, each of size $n + 1$, using the Frank, Normal, Clayton and Gumbel copulas. For each simulated sample, the first n pairs are used as data for our predictive method, the additional pair is considered as a future observation and is used to test the predictive performance of this method. Equations (2.9) and (2.10) in Section 2.5 are used to indicate the performance of the proposed method. In other words, the proposed method performs well if the two inequalities in equations (2.9) and (2.10) hold.

Based on previous example in Section 3.3.2, we conducted two types of simulation studies. First, in Section 3.4.1 we use auto-driven bandwidth selection where we let algorithm namely `npudistbw` in the R package `np` [49] choose the bandwidth. Secondly, in Section 3.4.2 we use manual bandwidth selection where we choose the value of bandwidth manually. In function `npudistbw`, a multivariate numerical

search algorithm uses direction set (Powell [78]) methods in multidimensions [49] to optimize the bandwidth. In the `np` package, the optimizer used is Powell's conjugate direction method, which requires the setting of initial values and search directions for bandwidths, and when restarting, random values for successive invocations [49].

3.4.1 `np` R package bandwidth selection

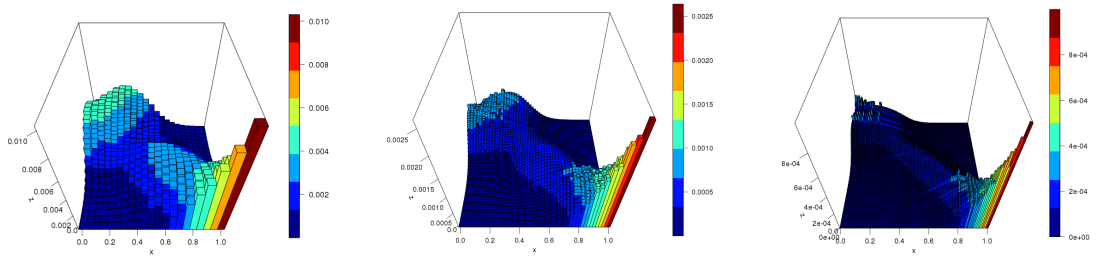
In this section, we let algorithm in the `np` R package choose the bandwidth using the normal reference rule-of-thumb and LSCV bandwidth selections with fixed and adaptive-nn bandwidths. We run $N = 10,000$ simulations for different sample sizes, $n = 20, 50, 100$, using the Normal, Frank, Clayton and Gumbel copulas with $\tau = -0.75, -0.50, -0.25, 0.25, 0.50, 0.75$. One can use any values of q, we choose the same values of q used in Section 2.5, i.e. $q = 0.25, 0.50, 0.75$. As discussed in Section 2.5, for $q \in (0, 1)$, the inverse values of the lower and upper survival functions of T_{n+1} in equations (2.5) and (2.6) are defined as in equations (2.7) and (2.8), respectively. In this simulation study, we show results from the Clayton and Frank copulas for both methods. We also repeat the simulation study for the Normal and Gumbel copulas, which leads to the same conclusions as Clayton and Frank copulas.

Table 3.8 shows the predictive performance of the proposed method with kernel-based copula, using normal reference rule-of-thumb bandwidth selection and fixed bandwidth for simulated data from the Clayton copula. Table 3.9 shows the corresponding bandwidth. The bandwidth values b_x and b_y, in Table 3.9, are average of the respective bandwidths over $10,000$ runs. In this section, θ which is given in the second column of each table, is the copula parameter value corresponding to the Kendall's tau given in the same table, as discussed in Section 2.2. In Table 3.8, θ is the Clayton copula parameter value. Table 3.8 shows that there are a few cases for which q is not contained in the interval $[p_1, p_2]$ especially for $\tau = -0.5$ and $\tau = -0.75$, and sample size, $n = 50$ and $n = 100$. These are highlighted by bold font numbers in the table. First of all, the data are simulated from the Clayton copula, so the simulated data obtained will exhibit greater dependence in the negative tail than in the positive tail. For example, consider the data simulated from $\tau = -0.75$, the data will have greater dependence at the large x and small y observation val-

ues of the data set as mentioned in Section 3.2. Then, using the normal reference rule-of-thumb bandwidth selection with fixed bandwidth, the probabilities h_{ij} are highly scattered at the right-front corner of the 3D-plot of the probabilities h_{ij} as shown in Figure 3.9 for different sample sizes. As we are interested on the sum of the two values in the bivariate data, the probabilities h_{ij} are tend to be included when calculating the lower and upper of the survival functions for t at the middle or diagonal of the 3D-plots. But, when t at the left-front corner (small x and small y observation values) and right-back corner (large x and large y observation values) of the 3D-plots, very small value of probabilities h_{ij} to be included when calculating the lower and upper of the survival functions. The results that the values $q = 0.25$ and $q = 0.75$ are not in the corresponding p_1 and p_2, are mostly for $\tau = -0.5$ and $\tau = -0.75$, and for $n = 20$, $n = 50$ and $n = 100$. For positive correlation, the values q are not in the corresponding p_1 and p_2 for $n = 100$ and two cases for $n = 50$. From Table 3.9, we can see that the average bandwidth of $10,000$ repetitions is smaller as n increases, which is a logical feature in estimation. One feature noticeable is that the bandwidth corresponding to negative τ is greater than positive τ. This might occur because the simulated data from the Clayton copula have greater dependence in the negative tail. From this table, we also see that, as the correlation decrease, the bandwidth values become larger regardless of positive or negative correlation. This feature reflects the closeness of the data to each other.

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	6.0000	0.25	0.2470	0.2987	**0.2516**	0.2733	**0.2566**	0.2671
		0.50	0.4818	0.5283	0.4951	0.5136	0.4975	0.5079
		0.75	0.7137	0.7612	0.7416	0.7612	**0.7516**	0.7609
0.5	2.0000	0.25	0.2398	0.2987	0.2475	0.2744	**0.2540**	0.2652
		0.50	0.4778	0.5356	0.4919	0.5130	0.4996	0.5113
		0.75	0.7150	0.7602	0.7450	0.7608	**0.7531**	0.7610
0.25	0.6667	0.25	0.2277	0.2937	0.2397	0.2691	**0.2537**	0.2673
		0.50	0.4821	0.5460	**0.5037**	0.5362	**0.5045**	0.5196
		0.75	0.7211	0.7728	0.7461	0.7681	0.7462	0.7573
-0.25	-0.6667	0.25	0.1873	0.2692	0.2234	0.2636	0.2336	0.2510
		0.50	0.4338	0.5600	0.4839	0.5420	0.4840	0.5133
		0.75	0.7274	0.8086	0.7499	0.7829	0.7479	0.7633
-0.5	-2.0000	0.25	0.1476	0.2509	0.1874	**0.2317**	0.2119	**0.2350**
		0.50	0.4097	0.6160	0.4536	0.5514	0.4827	0.5376
		0.75	**0.7577**	0.8563	**0.7735**	0.8155	**0.7725**	0.7956
-0.75	-6.0000	0.25	0.0626	**0.1743**	0.1098	**0.1694**	0.1426	**0.1745**
		0.50	0.3150	0.6770	0.4119	0.6090	0.4593	0.5729
		0.75	**0.8193**	0.9431	**0.8435**	0.8996	**0.8380**	0.8687

Table 3.8: Simulated data from Clayton copula; normal reference rule-of-thumb; fixed bandwidth



(a) $n = 20$ (b) $n = 50$ (c) $n = 100$

Figure 3.9: 3D-plot of probabilities h_{ij} for Clayton simulated data with normal reference rule-of-thumb, fixed bandwidth for $\tau = -0.75$

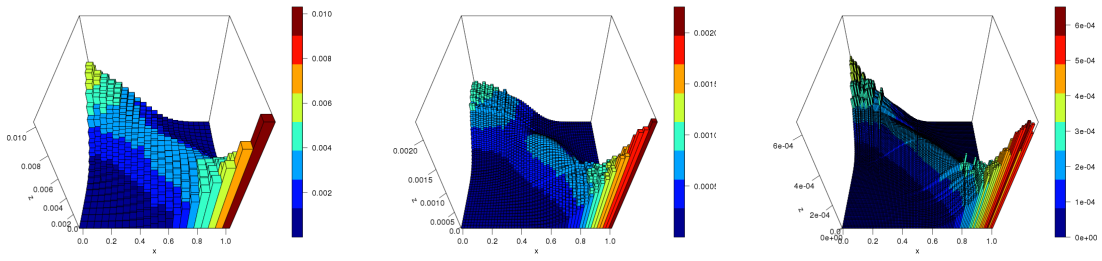
τ	θ	$n = 20$		$n = 50$		$n = 100$	
		b_X	b_Y	b_X	b_Y	b_X	b_Y
0.75	6.0000	0.0909	0.0910	0.0627	0.0627	0.0473	0.0473
0.5	2.0000	0.1628	0.1633	0.1144	0.1144	0.0874	0.0875
0.25	0.6667	0.2295	0.2294	0.1619	0.1611	0.1250	0.1248
-0.25	-0.6667	0.3432	0.2668	0.2279	0.1689	0.1702	0.1206
-0.5	-2.0000	0.3403	0.2321	0.2071	0.1288	0.1445	0.0863
-0.75	-6.0000	0.2367	0.1581	0.1254	0.0815	0.0816	0.0543

Table 3.9: Bandwidth for simulated data from Clayton copula; normal reference rule-of-thumb; fixed bandwidth

Table 3.10 shows the predictive performance of the proposed method, using normal reference rule-of-thumb bandwidth selection and fixed bandwidth for simulated data from the Frank copula. Table 3.11 shows the corresponding bandwidth. Table 3.10 shows that there are a few cases for which q is not contained in the interval $[p_1, p_2]$ which quite similar with the data simulated from the Clayton copula. But, the number of highlighted bold font number are less than Table 3.8. This happens because we simulate data from the Frank copula which is symmetric and as we applied the normal reference rule-of-thumb bandwidth selection and fixed bandwidth, the probabilities h_{ij} are quite symmetrically distributed as shown in Figure 3.10. However, for $\tau = 0.5$ and $\tau = 0.75$, and for $n = 50$ and $n = 100$, the q is not contained in the interval $[p_1, p_2]$. Table 3.11 shows the same relationship between bandwidth and sample sizes. As n increases, the average bandwidths are decreases, and as the strength of correlation become stronger, the average bandwidths tend to be smaller.

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	14.1385	0.25	0.2422	0.2909	0.2494	0.2702	**0.2515**	0.2621
		0.50	0.4778	0.5248	0.4871	0.5059	**0.5040**	0.5126
		0.75	0.7146	0.7611	0.7288	**0.7468**	0.7458	0.7574
0.5	5.7363	0.25	0.2478	0.2998	**0.2526**	0.2768	0.2484	0.2592
		0.50	0.4741	0.5236	0.4957	0.5169	0.4970	0.5066
		0.75	0.7063	0.7605	0.7377	0.7586	0.7391	**0.7495**
0.25	2.3719	0.25	0.2409	0.2972	0.2495	0.2735	**0.2517**	0.2648
		0.50	0.4691	0.5347	0.4876	0.5147	0.4983	0.5137
		0.75	0.7116	0.7690	0.7283	0.7514	0.7426	0.7542
-0.25	-2.3719	0.25	0.1927	0.2775	0.2190	0.2543	0.2339	0.2525
		0.50	0.4459	0.5656	0.4716	0.5289	0.4897	0.5183
		0.75	0.7279	0.8132	0.7441	0.7799	0.7501	0.7670
-0.5	-5.7363	0.25	0.1504	0.2629	0.1814	**0.2289**	0.2105	**0.2348**
		0.50	0.4174	0.5994	0.4603	0.5408	0.4810	0.5215
		0.75	**0.7529**	0.8626	**0.7589**	0.8058	**0.7654**	0.7907
-0.75	-14.1385	0.25	0.0467	**0.1852**	0.1033	0.1732	0.1445	**0.1863**
		0.50	0.3415	0.6675	0.4330	0.5827	0.4808	0.5574
		0.75	**0.8210**	0.9529	**0.8360**	0.9044	**0.8385**	0.8761

Table 3.10: Simulated data from Frank copula; normal reference rule-of-thumb; fixed bandwidth



(a) $n = 20$ (b) $n = 50$ (c) $n = 100$

Figure 3.10: 3D-plot of probabilities h_{ij} for Frank simulated data with normal reference rule-of-thumb, fixed bandwidth for $\tau = -0.75$

τ	θ	$n = 20$		$n = 50$		$n = 100$	
		b_X	b_Y	b_X	b_Y	b_X	b_Y
0.75	14.1385	0.0842	0.0843	0.0619	0.0620	0.0497	0.0497
0.5	5.7363	0.1422	0.1422	0.1030	0.1034	0.0815	0.0815
0.25	2.3719	0.2034	0.2035	0.1439	0.1437	0.1108	0.1106
-0.25	-2.3719	0.3085	0.3095	0.2010	0.2011	0.1464	0.1460
-0.5	-5.7363	0.2921	0.2925	0.1718	0.1716	0.1212	0.1213
-0.75	-14.1385	0.1985	0.1989	0.1096	0.1093	0.0768	0.0769

Table 3.11: Bandwidth for simulated data from Frank copula; normal reference rule-of-thumb; fixed bandwidth

Tables 3.12 and 3.14 show the predictive performance of the proposed method, using LSCV bandwidth selection with fixed bandwidth for simulated data from the Clayton and Frank copulas, respectively. The corresponding bandwidths are shown in Tables 3.13 and 3.15, respectively. Tables 3.12 and 3.14 show that there are quite many cases where q is not in the interval p_1 and p_2, mostly at $q = 0.25$ and $q = 0.75$ for negative τ. This happened due to the probabilities h_{ij} obtained, using the bandwidth selection and type of bandwidth used for these tables, are different. This can be shown by 3D-plots in Figures 3.11 and 3.12 for data simulated from the Clayton and Frank copulas, respectively. Tables 3.13 and 3.15 show that the average bandwidths for b_x and b_y are quite different because the bandwidth selection method that we used for these tables are LSCV, where the bandwidth is selected based on the smallest integrated squared error as discussed in Section 3.2. This reflects the trade-off between the bias of the estimator and its variance.

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	6.0000	0.25	0.2339	0.2846	0.2438	0.2653	0.2401	**0.2499**
		0.50	0.4744	0.5244	0.4984	0.5171	0.4944	0.5060
		0.75	0.7236	0.7698	0.7422	0.7621	0.7487	0.7572
0.5	2.0000	0.25	0.2357	0.2938	**0.2505**	0.2799	0.2450	0.2561
		0.50	0.4845	0.5375	**0.5002**	0.5222	0.4935	0.5073
		0.75	0.7231	0.7693	0.7462	0.7659	0.7411	0.7507
0.25	0.6667	0.25	0.2370	0.3017	0.2449	0.2687	0.2472	0.2604
		0.50	0.4811	0.5527	0.4931	0.5223	0.4981	0.5120
		0.75	0.7201	0.7705	0.7366	0.7562	0.7413	0.7525
-0.25	-0.6667	0.25	0.1786	0.2549	0.2066	**0.2426**	0.2218	**0.2399**
		0.50	0.4428	0.5702	0.4769	0.5379	0.4820	0.5137
		0.75	0.7487	0.8239	**0.7583**	0.7886	**0.7568**	0.7743
-0.5	-2.0000	0.25	0.1193	**0.2087**	0.1708	**0.2132**	0.2130	**0.2346**
		0.50	0.4065	0.6071	0.4551	0.5584	0.4946	0.5451
		0.75	**0.7929**	0.8818	**0.7895**	0.8292	**0.7867**	0.8072
-0.75	-6.0000	0.25	0.0515	**0.1607**	0.1203	**0.1805**	0.1615	**0.1998**
		0.50	0.3288	0.6952	0.4134	0.6149	0.4564	0.5758
		0.75	**0.8485**	0.9466	**0.8281**	0.8859	**0.8101**	0.8451

Table 3.12: Simulated data from Clayton copula; LSCV; fixed bandwidth

τ	θ	$n = 20$		$n = 50$		$n = 100$	
		b_X	b_Y	b_X	b_Y	b_X	b_Y
0.75	6.0000	0.0911	0.0912	0.0626	0.0626	0.0383	0.0383
0.5	2.0000	0.1635	0.1631	0.0953	0.0962	0.0750	0.0745
0.25	0.6667	0.2297	0.2286	0.1433	0.1410	0.1094	0.1090
-0.25	-0.6667	0.3432	0.2688	0.2134	0.1492	0.1549	0.1055
-0.5	-2.0000	0.3398	0.2317	0.1909	0.1106	0.1307	0.0751
-0.75	-6.0000	0.2376	0.1585	0.1129	0.0702	0.0735	0.0476

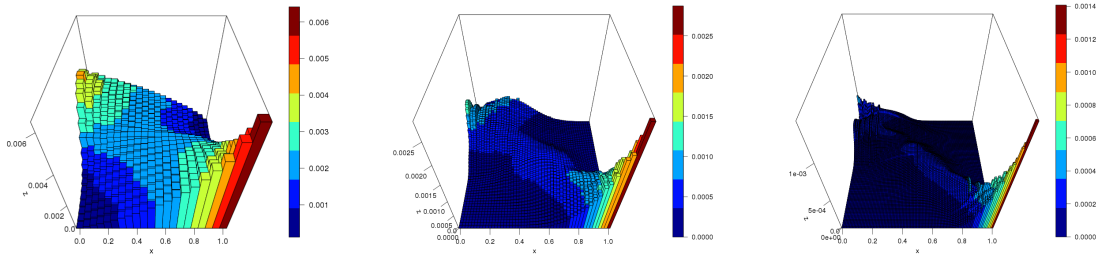
Table 3.13: Bandwidth for simulated data from Clayton copula; LSCV; fixed bandwidth

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	14.1385	0.25	0.2328	0.2814	0.2452	0.2647	0.2479	0.2585
		0.50	0.4753	0.5175	0.4935	0.5128	0.4959	0.5069
		0.75	0.7275	0.7767	0.7464	0.7659	0.7447	0.7548
0.5	5.7363	0.25	0.2383	0.2913	0.2421	0.2645	0.2464	0.2566
		0.50	0.4745	0.5257	0.4884	0.5112	0.4873	**0.4973**
		0.75	0.7159	0.7685	0.7399	0.7628	0.7376	**0.7482**
0.25	2.3719	0.25	0.2361	0.2957	0.2437	0.2682	0.2497	0.2610
		0.50	0.4727	0.5403	0.4851	0.5118	0.4960	0.5110
		0.75	0.7142	0.7726	0.7377	0.7625	0.7427	0.7537
-0.25	-2.3719	0.25	0.1784	0.2542	0.2182	0.2506	0.2321	0.2509
		0.50	0.4503	0.5766	0.4804	0.5272	**0.5019**	0.5271
		0.75	**0.7611**	0.8349	**0.7586**	0.7907	**0.7615**	0.7772
-0.5	-5.7363	0.25	0.1122	**0.2086**	0.1699	**0.2166**	0.1923	**0.2212**
		0.50	0.4098	0.5990	0.4649	0.5489	0.4763	0.5188
		0.75	**0.8023**	0.8925	**0.7864**	0.8318	**0.7745**	0.7976
-0.75	-14.1385	0.25	0.0424	**0.1592**	0.1164	**0.1941**	0.1619	**0.2098**
		0.50	0.3476	0.6672	0.4407	0.5908	0.4723	0.5531
		0.75	**0.8518**	0.9629	**0.8208**	0.8937	**0.8072**	0.8484

Table 3.14: Simulated data from Frank copula; LSCV; fixed bandwidth

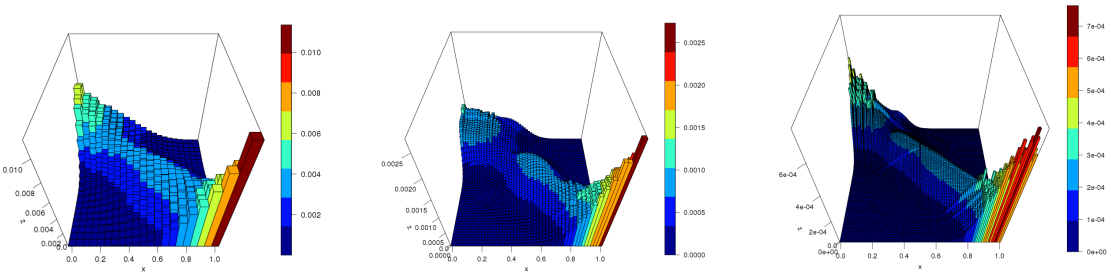
τ	θ	$n = 10$		$n = 50$		$n = 100$	
		b_X	b_Y	b_X	b_Y	b_X	b_Y
0.75	14.1385	0.0836	0.0837	0.0620	0.0620	0.0498	0.0497
0.5	5.7363	0.1423	0.1424	0.1031	0.1030	0.0817	0.0817
0.25	2.3719	0.2032	0.2035	0.1433	0.1435	0.1108	0.1106
-0.25	-2.3719	0.3094	0.3099	0.2006	0.2005	0.1463	0.1461
-0.5	-5.7363	0.2922	0.2929	0.1718	0.1719	0.1212	0.1211
-0.75	-14.1385	0.1982	0.1987	0.1096	0.1097	0.0769	0.0769

Table 3.15: Bandwidth for simulated data from Frank copula; LSCV; fixed bandwidth



(a) $n = 20$ (b) $n = 50$ (c) $n = 100$

Figure 3.11: 3D-plot of probabilities h_{ij} for Clayton simulated data with LSCV, fixed bandwidth for $\tau = -0.75$



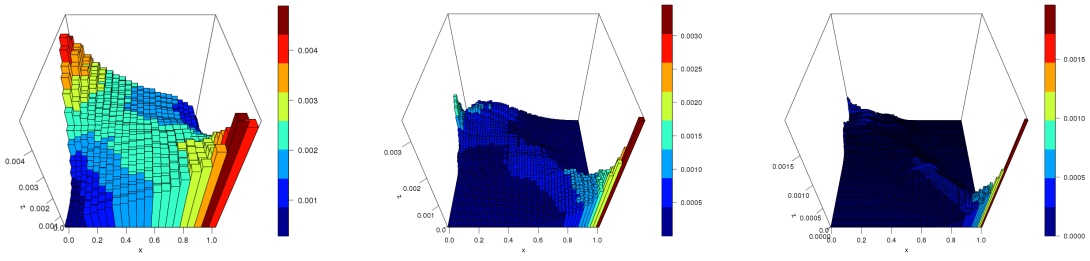
(a) $n = 20$ (b) $n = 50$ (c) $n = 100$

Figure 3.12: 3D-plot of probabilities h_{ij} for Frank simulated data with LSCV, fixed bandwidth for $\tau = -0.75$

Tables 3.16 and 3.18 show the predictive performance of the proposed method, using LSCV bandwidth selection with adaptive-nn bandwidth for simulated data from the Clayton and Frank copulas, respectively. These tables show that the results are similar with previous methods where there are many q not in the interval $[p_1, p_2]$ mostly for $q = 0.25$ and $q = 0.75$, for negative correlation. The 3D-plots of the probabilities h_{ij} for the adaptive-nn bandwidth are shown in Figures 3.13 and 3.14 for the Clayton and Frank copulas, respectively. Tables 3.17 and 3.19 show the average value of k_z out of $10,000$ repetitions for the adaptive-nn bandwidth for the Clayton and Frank copulas, respectively. These tables show that k_z is decreasing as the strength of the correlation get stronger for both negative and positive correlations. This feature reflects the adaptive-nn bandwidth that we used, whereby the bandwidth obtained is based on the minimum distance between the estimation point and its k-th closest neighbour. Therefore, as the data has a strong correlation regardless the sign of the correlation, smallest k_z is used to estimate the density and vice versa. Another noticeable feature, the k_z values obtained for negative correlation are bigger than the k_z values obtained for positive correlation. This feature has occurred due to the characteristic of the simulated data and the role of k_z, which plays a similar role to the bandwidth as mentioned in Section 3.2.

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	6.0000	0.25	0.2252	0.2751	0.2458	0.2667	0.2485	0.2594
		0.50	0.4712	0.5162	0.4910	0.5103	0.4979	0.5083
		0.75	0.7271	0.7725	0.7347	0.7545	0.7426	0.7530
0.5	2.0000	0.25	0.2296	0.2901	**0.2507**	0.2791	**0.2606**	0.2742
		0.50	0.4888	0.5426	0.4970	0.5206	**0.5073**	0.5193
		0.75	0.7268	0.7693	0.7436	0.7605	**0.7512**	0.7593
0.25	0.6667	0.25	0.2317	0.2973	**0.2532**	0.2796	**0.2600**	0.2740
		0.50	0.4797	0.5483	0.4997	0.5287	**0.5054**	0.5189
		0.75	0.7170	0.7685	0.7349	0.7556	0.7418	0.7508
-0.25	-0.6667	0.25	0.1642	**0.2397**	0.2054	**0.2335**	0.2259	**0.2439**
		0.50	0.4321	0.5566	0.4759	0.5270	0.4903	0.5223
		0.75	0.7462	0.8220	**0.7621**	0.7891	**0.7619**	0.7778
-0.5	-2.0000	0.25	0.1118	**0.1960**	0.1703	**0.2126**	0.2034	**0.2248**
		0.50	0.4017	0.6054	0.4584	0.5565	0.4752	0.5302
		0.75	**0.8029**	0.8854	**0.7946**	0.8370	**0.7811**	0.8042
-0.75	-6.0000	0.25	0.0503	**0.1498**	0.1226	**0.1875**	0.1722	**0.2107**
		0.50	0.3361	0.7001	0.4029	0.5985	0.4572	0.5759
		0.75	**0.8719**	0.9573	**0.8102**	0.8765	**0.8072**	0.8458

Table 3.16: Simulated data from Clayton copula; LSCV; adaptive-nn bandwidth



(a) $n = 20$ (b) $n = 50$ (c) $n = 100$

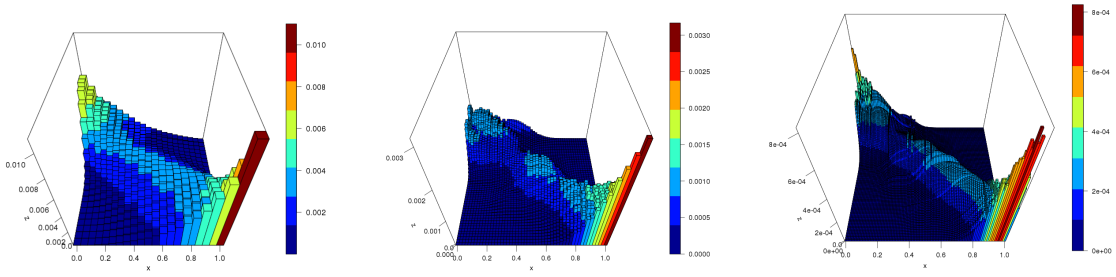
Figure 3.13: 3D-plot of probabilities h_{ij} for Clayton simulated data with LSCV, adaptive-nn bandwidth for $\tau = -0.75$

τ	θ	$n = 20$		$n = 50$		$n = 100$	
		k_x	k_y	k_x	k_y	k_x	k_y
0.75	6.0000	1.8961	2.0165	4.4392	4.5173	6.7678	6.6949
0.50	2.0000	4.0297	4.0229	8.8033	8.7824	13.3366	13.2996
0.25	0.6667	6.1116	6.1232	13.7028	13.5721	20.5234	20.4914
-0.25	-0.6667	9.5598	7.6125	20.5537	14.9094	27.1986	18.5107
-0.50	-2.0000	9.6201	6.6610	15.2807	9.3460	19.8181	12.7705
-0.75	-6.0000	6.9047	4.4894	8.5346	6.1043	11.4056	7.9370

Table 3.17: Bandwidth for simulated data from Clayton copula; LSCV; adaptive-nn bandwidth

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	14.1385	0.25	0.2286	0.2794	0.2405	0.2616	0.2467	0.2575
		0.50	0.4814	0.5268	0.4924	0.5107	0.4976	0.5090
		0.75	0.7247	0.7735	0.7382	0.7605	**0.7520**	0.7641
0.5	5.7363	0.25	0.2316	0.2784	0.2495	0.2726	**0.2543**	0.2649
		0.50	0.4739	0.5246	0.4922	0.5142	0.4940	0.5030
		0.75	0.7143	0.7697	0.7373	0.7580	0.7404	0.7522
0.25	2.3719	0.25	0.2390	0.2952	0.2482	0.2718	**0.2572**	0.2695
		0.50	0.4690	0.5337	0.4815	0.5112	**0.5034**	0.5183
		0.75	0.7035	0.7623	0.7265	0.7509	0.7489	0.7613
-0.25	-2.3719	0.25	0.1738	0.2510	0.2011	**0.2356**	0.2260	**0.2446**
		0.50	0.4406	0.5621	0.4776	0.5301	0.4913	0.5214
		0.75	**0.7532**	0.8358	**0.7632**	0.7962	**0.7663**	0.7858
-0.5	-5.7363	0.25	0.1057	**0.1976**	0.1681	**0.2138**	0.2131	**0.2378**
		0.50	0.4159	0.5999	0.4627	0.5455	0.4867	0.5278
		0.75	**0.8091**	0.8957	**0.7871**	0.8327	**0.7813**	0.8056
-0.75	-14.1385	0.25	0.0384	**0.1455**	0.1137	**0.1848**	0.1695	**0.2153**
		0.50	0.3513	0.6676	0.4269	0.5717	0.4746	0.5564
		0.75	**0.8665**	0.9670	**0.8107**	0.8879	**0.8101**	0.8529

Table 3.18: Simulated data from Frank copula; LSCV; adaptive-nn bandwidth



(a) $n = 20$ (b) $n = 50$ (c) $n = 100$

Figure 3.14: 3D-plot of probabilities h_{ij} for Frank simulated data with LSCV, adaptive-nn bandwidth for $\tau = -0.75$

τ	θ	$n = 20$		$n = 50$		$n = 100$	
		k_x	k_y	k_x	k_y	k_x	k_y
0.75	14.1385	1.9902	2.0998	4.5753	4.6184	7.4318	7.3959
0.5	5.7363	3.5318	3.5522	8.1875	8.1506	12.8683	12.8383
0.25	2.3719	5.4412	5.4190	12.0616	11.9634	18.2404	18.2444
-0.25	-2.3719	8.7456	8.7039	16.0160	15.9642	22.5893	22.5785
-0.5	-5.7363	8.3683	8.3336	12.4967	12.4804	17.1447	17.1081
-0.75	-14.1385	5.7864	5.7522	7.7915	7.8060	11.1956	11.1314

Table 3.19: Bandwidth for simulated data from Frank copula; LSCV; adaptive-nn bandwidth

Generally, from this simulation study, the proposed method seems to perform well ($q \in [p_1, p_2]$) for positive correlation regardless (mostly) of sample size, bandwidth selections and types of bandwidths. In the cases where q is not in the interval between p_1 and p_2, the q is quite close to p_1 or p_2. However, for negative correlation the proposed method does not perform so well, especially for strong negative correlation at quantiles q equal to 0.25 and 0.75. As discussed in Section 2.5, due to the fact that we are considering the events $T_{n+1} = X_{n+1} + Y_{n+1} > t$, and can be explained by considering the probabilities $h_{ij}(\hat{c})$ which are the key ingredients of our method for inference, the imprecision $p_2 - p_1$ is always greater for negative correlation than for positive correlation, and this effect is stronger for larger absolute values of the correlation.

Although the predictive performance of the proposed method is not good for negative correlation (especially strong negative correlation), a perhaps somewhat less expected feature of our method is seen when the sample size increases, which leads to the values of p_1 and p_2 to decrease or increase, respectively. This feature shows that the method might work well for negative correlation if we used small bandwidth for $n = 100$. We show this in next section, using the adaptive-nn bandwidth, we consider small values of k_z for $n = 100$. As mentioned in Section 3.2, the bandwidth, b controls how wide the probability mass is spreading and controls the smoothness and roughness of a density estimate. From this simulation study, we can see that the bandwidth decreases as n increases.

3.4.2 Manually selecting bandwidth

In order to investigate suitable bandwidths for prediction, we performed a simulation study with different values of k_z for the adaptive-nn method. As mentioned in Section 3.2, k_z is a smoothing parameter for the kernel estimate which controls the bandwidth values, and it therefore also controls the spread of probability mass around the observed data and the smoothness of the probabilities $h_{ij}(\hat{c})$.

We have run $N = 10,000$ simulations for $n = 20, 50, 100$ from the Normal, Frank, Clayton and Gumbel copulas, with $\tau = -0.75, -0.50, -0.25, 0.25, 0.50, 0.75$ and $q = 0.25, 0.50, 0.75$. For this simulation study we used the adaptive-nn bandwidth with Gaussian kernel for different values of $k_z = 1, 2, 3, 4$. The corresponding bandwidths for these k_z can be calculated by using formula given in Section 3.2. In this chapter, we show results of the simulation study for data simulated from the Clayton copula and some from Frank copula. We repeated the simulation study for the Normal and Gumbel copulas, the results obtained leads to the same conclusion as Clayton and Frank copulas.

Tables 3.20 - 3.23 show the results of the predictive performance of the proposed method for simulated data from the Clayton copula. These tables show that, for each value of k_z, there is at least one scenario for which q is not contained in $[p_1, p_2]$. As k_z increases, there are more scenarios for which q is not contained in $[p_1, p_2]$, especially for strong negative correlation. For $k_z = 1$, $k_z = 2$ and $k_z = 3$, the values

of q are mostly in the intervals $[p_1, p_2]$, even if they are not in the interval, they close to the intervals $[p_1, p_2]$. However, from Table 3.23, for $k_z = 4$, we can see that for $\tau = -0.5$ and $\tau = -0.75$, the values of q are not in the intervals $[p_1, p_2]$, especially for $q = 0.25$ and $q = 0.75$. This feature due to the characteristic of the Clayton copula discussed in Section 3.4.1. As the value of k_z increases, and n increases, the conditions for the probabilities h_{ij} mentioned in Section 2.3 are dissatisfied. This can be shown by 3D-plot of the probabilities h_{ij} for different k_z and sample sizes given in Figure 3.15. This figure show that the probabilities h_{ij} decreases as k_z and n increases, which shows the LSCV bandwidth selection with adaptive-nn bandwidth over-smooth the probabilities h_{ij}. In addition, as we interested on the sum events, calculating the lower and upper probabilities in equations (2.5) and (2.6) tend to include several more $h_{ij}(\hat{c})$ values in the latter than in the former, and for events $T_{n+1} > t$ these extra $h_{ij}(\hat{c})$ included in the upper probability tend to have the sum of their subscripts i and j about constant as explained in detail in Section 2.5. Hence, for positive correlation these extra $h_{ij}(\hat{c})$ tend to include few larger values for most values of t compared to negative correlation.

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	6.0000	0.25	0.2256	0.2733	0.2401	0.2627	0.2423	0.2536
		0.5	0.4714	0.5187	0.4956	0.5167	0.5000	0.5087
		0.75	0.7151	0.7611	0.7379	0.7580	**0.7506**	0.7597
0.5	2.0000	0.25	0.2231	0.2851	0.2433	0.2673	0.2494	0.2642
		0.5	0.4686	0.5206	0.4910	0.5097	0.4967	0.5066
		0.75	0.7229	0.7672	0.7465	0.7661	0.7446	0.7543
0.25	0.6667	0.25	0.2142	0.2782	0.2314	0.2570	0.2400	0.2554
		0.5	0.4599	0.5308	0.4862	0.5141	0.4951	0.5085
		0.75	0.7145	0.7686	0.7371	0.7567	**0.7523**	0.7620
-0.25	-0.6667	0.25	0.2047	0.2913	0.2377	0.2691	0.2438	0.2599
		0.5	0.4352	0.5614	0.4777	0.5331	0.4946	0.5235
		0.75	0.6999	0.7892	0.7340	0.7714	0.7474	0.7634
-0.5	-2.0000	0.25	0.1888	0.3088	0.2279	0.2783	0.2391	0.2660
		0.5	0.4065	0.5983	0.4573	0.5555	0.4775	0.5303
		0.75	0.6970	0.8157	0.7227	0.7733	0.7310	0.7567
-0.75	-6.0000	0.25	0.1223	0.3370	0.1894	0.2931	0.2246	0.2783
		0.5	0.3213	0.6914	0.4047	0.6101	0.4486	0.5699
		0.75	0.6784	0.8835	0.7172	0.8177	0.7332	0.7852

Table 3.20: Simulated data from Clayton copula; adaptive-nn; $k_z = 1$

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	6.0000	0.25	0.2171	0.2701	0.2427	0.2672	**0.2539**	0.2640
		0.5	0.4747	0.5267	0.4980	0.5179	0.4972	0.5076
		0.75	0.7221	0.7666	0.7469	0.7640	0.7413	0.7515
0.5	2.0000	0.25	0.2225	0.2786	0.2328	0.2600	0.2406	0.2523
		0.5	0.4706	0.5273	0.4848	0.5069	0.4935	0.5079
		0.75	0.7182	0.7640	0.7424	0.7591	**0.7538**	0.7634
0.25	0.6667	0.25	0.2157	0.2782	0.2363	0.2632	0.2457	0.2584
		0.5	0.4685	0.5397	0.4835	0.5155	0.4968	0.5137
		0.75	0.7261	0.7782	0.7366	0.7610	**0.7505**	0.7606
-0.25	-0.6667	0.25	0.2117	0.2917	0.2329	0.2698	0.2401	0.2588
		0.5	0.4365	0.5670	0.4653	0.5216	0.4794	0.5100
		0.75	0.7129	0.7983	0.7305	0.7693	0.7406	0.7593
-0.5	-2.0000	0.25	0.1757	0.2922	0.2232	0.2779	0.2385	0.2636
		0.5	0.3959	0.5982	0.4577	0.5587	0.4783	0.5347
		0.75	0.7112	0.8311	0.7314	0.7837	0.7408	0.7676
-0.75	-6.0000	0.25	0.0991	0.2843	0.1844	0.2828	0.2221	0.2744
		0.5	0.3096	0.6752	0.4121	0.6120	0.4521	0.5730
		0.75	0.7093	0.8974	0.7311	0.8257	0.7440	0.7889

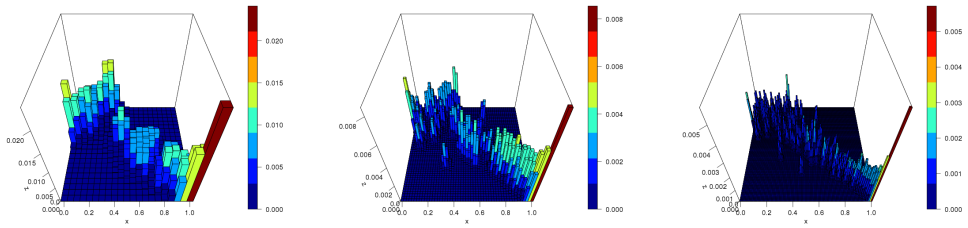
Table 3.21: Simulated data from Clayton copula; adaptive-nn; $k_z = 2$

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	6.0000	0.25	0.2365	0.2890	0.2397	0.2620	0.2431	0.2558
		0.5	0.4808	0.5263	0.4939	0.5138	0.4953	0.5062
		0.75	0.7199	0.7681	0.7407	0.7591	0.7426	0.7522
0.5	2.0000	0.25	0.2236	0.2859	0.2352	0.2620	0.2466	0.2606
		0.5	0.4753	0.5284	0.4852	0.5090	0.4916	0.5024
		0.75	0.7208	0.7683	0.7398	0.7590	0.7474	0.7565
0.25	0.6667	0.25	0.2246	0.2899	0.2303	0.2580	0.2452	0.2571
		0.5	0.4774	0.5438	0.4891	0.5193	0.4926	0.5090
		0.75	0.7223	0.7766	0.7426	0.7644	0.7425	0.7536
-0.25	-0.6667	0.25	0.1953	0.2720	0.2321	0.2668	0.2448	0.2598
		0.5	0.4245	0.5568	0.4722	0.5303	0.4877	0.5166
		0.75	0.7161	0.7986	0.7301	0.7620	0.7452	0.7630
-0.5	-2.0000	0.25	0.1547	0.2645	0.2130	0.2638	0.2403	0.2675
		0.5	0.3897	0.5925	0.4513	0.5489	0.4888	0.5435
		0.75	0.7249	0.8351	0.7283	0.7794	0.7488	0.7710
-0.75	-6.0000	0.25	0.0905	**0.2351**	0.1728	0.2646	0.2061	0.2530
		0.5	0.3256	0.6820	0.4110	0.6108	0.4509	0.5676
		0.75	**0.7723**	0.9183	**0.7516**	0.8363	0.7459	0.7970

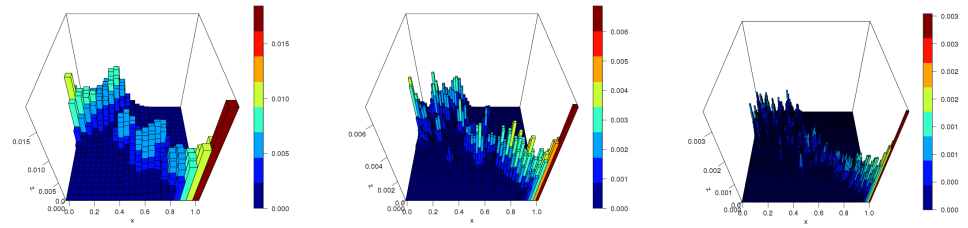
Table 3.22: Simulated data from Clayton copula; adaptive-nn; $k_z = 3$

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	6.0000	0.25	0.2376	0.2911	0.2460	0.2669	0.2415	0.2516
		0.5	0.4800	0.5313	0.4991	0.5198	0.4979	0.5082
		0.75	0.7193	0.7616	0.7455	0.7634	0.7418	0.7526
0.5	2.0000	0.25	0.2386	0.2961	0.2469	0.2759	0.2481	0.2607
		0.5	0.4797	0.5328	0.4965	0.5193	**0.5010**	0.5118
		0.75	0.7147	0.7610	0.7441	0.7601	0.7455	0.7555
0.25	0.6667	0.25	0.2201	0.2852	0.2391	0.2660	0.2445	0.2585
		0.5	0.4712	0.5425	0.4839	0.5146	0.4959	0.5099
		0.75	0.7181	0.7690	0.7397	0.7583	0.7472	0.7569
-0.25	-0.6667	0.25	0.1905	0.2680	0.2375	0.2773	0.2380	0.2549
		0.5	0.4378	0.5651	0.4870	0.5396	0.4843	0.5142
		0.75	0.7310	0.8116	0.7472	0.7807	0.7458	0.7621
-0.5	-2.0000	0.25	0.1402	**0.2419**	0.2112	0.2597	0.2357	0.2602
		0.5	0.3925	0.5947	0.4602	0.5554	0.4784	0.5366
		0.75	0.7478	0.8503	0.7381	0.7853	0.7459	0.7716
-0.75	-6.0000	0.25	0.0661	**0.1866**	0.1598	**0.2370**	0.2094	0.2582
		0.5	0.3320	0.6946	0.4005	0.6002	0.4530	0.5701
		0.75	**0.8269**	0.9386	**0.7632**	0.8482	**0.7565**	0.8036

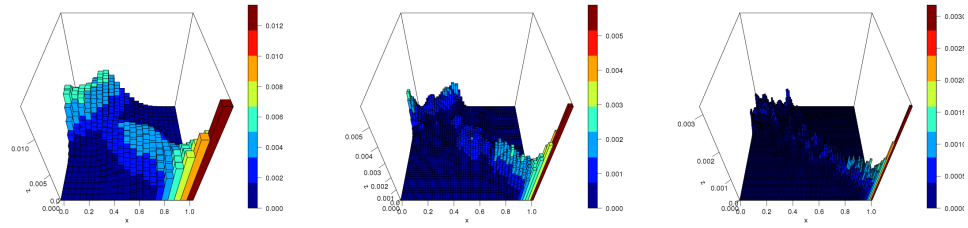
Table 3.23: Simulated data from Clayton copula; adaptive-nn; $k_z = 4$



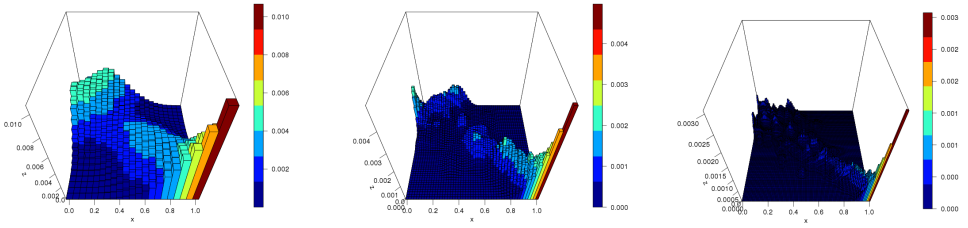
(a) $k_z = 1$, $n = 20$ (b) $k_z = 1$, $n = 50$ (c) $k_z = 1$, $n = 100$



(d) $k_z = 2$, $n = 20$ (e) $k_z = 2$, $n = 50$ (f) $k_z = 2$, $n = 100$



(g) $k_z = 3$, $n = 20$ (h) $k_z = 3$, $n = 50$ (i) $k_z = 3$, $n = 100$



(j) $k_z = 4$, $n = 20$ (k) $k_z = 4$, $n = 50$ (l) $k_z = 4$, $n = 100$

Figure 3.15: 3D-plot of probabilities h_{ij} for Clayton simulated data for $k_z = 1, 2, 3, 4$ and $\tau = -0.75$

For data simulated from the Frank copula, the performance of the method is shown in Tables 3.24 - 3.26, for $k_z = 1$, $k_z = 2$ and $k_z = 3$, respectively. For $k_z = 4$,

the results leads to similar conclusion as data simulated from the Clayton copula discussed above. We see that $q \in [p_1, p_2]$ for all repeated cases for $k_z = 1$ and $k_z = 2$. For $k_z = 3$, there are cases where q is not contained in intervals $[p_1, p_2]$, but the number of q not contained in the intervals $[p_1, p_2]$ are less than for the data simulated from the Clayton copula. Another noticeable feature from these tables is that the predictive performance of the proposed method (most cases) works well, with data simulated from the Frank copula compared to data simulated from the Clayton copula. This happened because of the characteristic of the copula itself. The 3D-plot of the probabilities h_{ij} for Frank copula is given in Figure 3.16. Figure 3.16 shows that the probabilities h_{ij} are symmetrically distributed due to the fact that we simulate data from the Frank copula. This figure shows that the probabilities h_{ij} decreases as k_z and n increases, which again, shows the LSCV bandwidth selection with adaptive-nn bandwidth over-smooth the probabilities h_{ij}.

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	14.1385	0.25	0.2229	0.2705	0.2427	0.2621	0.2413	0.2516
		0.5	0.4720	0.5224	0.4933	0.5145	0.4984	0.5074
		0.75	0.7251	0.7770	0.7478	0.7676	0.7466	0.7560
0.5	5.7363	0.25	0.2241	0.2787	0.2359	0.2592	0.2461	0.2576
		0.5	0.4766	0.5335	0.4884	0.5128	0.4975	0.5085
		0.75	0.7268	0.7783	0.7356	0.7594	0.7448	0.7576
0.25	2.3719	0.25	0.2226	0.2804	0.2423	0.2674	0.2438	0.2565
		0.5	0.4717	0.5353	0.4895	0.5150	0.4908	0.5035
		0.75	0.7237	0.7827	0.7330	0.7578	0.7422	0.7567
-0.25	-2.3719	0.25	0.2047	0.2924	0.2347	0.2698	0.2411	0.2582
		0.5	0.4354	0.5585	0.4738	0.5228	0.4916	0.5191
		0.75	0.7123	0.8003	0.7302	0.7674	0.7392	0.7566
-0.5	-5.7363	0.25	0.1827	0.3162	0.2233	0.2814	0.2410	0.2690
		0.5	0.4099	0.5957	0.4684	0.5494	0.4854	0.5279
		0.75	0.6951	0.8276	0.7295	0.7837	0.7386	0.7667
-0.75	-14.1385	0.25	0.1157	0.3505	0.1907	0.3010	0.2174	0.2763
		0.5	0.3468	0.6697	0.4299	0.5859	0.4601	0.5441
		0.75	0.6667	0.8918	0.7046	0.8185	0.7315	0.7861

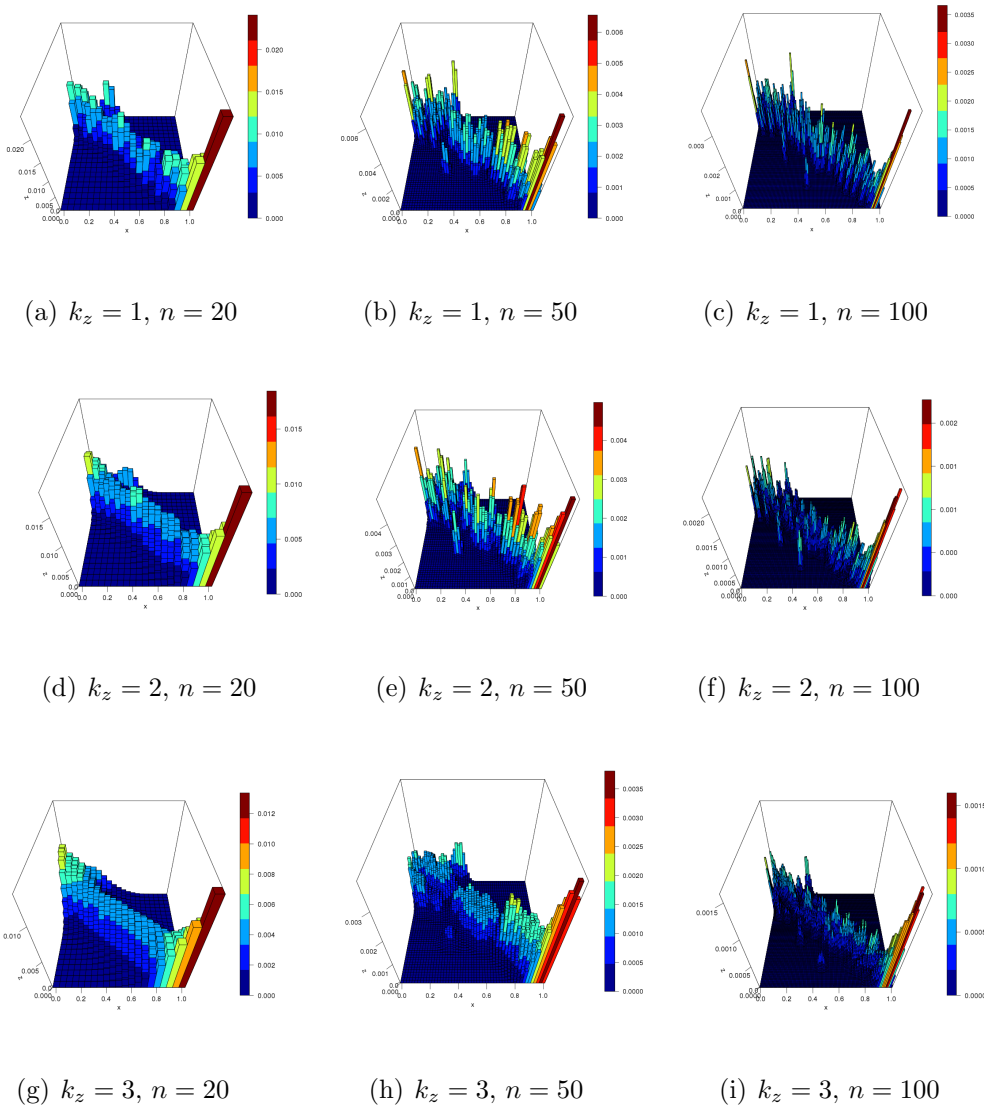
Table 3.24: Simulated data from Frank copula; adaptive-nn; $k_z = 1$

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	14.1385	0.25	0.2264	0.2717	0.2389	0.2593	0.2440	0.2543
		0.5	0.4836	0.5296	0.4919	0.5130	0.4937	0.5042
		0.75	0.7232	0.7754	0.7422	0.7642	0.7421	0.7520
0.5	5.7363	0.25	0.2318	0.2788	0.2378	0.2586	0.2461	0.2555
		0.5	0.4746	0.5239	0.4864	0.5068	0.4939	0.5036
		0.75	0.7202	0.7722	0.7338	0.7552	0.7456	0.7577
0.25	2.3719	0.25	0.2288	0.2878	0.2464	0.2688	0.2485	0.2604
		0.5	0.4776	0.5420	0.4925	0.5208	0.4927	0.5067
		0.75	0.7199	0.7734	0.7392	0.7647	0.7437	0.7555
-0.25	-2.3719	0.25	0.2037	0.2902	0.2319	0.2709	0.2411	0.2577
		0.5	0.4436	0.5713	0.4730	0.5250	0.4890	0.5167
		0.75	0.7152	0.7992	0.7328	0.7726	0.7409	0.7603
-0.5	-5.7363	0.25	0.1706	0.2948	0.2222	0.2773	0.2431	0.2715
		0.5	0.4065	0.5896	0.4655	0.5473	0.4984	0.5370
		0.75	0.7054	0.8305	0.7334	0.7862	0.7487	0.7714
-0.75	-14.1385	0.25	0.0993	0.3055	0.1819	0.2921	0.2236	0.2801
		0.5	0.3367	0.6632	0.4224	0.5828	0.4618	0.5441
		0.75	0.6954	0.9030	0.7211	0.8189	0.7308	0.7843

Table 3.25: Simulated data from Frank copula; adaptive-nn; $k_z = 2$

τ	θ	q	$n = 20$		$n = 50$		$n = 100$	
			p_1	p_2	p_1	p_2	p_1	p_2
0.75	14.1385	0.25	0.2322	0.2825	0.2453	0.2664	0.2479	0.2571
		0.5	0.4864	0.5381	0.4929	0.5127	0.4972	0.5076
		0.75	0.7270	0.7750	0.7441	0.7637	0.7437	0.7536
0.5	5.7363	0.25	0.2392	0.2927	0.2353	0.2560	0.2359	**0.2477**
		0.5	0.4817	0.5360	0.4915	0.5117	0.4924	0.5010
		0.75	0.7182	0.7714	0.7375	0.7588	0.7387	**0.7498**
0.25	2.3719	0.25	0.2206	0.2799	0.2452	0.2691	0.2479	0.2598
		0.5	0.4733	0.5323	0.4848	0.5119	0.4952	0.5097
		0.75	0.7145	0.7704	0.7357	0.7587	0.7466	0.7591
-0.25	-2.3719	0.25	0.1991	0.2782	0.2288	0.2663	0.2426	0.2621
		0.5	0.4288	0.5523	0.4806	0.5331	0.4843	0.5149
		0.75	0.7094	0.7979	0.7389	0.7739	0.7407	0.7602
-0.5	-5.7363	0.25	0.1617	0.2826	0.2171	0.2747	0.2392	0.2678
		0.5	0.4134	0.5905	0.4606	0.5395	0.4882	0.5311
		0.75	0.7219	0.8448	0.7276	0.7841	0.7450	0.7708
-0.75	-14.1385	0.25	0.0748	**0.2430**	0.1706	0.2705	0.2081	0.2610
		0.5	0.3389	0.6631	0.4231	0.5745	0.4592	0.5426
		0.75	**0.7582**	0.9316	0.7276	0.8341	0.7413	0.7962

Table 3.26: Simulated data from Frank copula; adaptive-nn; $k_z = 3$



(a) $k_z = 1$, $n = 20$ (b) $k_z = 1$, $n = 50$ (c) $k_z = 1$, $n = 100$

(d) $k_z = 2$, $n = 20$ (e) $k_z = 2$, $n = 50$ (f) $k_z = 2$, $n = 100$

(g) $k_z = 3$, $n = 20$ (h) $k_z = 3$, $n = 50$ (i) $k_z = 3$, $n = 100$

Figure 3.16: 3D-plot of probabilities h_{ij} for Frank simulated data for $k_z = 1, 2, 3$ and $\tau = -0.75$

From this simulation study, the proposed method can be considered to work quite well to giving $q \in [p_1, p_2]$, for positive and negative correlation, for the considered sample sizes. However, the performance of the proposed method depends on the k_z value and the distribution used for the simulation. As discussed above, the conditions for the probabilities h_{ij} mentioned in Section 2.3 are not met, as k_z increases or n increases. From this simulation study, it is suggested that the proposed method works best if we used $k_z = 1$, $k_z = 2$ or $k_z = 3$, regardless of sample size (for $n = 20, 50, 100$), the strength of the correlation and the copula families. The

proposed method also works well with $k_z = 4$ for $n \geq 50$, except for $\tau = -0.75$ in Table 3.23.

As known, k_z also controls the bandwidth for the adaptive-nn bandwidth. In this simulation study, we do not show the exact value of the bandwidth. However, as mentioned before in Section 3.2, the bandwidth decreases as k_z increases or n increases. From Section 3.4.1, the average bandwidths for data simulated from the Clayton copula are shown in Table 3.17 which give large values of k_z for most cases discussed. As we compare the average bandwidth obtained in Section 3.4.1 with the predictive performance in this section, the predictive performance of the proposed method work well when we have smaller values of k_z especially for the negative correlation.

In terms of imprecision, for corresponding cases with increasing n, the imprecision, reflected through the difference $p_2 - p_1$, decreases. This is logical from the perspective that more data allow more precise inferences, which is common in statistical methods using imprecise probabilities [2]. In addition, as the events of interest involve sum of the bivariate data, the imprecision of the proposed method is larger in case of negative correlation than for positive correlation, as discussed in Section 2.5.

Generally, based on simulation study in Sections 3.4.1 and 3.4.2, for the misspecified copula occurred in Chapter 2, the proposed method work quite well with kernel-based copula for larger n as discussed above. However, the performance of the proposed method depends on the bandwidth selections, types of bandwidths and the characteristic of the simulated data. The probabilities h_{ij} obtained rely on a trade-off between the value of k_z and the sample size. As our main interest is to use the nonparametric copula in order to solve the misspecified copula for larger sample size, further study is needed by using other types of nonparametric copulas in comparison to the results obtained in this chapter. One should also consider other types of dependence structures such as nonlinear. We left these as topics for future research.

3.5 Examples

In this section, we present two examples using the same data sets discussed in Section 2.6 in order to show the proposed method with kernel-based copula for real data sets.

3.5.1 Insurance example

Consider the insurance data set in example 2.6.1 with interest in the same event $T_{n+1} = X_{n+1} + Y_{n+1} > t$. We study the appropriate bandwidth to be used for these data in order to obtain good prediction for a future observation. Recall that in Section 3.3.2, we have used same data set in order to investigate which bandwidth selections and types of bandwidths to be used for analysing the predictive performance of the proposed method. However, based on simulation results in Section 3.4.2, we investigate more details which values of k_z to be used for prediction. We used $k_z = 1, 2, 3, 4$ as discussed in Section 3.4.2. The results are shown in Table 3.27 and the corresponding bandwidths are given in Table 3.28.

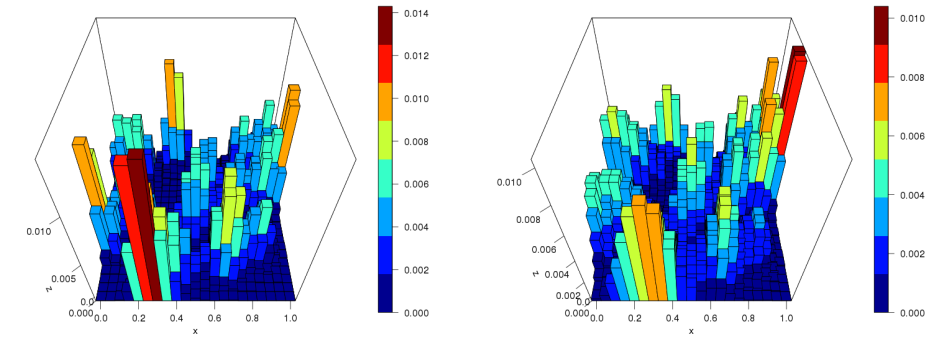
t in 1000s	$k_z = 1$		$k_z = 2$		$k_z = 3$		$k_z = 4$	
	$\underline{P}(T_{n+1} > t)$	$\overline{P}(T_{n+1} > t)$	$\underline{P}(T_{n+1} > t)$	$\overline{P}(T_{n+1} > t)$	$\underline{P}(T_{n+1} > t)$	$\overline{P}(T_{n+1} > t)$	$\underline{P}(T_{n+1} > t)$	$\overline{P}(T_{n+1} > t)$
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0.9089	0.9651	0.8993	0.9501	0.8956	0.9403	0.8914	0.9335
10	0.7084	0.7444	0.7158	0.7568	0.7286	0.7784	0.7366	0.7855
15	0.5605	0.6353	0.5661	0.6376	0.5917	0.6528	0.6053	0.6649
20	0.4998	0.5546	0.5016	0.5564	0.5130	0.5696	0.5230	0.5826
25	0.4536	0.5050	0.4567	0.5096	0.4527	0.5110	0.4564	0.5194
30	0.4157	0.4897	0.4167	0.4883	0.4098	0.4742	0.4083	0.4754
35	0.3340	0.4153	0.3367	0.4169	0.3363	0.4034	0.3345	0.3995
40	0.2953	0.3641	0.2944	0.3656	0.2925	0.3585	0.2899	0.3545
45	0.2702	0.3342	0.2681	0.3348	0.2650	0.3318	0.2638	0.3302
50	0.2637	0.3310	0.2602	0.3306	0.2491	0.3188	0.2433	0.3125
55	0.2006	0.2679	0.1997	0.2691	0.2092	0.2737	0.2093	0.2705
60	0.1944	0.2612	0.1909	0.2579	0.1934	0.2588	0.1924	0.2546
65	0.1846	0.2447	0.1804	0.2398	0.1749	0.2402	0.1728	0.2367
70	0.1395	0.2006	0.1371	0.2011	0.1430	0.2106	0.1458	0.2109
75	0.1311	0.1829	0.1278	0.1795	0.1323	0.1899	0.1342	0.1922
80	0.1203	0.1593	0.1171	0.1570	0.1150	0.1663	0.1152	0.1696
85	0.1025	0.1378	0.0992	0.1351	0.0949	0.1444	0.0947	0.1484

Table 3.27: NPI lower and upper probabilities; different values of k_z

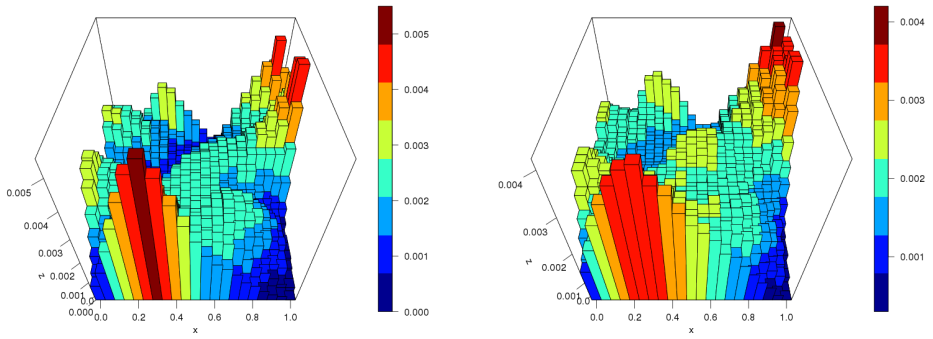
k_z	b for Loss	b for ALAE
1	0.1332	0.1332
2	0.2664	0.2664
3	0.3996	0.3996
4	0.5329	0.5329

Table 3.28: Bandwidth for Loss, x and ALAE, y

Table 3.27 shows the NPI lower and upper probabilities for the event $T_{n+1} > t$ for different values of k_z. There are many ways to explain the lower and upper probabilities obtained depending on the actual questions of interest. This table shows that the value of NPI lower and upper probabilities for the event $T_{n+1} > t$ are different at each t among the values of k_z considered. It is quite difficult to see the differences between the k_z from this table. However, these NPI lower and upper probabilities have been shown in Figure 3.7 in Section 3.3.2, which show that the NPI lower and upper probabilities become smooth as k_z increases. As mentioned in Section 3.4.2, as the k_z increases, the adaptive-nn bandwidth will not satisfy the conditions discussed in Section 2.3 where the sum of probabilities h_{ij} are not equal to 1.



(a) $k_x = k_y = 1$ (b) $k_x = k_y = 2$



(c) $k_x = k_y = 3$ (d) $k_x = k_y = 4$

Figure 3.17: 3D-plot of probabilities h_{ij} for adaptive-nn bandwidth with $k_z = 1, 2, 3, 4$ and $k_x = k_y$

Figure 3.17 shows 3D-plot of the probabilities h_{ij} for $k_z = 1, 2, 3, 4$. As seen in Figure 3.17, the probabilities h_{ij} are different for each value of k_z, which reflects the imprecision (the difference between the NPI upper and lower probabilities). Given smaller values of k_z, the probabilities h_{ij} are higher near the observation data. Figure 3.18 shows that the imprecision for $k_z = 1$ and $k_z = 2$ is not consistent (fluctuate up and down) for different values of t. While for $k_z = 3$ and $k_z = 4$, the imprecision is quite consistent with different values of t but it is not that much differences compared to $k_z = 1$ and $k_z = 2$. As discussed in Sections 2.6.1 and 3.4.2, this happens due to the sum event considered, which the imprecision is pretty similar through the main range of empirical distribution of the values $x_i + y_i$ due to positive correlation

between Loss and ALEA combined with interest in the sum of these quantities.

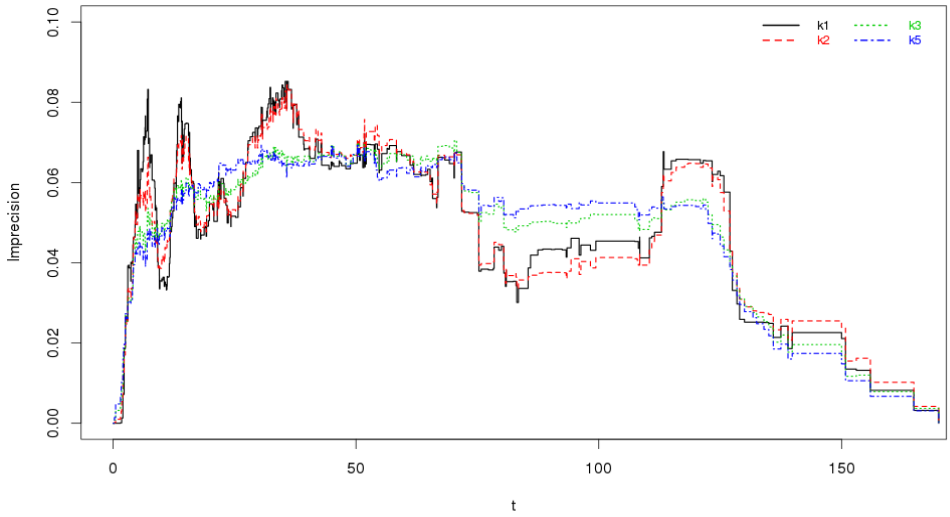


Figure 3.18: Imprecision for different values of k_z

In Table 3.28, the bandwidths increase as the values of k_z increase. This feature occurs due to the adaptive-nn bandwidth applied which it consider the kth nearest neighbour to estimate the density. The bandwidth of this data set is identical for loss and ALAE, due to same value of k_z used for both variables and the transform data used in estimating the density.

3.5.2 Body-Mass Index example

Consider the same data set and event of interest for the Body-Mass Index (BMI) as in example 2.6.2 in Section 2.6. Suppose that we are interested in the event that the next 11 year old girl has healthy weight, so, the event of interest $E(X_{n+1}, Y_{n+1})$ is that $\text{BMI}(X_{n+1}, Y_{n+1}) \in [14.08, 19.50)$. This example is different from the sum event that we have in example 3.5.1.

The lower and upper probabilities that resulting from our method for underweight, healthy weight, overweight and obese categories, using equations (2.3) and (2.4) in Section 2.4, are given in Table 3.29. For this table, we use the same algorithm from R package `np` [49] used in Section 3.4.1 to compute the bandwidths for height and weight. Table 3.29 shows the lower and upper probabilities for all

bandwidth selections and types of bandwidths considered in this example. The corresponding bandwidths are given in Table 3.30. For LSCV bandwidth selection with adaptive-nn bandwidth, the algorithm produced $k_h = 6$ and $k_w = 4$ for height and weight variables, respectively, and the corresponding bandwidths are given in Table 3.30.

	BMI$\in$	LSCV, Fixed bandwidth		Normal reference rule-of-thumb		LSCV, averaging adaptive-nn	
		$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
Underweight	[6.92,14.08)	0.0758	0.1486	0.0906	0.1607	0.0818	0.1504
Healthy weight	[14.08,19.50)	0.5910	0.7330	0.5745	0.7147	0.5892	0.7277
Overweight	[19.50,24.14)	0.1475	0.2519	0.1442	0.2512	0.1456	0.2519
Obese	[24.14,38.40)	0.0084	0.0437	0.0136	0.0505	0.0085	0.0449

Table 3.29: NPI lower and upper probabilities for different types of bandwidths

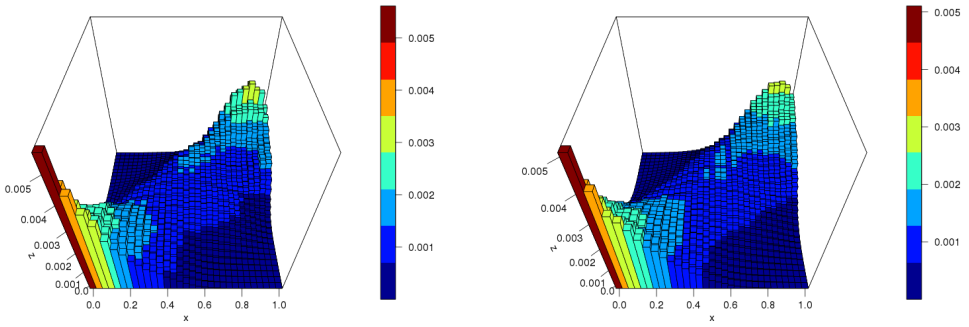
Types of bandwidths	b for heights	b for weights
LSCV, Fixed	0.1197	0.1073
Normal reference rule-of-thumb	0.1450	0.1450
LSCV, averaging adaptive-nn	0.7280	0.4854

Table 3.30: Bandwidth for height, h and weight, w

The 3D-plots in Figure 3.19 shows that the probabilities h_{ij} are quite similar for all methods applied, where the probabilities h_{ij} are higher at left-front corner of the 3D-plots compared to other corners. In addition, as the data set has a strong positive correlation, the probabilities h_{ij} are along the diagonal of the left-front corner to the right-back corner of the 3D-plots for all methods. However, the probabilities h_{ij} are quite different at certain values of (x_i, y_i). It should be emphasized that in this example we have different events from the sum event that we have above. Therefore, the direction (from left-front corner to right-back corner of the 3D-plots) of the probabilities h_{ij} to be included when calculating the lower and upper probabilities are different from example 3.5.1. The lower and upper probabilities obtained in Table 3.29 are reasonable and can be used in prediction. For example, from Table 3.29, the next eleven-year-old girl has healthy weight is at least 59.10% chances and at most 73.30% chances using the LSCV with fixed bandwidth.

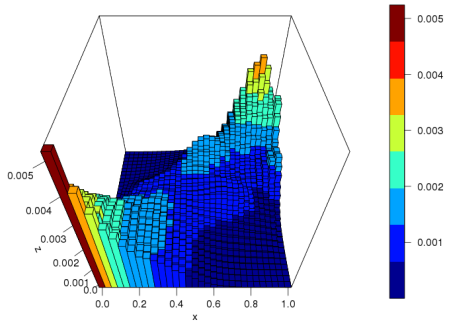
The bandwidth obtained are different among the methods used which can be

seen in Table 3.30, especially for adaptive-nn bandwidth. This occurred due to how the bandwidth selections and types of bandwidths chose the bandwidth, whereby the LSCV bandwidth selection is based on the minimum integrated squared error, and the adaptive-nn is based on the minimum distance from the observations to the nearest neighbour, as discussed in Section 3.2. We further investigate the adaptive-nn bandwidth for this example in order to study more details how the probabilities h_{ij} spread.



(a) LSCV and fixed bandwidth

(b) Normal reference rule-of-thumb



(c) LSCV and adaptive-nn bandwidth, $k_h = 6$ and $k_w = 4$

Figure 3.19: 3D-plot of probabilities h_{ij} for different bandwidth selections and types of bandwidths

Table 3.31 shows the lower and upper probabilities of the BMI event (for all categories) using different values of k_z using the adaptive-nn bandwidth. We used $k_z = 1, 2, 3, 4, 5$, and the corresponding bandwidth for height and weight are also given in Table 3.31. In this example, we consider equal value of k_z for heights and weights.

k_z	b		BMI$(X_{n+1}, Y_{n+1}) \in$	$\underline{P}$	$\overline{P}$	$\triangle(\cdot)$
1	0.1213	Underweight	$[6.92, 14.08)$	0.0290	0.1303	0.1012
		Healthy weight	$[14.08, 19.50)$	0.6031	0.7675	0.1644
		Overweight	$[19.50, 24.14)$	0.1685	0.2664	0.0979
		Obese	$[24.14, 38.40)$	0.0002	0.0349	0.0347
2	0.2427	Underweight	$[6.92, 14.08)$	0.0457	0.1303	0.0846
		Healthy weight	$[14.08, 19.50)$	0.6069	0.7596	0.1527
		Overweight	$[19.50, 24.14)$	0.1585	0.2611	0.1026
		Obese	$[24.14, 38.40)$	0.0017	0.0362	0.0346
3	0.3640	Underweight	$[6.92, 14.08)$	0.0578	0.1353	0.0775
		Healthy weight	$[14.08, 19.50)$	0.6044	0.7473	0.1429
		Overweight	$[19.50, 24.14)$	0.1570	0.2577	0.1006
		Obese	$[24.14, 38.40)$	0.0026	0.0379	0.0353
4	0.4854	Underweight	$[6.92, 14.08)$	0.0709	0.1423	0.0715
		Healthy weight	$[14.08, 19.50)$	0.6005	0.7394	0.1389
		Overweight	$[19.50, 24.14)$	0.1481	0.2523	0.1042
		Obese	$[24.14, 38.40)$	0.0049	0.0417	0.0368
5	0.6067	Underweight	$[6.92, 14.08)$	0.0782	0.1499	0.0717
		Healthy weight	$[14.08, 19.50)$	0.5908	0.7318	0.1411
		Overweight	$[19.50, 24.14)$	0.1450	0.2506	0.1056
		Obese	$[24.14, 38.40)$	0.0087	0.0449	0.0362

Table 3.31: NPI lower and upper probabilities for different values of k_z; adaptive-nn bandwidth

As in Section 2.6.2, we assume $x_0 = 1.25$, $x_{31} = 1.70$, $y_0 = 20$, $y_{31} = 60$, the minimum BMI index corresponding to $x_0 = 1.25$ and $y_0 = 20$ equal to 6.92, and the maximum BMI index corresponding to $x_{31} = 1.70$ and $y_{31} = 60$ is equal to 38.40. As mentioned in Section 2.6.2, choosing different values for x_0, x_{31}, y_0 and y_{31} will have an impact on the minimum and the maximum values of BMI, therefore will also affect on the lower and upper probabilities presented in Tables 3.29 and 3.31, but the impact is expected to be small. These assumed values might be based on

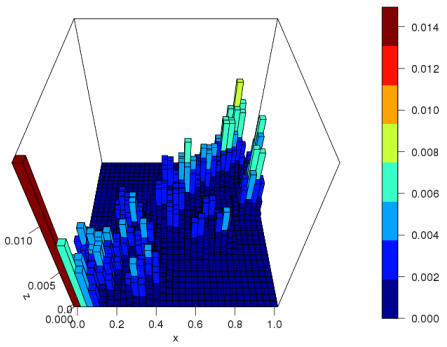
general information of the variables.

Table 3.31 shows that the lower and upper probabilities obtained for all categories are different, for different values of k_z. This feature occurred due to the different probabilities h_{ij} obtained for different values of k_z. This can be shown by using 3D-plots of the probabilities h_{ij} in Figure 3.20. As mentioned in Section 3.4.2, as k_z increases, the probabilities h_{ij} decrease, whereby the adaptive-nn bandwidth method over-smooth the probabilities h_{ij}. Therefore, this feature is affecting the lower and upper probabilities obtained in Table 3.31.

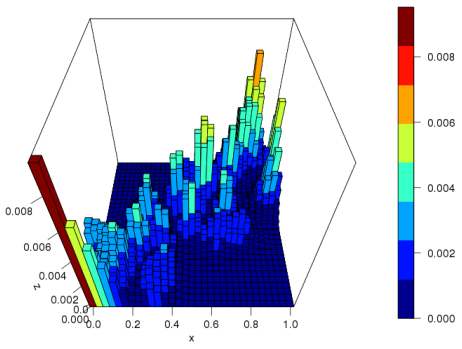
Based on the analysis in this example, by considering strong positive correlation and different event of interest (i.e. BMI), we suggest to used the proposed method with $k_z = 1, 2, 3, 4, 5$. The proposed method gives reasonable lower and upper probabilities for all categories for all values of k_z discussed above. This show that the proposed method works well with LSCV bandwidth selection with adaptive-nn bandwidth.

In Chapters 2 and 3, we used the same examples to illustrate the proposed method i.e. insurance and BMI examples. For insurance example, the event of interest is the total sum of the bivariate data. Based on the results discussed in Sections 2.6.1 and 3.5.1, the proposed method works well either using parametric copula or kernel-based nonparametric copula given the sample size. However, for LSCV bandwidth selection with adaptive-nn bandwidth, it was suggested to use adaptive-nn bandwidth with options $k_z = 1$, $k_z = 2$, $k_z = 3$ or $k_z = 4$ depending on the interest of study as discussed in Section 3.5.1.

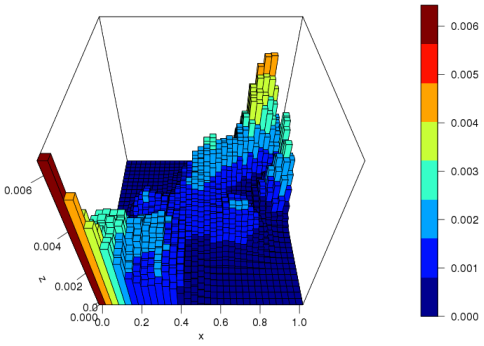
For the BMI example, the event of interest is different from the simulation studies in Sections 2.5, 3.4.1, 3.4.2 and insurance example. Although the event of interest is different, how the probabilities h_{ij} obtained are similar, including the strength of the correlation and the sample size, for both parametric and nonparametric copulas. From this example, the lower and upper probabilities are determined by including the probabilities h_{ij} or not, which also depends on the events of interest. As discussed in Sections 2.6.2 and 3.5.2, the proposed method works well using both parametric and kernel-based nonparametric copulas. But, it was suggested to use $k_z \leq 5$ for the LSCV bandwidth selection with adaptive-nn bandwidth because as we increase the



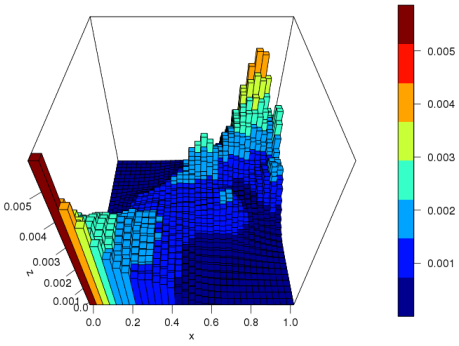
(a) $k_x = k_y = 1$



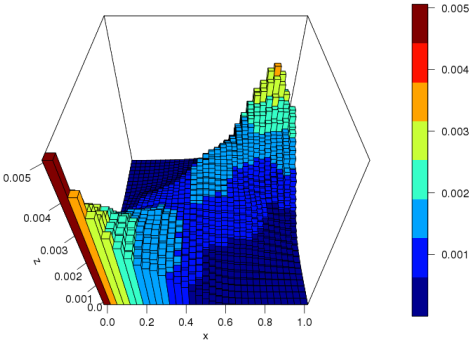
(b) $k_x = k_y = 2$



(c) $k_x = k_y = 3$



(d) $k_x = k_y = 4$



(e) $k_x = k_y = 5$

Figure 3.20: 3D-plots of probabilities h_{ij} for different values of k_z where $k_x = k_y$; adaptive-nn bandwidth

$k_z > 5$, the probabilities h_{ij} will not sum to 1 which dissatisfied the three conditions discussed in Section 2.3.

3.6 Concluding remarks

In this chapter we presented a new proposed method in Chapter 2 with the use of nonparametric copulas. The main research of this chapter is to use kernel-based copula method to overcome the misspecification problem occurred in Chapter 2 where for large data set, the method presented in this chapter can be used and leads to sensible inferences.

The probabilities h_{ij} are the key ingredients of our method for inference. With a kernel-copula based method, it gives more freedom to obtain the probabilities h_{ij}. However, there are three conditions on the probabilities h_{ij} to take into account as discussed in Section 2.3. Based on our study in this chapter, generally, the normal reference rule-of-thumb and the LSCV bandwidth selections for fixed and adaptive-nn types of bandwidths satisfied the conditions. The proposed method works well specifically for a positive correlation, regardless of sample size. However, the predictive performance of the proposed method does not work so well for negative correlation as discussed in Section 3.4.1 specifically for sum event of interest. As we investigate further the probabilities h_{ij} in Section 3.4.2, the proposed method works well using adaptive-nn bandwidth, regardless of the strength of the correlation but it depends on the value of k_z, where the probabilities h_{ij} obtained rely on a trade-off between the k_z and sample size.

As mentioned in Section 3.2, the standard kernel estimator of the copula density suffers from boundary biases and inconsistency due to unbounded densities. It should be emphasized that using the NPI on the marginals combined with the discretization of the copula, the problem does not occur in this research due to the transformations of variables that are used to estimate the densities, which is free of boundary bias. However, this topic should be studied in detail theoretically in terms of mathematical equations, and we left this question as a future research. Again, in this chapter, the major advantage of this presented method is its relatively

easy computations, as the use of NPI on the marginals combines naturally with the discretization of the copula. The kernel-based copulas considered in this chapter are implemented using command available in R. As long as suitable nonparametric copula estimation methods are available, these can be implemented in our method without any difficulties.

However, further study is needed for this chapter, in particular for use of other nonparametric copula methods, as discussed in Section 3.2, or other types of dependence structures such as nonlinear dependence structure of the bivariate data. The performance of the proposed method should be studied and investigated. We left these as topics for future research.

Chapter 4

NPI for combining diagnostic tests

4.1 Introduction

Measuring the accuracy of diagnostics tests is crucial in many application areas including medicine and health care. The Receiver Operating Characteristic (ROC) curve is a popular statistical tool for describing the performance of diagnostic tests. The area under the ROC curve (AUC) is often used as a measure of the overall performance of the diagnostics test [75]. It is increasingly clear that in medical settings, one test result is often not sufficient to serve as screening device for early detection of diseases [43, 100]. In addition, many researchers believe that a combination of test results will potentially lead to more sensitive screening rules for detecting diseases [67, 77]. Therefore in medical application, there is great interest in developing strategies for combining test results in order to increase the diagnostic accuracy. Usually [75], the objective function to be maximized is the area under the ROC curve (AUC).

Many researchers have discussed ways for combining test results, for example in [37, 76]. Often, linear combinations of the test results are used. For example, Su and Liu [91] derived an optimal linear combination that maximises the AUC when the test results for the non-diseased and diseased categories follow bivariate normal distributions. Pepe and Thompson [77] considered an empirical search of the optimal linear combination that maximises the Mann-Whitney U statistic of AUC, but this method is computationally complex as a search algorithm must be used. Liu et al.

[68] proposed a linear combination by combining the minimum and maximum values of the test results. This involves searching for a single coefficient that maximises the Mann-Whitney U statistic of AUC but not all test results are measured on the same scale [68]. Esteban et al. [37] proposed a step-by-step algorithm for estimating the parameter of a linear combination of the test results, which is close to the maximizing the AUC corresponding to the best linear combination. Kang et al. [57] proposed a nonparametric stepwise approach for the linear combination of the test results to search coefficient that maximises the Mann-Whitney U statistic of AUC. Both methods proposed by Esteban et al. [37] and Kang et al. [57] are computationally tractable. Recently, Yan et al. [99] proposed a combination method called pairwise approach, to maximize the AUC, by pairing one biomarker with the other biomarkers separately specifically for weak biomarkers ($0.50 < \text{AUC} < 0.70$).

All researchers mentioned above did not take dependence structures into account, such as using copula except Ghosh [43] and Sen [86]. Sen [86] presented the concept of copulas for multivariate distributions and dependence, and motivated the benefit of copulas via a number of applications including the design of clinical trials, microarray studies with survival endpoints and the analysis of dependent ROC curves. Ghosh [43] presented a binormal model for ROC curve estimation to accommodate multiple test results by considering the dependence using copulas. As mentioned by Bansal and Pepe [5], the dependence could be very important among the test results. They investigated the increment in the performance of measure accuracy that is possible by combining a novel continuous test result with a moderately performing standard continuous test result (AUC around 0.70 to 0.80) and found that an uncorrelated continuous test result with moderate performance on its own usually yields only minimally improved performance on the AUC [5]. The novel test result that has very poor performance on its own but is highly correlated with the standard test result, and a novel test result with poor ($\text{AUC} < 0.70$) to moderate performance that is highly correlated with the standard test result gives large improvements in the performance of measure accuracy [5].

The performance of AUC estimation is measured using a re-substitution method (use complete data set), as used often for example, by Su and Liu [91], Pepe and

Thomson [77], Pepe [76], Vexler at al. [95], Jin and Lu [54], Esteban at al. [37] and Liu et al. [68]. Re-substitution methods begin with finding the linear combination coefficient (let say, optimal coefficient, $\hat{\alpha}$) from a data set, then a total score is calculated by linearly combining the diagnostic test results using the optimal coefficient, $\hat{\alpha}$, which gives the maximal AUC value. Finally, the AUC was maximized based on the total score. This re-substitution method is usually overoptimistic for maximizing the diagnostic accuracy of future observations [28, 33, 53, 57]. So, the maximized AUC may perform well for the data set used but this is no guarantee for good performance for a future observation. Huang et al. [53] and Kang et al. [57] propose a leave-out one pair (LO1P) method, which to compares between the linear combination methods of the test results more fairly. Huang et al. [53] proved that the LO1P cross validation gives unbiased AUC maximized that associated with the combination coefficient (i.e. α).

Many articles have addressed the problem of finding the optimal linear combinations to maximise the AUC, as mentioned above. In this chapter, we introduce NPI for combining two diagnostic test results. First, by considering a weighted average of the two diagnostic test results without parametric copula, which directly applies the results of NPI for single diagnostic test [27]. Second we use NPI with a parametric copula introduced in Chapter 2, to combine two test results. NPI has been used for accuracy of the diagnostic tests with ordinal outcomes, with the inferences based on data for a disease group and non-disease group [35]. For accuracy of binary tests, NPI has been presented and discussed by Coolen-Maturi et al. [26], and for continuous test results in [27]. As NPI does not aim at inference for an entire population but instead explicitly considers a future observation, this provides an attractive alternative to standard methods [26].

We briefly discuss the basic concept of the empirical (distribution-free approach) and NPI-based ROC curves for a single test result in Section 4.2. We briefly discuss the empirical ROC curves for combining two diagnostics test results in order to optimize the diagnostic accuracy in Section 4.3. We present NPI for combining two diagnostic tests without copula including ROC curves and AUC in Section 4.4. In Section 4.5, we present the concepts of NPI for a weighted average of bivariate

continuous diagnostic test results, taking dependence structure into account using copulas for ROC curves and the AUC. We investigate the predictive performance of these approaches in Section 4.6 by simulation study. We present an example using data from the literature in Section 4.7. The chapter is finished with some concluding remarks in Section 4.8.

4.2 Receiver Operating Characteristic curve

The evaluation of the accuracy of diagnostic tests is important in medical applications where such tests are performed to detect diseases. Often, a diagnostic test yields more than one output value of test results. The diagnostic test results can take two values (binary test), or a value in a finite number of ordered categories (ordinal test), or real values (continuous test) [75]. There are several accuracy measures which vary depending on the type of diagnostic test results mentioned above, for example for continuous test results, Receiver Operating Characteristic (ROC) curve is often used [75]. In this chapter, we focus on the ROC curve as we consider the continuous test results. In addition to medical applications, ROC curves also play an important role in areas such as signal detection and machine learning [9], radiology [47], data mining [84] and credit scoring [7].

Let Y denote the result of a diagnostic test, assumed to be a continuous random quantity. Using a threshold ξ, the test result is assumed to be positive if $Y > \xi$, which indicates the disease, and negative if $Y \leq \xi$, where $\xi \in (-\infty, \infty)$. The sensitivity of a test is the probability of a positive test result for an individual with the disease, this is also known as the true positive fraction (TPF). The specificity is the probability of a negative test result for an individual without the disease. An accurate diagnostic test will have sensitivity and specificity both close to one. The false positive fraction (FPF) is the probability of a positive test result for an individual without the condition, hence, the specificity is equal to $1 - \text{FPF}$.

Let D denote the disease status, where $D = 1$ for the diseased group and $D = 0$ for the non-diseased group. Let Y^1 be used to denote the test result for the diseased group and Y^0 be used to denote the test result for the non-diseased group, let n_1

and n_0 be the numbers of individuals in the diseased and the non-diseased groups, respectively. The TPF and FPF can be written as

$$\text{TPF}(\xi) = P\left[Y^1 > \xi | D = 1\right] = S_1(\xi)$$
$$\text{FPF}(\xi) = P\left[Y^0 > \xi | D = 0\right] = S_0(\xi)$$

where $S_1(\xi)$ and $S_0(\xi)$ are the survival functions for the random quantities Y^1 and Y^0 for the diagnostic test results for the diseased and non-diseased groups, respectively.

The ROC curve is a graphical plot that illustrates the performance of diagnostic tests which yield ordinal or continuous results. The curve is created by plotting the TPF(ξ) against the FPF(ξ) at all possible threshold settings, ξ and can be defined as

$$\text{ROC} = \{(\text{FPF}(\xi), \text{TPF}(\xi)), \xi \in (-\infty, \infty)\} \tag{4.1}$$

The ROC curve depicts relative trade-offs between TPF(ξ) and FPF(ξ). A test is considered ideal if it completely separates the individuals with and without the disease for a particular threshold ξ, FPF(ξ) = 0 and TPF(ξ) = 1. For an extreme situation, a test has no ability to distinguish between the individuals with and without the disease if FPF(ξ) = TPF(ξ) for all thresholds ξ.

4.2.1 Empirical ROC curve

In this section, we briefly review the empirical method for the ROC curve. The ROC curve depends on the distributions of Y^1 and Y^0, however these distributions are usually unknown. The ROC curve for a diagnostic test with continuous results can be estimated by the nonparametric empirical method. This method is popular due to its flexibility to adapt fully to the available data, it yields the empirical ROC curve which we will use in Section 4.3, in particular to compare with the NPI method introduced in this thesis. Methods using assumed parametric distributions for both Y^1 and Y^0, together with methods for estimation of the parameters, are of course also used, but are less popular because these require strong assumptions about the forms of the distribution of the diagnostic test results [75]. More details on these methods can be found in [75].

Suppose that we have test data on n_1 individuals from a diseased group and n_0 individuals from a non-diseased group, denoted by $y_i^1, i = 1, ..., n_1$ and $y_i^0, i = 1, ..., n_0$, respectively. Assume that the two groups are fully independent, meaning that no information about any aspect related to one group contains information about any aspect of the other group. For the empirical ROC curve method, these observations per group are assumed to be realisations of random quantities that are identically distributed as Y^1 for the diseased group, and as Y^0 for the non-diseased group, with corresponding survival functions $S_1(y) = P[Y^1 > y]$ and $S_0(y) = P[Y^0 > y]$. The empirical estimator of the ROC is [75],

$$\widehat{\text{ROC}} = \left\{ \left(\widehat{\text{FPF}}(\xi), \widehat{\text{TPF}}(\xi) \right), \xi \in (-\infty, \infty) \right\} \tag{4.2}$$

with

$$\widehat{\text{TPF}}(\xi) = \hat{S}_1(\xi) = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{1}\{y_i^1 > \xi\} \tag{4.3}$$

$$\widehat{\text{FPF}}(\xi) = \hat{S}_0(\xi) = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{1}\{y_i^0 > \xi\} \tag{4.4}$$

where $\mathbf{1}\{A\}$ is the indicator function which is equal to 1 if A is true and 0 else, and where $\hat{S}_1(\xi)$ and $\hat{S}_0(\xi)$ are the empirical survival functions for Y^1 and Y^0, respectively.

To represent the accuracy of a diagnostic test or to compare two or more ROC curves, a single numerical value or summary may be useful in many cases [75]. A useful summary is the area under the ROC curve, AUC [75]. The AUC measures the overall performance of the diagnostic test. Higher values of AUC indicate more precise tests, with AUC = 1 for a perfect test, and AUC = 0.5 for uninformative tests. We can also write the ROC curve in equation (4.2) as $\text{ROC}(\cdot) = \{(t, ROC(t)), t \in (0, 1)\}$, where the ROC function maps t to $TPF(\xi)$, and ξ is the threshold corresponding to $FPF(\xi) = t$ [75]. Thus the AUC is [75]

$$\text{AUC} = \int_0^1 \text{ROC}(t)dt \tag{4.5}$$

The AUC is equal to the probability that the test results from a randomly selected pair of diseased and non-diseased subjects are correctly ordered [4], i.e.

$$\text{AUC} = P\left[Y^1 > Y^0\right] \tag{4.6}$$

Proof: We have

$$\begin{aligned}\text{AUC} &= \int_0^1 \text{ROC}(t)dt \\ &= \int_0^1 S_1(S_0^{-1}(t))dt \\ &= \int_{-\infty}^{\infty} S_1(y)dS_0(y) \\ &= \int_{-\infty}^{\infty} P[Y^1 > y]f_0(y)\end{aligned}$$

Using the change of variable from t to $y = S_0^{-1}$ in the second line and where f_0 denotes the probability density function for Y_0 in the third line. Thus by statistical independence of Y_1 and Y_0, we can write

$$\begin{aligned}\text{AUC} = &\int_{-\infty}^{\infty} P[Y^1 > y, Y^0 = y]dy \\ =&P[Y^1 > Y^0]\end{aligned}$$

□

The empirical estimator of the AUC is the well-known Mann-Whitney U statistic [75], which is defined as

$$\widehat{\text{AUC}} = \frac{1}{n_1 n_0} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \psi(y_i^1, y_j^0) \tag{4.7}$$

where

$$\psi(y_i^1, y_j^0) = \begin{cases} 1, & \text{if } y_i^1 > y_j^0 \\ \frac{1}{2}, & \text{if } y_i^1 = y_j^0 \\ 0, & \text{if } y_i^1 < y_j^0 \end{cases} \tag{4.8}$$

The empirical estimation $\widehat{\text{AUC}}$ value will be used in Section 4.3 for a weighted average of bivariate diagnostic test results.

4.2.2 NPI for ROC curve

In this section, we introduce Nonparametric Predictive Inference (NPI) for diagnostic accuracy, following Coolen-Maturi et al. [27]. The NPI method is different from the nonparametric empirical method as it is explicitly predictive, considering

a single next future observation given the past observations instead of aiming at estimation for an entire assumed underlying population. As mentioned in Section 1.2, in NPI the uncertainty is quantified by lower and upper probabilities for events of interest. The NPI lower and upper ROC curves, and the corresponding lower and upper AUC, have been derived by Coolen-Maturi et al. [27], corresponding to the assumptions $A_{(n_1)}$ for the diseased group and $A_{(n_0)}$ for the non-diseased group, where the inferences involve one further patient from each group.

Suppose that $\{Y_i^1, i = 1, ..., n_1, n_1 + 1\}$ and $\{Y_i^0, i = 1, ..., n_0, n_0 + 1\}$ are continuous and exchangeable random quantities from the diseased group and the non-diseased group, where $Y_{n_1+1}^1$ and $Y_{n_0+1}^0$ are the next future observations from the diseased and non-diseased groups following n_1 and n_0 observations, respectively. As explained in Section 4.2.1, we assume that both groups are fully independent. Let $y_1^1 < ... < y_{n_1}^1$ be the ordered observed values for n_1 individuals from the diseased group and $y_1^0 < ... < y_{n_0}^0$ the ordered observed values for n_0 individuals from the non-diseased group. For ease of notation, let $y_0^1 = y_0^0 = -\infty$ and $y_{n_1+1}^1 = y_{n_0+1}^0 = \infty$ and assume that there are no ties in the data. The NPI lower and upper survival functions for $Y_{n_1+1}^1$ and $Y_{n_0+1}^0$ are

$$\underline{\mathrm{TPF}}\,(\xi) = \underline{S}_1(\xi) = \underline{P}(Y_{n_1+1}^1 > \xi) = \frac{\sum_{i=1}^{n_1} \mathbf{1}\{y_i^1 > \xi\}}{n_1 + 1} \tag{4.9}$$

$$\overline{\mathrm{TPF}}\,(\xi) = \overline{S}_1(\xi) = \overline{P}(Y_{n_1+1}^1 > \xi) = \frac{\sum_{i=1}^{n_1} \mathbf{1}\{y_i^1 > \xi\} + 1}{n_1 + 1} \tag{4.10}$$

$$\underline{\mathrm{FPF}}\,(\xi) = \underline{S}_0(\xi) = \underline{P}(Y_{n_0+1}^0 > \xi) = \frac{\sum_{j=1}^{n_0} \mathbf{1}\{y_j^0 > \xi\}}{n_0 + 1} \tag{4.11}$$

$$\overline{\mathrm{FPF}}\,(\xi) = \overline{S}_0(\xi) = \overline{P}(Y_{n_0+1}^0 > \xi) = \frac{\sum_{j=1}^{n_0} \mathbf{1}\{y_j^0 > \xi\} + 1}{n_0 + 1} \tag{4.12}$$

where $\underline{P}$ and $\overline{P}$ are the NPI lower and upper probabilities [1]. These NPI lower and upper survival functions are used to derive the lower and upper FPF and TPF for the next future individual per group, for different threshold values ξ, which then are combined to derive the corresponding NPI lower and upper ROC curves. The NPI lower and upper survival functions are optimal bounds for all survival functions corresponding to $A_{(n)}$, they immediately lead to the optimal bounds for the TPF and FPF [22]. As the ROC combines the survival functions for the two groups, the

NPI lower and upper ROC curves are defined to be the optimal bounds for all such curves corresponding to any pair of survival functions $S_1(t)$ and $S_0(t)$ for $Y^1_{n_1+1}$ and $Y^0_{n_0+1}$ in between their respective NPI lower and upper survival functions as given by equations (4.9) - (4.12). As the ROC curve depends monotonously on the survival functions, it is easily seen that the optimal bounds, the NPI lower and upper ROC curves, are [27]

$$\underline{ROC} = \left\{ \left(\overline{FPF}(\xi), \underline{TPF}(\xi) \right), \xi \in (-\infty, \infty) \right\} \tag{4.13}$$

$$\overline{ROC} = \left\{ \left(\underline{FPF}(\xi), \overline{TPF}(\xi) \right), \xi \in (-\infty, \infty) \right\} \tag{4.14}$$

For all ξ, it can be seen that $\underline{\mathrm{FPF}}\,(\xi) \leq \widehat{\mathrm{FPF}}(\xi) \leq \overline{\mathrm{FPF}}\,(\xi)$ and $\underline{\mathrm{TPF}}\,(\xi) \leq \widehat{\mathrm{TPF}}(\xi) \leq \overline{\mathrm{TPF}}\,(\xi)$. This implies that the empirical ROC curve is bounded by the NPI lower and upper ROC curves [27].

Consider an event that the test result for the next future individual from the diseased group is greater than the test result for the next future individual from the non-diseased group, the NPI lower and upper probabilities for the event is defined as [27]

$$\underline{\mathrm{AUC}} = \underline{P}(Y^1_{n_1+1} > Y^0_{n_0+1}) = \frac{1}{(n_1+1)(n_0+1)} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \mathbf{1}(y^1_i \geq y^0_j) \tag{4.15}$$

$$\overline{\mathrm{AUC}} = \overline{P}(Y^1_{n_1+1} > Y^0_{n_0+1}) = \frac{1}{(n_1+1)(n_0+1)} \left[\sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \mathbf{1}\left\{ y^1_i \geq y^0_j \right\} + n_1 + n_0 + 1 \right] \tag{4.16}$$

The imprecision of the NPI lower and upper AUC (the difference between NPI upper AUC and the NPI lower AUC) depends only on the two sample sizes, n_1 and n_0 [27]. The empirical and NPI lower and upper ROC curves discussed here will be used in Sections 4.3 and 4.4.

4.3 Empirical method for combining two diagnostic tests

In this section, we briefly review methods used for combining two diagnostic tests in order to optimize the diagnostic accuracy by following Pepe and Thompson [77], who proposed an empirical approach which relates to our work in this chapter.

Let D be a disease status where $D = 1$ for diseased group and $D = 0$ for non-diseased group. Let X^D and Y^D be continuous random quantities of two diagnostic test results. Consider a weighted average of the two test results, $T^D(X^D, Y^D) = \alpha X^D + (1 - \alpha)Y^D$ where $\alpha \in [0, 1]$ and the coefficient α is chosen to maximize the diagnostic accuracy associated with the resultant composite score T^D. In this chapter, we focus on the area under the ROC curve (AUC) as the objective function following [77]. As discussed in Section 4.2, the ROC curve for a total or composite score such as T^D is defined as the set of points $\{(FPF(\xi), TPF(\xi)), \xi \in (-\infty, \infty)\}$ where $TPF(\xi) = P[T_i^1 > \xi | D = 1]$ and can be interpreted as the true positive rate associated with the positivity criterion $T > \xi$ and $FPF(\xi) = P[T_i^0 > \xi | D = 0]$, which similarly can be interpreted as the false positive rate at threshold ξ [77].

Suppose that we have two test results on each of n_1 individuals from a diseased group and n_0 individuals from a non-diseased group, denoted by $\{(x_i^1, y_i^1), i = 1, ..., n_1\}$ and $\left\{(x_j^0, y_j^0), j = 1, ..., n_0\right\}$, respectively. Consider a weighted average of the test results, $t_i^1 = \alpha x_i^1 + (1-\alpha) y_i^1$ for diseased group and $t_j^0 = \alpha x_j^0 + (1-\alpha) y_j^0$ for non-diseased group where $\alpha \in [0, 1]$. In line with equations (4.2) - (4.4), the empirical estimator of the ROC curve is

$$\widehat{\mathrm{ROC}} = \left\{ \left(\widehat{\mathrm{FPF}}(\xi), \widehat{\mathrm{TPF}}(\xi) \right), \xi \in (-\infty, \infty) \right\} \tag{4.17}$$

with

$$\widehat{\mathrm{TPF}}(\xi) = \hat{S}_1(\xi) = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{1}\{t_i^1 > \xi\} \tag{4.18}$$

$$\widehat{\mathrm{FPF}}(\xi) = \hat{S}_0(\xi) = \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbf{1}\{t_j^0 > \xi\} \tag{4.19}$$

where $\mathbf{1}\{A\}$ is the indicator function which is equal to 1 if A is true and 0 else, and where $\hat{S}_1(\cdot)$ and $\hat{S}_0(\cdot)$ are the empirical survival functions for T^1 and T^0, respectively.

The AUC of the T^D can be interpreted as a probability $P[T^1 \geq T^0]$ where T^1 and T^0 are composite scores for independent, randomly selected study units from the diseased and non-diseased groups, respectively [4]. As mentioned in Section 4.2.1, the empirical estimator of the AUC is the well-known Mann-Whitney U statistic [75], which is defined in equation (4.7). Alternatively, the AUC associated with

$T^D(X^D, Y^D)$ given by Pepe and Thomson [77] is

$$\widehat{\text{AUC}} = \frac{1}{n_1 n_0} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \mathbf{1}\left\{\alpha x_i^1 + (1-\alpha) y_i^1 \geq \alpha x_j^0 + (1-\alpha) y_j^0\right\} \tag{4.20}$$

An optimal coefficient, α_{opt} is defined by α that maximizes the AUC in equation (4.20) and can be denoted by $\hat{\alpha}$.

We should emphasize that the linear combination used by Pepe and Thomson in [77] is slightly different from our weighted average.

4.4 NPI without copula for combining two diagnostic tests

In this section, we present NPI for a weighted average of two diagnostic test results without copula where we directly apply the results presented in Section 4.2.2. Consider the same random quantity of diagnostics test results, X and Y in Section 4.3, let $X_{n_D+1}^D$ and $Y_{n_D+1}^D$ be the next future observation of the diagnostics test results and let $T_{n_D+1}^D = \alpha X_{n_D+1}^D + (1-\alpha) Y_{n_D+1}^D$ be the weighted average of the future two test results where $\alpha \in [0,1]$. Let $t_1^1 < ... < t_{n_1}^1$ be the ordered observed values for n_1 total of the two test results from the diseased group and $t_1^0 < ... < t_{n_0}^0$ be the ordered observed values for n_0 total of the two test results from the non-diseased group. For ease of notation, let $t_0^1 = t_0^0 = -\infty$ and $t_{n_1+1}^1 = t_{n_0+1}^0 = \infty$. We assume that there are no ties in the data (these can be dealt with by assuming that such observations differ by a very small amount, a common method to break ties in statistics [71]). From equations (4.9) - (4.12) in Section 4.2.2, the NPI lower and upper survival functions for $T_{n_1+1}^1$ and $T_{n_0+1}^0$ are

$$\underline{\text{TPF}}\,(\xi) = \underline{S}_1(\xi) = \underline{P}(T_{n_1+1}^1 > \xi) = \frac{\sum_{i=1}^{n_1} \mathbf{1}\{t_i^1 > \xi\}}{n_1 + 1} \tag{4.21}$$

$$\overline{\text{TPF}}\,(\xi) = \overline{S}_1(\xi) = \overline{P}(T_{n_1+1}^1 > \xi) = \frac{\sum_{i=1}^{n_1} \mathbf{1}\{t_i^1 > \xi\} + 1}{n_1 + 1} \tag{4.22}$$

$$\underline{\text{FPF}}\,(\xi) = \underline{S}_0(\xi) = \underline{P}(T_{n_0+1}^0 > \xi) = \frac{\sum_{j=1}^{n_0} \mathbf{1}\{t_j^0 > \xi\}}{n_0 + 1} \tag{4.23}$$

$$\overline{\text{FPF}}\,(\xi) = \overline{S}_0(\xi) = \overline{P}(T_{n_0+1}^0 > \xi) = \frac{\sum_{j=1}^{n_0} \mathbf{1}\{t_j^0 > \xi\} + 1}{n_0 + 1} \tag{4.24}$$

where $\underline{P}$ and $\overline{P}$ are the NPI lower and upper probabilities [1] and threshold, $\xi \in (-\infty, \infty)$. In line with equations (4.13) and (4.14) in Section 4.2.2, the ROC curve clearly depends monotonously on the survival functions, it is easily seen that the optimal bounds, which define to be the NPI lower and upper ROC curves, are

$$\underline{ROC} = \left\{ \left(\overline{FPF}(\xi), \underline{TPF}(\xi) \right), \xi \in (-\infty, \infty) \right\} \tag{4.25}$$

$$\overline{ROC} = \left\{ \left(\underline{FPF}(\xi), \overline{TPF}(\xi) \right), \xi \in (-\infty, \infty) \right\} \tag{4.26}$$

For all ξ, $\underline{\mathrm{FPF}}(\xi) \leq \widehat{\mathrm{FPF}}(\xi) \leq \overline{\mathrm{FPF}}(\xi)$ and $\underline{\mathrm{TPF}}(\xi) \leq \widehat{\mathrm{TPF}}(\xi) \leq \overline{\mathrm{TPF}}(\xi)$, this implies that the empirical ROC curve is bounded by the NPI lower and upper ROC curves [27].

In line with equation (4.6) and Section 4.3, we are interested in the NPI lower and upper probabilities for the event that the weighted average score for the future two test results from the diseased group is greater than the weighted average score for the future two test results from the non-diseased group. In line with equations (4.15) and (4.16) in Section 4.2.2 and Coolen-Maturi et al. [27], the NPI lower and upper combine AUC can be defined as

$$\begin{aligned}
\underline{\mathrm{AUC}} =& \underline{P}(T^1_{n_1+1} > T^0_{n_0+1}) \\
=& \frac{1}{(n_1+1)(n_0+1)} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \mathbf{1}(\alpha x_i^1 + (1-\alpha) y_i^1 \geq \alpha x_j^0 + (1-\alpha) y_j^0) \\
=& \frac{1}{(n_1+1)(n_0+1)} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \mathbf{1}(t_i^1 \geq t_j^0)
\end{aligned} \tag{4.27}$$

$$\begin{aligned}
\overline{\mathrm{AUC}} =& \overline{P}(T^1_{n_1+1} > T^0_{n_0+1}) \\
=& \frac{1}{(n_1+1)(n_0+1)} \left[\sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \mathbf{1}\left\{ \alpha x_i^1 + (1-\alpha) y_i^1 \geq \alpha x_j^0 + (1-\alpha) y_j^0 \right\} + n_1 + n_0 + 1 \right] \\
=& \frac{1}{(n_1+1)(n_0+1)} \left[\sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \mathbf{1}\left\{ t_i^1 \geq t_j^0 \right\} + n_1 + n_0 + 1 \right]
\end{aligned} \tag{4.28}$$

The optimal coefficients, α_{opt}'s that maximizes the lower and upper AUC in equations (4.27) and (4.28) can be denoted by $\hat{\alpha}_L$ and $\hat{\alpha}_U$, respectively.

4.5 NPI with parametric copula for bivariate diagnostic tests

In this section we present NPI for the weighted average of the two diagnostic tests to optimize the diagnostic accuracy with consideration of the dependence structure, using a copula as proposed in Chapter 2. As mentioned in Section 4.1, dependence is important when considering the combination of the bivariate test results, as it can influence the accuracy of detection of diseases [5]. From Section 2.3, we found that NPI with parametric copula for bivariate data is a straightforward method for prediction of a bivariate random quantity, where imprecision in the marginals provides robustness with regard to the assumed copula for small sample size. Hence, in this chapter, the proposed method in Chapter 2 can be used and considered to measure the accuracy of the bivariate diagnostic test results in order to increase the detection rate.

Consider a bivariate random quantity of diagnostic test results, (X, Y), let $(X^D_{n_D+1}, Y^D_{n_D+1})$ be the next future bivariate random quantity of diagnostic test results and let $T^D_{n_D+1} = \alpha X^D_{n_D+1} + (1-\alpha) Y^D_{n_D+1}$ be the weighted average of the future two test results where $\alpha \in [0, 1]$. For the diseased group, the lower probability for the event that the sum of the next future observations will exceed a particular threshold ξ is

$$\underline{S}^1_c(t) = \underline{P}(T^1_{n_1+1} > \xi) = \sum_{(i,l) \in L^1_t} h^1_{il}(\widehat{\theta}_1) \tag{4.29}$$

with $L^1_t = \{(i, l) : \alpha x^1_{i-1} + (1-\alpha) y^1_{l-1} > \xi\}$, and the corresponding upper probability is

$$\overline{S}^1_c(t) = \overline{P}(T^1_{n_1+1} > \xi) = \sum_{(i,l) \in U^1_t} h^1_{il}(\widehat{\theta}_1) \tag{4.30}$$

with $U^1_t = \{(i, l) : \alpha x^1_i + (1-\alpha) y^1_l > \xi\}$ where $\xi \in (-\infty, \infty)$, and $\underline{S}^1_c(t)$ and $\overline{S}^1_c(t)$ are the lower and upper survival functions for the sum of the next future observations, $T^1_{n_1+1}$ with considering copula denotes by subscript c. In line with equation (2.1) in Section 2.3, the probabilities $h^1_{il}(\widehat{\theta}_1)$ are defined as

$$h^1_{il}(\widehat{\theta}_1) = P_C(\widetilde{X}^1_{n_1+1} \in \left(\frac{i-1}{n_1+1}, \frac{i}{n_1+1}\right), \widetilde{Y}^1_{n_1+1} \in \left(\frac{l-1}{n_1+1}, \frac{l}{n_1+1}\right) | \widehat{\theta}_1) \tag{4.31}$$

for $i, l = 1, 2, \ldots, n_1 + 1$ where $P_C(\cdot|\hat{\theta}_1)$ represents the copula-based probability with estimated copula where $\hat{\theta}_1$ is a parameter value from parametric copula for diseased group.

For the non-diseased group, the lower probability for the event that the sum of the next future observations will exceed a particular threshold ξ is

$$\underline{S}_c^0(t) = \underline{P}(T_{n_0+1}^0 > \xi) = \sum_{(j,k)\in L_t^0} h_{jk}^0(\widehat{\theta_0}) \tag{4.32}$$

with $L_t^0 = \{(j,k) : \alpha x_{j-1}^0 + (1-\alpha) y_{k-1}^0 > \xi\}$, and the corresponding upper probability is

$$\overline{S}_c^0(t) = \overline{P}(T_{n_0+1}^0 > \xi) = \sum_{(j,k)\in U_t^0} h_{jk}^0(\widehat{\theta_0}) \tag{4.33}$$

with $U_t^0 = \{(j,k) : \alpha x_j^0 + (1-\alpha) y_k^0 > \xi\}$ where $\xi \in (-\infty, \infty)$, and $\underline{S}_c^0(t)$ and $\overline{S}_c^0(t)$ are the lower and upper survival functions for the sum of the next future observation, $T_{n_0+1}^0$. In line with equation (2.1) in Section 2.3, the probabilities $h_{jk}^0(\widehat{\theta_0})$ are defined as

$$h_{jk}^0(\widehat{\theta_0}) = P_C(\widetilde{X}_{n_0+1}^0 \in \left(\frac{j-1}{n_0+1}, \frac{j}{n_0+1}\right), \widetilde{Y}_{n_0+1}^0 \in \left(\frac{k-1}{n_0+1}, \frac{k}{n_0+1}\right) | \widehat{\theta_0}) \tag{4.34}$$

for $j, k = 1, 2, \ldots, n_0 + 1$ where $P_C(\cdot|\hat{\theta}_0)$ represents the copula-based probability with estimated copula where $\hat{\theta}_0$ is a parameter value from parametric copula for non-diseased group. Throughout this chapter, the subscript c is used to show the functions are considering the copula.

The NPI lower and upper survival functions from equations (4.29), (4.30), (4.32) and (4.33) are used to derive lower and upper FPF and TPF for the weighted average of the next future observation per group, for different threshold values ξ, and we combined to derive the corresponding NPI lower and upper ROC curves. In line with equations (4.21) - (4.24), the NPI lower and upper survival functions are optimal bounds for all survival functions corresponding to $A_{(n)}$ [22], which leads to the following optimal bounds for the TPF and FPF when considering the dependence structure

$$\underline{\mathrm{TPF}}_c(\xi) = \underline{S}_c^1(\xi) = \underline{P}(T_{n_1+1}^1 > \xi) = \sum_{(i,l)\in L_t^1} h_{il}^1(\widehat{\theta_1}) \tag{4.35}$$

$$\overline{\mathrm{TPF}}_c(\xi) = \overline{S}_c^1(\xi) = \overline{P}(T_{n_1+1}^1 > \xi) = \sum_{(i,l)\in U_t^1} h_{il}^1(\widehat{\theta_1}) \tag{4.36}$$

$$\underline{\mathrm{FPF}}_c(\xi) = \underline{S}_c^0(\xi) = \underline{P}(T_{n_0+1}^0 > \xi) = \sum_{(j,k)\in L_t^0} h_{jk}^0(\widehat{\theta_0}) \tag{4.37}$$

$$\overline{\mathrm{FPF}}_c(\xi) = \overline{S}_c^0(\xi) = \overline{P}(T_{n_0+1}^0 > \xi) = \sum_{(j,k)\in U_t^0} h_{jk}^0(\widehat{\theta_0}) \tag{4.38}$$

where $\underline{P}$ and $\overline{P}$ the are NPI lower and upper probabilities [1]. As the ROC combines the survival functions for the two groups, the NPI lower and upper ROC curves are again defined to be the optimal bounds for all such curves corresponding to any pair of survival functions $S_c^1(t)$ and $S_c^0(t)$ for $T_{n_1+1}^1$ and $T_{n_0+1}^0$ in between their respective NPI lower and upper survival functions, as given by equations (4.35) - (4.38). The ROC curve with copula clearly depends monotonously on the survival functions, it is easily seen that the optimal bounds, which are the NPI lower and upper ROC curves with copula, are

$$\underline{ROC}_c = \left\{ \left(\overline{FPF}_c(\xi), \underline{TPF}_c(\xi) \right), \xi \in (-\infty, \infty) \right\} \tag{4.39}$$

$$\overline{ROC}_c = \left\{ \left(\underline{FPF}_c(\xi), \overline{TPF}_c(\xi) \right), \xi \in (-\infty, \infty) \right\} \tag{4.40}$$

In order to optimize the diagnostic accuracy of the weighted average of the future two diagnostic test results, we maximize the area under ROC curve by finding the value of α such that $T_{n_D+1}^D = \alpha X_{n_D+1}^D + (1-\alpha) Y_{n_D+1}^D$ maximizes the AUC. For each block $B_{il}^1 = (x_{i-1}^1, x_i^1) \odot (y_{l-1}^1, y_l^1)$, generated by the observed data, let $t_{i-1,l-1}^1 = \alpha x_{i-1}^1 + (1-\alpha) y_{l-1}^1$ be the combined weighted value corresponding to the left-bottom of the block. And $t_{i,l}^1 = \alpha x_i^1 + (1-\alpha) y_l^1$ be the combined weighted value corresponding to the right-top of the block. The same can be defined for each block $B_{jk}^0 = (x_{j-1}^0, x_j^0) \odot (y_{k-1}^0, y_k^0)$, let $t_{j-1,k-1}^0 = \alpha x_{j-1}^0 + (1-\alpha) y_{k-1}^0$ be the combined weighted value corresponding to the left-bottom of the block, and $t_{j,k}^0 = \alpha x_j^0 + (1-\alpha) y_k^0$ be the combined weighted value corresponding to the right-top of the block. In line with equations (4.29) - (4.34), the NPI lower and upper probabilities AUC associated with the weighted average for the bivariate diagnostic

test results with parametric copula can directly be defined as

$$\begin{aligned}\underline{AUC}_c &= \underline{P}(T^1_{n_1+1} > T^0_{n_0+1}) \\ &= \sum_{i=1}^{n_1+1}\sum_{l=1}^{n_1+1} h^1_{il}(\hat{\theta}_1) \sum_{j=1}^{n_0+1}\sum_{k=1}^{n_0+1} \mathbf{1}\{t^0_{j,k} < t^1_{i-1,l-1}\} h^0_{jk}(\hat{\theta}_0)\end{aligned} \tag{4.41}$$

$$\begin{aligned}\overline{AUC}_c &= \overline{P}(T^1_{n_1+1} > T^0_{n_0+1}) \\ &= \sum_{i=1}^{n_1+1}\sum_{l=1}^{n_1+1} h^1_{il}(\hat{\theta}_1) \sum_{j=1}^{n_0+1}\sum_{k=1}^{n_0+1} \mathbf{1}\{t^0_{j-1,k-1} < t^1_{i,l}\} h^0_{jk}(\hat{\theta}_0)\end{aligned} \tag{4.42}$$

where $\mathbf{1}\{A\}$ is an indicator function which is equal to $\mathbf{1}$ if event A occurs and 0 else. The optimal coefficients, α_{opt}'s that maximizes the AUC in equations (4.41) and (4.42) can be denoted by $\hat{\alpha}^c_L$ and $\hat{\alpha}^c_U$, respectively.

Proof: The NPI lower probability for the event $T^1_{n_1+1} > T^0_{n_0+1}$ is derived as follows:

$$\begin{aligned}
P &= P(T^1_{n_1+1} > T^0_{n_0+1}) \\
&= \sum_{i=1}^{n_1+1}\sum_{l=1}^{n_1+1} P(T^0_{n_0+1} < \alpha X^1_{n_1+1} + (1-\alpha)Y^1_{n_1+1}, X^1_{n_1+1} \in (x^1_{i-1}, x^1_i), Y^1_{n_1+1} \in (y^1_{l-1}, y^1_l)) \\
&= \sum_{i=1}^{n_1+1}\sum_{l=1}^{n_1+1} P\left(T^0_{n_0+1} < \alpha X^1_{n_1+1} + (1-\alpha)Y^1_{n_1+1}, (X^1_{n_1+1}, Y^1_{n_1+1}) \in B^1_{il}\right) \\
&\geq \sum_{i=1}^{n_1+1}\sum_{l=1}^{n_1+1} h^1_{il}(\hat{\theta}_1) P(T^0_{n_0+1} < t^1_{i-1,l-1}) \\
&= \sum_{i=1}^{n_1+1}\sum_{l=1}^{n_1+1} h^1_{il}(\hat{\theta}_1) \sum_{j=1}^{n_0+1}\sum_{k=1}^{n_0+1} P\left(\alpha X^0_{n_0+1} + (1-\alpha)Y^0_{n_0+1} < t^1_{i-1,l-1}, (X^0_{n_0+1}, Y^0_{n_0+1}) \in B^0_{jk}\right) \\
&\geq \sum_{i=1}^{n_1+1}\sum_{l=1}^{n_1+1} h^1_{il}(\hat{\theta}_1) \sum_{j=1}^{n_0+1}\sum_{k=1}^{n_0+1} \mathbf{1}\{t^0_{j,k} < t^1_{i-1,l-1}\} h^0_{jk}(\hat{\theta}_0)
\end{aligned}$$

For the lower probability, we want to make the probability for the event $T^1_{n_1+1} > T^0_{n_0+1}$ as small as possible. To this end, the first inequality follows by putting the probability $h^1_{il}(\hat{\theta}_1)$ corresponding to the block B^1_{il} to the left-bottom of the block, for all $i, l = 1, \ldots, n_1 + 1$. Thus the corresponding combined weighted value is $t^1_{i-1,l-1} = \alpha x^1_{i-1} + (1-\alpha)y^1_{l-1}$. The second inequality follows by putting the probability $h^0_{jk}(\hat{\theta}_0)$ corresponding to the block B^0_{jk} to the right-top of the block, for all $j, k = 1, \ldots, n_0 + 1$, and the corresponding combined weighted value is

$t^0_{j,k} = \alpha x^0_j + (1-\alpha) y^0_k$.

The NPI upper probability for the event $T^1_{n_1+1} > T^0_{n_0+1}$ is derived as follows:

$$
\begin{aligned}
P &= P(T^1_{n_1+1} > T^0_{n_0+1}) \\
&= \sum_{i=1}^{n_1+1} \sum_{l=1}^{n_1+1} P(T^0_{n_0+1} < \alpha X^1_{n_1+1} + (1-\alpha) Y^1_{n_1+1}, X^1_{n_1+1} \in (x^1_{i-1}, x^1_i), Y^1_{n_1+1} \in (y^1_{l-1}, y^1_l)) \\
&= \sum_{i=1}^{n_1+1} \sum_{l=1}^{n_1+1} P\left(T^0_{n_0+1} < \alpha X^1_{n_1+1} + (1-\alpha) Y^1_{n_1+1}, (X^1_{n_1+1}, Y^1_{n_1+1}) \in B^1_{il}\right) \\
&\leq \sum_{i=1}^{n_1+1} \sum_{l=1}^{n_1+1} h^1_{il}(\hat{\theta}_1) P(T^0_{n_0+1} < t^1_{i,l}) \\
&= \sum_{i=1}^{n_1+1} \sum_{l=1}^{n_1+1} h^1_{il}(\hat{\theta}_1) \sum_{j=1}^{n_0+1} \sum_{k=1}^{n_0+1} P\left(\alpha X^0_{n_0+1} + (1-\alpha) Y^0_{n_0+1} < t^1_{i,l}, (X^0_{n_0+1}, Y^0_{n_0+1}) \in B^0_{jk}\right) \\
&\leq \sum_{i=1}^{n_1+1} \sum_{l=1}^{n_1+1} h^1_{il}(\hat{\theta}_1) \sum_{j=1}^{n_0+1} \sum_{k=1}^{n_0+1} \mathbf{1}\{t^0_{j-1,k-1} < t^1_{i,l}\} h^0_{jk}(\hat{\theta}_0)
\end{aligned}
$$

For the upper probability, we want to make the probability for the event $T^1_{n_1+1} > T^0_{n_0+1}$ as large as possible. To this end, the first inequality follows by putting the probability $h^1_{il}(\hat{\theta}_1)$ corresponding to the block B^1_{il} to the right-top of the block, for all $i, l = 1, \ldots, n_1 + 1$. Thus the corresponding combined weighted value is $t^1_{i,l} = \alpha x^1_i + (1-\alpha) y^1_l$. The second inequality follows by putting the probability $h^0_{jk}(\hat{\theta}_0)$ corresponding to the block B^0_{jk} to the left-bottom of the block, for all $j, k = 1, \ldots, n_0 + 1$, and the corresponding combined weighted value is $t^0_{j-1,k-1} = \alpha x^0_{j-1} + (1-\alpha) y^0_{k-1}$.

□

This proof has a similar structure as the proof of the NPI lower and upper probabilities for comparing two independent groups introduced by Coolen [16] and used by Maturi [71].

4.6 Predictive performance

In this section we analyse the performance of the proposed method including the empirical and NPI without copula methods. We will use the re-substitution and LO1P

simulation methods to evaluate the performance of the methods that we presented in Sections 4.3, 4.4 and 4.5. As mentioned in Section 4.1, the re-substitution method used complete data set in order to find the optimal coefficient which maximizes the AUC.

We have discussed three methods i.e. empirical AUC (distribution-free approach) by Pepe and Thompson [77], NPI AUC without copula and NPI AUC with parametric copula. For NPI AUC with copula method, we use the parametric copulas discussed in Section 2.2 (i.e. Clayton, Frank, Normal and Gumbel copulas). The simulation method for LO1P will be explained below. Generally, the results of the predictive performance in this simulation studies are based on $10,000$ bivariate simulated samples, which are simulated from bivariate normal distributions with different means and correlations for both groups.

Before we explain our procedures for the simulation study for all approaches, we should emphasize that the optimized coefficients, α obtained using all approaches defined in Sections 4.3, 4.4 and 4.5, for empirical method is $\hat{\alpha}$, for NPI without copulas are $\hat{\alpha}_L$ and $\hat{\alpha}_U$, and for NPI with copulas are $\hat{\alpha}_L^c$ and $\hat{\alpha}_U^c$.

For the LO1P simulation method, for each group, for each pair simulated sample size, n_D, the first $n_D - 1$ pairs will be used to find the optimal coefficient α which maximize the AUC value, and the remaining pair n_D from each group, which is considered as a future observation, will be used to test the prediction performance of this method. So, this method uses one observation pair from each group as the validation set and the remaining data as the training set. Using the training data set, considering the weighted average, $T^D_{n_D+1} = \alpha X^D_{n_D+1} + (1-\alpha)Y^D_{n_D+1}$ where $\alpha \in [0,1]$, we find the optimal coefficients for all approaches (i.e. $\hat{\alpha}$, $\hat{\alpha}_L$ and $\hat{\alpha}_U$, and $\hat{\alpha}_L^c$ and $\hat{\alpha}_U^c$). Then, using the optimal coefficient obtained from the training data set, we check the weighted average of the future pair observation for the diseased group is greater than the weighted average of the future pair observation for the non-diseased group or not.

4.6.1 Simulation Results

We have run four cases as defined in Table 4.1, and use the re-substitution and LO1P simulation methods. These cases were chosen based on how much the distributions of diseased and non-diseased groups overlap each other. For Case A, the difference between mean values of diseased and non-diseased for Y is larger than X. For Case B, the difference between mean values of diseased and non-diseased for X is larger than Y. We have similar differences between mean values of diseased and non-diseased for X and Y for Case C. While for Case D, we create this case based on available real data set in literature where the difference mean values for X is larger than Y between the groups. The variances for X and Y for all cases are equal to 1, for the diseased and non-diseased groups. The correlation between X and Y are equal to 0.5 for Cases A - C, for the diseased and non-diseased groups. For the Case D, the correlation between X and Y is equal to 0.14 for the diseased and -0.14 for the non-diseased groups.

Cases	μ_{X^1}	μ_{Y^1}	μ_{X^0}	μ_{Y^0}	ρ^1	ρ^0
Case A	3.00	1.50	2.50	0.50	0.50	0.50
Case B	0.40	0.20	0.00	0.00	0.50	0.50
Case C	0.40	1.00	0.20	1.20	0.50	0.50
Case D	0.44	0.22	-0.78	-0.40	0.14	-0.14

Table 4.1: Scenarios of Simulated Data

As discussed in Section 4.6, in each run of the simulation $10,000$, n_D normal bivariate samples are generated for both groups for Cases A to D. Sample of sizes $(n_1, n_0) = (10, 10), (20, 30), (30, 50), (50, 50), (90, 50)$ are generated for all cases discussed above. The data set is divided into a training (observation) data set and

a pair future observation as a validation data set for each group. For each pair of sample size, we use the re-substitution method (using complete data) and LO1P of simulation method mentioned in Section 4.6. We consider all parametric copulas discussed in Section 2.2 but, we give results from Clayton copula for the NPI with copula approach. As we discussed in Chapter 2, the parametric copulas used do not give many differences. All tables in this simulation study in this section present results from three approaches discussed in Sections 4.3, 4.4 and 4.5. Each table has results from the re-substitution and LO1P simulation methods for each pair of sample size. For the re-substitution method, each table shows the optimal coefficients for all approaches and the corresponding maximized AUC values. For the LO1P simulation method, each table shows the optimal coefficients for all approaches and the corresponding maximize AUC values, and the proportion of cases in which the weighted average of the future pair observation for the diseased group is greater than the weighted average of the future pair observation for the non-diseased group.

Table 4.2 shows the results of the re-substitution method and the LO1P simulations method for Case A. Both methods show that the proposed method, NPI with parametric copula, gives more weight to Y. The difference between mean values of the diseased and non-diseased groups for Y is greater than for X for all sample sizes except for $n_1 = 90$ and $n_0 = 50$. So, it seems that our method put more weight to the variable which provides the descent difference, for small sample sizes. It is always that the AUC value of the empirical method is in between the lower and upper AUC values for NPI without copula as discussed in Section 4.4. For the value in Table 4.2, the AUC value of the empirical method is also bounded by the lower and upper AUC values of NPI with parametric copula, for both simulation methods applied. The lower and upper AUC values for NPI without copula are nested within those for NPI with copula. This simulation do not show a meaningful improvement by including the copula into the NPI approach. This happens due to the fact that the data are simulated from bivariate normal distribution, so the dependence structure is linear and the copula does not has a great chance to take other aspects of dependence in the data.

For the LO1P simulation method, Table 4.2 shows that NPI with parametric

copula works well, for small sizes based on the proportion of the weighted average of the future pair observations for the diseased group is greater than the proportion of the weighted average of the future pair observations for the non-diseased group. However, as the sample size increases (i.e. $n_1, n_0 \geq 50$), NPI with parametric copula does not work so well. This feature reflects the result in Chapter 2 whereby for small sample sizes, the parametric copula work well with the proposed method. While, for larger sample size, this happens due to the misspecification of the copula used for this Case A. As discussed in Section 2.5, the proposed method provides robustness for the predictive inferences which depends on the parametric copulas used, for small sample sizes.

Method	n_1	n_0	Re-substitution method				LO1P simulation method					
			$\hat{\alpha}$		AUC		$\hat{\alpha}$		AUC			$P(T_f^1 > T_f^0)$
Empirical	10	10	0.3148		0.7762		0.3269		0.7773			0.7420
	20	30	0.2151		0.7679		0.2192		0.7679			0.7549
	30	50	0.1699		0.7650		0.1728		0.7652			0.7491
	50	50	0.1519		0.7635		0.1536		0.7636			0.7583
	90	50	0.1291		0.7625		0.1307		0.7625			0.7571
NPI without Copula			$\hat{\alpha_L}$	$\underline{AUC}$	$\hat{\alpha_U}$	$\overline{AUC}$	$\hat{\alpha_L}$	$\underline{AUC}$	$\underline{P}(T_f^1 > T_f^0)$	$\hat{\alpha_U}$	$\overline{AUC}$	$\overline{P}(T_f^1 > T_f^0)$
	10	10	0.3149	0.6415	0.3149	0.8151	0.3266	0.6296	0.7423	0.3266	0.8196	0.7422
	20	30	0.2151	0.7077	0.2151	0.7861	0.2192	0.7052	0.7549	0.2192	0.7869	0.7549
	30	50	0.1699	0.7258	0.1699	0.7770	0.1729	0.7249	0.7491	0.1729	0.7776	0.7491
	50	50	0.1519	0.7338	0.1519	0.7727	0.1536	0.7333	0.7582	0.1537	0.7729	0.7583
	90	50	0.1291	0.7393	0.1291	0.7697	0.1307	0.7390	0.7571	0.1307	0.7699	0.7571
NPI with Clayton Copula			$\hat{\alpha_L^c}$	$\underline{AUC}_c$	$\hat{\alpha_U^c}$	$\overline{AUC}_c$	$\hat{\alpha_L^c}$	$\underline{AUC}_c$	$\underline{P}(T_f^1 > T_f^0)$	$\hat{\alpha_U^c}$	$\overline{AUC}_c$	$\overline{P}(T_f^1 > T_f^0)$
	10	10	0.2540	0.5790	0.2446	0.8289	0.2646	0.5629	0.7457	0.2559	0.8345	0.7463
	20	30	0.1835	0.6789	0.1730	0.7936	0.1867	0.6752	0.7558	0.1763	0.7946	0.7574
	30	50	0.1600	0.7082	0.1516	0.7830	0.1613	0.7068	0.7493	0.1528	0.7836	0.7507
	50	50	0.1532	0.7195	0.1500	0.7776	0.1542	0.7187	0.7560	0.1509	0.7779	0.7568
	90	50	0.1429	0.7271	0.1442	0.7739	0.1433	0.7265	0.7541	0.1448	0.7741	0.7538

Table 4.2: LO1P and re-substitution simulation method for Case A

Table 4.3 shows the results for the re-substitution method and for the LO1P simulation method for Case B. For this case, the difference in mean values is larger for X than for Y, we note that that NPI with parametric copula puts more weight to X compared to other methods. So, as for Case A, our method seem to give some more weight to the variable which is most different between the two groups. For the LO1P simulation method, Table 4.3 shows that the method work well in the sense that the proportion of the weighted average of the future pair observations of the diseased group is greater than the proportion of the weighted average of the future pair observations of the non-diseased group, except for sample size $n_1 = 50$ and $n_0 = 50$.

Method	n_1	n_0	Re-substitution method				LO1P simulation method					
			$\hat{\alpha}$		AUC		$\hat{\alpha}$		AUC			$P(T_f^1 > T_f^0)$
Empirical	10	10	0.5882		0.6447		0.5830		0.6474			0.5939
	20	30	0.6615		0.6287		0.6577		0.6294			0.5958
	30	50	0.7018		0.6225		0.6994		0.6228			0.5963
	50	50	0.7295		0.6202		0.7277		0.6205			0.6005
	90	50	0.7555		0.6175		0.7529		0.6176			0.6009
NPI without Copula			$\hat{\alpha_L}$	$\underline{AUC}$	$\hat{\alpha_U}$	$\overline{AUC}$	$\hat{\alpha_L}$	$\underline{AUC}$	$\underline{P}(T_f^1 > T_f^0)$	$\hat{\alpha_U}$	$\overline{AUC}$	$\overline{P}(T_f^1 > T_f^0)$
	10	10	0.5883	0.5328	0.5883	0.7064	0.5829	0.5244	0.5939	0.5829	0.7144	0.5939
	20	30	0.6615	0.5795	0.6615	0.6578	0.6578	0.5780	0.5956	0.6578	0.6597	0.5957
	30	50	0.7018	0.5906	0.7018	0.6419	0.6994	0.5900	0.5964	0.6995	0.6427	0.5964
	50	50	0.7295	0.5961	0.7295	0.6350	0.7277	0.5959	0.6005	0.7277	0.6355	0.6005
	90	50	0.7554	0.5988	0.7554	0.6291	0.7529	0.5985	0.6009	0.7529	0.6294	0.6009
NPI with Clayton Copula			$\hat{\alpha_L^c}$	$\underline{AUC}_c$	$\hat{\alpha_U^c}$	$\overline{AUC}_c$	$\hat{\alpha_L^c}$	$\underline{AUC}_c$	$\underline{P}(T_f^1 > T_f^0)$	$\hat{\alpha_U^c}$	$\overline{AUC}_c$	$\overline{P}(T_f^1 > T_f^0)$
	10	10	0.6278	0.4863	0.6299	0.7301	0.6197	0.4745	0.5940	0.6217	0.7397	0.5949
	20	30	0.6994	0.5581	0.7012	0.6703	0.6990	0.5557	0.5963	0.7009	0.6726	0.5968
	30	50	0.7347	0.5773	0.7348	0.6508	0.7321	0.5764	0.5983	0.7321	0.6519	0.5984
	50	50	0.7570	0.5854	0.7545	0.6418	0.7562	0.5849	0.5990	0.7536	0.6424	0.5991
	90	50	0.7755	0.5897	0.7706	0.6344	0.7745	0.5892	0.6035	0.7698	0.6348	0.6040

Table 4.3: LO1P and re-substitution simulation method for Case B

Table 4.4 shows the results for the re-substitution method and for the LO1P simulation method for Case C. Both methods show that the proposed method gives more weight to X compared to other methods. In this case, our method seem to give some more weight to variable although we have equal differences between the two groups, but in different direction (i.e. we have positive difference for X and negative difference for Y). This feature occurs due to our $\alpha \in [0, 1]$. A general linear combination of the bivariate diagnostic test results will allow this scenario and this topic is left for future research. For the LO1P simulation method, table 4.4 shows that the proposed method work well if we compare the proportion of the weighted average of the future pair observations of the diseased group is greater than the proportion of the weighted average of the future pair observations of the non-diseased group, for all sample sizes. This shows that the proposed method correctly ordered the future observation(s).

Method	n_1	n_0	Re-substitution method				LO1P simulation method					
			$\hat{\alpha}$	AUC			$\hat{\alpha}$	AUC				
Empirical	10	10	0.6797	0.5658			0.6643	0.5678			0.5194	
	20	30	0.8122	0.5575			0.8074	0.5580			0.5268	
	30	50	0.8801	0.5553			0.8750	0.5554			0.5391	
	50	50	0.9153	0.5554			0.9127	0.5554			0.5448	
	90	50	0.9465	0.5547			0.9437	0.5547			0.5514	
NPI without Copula			$\hat{\alpha_L}$	$\underline{AUC}$	$\hat{\alpha_U}$	$\overline{AUC}$	$\hat{\alpha_L}$	$\underline{AUC}$	$\underline{P}(T_f^1 > T_f^0)$	$\hat{\alpha_U}$	$\overline{AUC}$	$\overline{P}(T_f^1 > T_f^0)$
	10	10	0.6797	0.4676	0.6797	0.6411	0.6642	0.4599	0.5195	0.6642	0.6499	0.5195
	20	30	0.8122	0.5138	0.8123	0.5922	0.8073	0.5124	0.5268	0.8074	0.5941	0.5268
	30	50	0.8801	0.5269	0.8801	0.5781	0.8750	0.5262	0.5391	0.8750	0.5789	0.5391
	50	50	0.9153	0.5338	0.9153	0.5726	0.9127	0.5334	0.5449	0.9127	0.5730	0.5449
	90	50	0.9465	0.5379	0.9465	0.5682	0.9437	0.5376	0.5514	0.9437	0.5685	0.5514
NPI with Clayton Copula			$\hat{\alpha_L^c}$	$\underline{AUC}_c$	$\hat{\alpha_U^c}$	$\overline{AUC}_c$	$\hat{\alpha_L^c}$	$\underline{AUC}_c$	$\underline{P}(T_f^1 > T_f^0)$	$\hat{\alpha_U^c}$	$\overline{AUC}_c$	$\overline{P}(T_f^1 > T_f^0)$
	10	10	0.7551	0.4329	0.7560	0.6754	0.7401	0.4225	0.5265	0.7425	0.6863	0.5271
	20	30	0.8810	0.4982	0.8817	0.6102	0.8747	0.4962	0.5343	0.8760	0.6129	0.5342
	30	50	0.9338	0.5169	0.9336	0.5903	0.9307	0.5160	0.5415	0.9306	0.5914	0.5414
	50	50	0.9582	0.5254	0.9578	0.5812	0.9569	0.5248	0.5510	0.9561	0.5818	0.5508
	90	50	0.9755	0.5300	0.9745	0.5739	0.9748	0.5296	0.5548	0.9738	0.5743	0.5548

Table 4.4: LO1P simulation method for Case C

Table 4.5 shows the results for the re-substitution method and for the LO1P simulation method for Case D. As mentioned earlier, this scenario is created based

on the real data that available in the literature and will be discussed in Section 4.7. Both methods show that the NPI with parametric copula method gives more weight to X compared to other methods for all sample sizes. So, as for Cases A and B, our method seems to give some more weight to the variable which is most different between the two groups. For the LO1P simulation method, the proposed method does not work so well based on the proportion of weighted average of the future observations for diseased is greater than non-diseased groups, is less compared to other methods except for sample size $n_1 = 90$ and $n_0 = 50$. However, these results do not show many differences.

Method	n_1	n_0	Re-substitution method				LO1P simulation method					
			$\hat{\alpha}$		AUC		$\hat{\alpha}$		AUC			$P(T_f^1 > T_f^0)$
Empirical	10	10	0.6206		0.8607		0.6132		0.8626			0.8204
	20	30	0.6609		0.8465		0.6616		0.8469			0.8290
	30	50	0.6681		0.8414		0.6685		0.8417			0.8292
	50	50	0.6714		0.8400		0.6717		0.8402			0.8292
	90	50	0.6697		0.8380		0.6702		0.8381			0.8304
NPI without Copula			$\hat{\alpha_L}$	$\underline{AUC}$	$\hat{\alpha_U}$	$\overline{AUC}$	$\hat{\alpha_L}$	$\underline{AUC}$	$\underline{P}(T_f^1 > T_f^0)$	$\hat{\alpha_U}$	$\overline{AUC}$	$\overline{P}(T_f^1 > T_f^0)$
	10	10	0.6207	0.7113	0.6208	0.8849	0.6129	0.6987	0.8204	0.6129	0.8887	0.8204
	20	30	0.6609	0.7802	0.6609	0.8585	0.6615	0.7778	0.8291	0.6615	0.8594	0.8291
	30	50	0.6681	0.7983	0.6681	0.8496	0.6685	0.7974	0.8292	0.6685	0.8501	0.8292
	50	50	0.6714	0.8074	0.6715	0.8462	0.6718	0.8069	0.8291	0.6718	0.8465	0.8291
	90	50	0.6697	0.8126	0.6697	0.8429	0.6702	0.8122	0.8305	0.6702	0.8431	0.8305
NPI with Clayton Copula			$\hat{\alpha_L^c}$	$\underline{AUC}_c$	$\hat{\alpha_U^c}$	$\overline{AUC}_c$	$\hat{\alpha_L^c}$	$\underline{AUC}_c$	$\underline{P}(T_f^1 > T_f^0)$	$\hat{\alpha_U^c}$	$\overline{AUC}_c$	$\overline{P}(T_f^1 > T_f^0)$
	10	10	0.6803	0.5984	0.6892	0.8871	0.6751	0.5795	0.8132	0.6826	0.8919	0.8125
	20	30	0.6939	0.7198	0.7047	0.8569	0.6954	0.7155	0.8298	0.7066	0.8579	0.8278
	30	50	0.6893	0.7563	0.6979	0.8474	0.6892	0.7544	0.8285	0.6981	0.8479	0.8279
	50	50	0.6887	0.7730	0.6940	0.8438	0.6890	0.7719	0.8283	0.6944	0.8441	0.8280
	90	50	0.6851	0.7839	0.6877	0.8407	0.6851	0.7831	0.8324	0.6877	0.8409	0.8326

Table 4.5: LO1P simulation method for Case D

It seems that for all cases discussed in this section, the simulation suggest that the proposed method works well especially, for small sample sizes. In term of imprecision, NPI with parametric copula seems to lead to more imprecision than NPI without copula, and as the sample size for each group increases, the imprecision is reduced.

As mentioned in Chapter 2, which discussed the NPI with parametric copula, the proposed method requires relatively easy computations, as the use of NPI on the marginals combines naturally with the discretization of the copula. However, the simulation study as represented here is time consuming, because of each simulation, the parameter of the copula must be estimated. This reduced our ability to perform many more simulations for different scenarios, which is left for future research.

The next step in this research is to explore the use of NPI with a nonparametric copula for this ROC curve scenario, and to investigate its performance. Due to time consuming this has not yet been done.

4.7 Example

In this section, an example is presented using a data set from the literature to illustrate the method proposed in this chapter. The data set considers diagnostic markers for pancreatic cancer and consists of 141 patients [98]; 90 pancreatic cancer patients and 51 control group patients with pancreatitis. Two serum markers were measured on these patients, the antigens CA125 and CA19-9 which are positively correlated [77]. To illustrate our approach, we have adjusted the data to avoid tied observations, as discussed in Section 1.2. Let antigen CA19-9 be the X variable and antigen CA125 be the Y variable. In this example, the data are transformed to a natural logarithmic scale as used by Pepe and Thompson in [77]. Then we standardize the data to have mean zero and variance one in order to assist in the interpretation of α as a relative weight of Y to X in the combination. The mean values for X are 0.44 for the diseased group and -0.78 for the non-diseased group, and the mean values for Y are 0.22 for the diseased group and -0.40 for the non-diseased group. The scatter plot of this data set is presented in Figure 4.1.

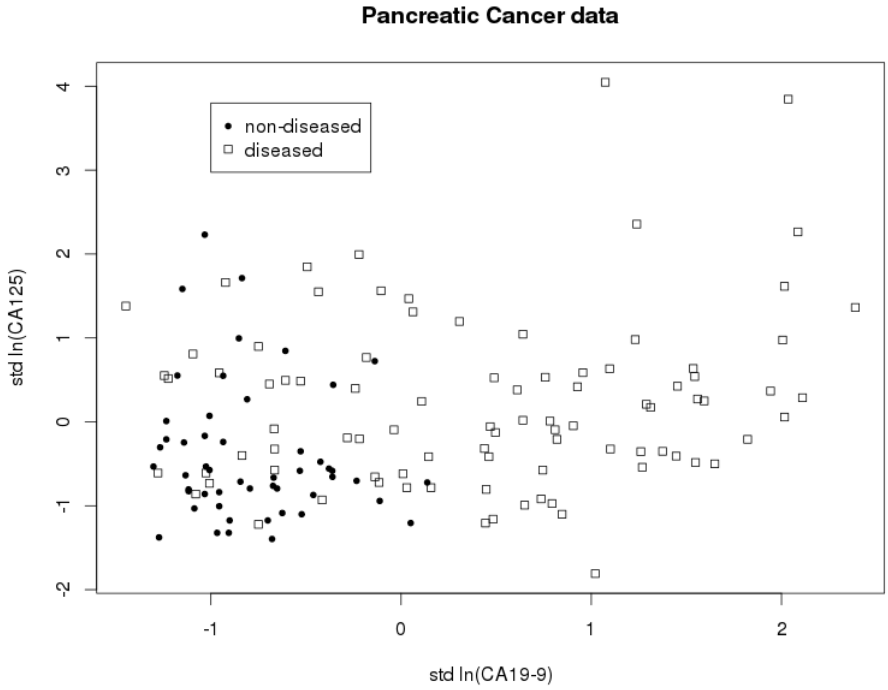


Figure 4.1: Scatter plot for pancreatic cancer data set

The empirical ROC curve and the NPI lower and upper ROC curves, for NPI without copula, are shown in Figure 4.2 for CA19-9 (the X variable), and in Figure 4.3 for CA125 (the Y variable), respectively. The corresponding AUC values are shown in Table 4.6, which shows that the AUC value using the empirical method, if only antigen CA19-9 is used, is 0.8614, and for antigen CA125, it is 0.7056. Using NPI without copula, the lower and upper AUC values for antigen CA19-9 are 0.8347 and 0.8648, respectively. For antigen CA125, these lower and upper AUC values are 0.6883 and 0.7130, respectively. These results illustrate the fact that the AUC value for the empirical method is always in between the lower and upper AUC values for NPI without copula, as shown in Sections 4.4.

Antigen	Empirical AUC	NPI without copula	
		Lower Prob	Upper Prob
CA19-9	0.8614	0.8347	0.8648
CA125	0.7056	0.6883	0.7130

Table 4.6: AUC values for empirical ROC and NPI without copula ROC for antigens CA19-9 and CA125

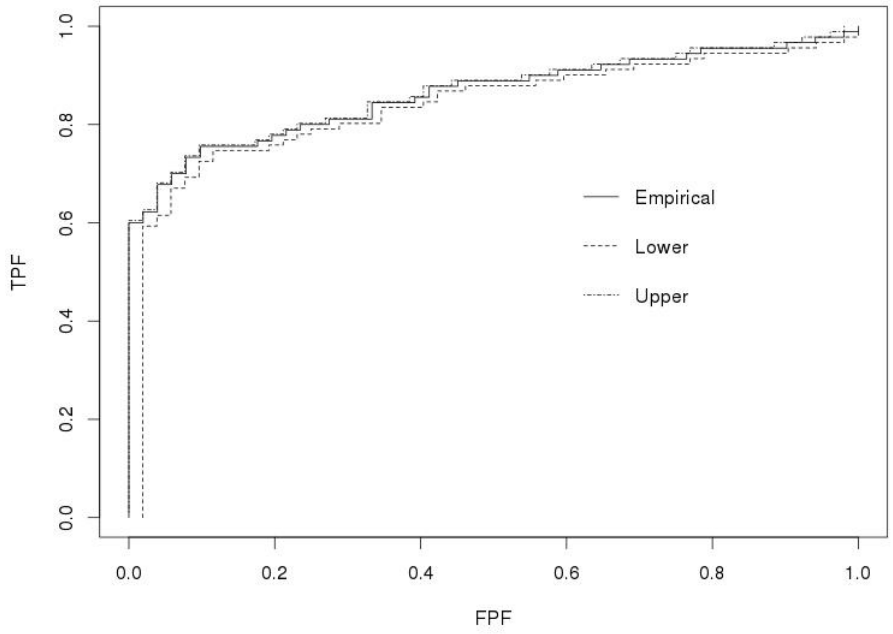


Figure 4.2: ROC curves for X, antigen CA19-9

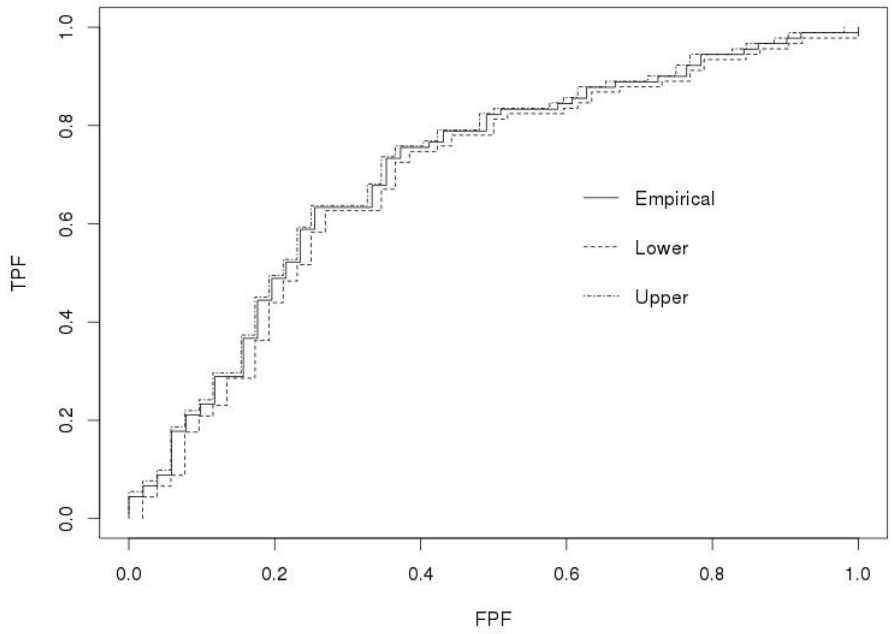


Figure 4.3: ROC curves for Y, antigen CA125

We consider the dependence structure by using parametric copulas, as before we use the Normal, Frank, Clayton and Gumbel copulas. It should be emphasized that any parametric copulas can be used. Consider a weighted average, $T^D = \alpha X^D + (1-\alpha)Y^D$ for empirical method, and $T^D_{n_D+1} = \alpha X^D_{n_D+1} + (1-\alpha)Y^D_{n_D+1}$ for NPI without copula and NPI with copula methods. The optimal coefficients and the corresponding AUC values for all methods are shown in Table 4.7.

	$\hat{\alpha}$		AUC	
Empirical ROC	0.7188		0.8939	
	$\hat{\alpha}_L$	$\underline{AUC}$	$\hat{\alpha}_U$	$\overline{AUC}$
NPI without Copula	0.7188	0.8671	0.7188	0.8971
	$\hat{\alpha}^c_L$	$\underline{AUC}_c$	$\hat{\alpha}^c_U$	$\overline{AUC}_c$
NPI with Normal Copula	0.7160	0.8306	0.7151	0.8896
NPI with Frank Copula	0.7077	0.8324	0.7077	0.8920
NPI with Clayton Copula	0.7066	0.8364	0.7061	0.8947
NPI with Gumbel Copula	0.7215	0.8301	0.7226	0.8880

Table 4.7: AUC values for different methods

For the empirical method, Table 4.7 shows that the optimal $\hat{\alpha}$ is 0.7188 and the corresponding maximum AUC is 0.8939. For NPI without copula, we get $\hat{\alpha}_L = \hat{\alpha}_U = 0.7188$ and the corresponding lower and upper AUC values are 0.8671 and 0.8971, respectively. For NPI with copula, we have different values of $\hat{\alpha}^c_L$, $\hat{\alpha}^c_U$ and the AUC values depending on the choice of copula. The Clayton copula gives the highest lower and upper AUC values compared to the other parametric copulas used, $\underline{\text{AUC}}_c = 0.8364$ and $\overline{\text{AUC}}_c = 0.8947$, with corresponding $\hat{\alpha}^c_L = 0.7066$ and $\hat{\alpha}^c_U = 0.7061$, respectively. This feature occurs due to the data set for diseased and non-diseased groups have a great dependence on the negative tails compared to positive tails, which is captured by the Clayton copula. This can be seen from Figure 4.1, where for each group, small x and y observation values are close to each other compared to large x and y observation values. The second highest of NPI lower and upper AUC values are achieved by the Frank copula and followed by Normal and Gumbel copulas as shown in Table 4.7.

By considering the weighted average in the combination of these two random

quantities, a quite large increment on AUC values for all approaches is achieved as compared to only one test results used. In terms of weighted values, we can see that the NPI with Gumbel copula puts more weight on X compared to the empirical method and NPI without copula, as the difference between mean values of the diseased and non-diseased groups for X greater than Y. We also saw this effect in the simulation study in Section 4.6. We show the ROC curves for all methods for the weighted average discussed above in Figure 4.4. The figure shows the ROC curves for the empirical method, NPI without copula and NPI with Clayton copula. This figure illustrates that the ROC curve for the empirical method is not always bounded by the lower and upper ROC curves for the NPI with parametric copula method.

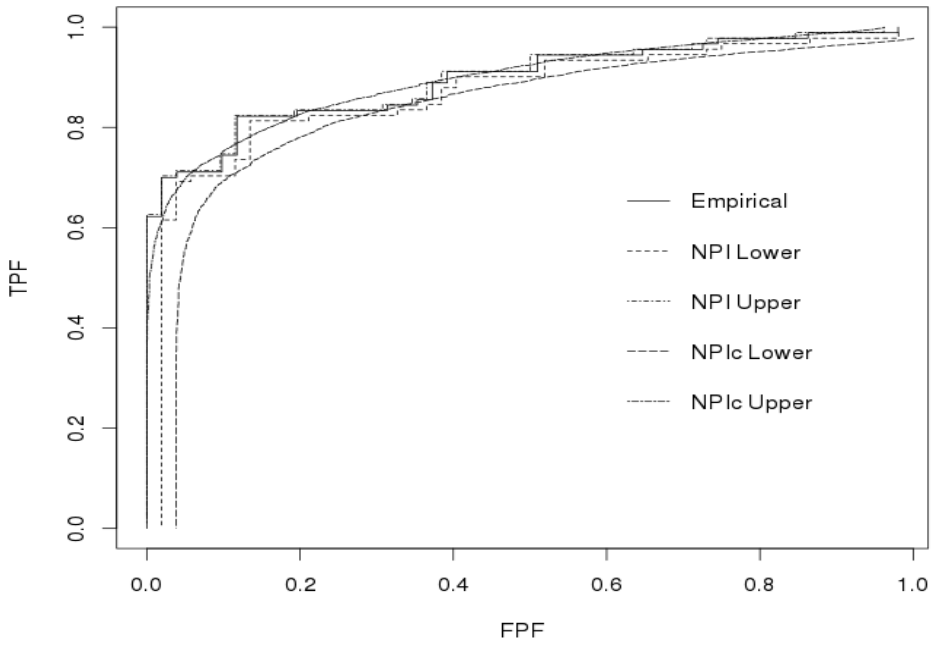


Figure 4.4: ROC curves for weighted average of antigen CA19-9 and CA125

4.8 Concluding remarks

This chapter presents an introduction of NPI for combining two diagnostic test results, aimed at maximizing the area under the ROC curve. We use NPI with

parametric copula, introduced in Chapter 2, and directly apply results in Section 4.2.2, to combine the two test results.

Based on the simulation study, it seems that NPI with parametric copula puts more weight on the variable for which the mean values differ most between the groups. When comparing the AUC values of the empirical method with the NPI without copula method, the AUC values of the empirical method are always in between the NPI lower and upper AUC values (without copula), this does not generally hold for NPI with copula, due to the effect of using a parametric copula and the lower and upper AUC's get associated with different optimized α.

The lower and upper AUC values for NPI without copula are nested within those for NPI with copula. Therefore, these simulations do not show a meaningful improvement by including the copula into the NPI approach. A likely reason for this is the fact that the data are simulated from bivariate normal distributions, so the dependence structure is linear. The use of a linear combination of the two variables may effectively deal with this linear dependence, hence the copula has no further opportunity to pick up other aspects of dependence in the data. We expect that the use of the copula, and particularly nonparametric copulas, in our method will make a positive difference to the ROC approach in this chapter if the underlying data have a nonlinear dependence structure. Due to time constraints for this research project we have not yet been able to investigate this, it is left as an important topic for future research.

In this work we limit the coefficient, $\alpha \in [0, 1]$. We might consider a general linear combination of the two variables and investigate the performance of the proposed method. This gives some more freedom and is likely to give better results in some cases. However, this was a first step to consider the method of NPI with copulas for such inference, we wished to keep the combination simple so that results could be easily interpreted, which a weighted average allows.

Chapter 5

Conclusions

This chapter provides a brief summary of the main results presented in this thesis and some important challenges for future research. In this thesis, we have presented Nonparametric Predictive Inference (NPI) combined with a copula for bivariate data. We discussed the performance of the proposed method with parametric copula and nonparametric copula, specifically kernel-based copula. We introduce NPI for combining two diagnostic test results, by considering a weighted average of the two diagnostic test results directly applying the results in Section 4.2.2, and use NPI with parametric copula as introduced in Chapter 2.

The method presented in this research has a novel aspect within statistical theory using imprecise probabilities. Traditionally, imprecision is used particularly on aspects for which one has relatively little information. Here, however, we use imprecision on the marginals but not on the copula, while the data tend to contain less information about the dependence structure than about the marginals. This is done as the imprecision on the marginals provides robustness with regard to the copula choice, for small to medium sample size, with the added benefit that the imprecise probability method used on the marginals is easy to implement and fits naturally to discretization of the copula. This idea, to add imprecision to the easier part of an inference in order to provide robustness for the harder part, and all together simplifying computation, promises to have wider applicability, for example in big data scenarios where fast computation is crucial. We will explore this idea in other settings in future research.

NPI with a nonparametric copula, specifically kernel-based copula, for bivariate data seems to work well for large data sets. However, the performance depends on the bandwidth selections and types of bandwidths. For each application, for different events of interest and sample size, one should perform a detailed study to investigate the appropriate bandwidth. For future research, the use of other nonparametric copula methods should be considered, combined with the NPI on the marginals. The performance of this proposed method should be studied and investigated. One may also consider other types of dependence structures such as nonlinear dependence structure.

We presented the application of the proposed method in this thesis to a real world scenario, where a combination of bivariate data is relevant. The new method that we introduced for weighted averaging of bivariate diagnostic test results can be used as an alternative to the classic empirical method. The use of nonparametric copula for the weighted average of bivariate diagnostics test results can be considered but the predictive performance of the weighted average in this thesis should be studied and investigated. We left this topic for next research. Many can be done in order to possibly improve the proposed method. We can allow wider general linear combination instead of only weighted average of the bivariate diagnostic test results. Equally important is to study the threshold which corresponds the optimize coefficient given from the proposed method. We left these topics for future research.

It should be emphasized that the attractive frequentist properties of NPI mentioned in Section 1.2, are not claimed to hold generally for the inferences presented in this thesis, due to the assumption of a parametric copula and nonparametric copula which is combined with NPI. If this model assumption would indeed reflect the true underlying data generating mechanism, then the method would adopt the attractive properties, but this, of course, would never be the case in practice. This study could be extended to many different ways of applications such as in wind energy, survival analysis, hydrology, and finance; by considering events in between of the bivariate random quantities and taking dependence between these quantities into account. The proposed method requires easy computations, as the use of NPI on the marginals combines naturally with the discretization of the copula. Hence,

the computational complexity is only with regard to the estimation of the copula parameter, which for the copulas considered in this thesis is a routine procedure for which standard software is available. However, if one requires fast computation, example for real-time predictions, there may be fast computational algorithm available or one could possible create these; this is a interesting topic for future research.

As mentioned before, we restricted attention to a single future observation. One may be interested in multiple future observations, in NPI the inter-dependence of such multiple future observations is taken into account [19]. It will be of interest to develop this bivariate method for multiple future observations. The bivariate method presented in this thesis can straightforwardly be generalized to multivariate data, where the curse of dimensionality [32, 85] implies that the number of data required to get meaningful inferences grows exponentially with the dimension of the data. Application to higher dimensional situations is an important topic for future research.

Bibliography

[1] Augustin T. and Coolen F.P.A. (2004). Nonparametric Predictive Inference and Interval Probability. *Journal of Statistical Planning and Inference*, 124(2), 251–272.

[2] Augustin T., Coolen F.P.A., de Cooman G. and Troffaes M.C.M. (2014). *Introduction to Imprecise Probabilities*. Chichester: Wiley.

[3] Baker R.M. and Coolen F.P.A. (2010). Nonparametric predictive category selection for multinomial data. *Journal of Statistical Theory and Practice*, 4(3), 509–526.

[4] Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4), 387–415.

[5] Bansal A. and Sullivan Pepe M. (2013). When does combining markers improve classification performance and what are implications for practice? *Statistics in Medicine*, 32(11), 1877–1892.

[6] Barnett V. and Lewis T. (1994). *Outliers in Statistical Data*. Chichester: Wiley, 3rd edition.

[7] Bellotti T. and Crook J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36(2), 3302–3308.

[8] Bowman A.W. and Azzalini A. (2004). *Applied Smoothing Techniques for Data Analysis*. Oxford: Clarendon Press.

[9] Bradley A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145–1159.

[10] Breiman L., Meisel W. and Purcell E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, 19(2), 135–144.

[11] Cacoullos T. (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18(1), 179–189.

[12] Charpentier A., Fermanian J.D. and Scaillet O. (2007). The estimation of copulas: Theory and practice. In *Copulas: From theory to application in finance*, (Editor) R. Jörn, pp. 35–62. London: Risk Books.

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

[14] Cherubini U., Luciano E. and Vecchiato W. (2004). *Copula Methods in Finance*. Chichester: John Wiley & Sons.

[15] Clayton D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65, 141–151.

[16] Coolen F.P.A. (1996). Comparing two populations based on low stochastic structure assumptions. *Statistics & Probability Letters*, 29(4), 297–305.

[17] Coolen F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, 36(4), 349–357.

[18] Coolen F.P.A. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15(1-2), 21–47.

[19] Coolen F.P.A. (2011). Nonparametric predictive inference. In *International Encyclopedia of Statistical Science*, (Editor) M. Lovric, pp. 968–970. Berlin, Heidelberg: Springer Berlin Heidelberg.

[20] Coolen F.P.A. and Augustin T. (2005). Learning from multinomial data: A nonparametric predictive alternative to the Imprecise Dirichlet Model. In *ISIPTA'05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications.*, (Editors) F. Cozman, R. Nau and T. Seidenfeld, volume 5, pp. 125–134. SIPTA.

[21] Coolen F.P.A. and Augustin T. (2009). A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories. *International Journal of Approximate Reasoning*, 50(2), 217–230.

[22] Coolen F.P.A., Coolen-Schrijner P. and Yan K.J. (2002). Nonparametric predictive inference in reliability. *Reliability Engineering & System Safety*, 78(2), 185–193.

[23] Coolen F.P.A., Troffaes M.C. and Augustin T. (2011). Imprecise probability. In *International Encyclopedia of Statistical Science*, (Editor) M. Lovric, pp. 645–648. Berlin, Heidelberg: Springer Berlin Heidelberg.

[24] Coolen F.P.A. and Yan K.J. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126(1), 25–54.

[25] Coolen-Maturi T., Coolen F.P. and Muhammad N. (2016). Predictive Inference for Bivariate Data: Combining Nonparametric Predictive Inference for Marginals with an Estimated Copula. *Journal of Statistical Theory and Practice*, (just-accepted).

[26] Coolen-Maturi T., Coolen-Schrijner P. and Coolen F.P.A. (2012). Nonparametric predictive inference for binary diagnostic tests. *Journal of Statistical Theory and Practice*, 6(4), 665–680.

[27] Coolen-Maturi T., Coolen-Schrijner P. and Coolen F.P.A. (2012). Nonparametric predictive inference for diagnostic accuracy. *Journal of Statistical Planning and Inference*, 142(5), 1141 – 1150.

[28] Copas J. and Corbett P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika*, 89(2), 315–331.

[29] Cortes C. and Vapnik V. (1995). Support-vector networks. *Machine learning*, 20(3), 273–297.

[30] De Finetti B. (1974). *Theory of Probability: A Critical Introductory Treatment*. London: Wiley.

[31] Deheuvels P. (1980). Non parametric tests of independence. In *Statistique non Paramétrique Asymptotique: Actes des Journées Statistiques, Rouen, France, Juin 1979*, (Editor) J.P. Raoult, pp. 95–107. Berlin, Heidelberg: Springer Berlin Heidelberg.

[32] Donoho D.L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pp. 1–32. Http://statweb.stanford.edu/ donoho/Lectures/CBMS/Curses.pdf.

[33] Efron B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382), 316–331.

[34] Elkhafifi F.F. (2012). *Nonparametric Predictive Inference for Ordinal Data and Accuracy of Diagnostic Tests*. Ph.D. thesis, Durham University, Durham, UK. Available from www.npi-statistics.com.

[35] Elkhafifi F.F. and Coolen F.P.A. (2012). Nonparametric Predictive Inference for accuracy of ordinal diagnostic tests. *Journal of Statistical Theory and Practice*, 6(4), 681–697.

[36] Embrechts P., Lindskog F. and Mcneil A. (2003). Modelling dependence with copulas and applications to risk management. In *Handbook of Heavy Tailed Distributions in Finance*, (Editor) S.T. Rachev, volume 1, pp. 329 – 384. Amsterdam: North-Holland.

[37] Esteban L.M., Sanz G. and Borque A. (2011). A step-by-step algorithm for combining diagnostic tests. *Journal of Applied Statistics*, 38(5), 899–911.

[38] Frank M.J. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. *Aequationes Mathematicae*, 19, 194–226.

[39] Frees E.W. and Valdez E.A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2, 1–25.

[40] Geenens G., Charpentier A. and Paindaveine D. (2014). Probit transformation for nonparametric kernel estimation of the copula density. *arXiv preprint arXiv:1404.4414*. Http://arxiv.org/abs/1404.4414.

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

[42] Genest C., Ghoudi K. and Rivest L.P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3), 543–552.

[43] Ghosh D. (May 2004). Semiparametic models and estimation procedures for binormal ROC curves with multiple biomarkers. *The University of Michigan Department of Biostatistics Working Paper Series*. Http://biostats.bepress.com/umichbiostat/paper39.

[44] Gijbels I. and Mielniczuk J. (1990). Estimating the density of a copula function. *Communications in Statistics-Theory and Methods*, 19(2), 445–464.

[45] Gumbel E.J. (1960). Distributions des valeurs extremes en plusieurs dimensions. *Publications de l'Institute de Statistíque de l'Université de Paris*, 9, 171–173.

[46] Hand D.J., Daly F., Lunn A.D., McConway K.J. and Ostrowski E. (1994). *A Handbook of Small Data Sets*. London: Chapman & Hall.

[47] Hanley J.A. and McNeil B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3), 839–843.

[48] Hawkins D.M. (1980). *Identification of Outliers*, volume 11. London: Chapman & Hall.

[49] Hayfield T. and Racine J.S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 1–32. R package version 0.60-2 (https://cran.r-project.org/web/packages/np/np.pdf).

[50] Hill B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, pp. 677–691.

[51] Hill B.M. (1988). De Finetti's theorem, induction, and A_n, or Bayesian nonparametric predictive inference (with discussion). In *Bayesian Statistics 3*, (Editors) J.M. Bernando, M.H. DeGroot, D.V. Lindley and A. Smith, pp. 211–241. Oxford University Press.

[52] Hofmann T., Schölkopf B. and Smola A.J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, pp. 1171–1220.

[53] Huang X., Qin G. and Fang Y. (2011). Optimal combinations of diagnostic tests based on AUC. *Biometrics*, 67(2), 568–576.

[54] Jin H. and Lu Y. (2009). The optimal linear combination of multiple predictors under the generalized linear models. *Statistics & Probability Letters*, 79(22), 2321–2327.

[55] Joe H. (1997). *Multivariate Models and Multivariate Dependence Concepts*, volume 73. New Jersey: Chapman & Hall.

[56] Joe H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2), 401–419.

[57] Kang L., Liu A. and Tian L. (2013). Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Statistical Methods in Medical Research*. SAGE Publications.

[58] Kim G., Silvapulle M.J. and Silvapulle P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, 51(6), 2836 – 2850.

[59] Klugman S.A., Panjer H.H. and Willmot G.E. (2012). *Loss Models: From Data to Decisions*, volume 715. New Jersey: John Wiley & Sons.

[60] Koita A., Daucher D. and Fogli M. (2013). Multidimensional risk assessment for vehicle trajectories by using copulas. In *ICOSSAR*, p. 7. France. Https://hal.archives-ouvertes.fr/hal-00865839/document.

[61] Kojadinovic I. and Yan J. (2010). Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insurance: Mathematics and Economics*, 47(1), 52–63.

[62] Lawless J.F. and Fredette M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92, 529–542.

[63] Li Q. and Racine J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2), 266–292.

[64] Li Q. and Racine J.S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

[65] Li X., Mikusiński P., Sherwood H. and Taylor M.D. (1997). On Approximation of Copulas. In *Distributions with given Marginals and Moment Problems*, (Editors) V. Beneš and J. Štěpán, pp. 107–116. Dordrecht: Springer Netherlands.

[66] Lin J. and Wu X. (2015). Smooth tests of copula specifications. *Journal of Business & Economic Statistics*, 33(1), 128–143.

[67] Liu A., Schisterman E.F. and Zhu Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, 24(1), 37–47.

[68] Liu C., Liu A. and Halabi S. (2011). A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, 30(16), 2005–2014.

[69] Loader C.R. (1999). Bandwidth selection: classical or plug-in? *Annals of Statistics*, pp. 415–438.

[70] Loftsgaarden D.O. and Quesenberry C.P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3), 1049–1051.

[71] Maturi T.A. (2010). *Nonparametric Predictive Inference for Multiple Comparisons*. Ph.D. thesis, Durham University, Durham, UK. Available from www.npi-statistics.com.

[72] Muhammad N., Coolen F.P.A. and Coolen-Maturi T. (2015). Predictive inference for bivariate data with nonparametric copula. *AIP Conference Proceedings*. To appear.

[73] Nelsen R.B. (2007). *An introduction to copulas*. New York: Springer Science & Business Media.

[74] Parzen E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3), 1065–1076.

[75] Pepe M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

[76] Pepe M.S., Cai T. and Longton G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1), 221–229.

[77] Pepe M.S. and Thompson M.L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2), 123–140.

[78] Powell M.J. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2), 155–162.

[79] Purcaru O. (2003). Semi-parametric Archimedean copula modelling in actuarial science. *Insurance, Mathematics and Economics*, 33, 419–420.

[80] Rank J. (2007). *Copulas: From Theory to Application in Finance*. London: Risk Books.

[81] Rudemo M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pp. 65–78.

[82] Scaillet O. and Fermanian J.D. (2002). Nonparametric estimation of copulas for time series. *FAME (Financial Asset Management and Engineering) Research Paper*, (57).

[83] Schepsmeier U., Stoeber J. and Brechmann E.C. (2013). *VineCopula: Statistical inference of vine copulas*. R package version 1.1-1 (http://www2.uaem.mx/r-mirror/web/packages/VineCopula/VineCopula.pdf).

[84] Schultz M., Eskin E., Zadok E. and Stolfo S. (2001). Data mining methods for detection of new malicious executables. In *Security and Privacy, 2001. S P 2001. Proceedings. 2001 IEEE Symposium on*, pp. 38–49.

[85] Scott D.W. (2009). *Multivariate Density Estimation: Theory, Practice, and Visualization*, volume 383. New York: John Wiley & Sons.

[86] Sen K.O.P.K. (2003). Copulas: Concepts and novel applications. *Metron*, 61(3), 323–353.

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

[88] Silverman B.W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26. London: CRC press.

[89] Sklar A.W. (1959). Fonctions de répartitionà n-dimension et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

[90] Staniswalis J., Messer K. and Finston D. (1991). Kernel estimators for multivariate smoothing. Technical report, Department of Statistics, Stanford University, Stanford California. Https://statistics.stanford.edu/sites/default/files/OLK

[91] Su J.Q. and Liu J.S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424), 1350–1355.

[92] Tang X.S., Li D.Q., Zhou C.B. and Zhang L.M. (2013). Bivariate distribution models using copulas for reliability analysis. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 227(5), 499–512.

[93] Trivedi P.K. and Zimmer D.M. (2005). Copula Modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1, 1–111.

[94] Tsukahara H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3), 357–375.

[95] Vexler A., Liu A. and Schisterman E.F. (2006). Efficient design and analysis of biospecimens with measurements subject to detection limit. *Biometrical Journal*, 48(5), 780–791.

[96] Wand M.P. and Jones M.C. (1994). *Kernel Smoothing*, volume 60. CRC Press.

[97] Wen K. and Wu X. (2015). Transformation-kernel estimation of the copula density. *Working Paper, Department of Agricultural Economics, Texas A&M University.* Http://agecon2.tamu.edu/people/faculty/wu-ximing/agecon2/public/copula.pdf.

[98] Wieand S., Gail M.H., James B.R. and James K.L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76(3), 585–592.

[99] Yan L., Tian L. and Liu S. (2015). Combining large number of weak biomarkers based on AUC. *Statistics in Medicine*, 34(29), 3811–3830.

[100] Zou K.H., Liu A., Bandos A.I., Ohno-Machado L. and Rockette H.E. (2011). *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. Boca Raton: CRC Press.