Kendall's tau for multivariate non-continuous data

Kuresevic, Ida

Award date:
2022

Department of Mathematics and Computer Science

Kendall's tau for multivariate non-continuous data

Bachelor Thesis

Ida Kuresevic

Supervisors:
dr. Elisa Perrone
dr. Zhuozhao Zhan

Eindhoven, August 2022

Contents

Chapter 1

Introduction

1.1 Background

With the vast amount of information and thus data that is available for collection and analysis in the world, we are often hoping to extract some information from it. This can help with the understanding and prediction of the behaviour of the gathered data. One such pattern which can be researched is the association of random vectors. Association can tell us whether and how two variables are related. That is, two variables can be positively or negatively associated. For instance, human height and weight are two strongly positively associated measures as people tend to be heavier the taller they are. Contrary to this, comparing lung capacity in smokers and non-smokers shows that the *more* a person smokes, the *lower* their lung capacity becomes. This is an example of negative association.
The strength of association is expressed by a correlation coefficient which usually ranges between -1 and 1. The closer the correlation coefficient is to 1 the stronger positively associated the random variables are. The same holds for negative values and negative association. Additionally, if the coefficient is close to 0, one can assume that the random variables are independent. For instance, frequency of smoking and height can be considered independent, as it does not hold (in the adult population) that taller people smoke any more or less than shorter people.
There are multiple different measures of association. Among the most widely used association measures are the Pearson correlation coefficient, Spearman's rho and Kendall's tau. The focus of this thesis is the rank correlation coefficient proposed by Kendall and first defined in his 1938 paper [1]. We chose Kendall's tau as it is the least researched out of the three coefficients above. Furthermore, it is not limited to the analysis of normally distributed random variables like Pearson's coefficient. Kendall's tau also measures monotonic relationships like Spearman's rho, rather than linear ones like Pearson's correlation. However, compared to Spearman's rho, Kendall's tau is more robust and often preferred.
An example of where and how Kendall's tau can be useful follows below.

1.1.1 Motivating example

In the pharmaceutical industry, a type of unit called a cleanroom is used for medicine production. These rooms are only allowed to have a very low number of airborne particles [2]. This needs to be the case for the medicinal products to be authorized and distributed. To verify that the necessary conditions are met the particle frequency is closely monitored. The measuring machines inside the cleanrooms relay the number and size of particles present at different times. In particular, the data will consist of particle counts for multiple size intervals over time [2]. The information retained at any specific time point corresponds to one sample. One can note that as count data is used, the random variable is not continuous. Also, depending on how many size intervals are considered the data is not bivariate but rather has more than two variables.

In order to meet the conditions of the clean rooms, it needs to be checked for every count that it does not exceed the allowed number at any time. There are several possible ways to do this. One could simply look at the measurements taken by the monitoring machines, however, the problem with this is the machines' efficiency. The machines which are currently used, take nearly a week before they can return any measurements. This delays the distribution of medical products significantly. Recently, newer versions are appearing on the market which can return near real-time results. However, these machines are not able to count viable particles, that is particles which contain living microorganisms and are those affecting the product authorization. Another method is to model the number of particles or their size distribution. This proves to be difficult as the particles detected are a mixture of different types of particles over a large size range. It is also not ideal to fit the data to a continuous distribution as was shown in [2].
A different approach to gain information on this particle behaviour could be to see if the frequencies of differently sized particles relate by calculating a correlation coefficient such as Kendall's tau.

1.2 Problem Statement

The problem in cases like the above is that, originally, Kendall's tau is defined to work only on bivariate continuous random variables. However, as is shown in the examples, it can be useful to also find Kendall's tau for other types of random variables. Thus the thesis aims to answer the following main research question:

Which methods can be identified to compute Kendall's tau for multivariate non-continuous random variables and how do they perform?

As already noted, Kendall's tau is defined for bivariate continuous data. In particular, Kendall's tau is defined by using concordance. The latter is a measure which allows classifying whether random variables are associated. More on this can be found in Chapter 2. This definition of concordance is only meant for bivariate continuous random variables. So through this, the application of Kendall's tau is also limited.
However, as Kendall's tau is a popular measure of association, there already exist several expansions and adaptations for it. These provide a value for Kendall's tau for random variables which are not necessarily continuous or bivariate. Some examples of this are Kendall's own adaptation τ_b which accounts for ties [3] or Pimentel's suggestion for zero-inflated random variables [4] and more accurate extension based on Pimentel's, such as τ_{PHZ} [5]. For the multivariate continuous case, some options are taking the average of pairwise taus [6] or Joe's extension from his 1990 paper [7].
In this thesis, we analyse these different methods of extending Kendall's tau and define new methods for the calculation of Kendall's tau for multivariate random variables with ties. In particular, zero-inflated random variables are taken into consideration. Without loss of generality, we will focus on the trivariate case. Thus, we assume there are three non-continuous random variables for which their overall correlation/association wants to be determined. These three random variables (X_1, X_2, X_3) can create three different pairs (X_1 & X_2, X_1 & X_3, X_2 & X_3). For these three pairs, it is possible to calculate the "pairwise" bivariate Kendall's tau. The first part to look at is how these three different pairwise Kendall's taus relate to the adapted multivariate Kendall's tau found from all three random variables simultaneously. As an intermediary step one can start with the continuous case. Here Joe's expansion for the multivariate tau as well as the average were chosen to be compared to. The second step is to see how ties in a variable affect the value of correlation in the multivariate case. That is whether having a tie might lower the resulting value of Kendall's tau compared to not having the tie for example. We also look at whether it is possible to discern different types of ties by their effects on the resulting value. This entails both the frequency and position of ties in the random variable. The outcome of these two points of interest then allows us to showcase how one could calculate a correlation coefficient such as Kendall's tau from trivariate random vectors containing ties.

In order to reach the main research question, some sub-questions were formulated to help along the way and stay on track.

- How does the matrix of pairwise Kendall's taus for a continuous multivariate random vector relate to existing methods of computations for multivariate Kendall's tau?
- How does the introduction of ties influence the value of Kendall's tau? Do these effects differ for various types of ties?
- How can one define Kendall's tau for multivariate random variables with ties?
- How do these methods perform in a simulated setting?

We outline the structure of the thesis. In Chapter 2, the main definitions and notation used throughout the thesis are introduced, including Kendall's tau and some of its relevant extensions. In Chapter 3, the behaviour of these extensions is analysed to determine how the problems of multivariate and non-continuous random variables are approached separately. With this information, suggestions of methods which could be used to calculate Kendall's tau for non-continuous multivariate random variables are given. Next, these suggested methods are evaluated in Chapter 4 and applied on synthetic data in Chapter 5. Finally, we summarize our findings and discuss possible future directions in Chapter 6.

Chapter 2

Basic Concepts

2.1 Notation

We introduce the notation used throughout the paper.
Let $X = (X_1, X_2, ..., X_d)$ be a random vector with cumulative distribution function (cdf) $F = (F_1, F_2, ..., F_d)$. In particular, X_i is a random variable with marginal cdf F_i, for all $i = 1, ..., d$. If there are n samples from this random vector the j-th sample is denoted by $X^{(j)} = (X_{j1}, ..., X_{jd})$, for $j = 1, ..., n$. Kendall's tau is generally denoted by τ but to differentiate the results obtained from the various estimators, a subscript is added.

2.2 Concordance

As mentioned in the introduction, Kendall's tau is defined by the concept of concordance.
Let $X = (X_1, X_2)$ be a pair of continuous random variables. Two observations from this random vector $X^{(1)} = (X_{11}, X_{12})$ and $X^{(2)} = (X_{21}, X_{22})$ are said to be concordant if either

$$(X_{11} < X_{21} \text{ and } X_{12} < X_{22}) \quad \text{OR} \quad (X_{11} > X_{21} \text{ and } X_{12} > X_{22}).$$

Otherwise they are discordant, that is if

$$(X_{11} < X_{21}, \text{ but } X_{12} > X_{22}) \quad \text{OR} \quad (X_{11} > X_{21}, \text{ but } X_{12} < X_{22}).$$

This property can also be formulated in terms of the sign function,

$$\text{sgn}(x) := \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Then the pair of observations is concordant if $\text{sgn}(X_{11} - X_{21}) = \text{sgn}(X_{12} - X_{22})$. By the same logic it is discordant if $\text{sgn}(X_{11} - X_{21}) = -\text{sgn}(X_{12} - X_{22})$.
This concept can also be visualized in a two-dimensional graph (figure 2.1). Imagine any sample observation $X^{(j)} = (X_{j1}, X_{j2})$ on the plane. One can draw two lines parallel to the axes which pass through that point, effectively dividing the plane into four parts. Now, any other observation that is either in the upper right or lower left part is concordant with that point. If it is in one of the other two regions, they are discordant.

Consider the following example with the three observations

$$X^{(1)} = (X_{11}, X_{12}) = (5, 3),\ X^{(2)} = (X_{21}, X_{22}) = (7, 4),\ X^{(3)} = (X_{31}, X_{32}) = (3, 8).$$

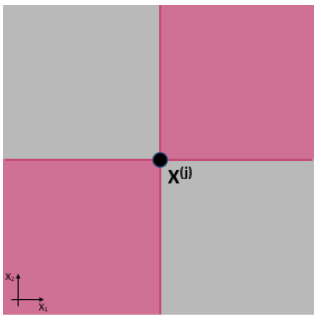

Figure 2.1: In relation to $X^{(j)}$, the points in the pink area are concordant, whilst the ones in the grey part are discordant.

We can see that $X^{(1)}$ and $X^{(2)}$ are concordant as

$$X_{11} = 5 < 7 = X_{21} \text{ and } X_{12} = 3 < 4 = X_{22} \quad \text{or alternatively} \quad \text{sgn}(5-7) = -1 = \text{sgn}(3-4).$$

On the other side, $X^{(1)}$ and $X^{(3)}$ are discordant as

$$X_{11} = 5 > 3 = X_{31}, \text{ but } X_{12} = 3 < 8 = X_{32} \quad \text{or alternatively} \quad \text{sgn}(5-3) = 1 = -\text{sgn}(3-8).$$

Similarly, $X^{(2)}$ and $X^{(3)}$ are also discordant.

2.3 Kendall's Tau

Kendall's tau is a measure of association based on concordance and discordance. More specifically, it is defined as the difference between the probabilities of concordance and discordance. That is, the probabilities that any given pair, sampled from the distribution of the considered random vector, is concordant or discordant. As these probabilities lie between zero and one, the resulting value for Kendall's tau is between -1 and 1, for continuous bivariate random variables. The higher the value is, the stronger the concordance is. Conversely, if it is negative, the variable is more discordant. The definition for Kendall's tau can be written as follows.

$$\tau = \mathbb{P}(\text{concordance}) - \mathbb{P}(\text{discordance})$$

The sample version can thus be found by counting how many of all the possible pairs which can be formed are concordant [1]. Assuming there are n observations, one needs to check $\binom{n}{2}$ distinct pairs. It is given by

$$\tau = \frac{(\#\text{ of concordant pairs}) - (\#\text{ of discordant pairs})}{(\text{total }\#\text{ of pairs})} = \frac{C-D}{P}$$

where C and D are the number of concordant and discordant pairs respectively, and P is the total number of possible pairs $\binom{n}{2}$.

Consider the extended example from before, which can be seen in Table 2.1. There are a total of $P = \binom{5}{2} = 10$ distinct pairs which need to be checked for concordance.
We have already determined that the first pair is concordant with the second and discordant with the third. One can also see that it is concordant with the fourth and fifth as well. Doing this comparison for each pair we find that out of the ten pairs seven are concordant and three are discordant. So Kendall's tau is equal to $\tau = (C-D)/P = (7-3)/10 = 0.4$. Thus the imagined data here is more concordant than discordant.

$X^{(j)}$	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$
X_{j1}	5	7	3	8	9
X_{j2}	3	4	8	6	9

Table 2.1: First example

2.3.1 Non-continuous random variables

The definition above only works for continuous random variables where the probability of repeated values, that is ties, is zero. If ties are present in a comparison of pairs, that pair cannot be classified as either concordant or discordant. This limits the use of Kendall's tau as often the probability of ties is not zero. Kendall thus adapted his definition of the measure for application to data including ties [3]. The adapted tau is given in the following expression where C, D and P are the same as above.

$$\tau_b = \frac{C - D}{\sqrt{(P - T_1)(P - T_2)}}$$

where

$$T_1 = \sum_i \frac{t_{i*}(t_{i*} - 1)}{2} \quad , \quad T_2 = \sum_i \frac{t_{*i}(t_{*i} - 1)}{2}$$

with t_{i*} the number of ties in the i-th group of ties for the first quantity and t_{*i} the number of ties in the i-th group of ties for the second quantity.
In the above method, it can be seen that the numerator did not change from the original definition. It once again contains the difference between the number of concordant and discordant pairs. However, the denominator changes; one does not divide by the total number of possible pairs anymore. Instead, for both quantities, only the number of unique pairs is counted. The denominator is then given by an approximate middle ground of unique pairs for both values. This middle value is the square root of their products. Note that this does not take into account whether a pair has a tie in only one or both values.

Consider the slightly adapted example from above (Table 2.2).

$X^{(j)}$	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$
X_{j1}	5	7	3	9	9
X_{j2}	3	4	9	6	9

Table 2.2: Second example

In this example we have $C = 5, D = 3, P = 10$. Furthermore, $T_1 = T_2 = 1$. This is because in the first row, i.e. for the first quantity, there is only one group of ties - the nines - and there are two of them so $T_1 = \frac{t_{1*}(t_{1*}-1)}{2} = \frac{2(2-1)}{2} = 1$. It works the same for T_2. So $\tau = (5-3)/\sqrt{(10-1)(10-1)} = 0.\bar{2}$.

Zero-inflated random variables
A particular type of data which has a significant amount of ties is zero-inflated random variables. For random variables stemming from such distributions, many observations will be equal to zero, creating ties. With such a significant amount of ties, τ_b ceases to give accurate results. This is due to the fact that all pairs are treated equal, no matter in how many values they are tied. There are several extensions for Kendall's tau which concentrate on this issue, in particular from Pimentel [4], [8]. For this project we consider the most recent extension on Pimentel's work [5]. The measure is defined by

$$\tau_{PHZ} = p_{11}^2 \tau_{11} + 2(p_{00}p_{11} - p_{01}p_{10}) + 2p_{11}[p_{10}(1 - 2p_1^* - p_1^\dagger) + p_{01}(1 - 2p_2^* - p_2^\dagger)].$$

In general, the subscripts of the different probabilities p reference the states of the two random variables. Their definitions are as follows (for $i \neq j$)

$$\begin{aligned} p_{00} &= \mathbb{P}(X_1 = 0, X_2 = 0), & p_{10} &= \mathbb{P}(X_1 > 0, X_2 = 0), \\ p_{01} &= \mathbb{P}(X_1 = 0, X_2 > 0), & p_{11} &= \mathbb{P}(X_1 > 0, X_2 > 0). \\ p_1^* &= \mathbb{P}(X_{j1} > X_{i1} | X_{j2} = 0 \wedge X_{i2} > 0), & p_2^* &= \mathbb{P}(X_{j2} > X_{i2} | X_{j1} = 0 \wedge X_{i1} > 0). \\ p_1^\dagger &= \mathbb{P}(X_{j1} = X_{i1} | X_{j2} = 0 \wedge X_{i2} > 0), & p_2^\dagger &= \mathbb{P}(X_{j2} = X_{i2} | X_{j1} = 0 \wedge X_{i1} > 0). \end{aligned}$$

Finally, we have τ_{11}, which is equal to the standard Kendall's tau for the reduced data set where both X_1 and X_2 are not zero. If margins have ties away from zero as well τ_b can be used here. In the definition given above one can see that the calculation of Kendall's tau is done in two parts. In the first, the leading term, the standard Kendall's tau is calculated for the values which are away from zero, and thus less likely to be tied. In the second part, the estimator accounts for the ties in zero but also within the margins ($p^\dagger$).

2.3.2 True Kendall's tau for count data

In the previous section, two extensions which estimate Kendall's tau for bivariate count data are described. However, with some additional underlying knowledge on the random variables for which Kendall's tau needs to be calculated one can find the true value with the definition from Nikoloulopoulos and Karlis [9]. This definition is based on the true tau definition for the original Kendall's tau and is best given in terms of copulas so this concept is first described below.

Copulas are a type of function which connect the marginal distribution functions of multiple random variables to their joint distribution function. Let there be the random vector $X = (X_1, X_2)$, with marginal cdf's F_1 and F_2, and joint cdf H. Sklar's theorem [10], which can be found in Nelsen's book [11], then states the following.

Theorem (Sklar) *Given a joint distribution function $H : \mathbb{R} \times \mathbb{R} \to [0, 1]$, having marginal distribution functions F_1 and F_2, there exists a copula C which connects H to F_1 and F_2 via*

$$\text{for all } x_1, x_2 \in \mathbb{R} \qquad H(x_1, x_2) = C(F_1(x_1), F_2(x_2)).$$

The copula C is (uniquely) determined on Ran $F_1 \times$ Ran F_2, where

$$\text{Ran } F_1 := \{y \in [0, 1] : \exists x \in \mathbb{R}, F_1(x) = y\}.$$

The above also holds for more than two random variables. In Nelsen's book the following statement is proven [12]. Let there be the vectors $X^{(1)} = (X_{11}, X_{12})$ and $X^{(2)} = (X_{21}, X_{22})$ with copulas C_1 and C_2, respectively. Then one can define Q to be the difference between the probabilities of concordance and discordance of $X^{(1)}$ and $X^{(2)}$. So it is given by

$$Q = \mathbb{P}((X_{11} - X_{21})(X_{12} - X_{22}) > 0) - \mathbb{P}((X_{11} - X_{21})(X_{12} - X_{22}) < 0).$$

It is shown that the above expression is equal to

$$Q = Q(C_1, C_2) = 4 \int \int_{I^2} C_2(u, v) dC_1(u, v) - 1.$$

By further theorems it is then shown that the original Kendall's tau for $X = (X_1, X_2)$ with copula C can also be written as [12]

$$\tau = Q(C, C) = 4 \int \int_{I^2} C(u, v) dC(u, v) - 1.$$

This holds as any sample $X^{(i)}$ that stems from the distribution of X has the same copula C. However, once again due to the addition of ties in the discrete case, this definition is not applicable.

This is because, as Sklar's theorem already states, the copula is not uniquely determined in the non-continuous case. Instead, Nikoloulopoulos and Karlis offer an alternative definition for the discrete case in their 2009 paper [9]. Let once again X_1 and X_2 be two discrete random variables with marginal cdf's F_i and pmf's f_i for $i = 1, 2$. Also let the associated copula be denoted by C. Their definition of Kendall's tau for X_1 and X_2 is given by

$$\tau(X_1, X_2) = \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} h(x_1, x_2)[4C(F_1(x_1 - 1), F_2(x_2 - 1)) - h(x_1, x_2)] + \sum_{x_1=0}^{\infty} (f_1^2(x_1) + f_2^2(x_1)) - 1$$

where we have that

$$h(x_1, x_2) = C(F_1(x_1), F_2(x_2)) - C(F_1(x_1 - 1), F_2(x_2)) - C(F_1(x_1), F_2(x_2 - 1)) + C(F_1(x_1 - 1), F_2(x_2 - 1))$$

is the joint probability mass function of X_1 and X_2.

2.3.3 Multivariate random vectors

Furthermore, whereas Kendall did give an adaptation of his association measure to account for ties, he did not suggest an expansion for multivariate random variables. This also partially stems from the fact that concordance is only defined for bivariate samples. The most prevalent and intuitive way to define concordance for multivariate random variables is to say that two samples are concordant if the same inequality holds for *all* their components, i.e. either $X_{1i} > X_{2i}$ or $X_{1i} < X_{2i}$ for all $i = 1, 2, ..., d$, where d is the dimension of the random vector [13].
This creates an imbalance between the probability of concordance and discordance. Recall that previously it was equally likely for a random pair to be concordant and discordant, as could be seen in the visualization graph (figure 2.1). Now consider only one dimension higher of a trivariate distribution with $X = (X_1, X_2, X_3)$. Two samples $X^{(i)}$ and $X^{(j)}$ are considered concordant only in the two following cases

- $X_{i1} < X_{j1} \quad \wedge \quad X_{i2} < X_{j2} \quad \wedge \quad X_{j3} < X_{i3}$
- $X_{i1} > X_{j1} \quad \wedge \quad X_{i2} > X_{j2} \quad \wedge \quad X_{j3} > X_{i3}$

However, this means that they are discordant for all the remaining six possible combinations.

- $X_{i1} < X_{j1} \quad \wedge \quad X_{i2} < X_{j2} \quad \wedge \quad X_{j3} > X_{i3}$
- $X_{i1} < X_{j1} \quad \wedge \quad X_{i2} > X_{j2} \quad \wedge \quad X_{j3} < X_{i3}$
- $X_{i1} > X_{j1} \quad \wedge \quad X_{i2} < X_{j2} \quad \wedge \quad X_{j3} < X_{i3}$
- $X_{i1} > X_{j1} \quad \wedge \quad X_{i2} > X_{j2} \quad \wedge \quad X_{j3} < X_{i3}$
- $X_{i1} > X_{j1} \quad \wedge \quad X_{i2} < X_{j2} \quad \wedge \quad X_{j3} > X_{i3}$
- $X_{i1} < X_{j1} \quad \wedge \quad X_{i2} > X_{j2} \quad \wedge \quad X_{j3} > X_{i3}$

Once again the regions of concordance and discordance can be visualized in a graph, this time three-dimensional (figure 2.2). Already the likelihood of a random triplet being discordant to $X^{(j)}$ is three times as high as being concordant. Naturally, this imbalance only increases with higher dimensions.

This shows that multivariate expansions, such as the two methods described below, are not as trivial as they might seem.

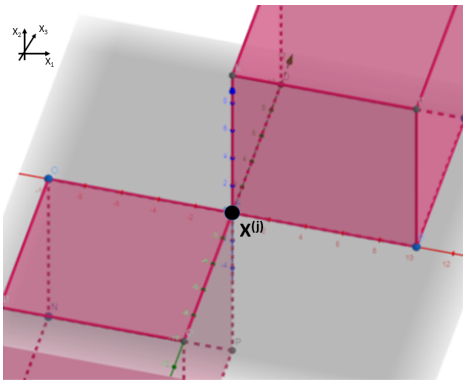

Figure 2.2: In relation to $X^{(j)}$, the points in the pink area are concordant, whilst the ones in the grey part are discordant.

Averaging over Pairwise taus

One possible extension for a random vector $X = (X_1, ..., X_d)$ is to compute Kendall's tau for every combination of two of these random variables. This can be used to construct Kendall's matrix K with the entries being the pairwise taus $k_{i,j} = \tau(X_i, X_j), 1 \leq i, j \leq d$. As properties of the matrix we find that it is symmetric as indeed $\tau(X_i, X_j) = \tau(X_j, X_i)$. Furthermore, the values along the main diagonal equal to one as two equal samples are naturally completely concordant, so $\tau(X_i, X_i) = 1$ (at least in the continuous case).
The matrix is thus given by

$$K = \begin{pmatrix} 1 & \tau(X_1, X_2) & \cdots & \tau(X_1, X_d) \\ \tau(X_1, X_2) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tau(X_{d-1}, X_d) \\ \tau(X_1, X_d) & \dots & \tau(X_{d-1}, X_d) & 1 \end{pmatrix}$$

This method gives some information but it would also be useful to get an overall value for the entire multivariate random vector. There are only a couple of conclusions which seem straightforward to assume about the multivariate tau when looking at K. First of all, if all values are positive, that is every possible pair of random variables is concordant to one another, then the multivariate tau should also be positive. The same holds if they are all - apart from the diagonal - negative, i.e. discordant.

The most straightforward method for using this matrix of pairwise Kendall's taus is to take the average of all distinct pairwise taus [6]. For a d-variate random vector there are $\binom{d}{2}$ distinct pairs that can be formed from the random variables. Thus the average is as usual the sum of all pairwise taus divided by the number of possible pairs:

$$\tau_{avg}(X_1, ..., X_d) = \frac{1}{\binom{d}{2}} \sum_{1 \leq i < j \leq d} \tau(X_i, X_j)$$

For the sample version, one can simply substitute the bivariate Kendall's tau τ in the formula with any of the desired estimates, depending on the type of random variable.

For a trivariate random vector, this can be applied as follows.

$$d = 3. \qquad \tau_{avg}(X_1, X_2, X_3) = \frac{1}{\binom{3}{2}} \sum_{1 \leq i < j \leq 3} \tau(X_i, X_j) = \frac{1}{3}(\tau(X_1, X_2) + \tau(X_1, X_3) + \tau(X_2, X_3))$$

Joe's extension

Joe defines a multivariate version of Kendall's tau along with giving a sample version [7]. The definition for Joe's multivariate tau for $X = (X_1, ..., X_d)$ is given by

$$\tau_J(X^{(i)}, X^{(j)}) = \sum_{k=[(d+1)/2]}^{d} w_k \mathbb{P}((X_{i1} - X_{j1}, ..., X_{id} - X_{jd}) \in B_{k,d-k})$$

Above, the square brackets denote rounding down of the value contained. Furthermore, $B_{k,d-k}$ is the set of vectors of length d for which k terms are positive and the remaining $d-k$ are negative, or vice versa. In mathematical terms

$$B_{k,d-k} = \{(a_1, ..., a_d) \in \mathbb{R}^n | (k \times a_i < 0 \wedge d - k \times a_i > 0) \vee (k \times a_i > 0 \wedge d - k \times a_i < 0)\}$$

This is essentially separating the probabilities of two samples being concordant ($k = d$), from the probabilities for different types of discordant pairs. Note as the set is "symmetric", i.e. $B_{k,d-k} = B_{d-k,k}$, the sum only needs to start at the halfway point $[(d+1)/2]$ in order to avoid double counting.
The coefficients are defined as

$$w_k = \sum_{l=2}^{d} \beta_l (c_{lk} - \frac{d_{lk}}{2^{l-1} - 1}), \qquad \text{with} \quad c_{lk} = \binom{k}{l} + \binom{d-k}{l}, \quad d_{lk} = \binom{d}{l} - c_{lk}$$

and β_l constants. Furthermore, there are several properties which should hold for w_k and can be used to calculate the different β's:

(i) $\quad w_k = w_{d-k}$

(ii) $\quad w_d \geq w_{d-1} \geq ... \geq w_{d'}$, where $d' = \lfloor \frac{d+1}{2} \rfloor$

(iii) $\quad w_0 = w_d = 1$

(iv) $\quad \begin{cases} \sum_{k=d'}^{d} w_k \binom{d}{k} = 0 & \text{, if } d \text{ is odd} \\ 2\sum_{k=d'+1}^{d} w_k \binom{d}{k} + w_{d'} \binom{d}{d'} = 0 & \text{, if } d \text{ is even} \end{cases} \qquad or \ \sum_{k=0}^{d} w_k \binom{d}{k} = 0 \quad [14]$

The sample version of Joe's estimator is given by

$$T_J(X^{(i)}, X^{(j)}) = \sum_{k=[(d+1)/2]}^{d} w_k \mathbb{P}(X^{(i)} - X^{(j)} \in B_{k,d-k}) = [\frac{2}{n(n-1)}] \sum_{k=[(d+1)/2]}^{d} w_k \sum_{i<j} \mathbf{1}_{B_{k,d-k}}\{X^{(i)} - X^{(j)}\}$$

Where $\mathbf{1}_{B_{k,d-k}}\{X^{(i)} - X^{(j)}\}$ is the indicator counting all instances of differences which are part of the different sets $B_{k,d-k}$.

In the trivariate case, one can easily find that $w_3 = 0$ and $w_2 = -\frac{1}{3}$ and thus the trivariate Joe's tau is given by

$$\begin{aligned} d = 3. \qquad \tau_J(X^{(i)}, X^{(j)}) &= \sum_{k=2}^{3} w_k \mathbb{P}((X_{i1} - X_{j1}, X_{i2} - X_{j2}, X_{i3} - X_{j3}) \in B_{k,d-k}) \\ &= (-\beta_2 - \frac{\beta_3}{3}) \mathbb{P}((X_{11} - X_{21}, X_{12} - X_{22}, X_{13} - X_{23}) \in B_{2,1}) \\ &\quad + (3\beta_2 + \beta_3) \mathbb{P}((X_{11} - X_{21}, X_{12} - X_{22}, X_{13} - X_{23}) \in B_{3,0}) \\ &= -\frac{1}{3} \mathbb{P}(\text{discordance}) + \mathbb{P}(\text{concordance}) \end{aligned}$$

Chapter 3

Methods

In this section, we first take a look at the relationship between the pairwise and multivariate extensions of Kendall's tau. Then the effect of ties on the resulting value of Kendall's tau is analysed. Finally, a suggestion is given where the two observations are combined in order to be able to compute Kendall's tau for the trivariate case with ties.

3.1 Pairwise and Multivariate taus

The connection is very clear for the averaging definition, where the multivariate Kendall's tau is defined as the mean of every possible pairwise Kendall's tau. It can be shown, however, that Joe's multivariate extension definition is closely related to the pairwise average definition.

Proposition *Let $X = (X_1, X_2, X_3)$ be a continuous random vector. The average multivariate tau is given by τ_{avg}, and Joe's expansion for the multivariate case is given by τ_J. It holds for any X that $\tau_{avg}(X) = \tau_J(X)$.*

The statement above can be shown by extending the expression for the average pairwise tau and group the terms up differently. This process is outlined below.
For ease of reading, we define the following additional notation. Let $s = s_1...s_d$ be a sequence of binary digits attached to a set $\{\mathbf{1}(X_{11} < X_{21}), \ldots, \mathbf{1}(X_{1d} < X_{2d})\}$. In the trivariate case $s = 110$ would, for instance, denote the set $\{X_{11} < X_{21}, X_{12} < X_{22}, X_{13} > X_{23}\}$. So $s = 111$ and $s = 000$ would denote a set of concordant pairs. Correspondingly, we define $p_s = p_{s_1 s_2 \ldots s_d}$ as the probability associated with the set. Thus, $p_{110} = \mathbb{P}(X_{i1} < X_{j1}, X_{i2} < X_{j2}, X_{i3} > X_{j3})$ and $p_{111} + p_{000}$ is equal to the probability of concordance of the observations. Finally, we can also replace a binary value with an asterisk if we do not consider it, for example p_{0*0} for $\mathbb{P}(X_{i1} > X_{j1}, X_{i3} > X_{j3})$.
We start from the definition of the average multivariate tau for $X = (X_1, X_2, X_3)$.

$$\tau_{avg}(X) = [\tau(X_1, X_2) + \tau(X_1, X_3) + \tau(X_3, X_2)]/3$$

Recall that in the bivariate case $\tau(X_i, X_j) = \mathbb{P}(concordance) - \mathbb{P}(discordance)$, which can now be written as $\tau(X) = (p_{11} + p_{00}) - (p_{10} + p_{01})$.
Now one can take each of these terms and expand it to the trivariate case as follows, in the example of $\tau(X_1, X_2)$

$$p_{11*} = p_{111} + p_{110}, \quad p_{00*} = p_{001} + p_{000}, \quad p_{10*} = p_{101} + p_{100}, \quad p_{01*} = p_{011} + p_{010}.$$

Doing this expansion for all three pairwise taus and adding them together gives

$$\begin{array}{cr l} & p_{111}+p_{000}+p_{110}+p_{001}-p_{101}-p_{100}-p_{011}-p_{010} & \text{for } \tau(X_1,X_2) \\ + & p_{111}+p_{000}+p_{101}+p_{010}-p_{110}-p_{100}-p_{011}-p_{001} & \text{for } \tau(X_1,X_3) \\ + & p_{111}+p_{000}+p_{011}+p_{100}-p_{110}-p_{010}-p_{101}-p_{001} & \text{for } \tau(X_2,X_3) \\ \hline = & 3(p_{111}+p_{000})-p_{110}-p_{001}-p_{101}-p_{100}-p_{011}-p_{010} & \end{array}$$

Note that the first two terms give the probability of concordance and the last six the probability of discordance for trivariate data. Thus dividing by three to get the average gives us that

$$\tau_{avg}(X) = \mathbb{P}(\text{Concordance}) - \frac{1}{3}\mathbb{P}(\text{Discordance})$$

This is precisely the expression which was found for Joe's extension in the trivariate case (Section 2.3.3) so indeed $\tau_{avg}(X) = \tau_J(X)$.

Taking a step back to the additional notation which was defined before the outline of the proof. A different interpretation of it would have immediately pointed us to Joe's definition. If one defines $S_k = \{s : s_1 + ... + s_d = k\}$ the set of pairs where their difference is negative k times, one can easily see the connection to Joe's $B_{k,d-k}$ used in his calculations. Namely $B_{k,d-k} = S_k + S_{d-k}$. One can also define the probability $p_{S_k} = \sum_{s \in S_k} p_s$. As we know that $w_0 = w_3 = 1$ and $w_1 = w_2 = -\frac{1}{3}$ it holds that

$$\begin{aligned} \tau_{avg}(X) &= \mathbb{P}(\text{Concordance}) - \frac{1}{3}\mathbb{P}(\text{Discordance}) \\ &= (p_{S_3} + p_{S_0}) - \frac{1}{3}(p_{S_2} + p_{S_1}) \\ &= \sum_{k=0}^{3} w_k p_{S_k} \\ &= \sum_{k=[(3+1)/2]}^{3} w_k (p_{S_k} - p_{S_{d-k}}) \\ &= \tau_J(X) \end{aligned}$$

Finally, we can make one remark about the intuitive logic of this result. We recall the graph representing the regions of concordance and discordance in three dimensions (figure 2.2). There, the probability of a point being concordant to another point is three times smaller than being discordant. So in some sense, one can see it as re-balancing these probabilities by dividing the probability of discordance by three.

3.2 Effect of ties

In his paper, Kendall notes that the rank of ties can be given a value corresponding to the average of their collective ranks [3]. An example of this can be seen below (Table 3.1). Note, in the second quantity the samples three to six are all tied, so their rank is equal to $\frac{1}{4}(3+4+5+6) = 4.5$.

This representation of ties as their average is not a necessary condition. It does not make a difference how these ties are represented as long as they all hold the same value. For example, instead of 4.5, the tied ranks in the second quantity could be given any value between 3 and 6. No matter what the chosen value is, the number of concordant and discordant pairs would not

$X^{(j)}$	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$	$X^{(4)}$	$X^{(5)}$	$X^{(6)}$	$X^{(7)}$	$X^{(8)}$	$X^{(9)}$	$X^{(10)}$
X_{j1}	1	2.5	2.5	4.5	4.5	6.5	6.5	8	9.5	9.5
X_{j2}	1	2	4.5	4.5	4.5	4.5	8	8	8	10

Table 3.1: Example 3

change. This is because concordance only looks at whether a value is larger than another and not by how much exactly.
We can once again recall the definition of Kendall's τ_b and evaluate it for the above example. We find $C = 33$, $D = 0$, and $P = \binom{10}{2} = 45$. Furthermore, in the first quantity there exist 4 groups of ties with 2 elements each so $T_1 = 4\frac{2(2-1)}{2} = 4$. The second quantity has two groups, one with 4 elements and one with three, so $T_2 = \frac{4(4-1)}{2} + \frac{3(3-1)}{2} = 9$. Finally, one can find that this sample is relatively strongly concordant, which makes sense as there are no discordant pairs

$$\tilde{\tau}_B = \frac{C - D}{\sqrt{(P - T_1)(P - T_2)}} = \frac{33}{\sqrt{41 * 36}} = 0.859$$

Already in this example, we can notice an effect of ties on the result. In the given sample there were no discordant pairs at all. In the continuous case, this would mean that all pairs are concordant and so we would get $\tau = 1$. However, because of the ties, which do not contribute to the count on the numerator, this maximal value is not reached. We do note that the denominator is slightly adjusted according to the number of ties otherwise the result would be even smaller.

3.2.1 Ties in bivariate data

Let us now consider different "types" of ties, to see whether these might have a different impact on the result. In the bivariate case, there are three groups that we can define. In the square brackets an example of such a case is given from the example above (Table 3.1).

- $X_{i1} = X_{j1}$, but $X_{i2} \neq X_{j2}$. $[X^{(2)}, X^{(3)}]$
- $X_{i1} \neq X_{j1}$, but $X_{i2} = X_{j2}$. $[X^{(3)}, X^{(4)}]$
- $X_{i1} = X_{j1}$, but $X_{i2} = X_{j2}$. $[X^{(4)}, X^{(5)}]$

Note that in the last case it is not necessary to also have $X_{i1} = X_{i2}$ as is coincidentally the case in the given example.
Let us first look at this last case where both quantities are tied and call such pairs *completely tied*. If one would have all the observations in a sample tied, we could not get a result with the definition of τ_b - due to division by zero. Recall, however, the formal definition of Kendall's tau as the difference between the probabilities of concordance and discordance. We can note that both of these are equal, and possibly 0. In either case, as every observation is equal the samples are not particularly concordant nor discordant, the result should thus be equal to 0.
Next, if all but one observation are completely tied and we assume, without loss of generality, that the last observation is concordant to the tied samples, Kendall's τ_b will be equal to $\frac{1}{\binom{n}{2}-\binom{n-1}{2}} = \frac{1}{n-1}$. The largest value this could take is 1, if $n = 2$. This is trivial as that just gives a concordant pair and no ties. However, as it is usually desired to have a large sample size n, this value quickly goes to 0 as well.
From a visualization perspective, this case would be the point where all four regions meet in the two-dimensional graph (figure 2.1).
Next, one of the first two cases is considered, where only one quantity is tied - it does not matter which one. One needs to note that the entire pair becomes irrelevant to the concordance concept. The tied quantity because it is tied, and can thus not be labelled as increasing or decreasing. But the other quantity also does not have a meaning by itself, as the concept of concordance cannot

exist in one dimension.
Once again if all values are tied in one quantity we will get $\tau_b = 0$, as $C = D = 0$. However, in the case where all but one sample are tied in a value, the result is different. As the second quantity is not tied in any points, the samples can be either concordant or discordant to the last observation. If we look at the extreme case where they are *all* concordant one gets $\tau_b = \frac{n-1}{\sqrt{(n-1)\binom{n}{2}}}$. This expression once again goes to zero, however as one might expect it happens slower than if both quantities are tied. For example, if $n = 1000$ then with complete ties one gets $\tau_b = 1/999 \approx 0.001$, whereas with these partial ties in one quantity one gets $\tau_b = \sqrt{5}/50 \approx 0.0447$.
Once again from a visualization point of view, this case would be the two perpendicular lines forming the borders between the concordant and discordant regions (figure 2.1).

3.2.2 Multivariate cases

Having looked at the bivariate case, we can take the next step by adding one additional random variable, and seeing if the behaviour stays the same and what is possibly added.
For this, we can start by recalling the three-dimensional graph representing regions of concordance with respect to a point (figure 2.2). We are interested in the parts where the two different regions meet. There is once again the point in which we have a complete tie as could also be found in the bivariate case. Similarly, the lines parallel to the axes, correspond to pairs which are tied in two values. The last value can once again not be ordered as concordant nor discordant as it acts alone. This also resembles the situation in the bivariate case.
The new type of tie is represented by the plane created between two regions. Here only one value is tied, and the other two are not. This is interesting because with two non-tied values we can attach a sense of concordance to the pair. We keep this in mind for the next section.

As a final remark, we can make a note about higher dimensions as well. Consider a d-variate case where we have two observations with k tied values. We can determine the multivariate concordance of the $d - k$ non-tied values. However, we cannot forget to account for the k ties, which should move the concordance result closer to zero.
In the next section, the new information attained up to here is combined to define some methods of calculating Kendall's tau for multivariate non-continuous data. We keep this last remark as verification of the validity of our method.

3.3 Expansion suggestions

We now suggest five different expansions to Kendall's tau which might perform well on multivariate non-continuous random variables according to the findings above.
In the first part of this chapter (Section 3.1) we looked at two methods, namely taking the average of pairwise taus and Joe's method. In the trivariate continuous case, we have found those to be equal. However, as the random variables which we will be looking at are not continuous, both methods should still be considered separately. The methods are not meant to be used for calculating Kendall's tau for non-continuous random variables. So we use the information gained in the second section as well as the methods presented in the previous chapter (Section 2.3) to suggest adaptations of these multivariate methods to the non-continuous case.
Let us first consider the method of taking averages. The result of this method varies based on the definition for Kendall's tau that is used to calculate the pairwise taus. When introduced in the paper, the original Kendall's tau was used. This was because we were mostly concentrating on solving only the problem of multivariate random variables. The original Kendall's tau naturally gives the most precise results for continuous random variables, so this was the best choice. Now that we are looking at non-continuous random variables, a different choice of method for calculating Kendall's tau might be better.
For one, τ_b can be used. It takes into account the ties that the original definition of Kendall's tau

cannot.

$$\tau_{avg(b)}(X) = \frac{1}{\binom{3}{2}} \sum_{1 \leq i < j \leq 3} \tau_b(X_i, X_j)$$

Similarly, one can use τ_{PHZ}. The generated data in our simulation will be zero-inflated, so one can expect this estimator to perform better than τ_b as it was specifically designed for zero-inflated distributions.

$$\tau_{avg(PHZ)}(X) = \frac{1}{\binom{d}{2}} \sum_{1 \leq i < j \leq d} \tau_{PHZ}(X_i, X_j)$$

Alternatively, instead of changing the calculation of tau to be able to deal with the introduction of ties, we can manipulate the data through jittering to resolve the ties at random. This gives us once again a continuous data set on which we can take the average of the original Kendall's taus. Although no method for non-continuous random variables is used, the behaviour of the suggestion will still be similar to what was found in the previous section. Namely, triplets containing one tie are not completely ignored. Instead, if the two non-tied values are concordant, they still contribute partly to the concordance count. This holds as with jittering the tie is randomly resolved. So it is roughly equally likely for the triplet to be classified as concordant or discordant. Through this, the partially concordant triplets add to the count of concordance but not as much as fully concordant triplets. Let Y be the jittered data set X, then we define this method as

$$\tau_{avg(jitter)}(X) = \frac{1}{\binom{d}{2}} \sum_{1 \leq i < j \leq d} \tau(Y_i, Y_j)$$

The two final methods are based on Joe's expansion. As given in the definition this method is also meant for continuous data. There is nothing straightforward that one can change about the method as was done above for the averaging. So we will try applying Joe's original definition τ_J. Here ties will be simply ignored. So if any ties are present in a comparison between two triplets this is not counted towards concordance nor discordance no matter how the other two values might relate. Naturally, we expect this not to perform very well, as nothing was adapted to accommodate the non-continuous random variables.
Finally, instead of changing the method, we can adapt the data as already suggested above. By jittering the data the ties are broken up and the data returns to a quasi-continuous distribution. This would allow Joe's method to perform much more accurately.

$$\tau_{J(jitter)}(X) = \tau_J(Y)$$

Chapter 4

Simulation

In this section, we first describe the simulation settings for the two simulations that are done. We line out and justify which parameters are being used in both simulations and what their purpose is. Finally, we take a look at the results and interpret them.

4.1 Simulation Setting

For all simulations, the sample size is set to be 300 and the number of simulations to be 1000.

4.1.1 Proof of Concept

In the first step, data from a trivariate continuous distribution is generated. Then Kendall's tau is calculated both with the averaging and Joe's method from the data. The results should show the accuracy of those methods compared to the true value of Kendall's tau. This comparison can allow us to set an acceptable margin of error for the second simulation. It should also further prove the proposition from Section 3.1 stating that the averaging and Joe's method are equal.
In this simulation data is generated from copulas. Two different copulas were used, the Frank and the Gaussian copula. The parameters of the copulas are closely related to the value of Kendall's tau for that distribution. In particular, these copulas and relations are shown below. First, the Frank copula, which takes one parameter θ, is given in the trivariate case by

$$C_\theta^{Fr}(u_1, u_2, u_3) = -\frac{1}{\theta}\log\left(\frac{\prod_{i=1}^{3}(\exp(-\theta u_i) - 1)}{\exp(-\theta) - 1}\right) \qquad , -\infty < \theta < \infty.$$

Kendall's tau is equal to

$$\tau = 1 - \frac{4}{\theta}\left(1 - \frac{1}{\theta}\int_0^\infty \frac{t}{exp(t) - 1}dt\right).$$

Second, the Gaussian copula takes a non-negative definite matrix R. The $d \times d$ matrix R can correspond to the correlation matrix with Pearson's correlation. The copula, as well as the corresponding formula for the pairwise Kendall's tau, are given below.

$$C_R^{Ga}(u_1, u_2, u_3) = \Phi_R(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3)),$$

where Φ^{-1} is the inverse cdf of a standard normal distribution and Φ_R is the multivariate cdf of a multivariate normal distribution.

$$\tau(X_i, X_j) = \frac{2}{\pi}\arcsin(r_{ij}),$$

where r_{ij} is the value in the i-th column and j-th row of the matrix R. That is, it corresponds to Pearson's correlation coefficient between the two considered variables.

So for the generation of data, we need to choose which true tau the random variable should have. Once this is determined the parameters for the copulas can be determined from the inverse of the formulas above. The simulation is run with multiple different true Kendall's taus to verify that the results hold in general and not only for one particular case. As the Frank copula only takes one parameter so there will only be three different simulations. Kendall's tau will be equal to 0.8, 0.5, and 0.2 for the three situations. The Gaussian copula can take pairwise parameters, so we choose to analyse different combinations of these pairwise taus. More precisely, we have the following seven situations (Table 4.1).

τ_{12}	τ_{13}	τ_{23}	τ_{avg}
0.8	0.8	0.8	0.8
0.8	0.8	0.5	0.7
0.8	0.8	0.2	0.6
0.8	0.5	0.2	0.5
0.8	0.2	0.2	0.4
0.5	0.2	0.2	0.3
0.2	0.2	0.2	0.2

Table 4.1: True values of tau used

For the result comparison, two different performance measures are used. In particular, the bias and the mean squared error (MSE). They are calculated as follows

$$\text{bias} = \frac{1}{s}\sum_{i=1}^{s}(\tilde{\tau} - \tau) \qquad \text{MSE} = \frac{1}{s}\sum_{i=1}^{s}(\tilde{\tau} - \tau)^2$$

Above s is the number of simulations, so it is equal to 1000. The estimand for which these measures are being calculated is denoted by $\tilde{\tau}$. The true value is simply τ.

4.1.2 Non-continuous data

In the second step, we test the newly suggested methods defined in the previous chapter (Section 3.3). For this multivariate non-continuous data needs to be generated. Along the five suggested methods also the true Kendall's tau is calculated. We observe how they compare to one another in the different scenarios.
The non-continuous data is generated according to a zero-inflated Poisson distribution. That is the marginals of the three variables are given by

$$F_{X_i}(x) = p_{0_i} + (1 - p_{0_i}) \cdot \frac{\lambda_i^x e^{-\lambda_i}}{x!} \qquad \text{for } i \in \{1, 2, 3\}$$

Where p_{0_i} is the probability of zero-inflation of the i-th variable, and λ_i is its parameter for the Poisson distribution. The parameter lambda is fixed to $\lambda_i = 2$ for all i. The three marginals along with a Pearson correlation matrix are used to generate the required non-continuous data, based on the Gaussian copula (as given above). The simulations will vary in the probability of zero-inflation for the three variables as well as the correlation matrix. In total, 12 situations will be analysed (Table 4.2). Below p_i is the zero-probability of the i-th variable, whereas ρ_{ij} is Pearson's correlation coefficient between the i-th and j-th random variable.
In the non-continuous case, the true value of Kendall's tau is the average of the pairwise taus calculated with the definition from Nikoloulopoulos and Karlis' paper as given in Section 2.3.2.
In this simulation, only the Mean-squared error is used.

4.1.3 Implementation

The entirety of the simulation was done in `R` [15], and the source code for it can be found in appendix A.2. Throughout we use several different methods for calculations of τ. The original

situation	zero-inflation	ρ_{12}	ρ_{13}	ρ_{23}
A1	$p_1 = 0$ $p_2 = 0$ $p_3 = 0.5$	0.8	0.5	0.2
A2		0.8	0.2	0.2
A3		0.5	0.2	0.2
A4		0.2	0.2	0.2
B1	$p_1 = 0.5$ $p_2 = 0.5$ $p_3 = 0.5$	0.8	0.5	0.2
B2		0.8	0.2	0.2
B3		0.5	0.2	0.2
B4		0.2	0.2	0.2
C1	$p_1 = 0.2$ $p_2 = 0.5$ $p_3 = 0.8$	0.8	0.5	0.2
C2		0.8	0.2	0.2
C3		0.5	0.2	0.2
C4		0.2	0.2	0.2

Table 4.2: Parameters used in non-continuous case

pairwise Kendall's tau, as well as τ_b, already exist as a function in `R`. If called on a multivariate data set it will return Kendall's matrix. The source code for τ_{PHZ} was provided in advance along with the corresponding paper [5]. Joe's estimate needed to be manually implemented, which was mostly a matter of counting concordant and discordant pairs. The jittering was also done with a pre-existent function in `R`. Finally, the true value of tau was computed with Karlis and Nikoloulopoulos' definition, the implementation for which was also available in [5]. The Gaussian copula is used for calculation of the treu value because this is the underlying copula to the "GenOrd" package used for data generation. As no closed form exists for the multivariate normal cdf it was approximated by the "pmvnorm" function.
For the generation of data two different packages were used "copula" and "GenOrd". The first was used to generate multivariate continuous data. As tau can be found from the parameters of the copulas, it is possible to reverse engineer the parameter to generate data according to the desired tau. Both the Gaussian and Frank copula were used. The second package helps generate multivariate discrete data. By defining the marginal cdf of the different random variables and giving a possible Pearson correlation matrix, the package can generate the type of data we want. To manipulate the number of ties in our distribution the marginals can simply be adjusted accordingly.

4.2 Results

4.2.1 Continuous data

As mentioned above we use two different copulas here to set a baseline for the accuracy of our estimators. In the table below the true values τ can be seen compared to the estimations $\tilde{\tau}$ (Table 4.5). The first part contains the results from the simulation with Frank's copula. Below the double line, the results of the simulations with the Gaussian copula can be seen.
The first result that we can recognize is what has been established in Section 3.1. Indeed, the estimation found through averaging the pairwise taus is the same as Joe's estimator.
Next, we note that the estimations for the simulation with Frank's copula are very accurate with a bias below 5e-04 and MSE below 7e-04. This is more than acceptable. The estimations for the simulations with the Gaussian copula are also good. However, an outlier can be seen in the third case where we have that $\tau_{12} = \tau_{13} = 0.8$ and $\tau_{23} = 0.2$. Here the estimations are significantly different with the estimation of the pairwise taus reaching only 0.6 instead of 0.8. This is also reinforced by the fact that the estimated parameters in this situation are significantly off of the true parameters (see Appendix A.1, Table A.2). However, this can be explained logically as when the first value is strongly correlated ($\tau = 0.8$) to both the second and the third, then one would

assume that the second and third are also significantly correlated.
We thus decide to only consider the last four scenarios, omitting the first three, in the non-continuous simulation.

Copula	τ_{12}	τ_{13}	τ_{23}	τ_{avg}	$\tilde{\tau}_{12}$	$\tilde{\tau}_{13}$	$\tilde{\tau}_{23}$	$\tilde{\tau}_{avg}$	$\tilde{\tau}_J$	bias	MSE
Frank	0.8	0.8	0.8	0.8	0.7997965	0.8001606	0.800878	0.8002784	0.8002784	0.000027838	6.92E-05
	0.5	0.5	0.5	0.5	0.5005309	0.500799	0.5000178	0.5004492	0.5004492	0.000449231	4.30E-04
	0.2	0.2	0.2	0.2	0.1997115	0.2007654	0.1992766	0.1999178	0.1999178	-8.21702E-05	6.53E-04
Gaussian	0.8	0.8	0.8	0.8	0.7997388	0.80016	0.7999708	0.7999565	0.7999565	-4.34634E-05	0.000121177
	0.8	0.8	0.5	0.7	0.7506731	0.7515001	0.5021732	0.6681155	0.6681155	-0.031884504	0.001376388
	0.8	0.8	0.2	0.6	0.6092272	0.6110449	0.220272	0.4801813	0.4801813	-0.119818655	0.014981453
	0.8	0.5	0.2	0.5	0.7252105	0.4798441	0.2050546	0.4700364	0.4700364	-0.029963612	0.001522015
	0.8	0.2	0.2	0.4	0.7995289	0.2012117	0.2009888	0.4005765	0.4005765	0.00057647	0.000647958
	0.5	0.2	0.2	0.3	0.4993807	0.2013427	0.2007323	0.3004852	0.3004852	0.000485232	0.000698895
	0.2	0.2	0.2	0.2	0.1996363	0.2012722	0.2004088	0.2004391	0.2004391	0.000439093	0.000707895

Table 4.3: True Kendall's tau and its estimates in the continuous cases

4.2.2 Non-continuous data

First, calculate the true pairwise taus τ_{ij} and their averages τ_{avg} in the different situations. Using Nikoloulopoulos and Karlis' definition, we find the following values.

situation	τ_{12}	τ_{13}	τ_{23}	τ_{avg}
A1	0.55083	0.28669	0.11484	0.317452
A2	0.55083	0.11484	0.11484	0.260166
A3	0.31982	0.11484	0.11484	0.183162
A4	0.12499	0.11484	0.11484	0.118222
B1	0.43642	0.25945	0.10517	0.267017
B2	0.43642	0.10517	0.10517	0.21559
B3	0.25945	0.10517	0.10517	0.1566
B4	0.10517	0.10517	0.10517	0.105172
C1	0.47981	0.19783	0.07563	0.251088
C2	0.47981	0.08306	0.07563	0.212831
C3	0.2845	0.08306	0.07563	0.147729
C4	0.11449	0.08306	0.07563	0.091056

Table 4.4: True Kendall's tau in the different situations

In the table below, the comparison of the multivariate taus can be seen with the mean squared error. It is clear from the results that the pairwise average of the τ_b performs the worst along with Joe's method where the ties are ignored. However, it might be interesting to note that the former always gives a higher result whereas the latter always gives a lower result than the true tau. As could have been predicted the average of the original pairwise Kendall's taus on jittered data nearly gives the same result as the original Joe's method applied on jittered data. The very small differences can be attributed to the slight variations in data that appear due to jittering. Both methods give accurate estimates of the true tau. Especially, when there is no or close to no zero-inflation. Of the two methods, averaging is preferred as it has lower variation and thus mean squared error. Finally, there is the averaging of τ_{PHZ}. This estimator does not perform as well when there is no zero-inflation. Although it still performs significantly better than the first two methods. With higher zero-inflation, this method becomes more accurate. Even though, it is not as exact as the last two methods its variation and MSE are significantly lower. This is at times desirable of course, so this method should not be discarded.

situation	τ_{avg}	$\widetilde{\tau}_{avg(B)}$	MSE	$\widetilde{\tau}_{avg(PHZ)}$	MSE	$\widetilde{\tau}_{avg(jitter)}$	MSE	$\widetilde{\tau}_{J(jitter)}$	MSE	$\widetilde{\tau}_{J}$	MSE
A1	0.317452	0.422881	0.01198	0.356947	0.0021	0.31978	0.00064	0.318836	0.00614	0.242542	0.00062
A2	0.260166	0.33977	0.00731	0.295877	0.00182	0.26031	0.00064	0.260241	0.00564	0.188443	0.00066
A3	0.183162	0.244535	0.00472	0.201956	0.00094	0.18364	0.00064	0.184268	0.0025	0.13773	0.00064
A4	0.118222	0.162205	0.00284	0.126592	0.00065	0.11876	0.00063	0.11934	0.00099	0.093124	0.0006
B1	0.267017	0.420576	0.02472	0.270771	0.0006	0.26458	0.00071	0.264746	0.00433	0.205946	0.00073
B2	0.21559	0.339983	0.01662	0.219349	0.00059	0.21411	0.00077	0.213901	0.00369	0.159797	0.00074
B3	0.1566	0.248998	0.00965	0.158532	0.0005	0.15761	0.00067	0.15645	0.00209	0.11605	0.00066
B4	0.105172	0.16741	0.00498	0.106136	0.00047	0.10465	0.00067	0.105167	0.00123	0.075949	0.00065
C1	0.251088	0.420909	0.0298	0.260529	0.00048	0.25241	0.00079	0.252031	0.01537	0.129248	0.00065
C2	0.212831	0.343603	0.01823	0.222256	0.00046	0.21407	0.00073	0.214141	0.01207	0.10484	0.00071
C3	0.147729	0.250244	0.01162	0.15131	0.00037	0.14888	0.0007	0.148689	0.00487	0.080248	0.00065
C4	0.091056	0.169175	0.00724	0.09253	0.00038	0.09178	0.00069	0.091763	0.00131	0.058644	0.00064

Table 4.5: True Kendall's tau and its estimates in the non-continuous cases

Chapter 5

Empirical Data Analysis

With the information found in the previous chapter, we can now apply the methods to synthetic data. This data was provided by a third party so no information is known about its at time of analysis. It is based, however, on the real-world measurements of pharmaceutical cleanrooms mentioned in the introduction [2]. This is count data with six different variables.

Before applying the methods for the calculation of Kendall's tau, we analyse the synthetic data. Some basic knowledge on the data is given by the summary statistics of the different random variables.

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
X_1	0	0	0	0.3261	1	3
X_2	0	1	1	1.469	2	7
X_3	0	4	5	5.659	7	17
X_4	0	1	1	1.45	2	6
X_5	0	0	1	0.7994	1	5
X_6	0	0	0	0.2289	0	3

Table 5.1: Summary statistics of synthetic data

For better visualisation and understanding of the summary, the plots in Figure 5.1 can be consulted. For more precise information on the frequency of values for the different random variables the Appendix A.1 can be consulted (Table A.3). From both the plots and the summary statistics table one can see that the X_1 and X_6 are very similar. They have a predominantly high amount of zero values, although higher for X_6. Similarly X_2 and X_4 resemble each other as well. They seem to follow a positively skewed distribution with again zero values being one of the more frequent ones. X_5 also seems to have a positive skew. However, it is shifted to the left compared to the last two. This makes the most frequent value zero for it. Finally, X_3 has by far the lowest frequency of zero values. It also resembles a right-skewed distribution but moved more to the right than the above.

From here on out we can estimate the probability mass functions for the random variables. Recall that this data set is created to imitate measurements from cleanrooms as described in the introduction (Section 1). Thus we can reasonably assume that the random variables follow a zero-inflated Poisson distribution [2]. It is thus also possible to estimate the zero-inflation p_0 as well as the rate λ. We can find the following results (Table 5.2).
From these results, it is possible to see that the lambdas are predominantly small, similar to the data generated in the simulation of the non-continuous data. Only the third variable has a larger rate of approximately 5.68. Furthermore, the zero-inflation is very low to non-existent in most cases which only leaves a Poisson distribution. Here the outlier is the sixth random variable for

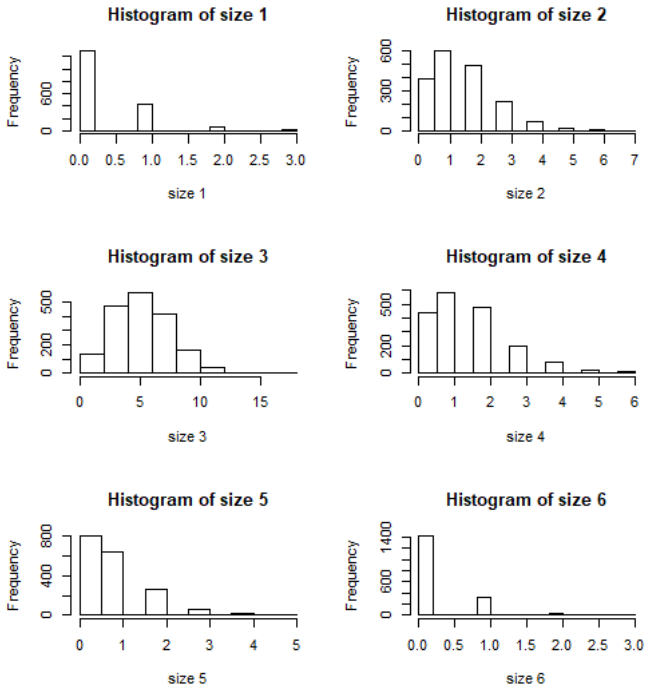

Figure 5.1: Histograms of the six variables

Variable	p_0	λ
X_1	0	0.3230258
X_2	0	1.4213035
X_3	0.003152994	5.6767878
X_4	0	1.4497082
X_5	0.033944073	0.8275343
X_6	0.112328687	0.2578532

Table 5.2: Parameters assuming the variables follow a ZIP distribution

which we find $p_0 = 0.112$.

From the results obtained in Section 4.2.2 we expect $\tilde{\tau}_{avg(jitter)}$, $\tilde{\tau}_{J(jitter)}$, and $\tilde{\tau}_{avg(PHZ)}$ to perform well. In particular, as the zero-inflation is so low, the methods with jittering are expected to give slightly more trusted results. It is important to note that a six-variate random vector is considered. Thus, the equality between τ_{avg} and τ_J which was proven for the trivariate case in Section 3.1 does not necessarily hold anymore. As this result was not explored for higher dimensions, we will be able to see by the values if it might still hold or not. Although, we expect that $\tilde{\tau}_{avg(b)}$ and $\tilde{\tau}_J$ do not give very accurate results we still compute them to see how they compare to the more trusted results.

Below one can first find the three Kendall's matrices which can be generated.

$$\tilde{K}_b = \begin{bmatrix} 1.000 & 0.000 & -0.031 & 0.005 & -0.011 & -0.032 \\ 0.000 & 1.000 & -0.025 & 0.000 & -0.015 & 0.025 \\ -0.031 & -0.025 & 1.000 & -0.015 & -0.021 & 0.021 \\ 0.005 & 0.000 & -0.015 & 1.000 & 0.007 & -0.015 \\ -0.011 & -0.015 & -0.021 & 0.007 & 1.000 & -0.011 \\ -0.032 & 0.025 & 0.021 & -0.015 & -0.011 & 1.000 \end{bmatrix}$$

$$\tilde{K}_{PHZ} = \begin{bmatrix} 1.000 & 0.001 & -0.021 & 0.003 & -0.006 & -0.012 \\ 0.001 & 1.000 & -0.027 & 0.002 & -0.014 & 0.013 \\ -0.021 & -0.027 & 1.000 & -0.014 & -0.013 & 0.010 \\ 0.003 & 0.002 & -0.014 & 1.000 & 0.003 & -0.009 \\ -0.006 & -0.014 & -0.013 & 0.003 & 1.000 & -0.006 \\ -0.012 & 0.013 & 0.010 & -0.009 & -0.006 & 1.000 \end{bmatrix}$$

$$\tilde{K}_{jitter} = \begin{bmatrix} 1.000 & -0.003 & -0.021 & 0.008 & -0.025 & -0.017 \\ -0.003 & 1.000 & -0.021 & -0.004 & -0.018 & 0.014 \\ -0.021 & -0.021 & 1.000 & -0.008 & -0.019 & 0.017 \\ 0.008 & -0.004 & -0.008 & 1.000 & 0.003 & 0.008 \\ -0.025 & -0.018 & -0.019 & 0.003 & 1.000 & 0.010 \\ -0.017 & 0.014 & 0.017 & 0.008 & 0.010 & 1.000 \end{bmatrix}$$

As one can see the values all float around zero (except the diagonals). This indicates no prevalent concordance or discordance between the individual pairs. We know that the averages will also be close to zero. The multivariate estimates calculated with the five different methods give the following values.

$\tilde{\tau}_{avg(B)}$	$\tilde{\tau}_{avg(PHZ)}$	$\tilde{\tau}_{avg(jitter)}$	$\tilde{\tau}_{J(jitter)}$	$\tilde{\tau}_J$
-0.007874179	-0.005904319	-0.005123999	-0.1997307	-0.1997712

Table 5.3: Estimated Kendall's taus

As expected, one can see that the values which are found by taking the average are close to zero. They are additionally all slightly but insignificantly negative. We find that $\tau_{avg(PHZ)}$ and $\tau_{avg(jitter)}$, are within 0.0008 of each other. Even $\tau_{avg(b)}$ gives a very similar value to the other averages with a difference of only 0.002.
However, unlike in the simulation with trivariate data, $\tau_{J(jitter)}$ does not give the same result as $\tau_{avg(jitter)}$. The resulting value for Kendall's tau with Joe's method is nearly -0.2. This would denote significant correlation and, more precisely, discordance. Surprisingly, Joe's original method τ_J does not differ significantly from $\tau_{J(jitter)}$.

Clearly, in higher dimensions, the results found through averaging and Joe's method are not in agreement anymore. However, the methods which were determined to be imprecise in the previous simulation, τ_J and $\tau_{avg(b)}$, do not differ as much from the other results.
With the data analysis being concluded, we have been informed that it follows a multinomial distribution. Thus, one would expect certain correlations and so Joe's method might give a better result.

Chapter 6

Conclusion and Discussion

In the thesis, we have looked at already existing extensions for Kendall's tau which are used for multivariate or non-continuous random variables. Then we have defined several possible methods for calculating Kendall's tau for multivariate and non-continuous random variables. Finally, we tested and verified the methods in a simulation and data analysis setting.

There are two extensions which were considered for multivariate random variables. Namely, taking the average of pairwise taus and Joe's extension. For these, we have found that they are equal to one another in the trivariate case. However, in higher dimensions, this equality does not necessarily hold as was seen in the empirical data analysis.
For the existing methods for calculating Kendall's tau from data with ties we have looked at τ_b and τ_{PHZ}. Different types of ties were identified and their effect was looked at. In general, our findings can be summarized as follows. In cases where a sample pair has ties but the remaining non-tied values are concordant, the pairs should still contribute to the result. The question of how much they should contribute remains unanswered. However, with the methods used in the simulation, it is possible to make a reasonable assumption. First of all, as Joe's method simply ignores any sample which contains a tie, it does not perform well, as expected. It will always give a lower value than the true Kendall's tau as it does not account for these partially concordant pairs at all. Kendall's τ_b does partially account for these. However, it seems the contribution size is wrong as it still significantly differs from the true Kendall's tau. On the other hand, the average of pairwise taus from jittered data as well as τ_{PHZ} seems to do this better. It is easier to interpret what happens with the jittering. Namely, as the ties are randomly resolved, about 50% of the partially concordant pairs contribute to the result. This holds for the trivariate case.
As already given in the results of the simulation (Section 4.2.2), it was found that in the trivariate case $\tau_{J(jitter)}$, $\tau_{avg(jitter)}$, and $\tau_{avg(PHZ)}$ perform the best. With the latter not being as precise when there is less zero-inflation but having a lower MSE in general. The other two methods $\tau_{avg(b)}$ and τ_J do not give very accurate results. Furthermore, in the empirical data analysis which consisted of a six-variate random vector, there was a significant difference. While $\tau_{avg(jitter)}$ and $\tau_{avg(PHZ)}$ still gave similar results to one another, $\tau_{J(jitter)}$ differs significantly. This is as mentioned above because the equality between τ_J and τ_{avg} is lost.

Apart from the results, let us also take a quick look at the approach used throughout the paper. First, the suggested methods in Section 3.3 are not built from the ground up. They are all based on already known extensions from different papers introduced in Section 2. This saved us some work on showing why they are relevant to the research question.
In the first part of the simulation, which was used as a proof of concept, we showed what was expected. However, we have also identified some cases which we decided to omit from the second part of the simulation as we were able to show that generating such data is not precise enough, because it forces it into an impossible shape.
In the second part of the simulation, we look at several different situations with various zero-

inflation and correlation coefficients. In this part, we only look at the trivariate case and do not vary the value of the parameter λ. As the methods are later applied to synthetic data which consists of six random variables, it could have been of use to see if the results found still hold in higher dimensions. Furthermore, one could have also looked at different rates for the Poisson distribution to test whether that would affect anything as this would affect the frequency of zero samples.

For future endeavours, the method of jittering could be looked at in more depth. There is not much literature yet on the efficiency of jittering in calculating Kendall's tau compared to direct methods which use non-continuous data. Jittering is often an already implemented function in computation software so it would serve as an easy and quick calculation method for Kendall's tau. Furthermore, the relationship between Joe's extension and the averaging over pairwise taus could be looked at in higher dimensions. In general, the theoretical, as well as practical behaviour of the methods beyond three dimensions, can be explored further to verify the results found here.

Bibliography

[1] M. G. Kendall. "A New Measure of Rank Correlation". In: *Biometrika* 30.1/2 (June 1938), p. 81. ISSN: 00063444. DOI: 10.2307/2332226. URL: https://www.jstor.org/stable/2332226?origin=crossref.

[2] Stephan A W Van Driel. *Particle Size Statistics with Grouped and Truncated Data Measured by the Biotrak Real Time Viable Particle Counter in Cleanrooms of Grade C and D*. Tech. rep. MSD - CENTER FOR MATHEMATICAL SCIENCES, Sept. 2018.

[3] M. G. Kendall. "The Treatment of Ties in Ranking Problems". In: *Biometrika* 33.3 (Nov. 1945), p. 239. ISSN: 00063444. DOI: 10.2307/2332303. URL: https://www.jstor.org/stable/2332303?origin=crossref.

[4] R.S. Pimentel. "Kendall's Tau and Spearman's Rho for Zero-Inflated Data". PhD thesis. Kalamazoo, Michigan: Western Michigan University, 2009.

[5] Elisa Perrone, Edwin R. van den Heuvel and Zhuozhao Zhan. "Kendall's tau estimator for bivariate zero-inflated count data". In: *arXiv* (Aug. 2022). DOI: 10.48550/arXiv.2208.03155. URL: http://arxiv.org/abs/2208.03155.

[6] M. D. Taylor. "Multivariate measures of concordance". In: *Annals of the Institute of Statistical Mathematics* 59.4 (Nov. 2007), pp. 789–806. ISSN: 0020-3157. DOI: 10.1007/s10463-006-0076-2. URL: http://link.springer.com/10.1007/s10463-006-0076-2.

[7] Harry Joe. "Multivariate concordance". In: *Journal of Multivariate Analysis* 35.1 (1990), pp. 12–30. ISSN: 10957243. DOI: 10.1016/0047-259X(90)90013-8.

[8] Ronald S. Pimentel, Magdalena Niewiadomska-Bugaj and Jung Chao Wang. "Association of zero-inflated continuous variables". In: *Statistics and Probability Letters* 96 (Jan. 2015), pp. 61–67. ISSN: 01677152. DOI: 10.1016/J.SPL.2014.09.002.

[9] Aristidis K. Nikoloulopoulos and Dimitris Karlis. "Modeling Multivariate Count Data Using Copulas". In: *Communications in Statistics - Simulation and Computation* 39.1 (Dec. 2009), pp. 172–187. ISSN: 0361-0918. DOI: 10.1080/03610910903391262.

[10] A. Sklar. "Fonctions de Répartition à n Dimensions et Leurs Marges". In: *Publications de l'Institut Statistique de l'Université de Paris* 8 (1959), pp. 229–231.

[11] Roger B. Nelsen. *An Introduction to Copulas*. 2nd ed. Springer Science & Business Media, June 2007. ISBN: 978-0-387-28659-4. DOI: 10.1007/0-387-28678-0.

[12] Roger B. Nelsen. "Concordance and Copulas: A Survey". In: *Distributions With Given Marginals and Statistical Modelling* (2002), pp. 169–177. DOI: 10.1007/978-94-017-0061-0{_}18. URL: https://link.springer.com/chapter/10.1007/978-94-017-0061-0_18.

[13] Mhamed Mesfioui and Jean François Quessy. "Concordance measures for multivariate non-continuous random vectors". In: *Journal of Multivariate Analysis* 101.10 (Nov. 2010), pp. 2398–2410. ISSN: 0047-259X. DOI: 10.1016/J.JMVA.2010.06.011.

[14] Irène Gijbels, Vojtěch Kika and Marek Omelka. "On the specification of multivariate association measures and their behaviour with increasing dimension". In: *Journal of Multivariate Analysis* 182 (Mar. 2021), p. 104704. ISSN: 0047259X. DOI: 10.1016/j.jmva.2020.104704. URL: https://linkinghub.elsevier.com/retrieve/pii/S0047259X20302852.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020.

Appendix A

Simulations

A.1 Results

Parameter estimates Frank copula.

true p	estimated p	variance	lower CI	upper Ci
18.1915397	18.2782852	7.15E-01	16.7116007	20.0106222
5.7362827	5.7655652	1.43E-01	5.0249844	6.520249
1.8608838	1.8668604	6.50E-02	1.3807637	2.3684463

Table A.1: Parameter estimates Frank copula

Parameter estimates Gaussian copula.

situation	true p	estimated p	variance	lower CI	upper Ci
1	0.9510565	0.9507361	3.95E-05	0.9372551	0.9620874
	0.9510565	0.9509394	3.94E-05	0.9383452	0.9622539
	0.9510565	0.9508479	3.94E-05	0.9382537	0.9617178
2	0.9510565	0.9239881	9.44E-05	0.9032664	0.9414158
	0.9510565	0.9245063	8.61E-05	0.9040785	0.9411095
	0.7071068	0.7088093	9.96E-04	0.6447739	0.7637881
3	0.9510565	0.8168741	4.73E-04	0.7702453	0.8570001
	0.9510565	0.818561	4.30E-04	0.7753575	0.8571823
	0.309017	0.3385537	3.07E-03	0.2282077	0.4398653
4	0.9510565	0.9079363	1.35E-04	0.8820768	0.9288206
	0.7071068	0.6836711	1.09E-03	0.6153643	0.7465624
	0.309017	0.3160118	3.11E-03	0.2069404	0.4175271
5	0.9510565	0.9506289	4.06E-05	0.9373473	0.9617758
	0.309017	0.3102943	3.09E-03	0.1976797	0.4160961
	0.309017	0.3099718	3.04E-03	0.2031028	0.4135227
6	0.7071068	0.7056928	1.04E-03	0.638876	0.7641858
	0.309017	0.3104762	3.17E-03	0.1986529	0.4200746
	0.309017	0.3095909	3.02E-03	0.2044143	0.4107018
7	0.309017	0.3079238	3.24E-03	0.1955575	0.41135
	0.309017	0.3103633	3.22E-03	0.1971063	0.417543
	0.309017	0.3091081	3.02E-03	0.2015062	0.4124414

Table A.2: Parameter estimates Gaussian copula

Frequency of values for the different random variables.

Variable X_i	X_1	X_2	X_3	X_4	X_5	X_6
$P(X_i = 0)$	0.72	0.219444444	0.004444444	0.240555556	0.451666667	0.797222222
$P(X_i = 1)$	0.23777778	0.335	0.01666667	0.32277778	0.35888889	0.18
$P(X_i = 2)$	0.03833333	0.27444444	0.05222222	0.26444444	0.14333333	0.01944444
$P(X_i = 3)$	0.003888889	0.119444444	0.117222222	0.111111111	0.032777778	0.003333333
$P(X_i = 4)$	0	0.03611111	0.145	0.04444444	0.01111111	0
$P(X_i = 5)$	0	0.011666667	0.175	0.012777778	0.002222222	0
$P(X_i = 6)$	0	0.002777778	0.140555556	0.003888889	0	0
$P(X_i = 7)$	0	0.001111111	0.136111111	0	0	0
$P(X_i = 8)$	0	0	0.09722222	0	0	0
$P(X_i = 9)$	0	0	0.06	0	0	0
$P(X_i = 10)$	0	0	0.02777778	0	0	0
$P(X_i = 11)$	0	0	0.01111111	0	0	0
$P(X_i = 12)$	0	0	0.01	0	0	0
$P(X_i = 13)$	0	0	0.002222222	0	0	0
$P(X_i = 14)$	0	0	0.001111111	0	0	0
$P(X_i = 15)$	0	0	0.001111111	0	0	0
$P(X_i = 16)$	0	0	0.001666667	0	0	0
$P(X_i = 17)$	0	0	0.000555556	0	0	0

Table A.3: Frequency of values for the random variables

A.2 Source code

```
# Simulation - CONTINUOUS CASE


# SETTING UP LIBRARIES
library(Metrics)
library(boot)
library(Rmisc)
library(utils)
library(copula)
library(here)

library(mvtnorm)
library(Matrix)
library(MASS)
library(GenOrd)



# DEFINE PARAMETERS
n <- 300  #nr of observations (100, 500, 1000)
sim <- 1000 #nr of simulations, from Morris (2019)
pair_tau <- c(0.8, 0.5, 0.2)  # possible pairwise taus. [negative values?]

combinate <- data.frame(matrix(nrow=0,ncol=3)) # different triples of pairwise taus
    considered
colnames(combinate) <- c("tau12", "tau13", "tau23")
for (i in 1:3){
  combinate <- rbind(combinate, data.frame(tau12=pair_tau[[1]],tau13=pair_tau[[1]],
      tau23=pair_tau[[i]]))
}
for (j in 2:3){
```

```
  combinate <- rbind(combinate, data.frame(tau12=pair_tau[[1]],tau13=pair_tau[[j]],
      tau23=pair_tau[[3]]))
}
for (k in 2:3){
  combinate <- rbind(combinate, data.frame(tau12=pair_tau[[k]],tau13=pair_tau[[3]],
      tau23=pair_tau[[3]]))
}

# DATAFRAMES
performance <- data.frame(matrix(ncol=18, nrow=0))
colnames(performance) <- c("size", "tau12","tau13","tau23","avgtau", "esttau12","
    esttau13","esttau23","estavgtau","estjoetau",
                           "bias","SEbias","empSE","SEempSE","MSE","SEMSE","cover",
                              "SEcover")
parameters <- data.frame(matrix(ncol=7, nrow=0))
colnames(parameters) <- c("size", "name","true","estavg","var", "CIlow","CIup")



# SIMULATION - Frank copula

for (p in 1:3){
  set.seed(9)

  # get copula parameters from taus
  pairtau <- c(pair_tau[p],pair_tau[p], pair_tau[p])
  truetau <- pair_tau[p]
  param <- iTau(frankCopula(), pair_tau[p])

  # temporary dataframes
  Rtau <- c()   #contain all results from cor(x, kendall)
  avgtau <- c() #contain the avg of estimated pairwise taus
  joetau <- c() #contain joe's sample version of multivariate tau
  estimates <- data.frame(matrix(ncol=12, nrow=0))
  colnames(estimates) <- c("Simulation", "meanx1", "meanx2", "meanx3", "varx1", "
      varx2", "varx3", "tau12","tau13","tau23", "avgtau", "joetau")
  estparam <- data.frame(matrix(ncol=2, nrow=0))
  colnames(estparam) <- c("Simulation", "estpar")

  for (i in 1:sim){
    # create copula and dataset
    cop <- frankCopula(param, dim=3)
    x <- rCopula(n, cop)

    # find some temporary values of interest
    tempK <- cor(x, method="kendall")
    tempRtau <- c(tempK[[1,2]], tempK[[1,3]], tempK[[2,3]])
    Rtau <- c(Rtau, tempRtau)
    tempavgtau <- (tempRtau[[1]]+tempRtau[[2]]+tempRtau[[3]])/3
    avgtau <- c(avgtau, tempavgtau)
    tempjoetau <- sample_joe(x)
    joetau <- c(joetau, tempjoetau)
    estimates <- rbind(estimates, data.frame(Simulation=i, meanx1=mean(x[,1]),
        meanx2=mean(x[,2]),meanx3=mean(x[,2]),
                                             varx1=var(x[,1]),varx2=var(x[,2]),
                                                 varx3=var(x[,2]), tau12=tempRtau
                                                 [[1]],
                                            tau13=tempRtau[[2]],tau23=tempRtau
                                                [[3]], avgtau=tempavgtau, joetau=
                                               tempjoetau))

    tempfit <- fitCopula(frankCopula(dim=3), x, method="itau")
    tempparam <- tempfit@copula@parameters
    estparam <- rbind(estparam, data.frame(Simulation=i, estpar=tempparam))
  }

  # update main dataframes
```

```
  perf <- perf_measures(estimates, truetau, sim, n)
  tempperf <- data.frame(size=n, tau12=pairtau[1], tau13=pairtau[2], tau23=pairtau
      [3], avgtau=truetau,
                         esttau12=mean(estimates$tau12), esttau13=mean(estimates$
                             tau13), esttau23=mean(estimates$tau23),
                         estavgtau=mean(estimates$avgtau), estjoetau=mean(estimates
                             $joetau), bias=perf[[1]], SEbias=perf[[2]],
                         empSE=perf[[3]], SEempSE=perf[[4]], MSE=perf[[5]], SEMSE=
                             perf[[6]], cover=perf[[7]], SEcover=perf[[8]])
  performance <- rbind(performance, tempperf)

  parameters <- rbind(parameters, data.frame(size=n, name="p", true=param,estavg=
      mean(estparam[,2]),var=var(estparam[,2]),
                                             CIlow=quantile(estparam[,2], 0.025)
                                                 [[1]],CIup=quantile(estparam[,2],
                                                 0.975)[[1]]))

}


# SIMULATION - Gaussian copula

for (p in 1:7){ # loop for all 7 combinations of pairwise taus
  set.seed(9)

  # get copula parameters from taus
  pairtau <- combinate[p,]
  truetau <- (sum(combinate[p,])/3)
  param <- c(iTau(normalCopula(), combinate[p,1]), iTau(normalCopula(), combinate[p
      ,2]), iTau(normalCopula(), combinate[p,3]))

  # temporary dataframes
  Rtau <- c()   #contain all results from cor(x, kendall)
  avgtau <- c() #contain the avg of estimated pairwise taus
  joetau <- c() #contain joe's sample version of multivariate tau
  estimates <- data.frame(matrix(ncol=12, nrow=0))
  colnames(estimates) <- c("Simulation", "meanx1", "meanx2", "meanx3", "varx1", "
      varx2", "varx3", "tau12","tau13","tau23", "avgtau", "joetau")
  estparam <- data.frame(matrix(ncol=4, nrow=0))
  colnames(estparam) <- c("Simulation", "estp1", "estp2", "estp3")

  for (i in 1:sim){
    # create copula and dataset
    cop <- normalCopula(param, dim=3, dispstr="un")
    x <- rCopula(n, cop)

    # find some temporary values of interest
    tempK <- cor(x, method="kendall")
    tempRtau <- c(tempK[[1,2]], tempK[[1,3]], tempK[[2,3]])
    Rtau <- c(Rtau, tempRtau)
    tempavgtau <- (tempRtau[[1]]+tempRtau[[2]]+tempRtau[[3]])/3
    avgtau <- c(avgtau, tempavgtau)
    tempjoetau <- sample_joe(x)
    joetau <- c(joetau, tempjoetau)
    estimates <- rbind(estimates, data.frame(Simulation=i, meanx1=mean(x[,1]),
        meanx2=mean(x[,2]),meanx3=mean(x[,2]),
                                             varx1=var(x[,1]),varx2=var(x[,2]),
                                                 varx3=var(x[,2]), tau12=tempRtau
                                                 [[1]],
                                            tau13=tempRtau[[2]],tau23=tempRtau
                                                [[3]], avgtau=tempavgtau, joetau=
                                               tempjoetau))

    tempfit <- fitCopula(normalCopula(dim=3, dispstr="un"), x, method="itau")
    tempparam <- tempfit@copula@parameters
    estparam <- rbind(estparam, data.frame(Simulation=i, estp1=tempparam[1],estp2=
        tempparam[2],estp3=tempparam[3]))
```

```
  }

  # update main dataframes
  perf <- perf_measures(estimates, truetau, sim, n)
  tempperf <- data.frame(size=n, tau12=pairtau[1], tau13=pairtau[2], tau23=pairtau
      [3], avgtau=truetau,
                         esttau12=mean(estimates$tau12), esttau13=mean(estimates$
                             tau13), esttau23=mean(estimates$tau23),
                         estavgtau=mean(estimates$avgtau), estjoetau=mean(estimates
                             $joetau), bias=perf[[1]], SEbias=perf[[2]],
                         empSE=perf[[3]], SEempSE=perf[[4]], MSE=perf[[5]], SEMSE=
                             perf[[6]], cover=perf[[7]], SEcover=perf[[8]])
  performance <- rbind(performance, tempperf)

  parms <- c("tau12", "tau13", "tau23")
  for (p in 1:3){
    parameters <- rbind(parameters, data.frame(size=n, name=parms[p], true=param[p
        ],estavg=mean(estparam[,(p+1)]),var=var(estparam[,p+1]),
                                               CIlow=quantile(estparam[,p+1],
                                                   0.025)[[1]],CIup=quantile(
                                                   estparam[,p+1], 0.975)[[1]]))
  }


}



# USED FUNCTIONS

# Performance calculation
perf_measures <- function(data, truetau, size, n){
  mean = mean(data$avgtau)
  var = var(data$avgtau)
  bias = sum(data$avgtau - truetau)/size
  SEbias = sqrt(var/size)
  empSE = sqrt(var)
  SEempSE = empSE/sqrt(2*(size-1))
  MSE = sum((data$avgtau - truetau)^2)/size
  SEMSE = sqrt(sum(((data$avgtau - truetau)^2-MSE)^2)/(size*(size-1)))
  CIlow = data$avgtau - qnorm(1-(0.05/2))*sqrt(var/n)
  CIup = data$avgtau + qnorm(1-(0.05/2))*sqrt(var/n)
  coverage = sum(CIlow <= truetau & truetau <= CIup)/size
  SEcoverage = sqrt(coverage*(1-coverage)/size)
  return(c(bias, SEbias, empSE, SEempSE, MSE, SEMSE, coverage, SEcoverage))
}


# joe's trivariate tau coording to his 1990 paper
sample_joe <- function(x){
  w2 <- -1/3  #precalculated w's
  w3 <- 1
  Iconc <- 0  #indicator for concordant triples counted
  Idisc <- 0  #indicator for discordant triples counted
  n <- nrow(x)
  for (i in 1:(n-1)){
    x1 = x[i,]
    tempconc <- sum( (x1[1]>x[(i+1):n,1] & x1[2]>x[(i+1):n,2] & x1[3]>x[(i+1):n,3])
        |
                       (x1[1]<x[(i+1):n,1] & x1[2]<x[(i+1):n,2] & x1[3]<x[(i+1):n
                           ,3]))
    Iconc <- Iconc + tempconc
    Idisc <- Idisc + (n-i-tempconc)
  }
  joetau <- w2 * 2/(n*(n-1)) * Idisc + w3 * 2/(n*(n-1)) * Iconc
  return(joetau)
```

```
}
```

```
# SIMULATION - NON-CONTINUOUS CASE


# SETTING UP LIBRARIES
library(Metrics)
library(boot)
library(Rmisc)
library(utils)
library(copula)
library(here)

library(mvtnorm)
library(Matrix)
library(MASS)
library(GenOrd)



# DEFINE PARAMETERS
n <- 300  #nr of observations (100, 500, 1000)
sim <- 1000 #nr of simulations, from Morris (2019) *

pair_tau <- c(0.8, 0.5, 0.2)  # possible correlation values
combinate <- data.frame(matrix(nrow=0,ncol=3)) # different triples of pairwise taus
    considered
colnames(combinate) <- c("tau12", "tau13", "tau23")
for (i in 1:3){
  combinate <- rbind(combinate, data.frame(tau12=pair_tau[[1]],tau13=pair_tau[[1]],
      tau23=pair_tau[[i]]))
}
for (j in 2:3){
  combinate <- rbind(combinate, data.frame(tau12=pair_tau[[1]],tau13=pair_tau[[j]],
      tau23=pair_tau[[3]]))
}
for (k in 2:3){
  combinate <- rbind(combinate, data.frame(tau12=pair_tau[[k]],tau13=pair_tau[[3]],
      tau23=pair_tau[[3]]))
}


# make-shift zip
marginals <- function(p0, lam){
  # RV1
  p01 <- p0[1]
  lam1 <- lam[1]
  pois1 <- ppois(0:qpois(0.99, lam1), lam1)
  m1 <- p01 + pois1*(1-p01)
  # RV2
  p02 <- p0[2]
  lam2 <- lam[2]
  pois2 <- ppois(0:qpois(0.99, lam2), lam2)
  m2 <- p02 + pois2*(1-p02)
  # RV3
  p03 <- p0[3]
  lam3 <- lam[3]
  pois3 <- ppois(0:qpois(0.99, lam3), lam3)
  m3 <- p03 + pois3*(1-p03)
  #marginals
  marge <- list( m1,  m2,  m3)
  return(marge)
}

# CHECKS for cor matrix
corrcheck(marge)
# for lam=2, p01=0.0, p02=0.0, p03=0.5    =>    -0.89 < cor12 < 1.00, -0.74 < cor13
```

```
    < 0.90, -0.74 < cor23 < 0.90
# for lam=2, p01=0.5, p02=0.5, p03=0.5    =>    -0.50 < cor12 < 1.00, -0.50 < cor13
    < 1.00, -0.50 < cor23 < 1.00
# for lam=2, p01=0.2, p02=0.5, p03=0.8    =>    -0.70 < cor12 < 0.92, -0.42 < cor13
    < 0.76, -0.27 < cor23 < 0.84

for (p in c(4,5,6,7)){
  print(p)
  comb <- c(combinate[p,1], combinate[p,2], combinate[p,3])
  sig <- matrix(c(1.0,comb[[1]],comb[[2]],
                  comb[[1]],1.0,comb[[3]],
                  comb[[2]],comb[3],1.0),
                3, 3, byrow=TRUE)
  #print(det(sig))  #NOT combinate[3]
  x0<- ordsample(n*sim, marge3, sig)
}
# all => not 3
# for lam=2, p01=0.0, p02=0.0, p03=0.5    =>    ok for all
# for lam=2, p01=0.5, p02=0.5, p03=0.5    =>    ok for all
# for lam=2, p01=0.2, p02=0.5, p03=0.8    =>    only 4, 5, 6, 7


# ------------------------------------------------------------------


# SIMULATION FUNCTIONS
# simulate 1 run
sim1run <- function(x0, i, n){
  x <- x0[((i-1)*n+1):(i*n),]

  # compute values of interest
  # - tau-b
  tempK <- cor(x, method="kendall")
  tempbtau <- c(tempK[[1,2]], tempK[[1,3]], tempK[[2,3]])
  tempavgbtau <- (tempbtau[[1]]+tempbtau[[2]]+tempbtau[[3]])/3
  # - tau-PHZ
  tempPHZtau <- c(calculate_phz_2022(x[,1], x[,2]), calculate_phz_2022(x[,1], x
      [,3]), calculate_phz_2022(x[,2], x[,3]))
  tempavgPHZtau <- (tempPHZtau[[1]]+tempPHZtau[[2]]+tempPHZtau[[3]])/3
  # - Joe
  tempjoeTtau <- sample_joe_noties(x)
  tempjoeJtau <- sample_joe_jitter(x)

  # save values in dataframe
  tempestimates <- data.frame(Simulation=i, btau12=tempbtau[[1]],btau13=tempbtau
      [[2]],btau23=tempbtau[[3]],
                              avgbtau=tempavgbtau,tau12=tempPHZtau[[1]],tau13=
                                  tempPHZtau[[2]], tau23=tempPHZtau[[3]],
                              avgtau=tempavgPHZtau,joeTtau=tempjoeTtau, joeJtau=
                                  tempjoeJtau)

  # save info on generated data
  estcor <- cor(x, method="pearson")
  tempestparam <- data.frame(Simulation=i, est1p0=sum(x[,1]==0)/n, est2p0=sum(x
      [,2]==0)/n, est3p0=sum(x[,3]==0)/n,
                                                  estlam1=mean(split(x[,1], x[,1]==0)$`
                                                      FALSE`), estlam2=mean(split(x[,2],
                                                       x[,2]==0)$`FALSE`), estlam3=mean(
                                                       split(x[,3], x[,3]==0)$`FALSE`),
                                                  estcor12=estcor[1,2],estcor13=estcor
                                                      [1,3],estcor23=estcor[2,3])

  # checks for exceptions
  tempix01 <- ifelse(c(sum((x[,1] == 0) & (x[,2] > 0)), sum((x[,1] == 0) & (x[,3] >
       0)), sum((x[,2] == 0) & (x[,3] > 0)))==0, 1,0)
  tempix11 <- ifelse(c(sum((x[,1] > 0) & (x[,2] > 0)), sum((x[,1] > 0) & (x[,3] >
      0)), sum((x[,2] > 0) & (x[,3] > 0)))==0, 1,0)
```

```
  tempix10 <- ifelse(c(sum((x[,1] > 0) & (x[,2] == 0)), sum((x[,1] > 0) & (x[,3] ==
      0)), sum((x[,2] > 0) & (x[,3] == 0)))==0, 1,0)
  tempcount <- list(tempix01, tempix11, tempix10)

  return(list(tempestimates, tempestparam, tempcount))

}

# simulation for fixed values
sim1simulation <- function(p0, lam, p, n, sim){
  set.seed(9)

  # temporary data
  estimates <- data.frame(matrix(ncol=11, nrow=0))
  colnames(estimates) <- c("Simulation", "btau12","btau13","btau23", "avgbtau", "
      tau12","tau13","tau23", "avgtau", "joeTtau", "joeJtau")
  estparam <- data.frame(matrix(ncol=10, nrow=0))
  colnames(estparam) <- c("Simulation", "est1p0","est2p0","est3p0","estlam1","
      estlam2","estlam3","estcor12","estcor13","estcor23")

  countix10 <- c(0,0,0)
  countix11 <- c(0,0,0)
  countix01 <- c(0,0,0)

  # generate data
  marge <- marginals(p0, lam)
  comb <- c(combinate[p,1], combinate[p,2], combinate[p,3])
  sig <- matrix(c(1.0,comb[[1]],comb[[2]],
                  comb[[1]],1.0,comb[[3]],
                  comb[[2]],comb[3],1.0),
                3, 3, byrow=TRUE)
  x0<- ordsample(n*sim, marge, sig)-1

  # simulate runs
  for (i in 1:sim){
    returns <- sim1run(x0, i, n)
    estimates <- rbind(estimates, returns[[1]])
    estparam <- rbind(estparam, returns[[2]])
    countix01 <- countix01 + returns[[3]][[1]]
    countix11 <- countix11 + returns[[3]][[2]]
    countix10 <- countix10 + returns[[3]][[3]]
  }

  # save simulation data
  # for performance
  tempempSE <- c(sqrt(var(estimates$avgbtau)),sqrt(var(estimates$avgtau)),sqrt(var(
      estimates$joeTtau)),sqrt(var(estimates$joeJtau)))
  tempSEempSE <- c(tempempSE[[1]]/sqrt(2*(sim-1)), tempempSE[[2]]/sqrt(2*(sim-1)),
      tempempSE[[3]]/sqrt(2*(sim-1)), tempempSE[[4]]/sqrt(2*(sim-1)))
  tempperf <- data.frame(size=n,  estbtau12=mean(estimates$btau12), estbtau13=mean(
      estimates$btau13), estbtau23=mean(estimates$btau23),
                         estavgbtau=mean(estimates$avgbtau), esttau12=mean(
                             estimates$tau12), esttau13=mean(estimates$tau13),
                         esttau23=mean(estimates$tau23), estavgtau=mean(estimates$
                             avgtau), estjoeTtau=mean(estimates$joeTtau),
                             estjoeJtau=mean(estimates$joeJtau),
                         empSEb=tempempSE[1], SEempSEb=tempSEempSE[1], empSE=
                             tempempSE[2], SEempSE=tempSEempSE[2], empSEjT=
                             tempempSE[3], SEempSEjT=tempSEempSE[3],
                         empSEjJ=tempempSE[4], SEempSEjJ=tempSEempSE[4])

  # for parameters
  tempparam <- data.frame(size=n, name="p01", true=p0[1],estavg=mean(estparam$
      est1p0),var=var(estparam$est1p0),
                                                  CIlow=quantile(estparam$est1p0, 0.025)
                                                      [[1]],CIup=quantile(estparam$
                                                      est1p0, 0.975)[[1]])
```

```
  tempparam <- rbind(tempparam, data.frame(size=n, name="p02", true=p0[2],estavg=
      mean(estparam$est2p0),var=var(estparam$est2p0),
                                           CIlow=quantile(estparam$est2p0, 0.025)
                                               [[1]],CIup=quantile(estparam$
                                               est2p0, 0.975)[[1]]))
  tempparam <- rbind(tempparam, data.frame(size=n, name="p03", true=p0[3],estavg=
      mean(estparam$est3p0),var=var(estparam$est3p0),
                                           CIlow=quantile(estparam$est3p0, 0.025)
                                               [[1]],CIup=quantile(estparam$
                                               est3p0, 0.975)[[1]]))

  tempparam <- rbind(tempparam, data.frame(size=n, name="lam1", true=lam[1],estavg=
      mean(estparam$estlam1),var=var(estparam$estlam1),
                                           CIlow=quantile(estparam$estlam1,
                                               0.025)[[1]],CIup=quantile(estparam
                                               $estlam1, 0.975)[[1]]))
  tempparam <- rbind(tempparam, data.frame(size=n, name="lam2", true=lam[2],estavg=
      mean(estparam$estlam2),var=var(estparam$estlam2),
                                           CIlow=quantile(estparam$estlam2,
                                               0.025)[[1]],CIup=quantile(estparam
                                               $estlam2, 0.975)[[1]]))
  tempparam <- rbind(tempparam, data.frame(size=n, name="lam3", true=lam[3],estavg=
      mean(estparam$estlam3),var=var(estparam$estlam3),
                                           CIlow=quantile(estparam$estlam3,
                                               0.025)[[1]],CIup=quantile(estparam
                                               $estlam3, 0.975)[[1]]))

  tempparam <- rbind(tempparam, data.frame(size=n, name="cor12", true=sig[1,2],
      estavg=mean(estparam$estcor12),var=var(estparam$estcor12),
                                           CIlow=quantile(estparam$estcor12,
                                               0.025)[[1]],CIup=quantile(estparam
                                               $estcor12, 0.975)[[1]]))
  tempparam <- rbind(tempparam, data.frame(size=n, name="cor13", true=sig[1,3],
      estavg=mean(estparam$estcor13),var=var(estparam$estcor13),
                                           CIlow=quantile(estparam$estcor13,
                                               0.025)[[1]],CIup=quantile(estparam
                                               $estcor13, 0.975)[[1]]))
  tempparam <- rbind(tempparam, data.frame(size=n, name="cor23", true=sig[2,3],
      estavg=mean(estparam$estcor23),var=var(estparam$estcor23),
                                           CIlow=quantile(estparam$estcor23,
                                               0.025)[[1]],CIup=quantile(estparam
                                               $estcor23, 0.975)[[1]]))

  counts <- list(countix01, countix11, countix10)

  return(list(x0, estimates, estparam, tempperf, tempparam, counts))
}



# SIMULATION
# main dataframes set-up
performance <- data.frame(matrix(ncol=16, nrow=0))
colnames(performance) <- c("size", "estbtau12","estbtau13","estbtau23","estavgbtau"
    ,
                           "esttau12","esttau13","esttau23","estavgtau", "
                               estjoeTtau","estjoeJtau",
                           "empSEb","SEempSEb", "empSE","SEempSE", "empSEjT","
                               SEempSEjT", "empSEjJ","SEempSEjJ")
parameters <- data.frame(matrix(ncol=7, nrow=0))
colnames(parameters) <- c("size", "name","true","estavg","var", "CIlow","CIup")

results <- vector(mode = "list", length = 12)
probs <- list(c(0,0,0.5),c(0.5,0.5,0.5),c(0.2,0.5,0.8))
i <- 0
pb <- winProgressBar(min = 0,max = 12)
for (prob in probs){
```

```
  for (p in 4:7){
    tempresults <- sim1simulation(prob, c(2,2,2), p, n, sim)
    results[[4*i+(p-3)]] <- tempresults
    setWinProgressBar(pb, 4*i+(p-3))
  }
  i <- i+1
}
close(pb)

taujitterlist <- data.frame()
for (j in 1:12){
  taujitter <- data.frame()
  for (i in 1:sim){
    x <- results[[j]][[1]][((i-1)*n+1):(i*n),]

    # compute values of interest
    # - tau-b
    tempK <- cor(jitter(x), method="kendall")
    tempbtau <- c(tempK[[1,2]], tempK[[1,3]], tempK[[2,3]])
    tempavgbtau <- (tempbtau[[1]]+tempbtau[[2]]+tempbtau[[3]])/3

    taujitter <- rbind(taujitter, data.frame(tempavgbtau))

  }
  print(j)
  taujitterlist <- rbind(taujitterlist, data.frame(mean(taujitter$tempavgbtau)))
}



results <- sim1simulation(c(0,0,0.5), c(2,2,2), 4, n, sim)

d <- vector(mode = "list", length = 3)
print(d)
d[[1]] <- list(2,4,1)
d[[2]] <- 7
d[[3]] <- c(5,2)
print(d)

d[[1]]
# USED FUNCTIONS

# joe's trivariate tau coording to his 1990 paper
sample_joe_noties <- function(x){
  w2 <- -1/3  #precalculated w's
  w3 <- 1
  Iconc <- 0  #indicator for concordant triples counted
  Idisc <- 0  #indicator for discordant triples counted
  n <- nrow(x)
  for (i in 1:(n-1)){
    x1 = x[i,]
    tempconc <- sum( (x1[1]>x[(i+1):n,1] & x1[2]>x[(i+1):n,2] & x1[3]>x[(i+1):n,3])
          |
                      (x1[1]<x[(i+1):n,1] & x1[2]<x[(i+1):n,2] & x1[3]<x[(i+1):n
                          ,3]))
    temptie <- sum(x1[1]==x[(i+1):n,1] | x1[2]==x[(i+1):n,2] | x1[3]==x[(i+1):n,3])
         #ignore ties
    Iconc <- Iconc + tempconc
    Idisc <- Idisc + (n-i-tempconc- temptie)
  }
  joetau <- w2 * 2/(n*(n-1)) * Idisc + w3 * 2/(n*(n-1)) * Iconc
  return(joetau)
}

sample_joe_jitter <- function(x){
  x <- jitter(x)
  w2 <- -1/3  #precalculated w's
```

```
  w3 <- 1
  Iconc <- 0  #indicator for concordant triples counted
  Idisc <- 0  #indicator for discordant triples counted
  n <- nrow(x)
  for (i in 1:(n-1)){
    x1 = x[i,]
    tempconc <- sum( (x1[1]>x[(i+1):n,1] & x1[2]>x[(i+1):n,2] & x1[3]>x[(i+1):n,3])
        |
                        (x1[1]<x[(i+1):n,1] & x1[2]<x[(i+1):n,2] & x1[3]<x[(i+1):n
                            ,3]))
    Iconc <- Iconc + tempconc
    Idisc <- Idisc + (n-i-tempconc)
  }
  joetau <- w2 * 2/(n*(n-1)) * Idisc + w3 * 2/(n*(n-1)) * Iconc
  return(joetau)
}
```

```
# Synthetic data analysis


# SETTING UP LIBRARIES
library(Metrics)
library(boot)
library(Rmisc)
library(utils)
library(copula)
library(here)

library(mvtnorm)
library(Matrix)
library(MASS)
library(GenOrd)


# SET UP DATA
syntheticdata <- read_excel("C:/Users/20191258/Downloads/syntheticdata.xls")

# reduce to three sizes
trivariatedata <- data.frame(matrix(ncol=3, nrow=1800))
colnames(trivariatedata) <- c("sizeA", "sizeB", "sizeC")
trivariatedata$sizeA <- syntheticdata$size1 + syntheticdata$size2
trivariatedata$sizeB <- syntheticdata$size3 + syntheticdata$size4
trivariatedata$sizeC <- syntheticdata$size5 + syntheticdata$size6


# DATA ANALYSIS

# quick summary for some info
summary(syntheticdata[2:7])

# find the percentage wise frequency of values ("estimated pmf")
n <- nrow(syntheticdata)
maxval <- 17 #from summary
probs <- data.frame(matrix(nrow=6, ncol=19))
for (s in 1:6){
  temppro <-c(s)
  for (i in 0:maxval){
    temppro <- c(temppro, sum(syntheticdata[[s+1]]==i)/n)
  }
  probs[s,] <- temppro
}

# assuming zip we can estimate lambda and p_o
# mean = (1-p0)lam
# var = (1-p0)lam(1+p0 lam)
params <- data.frame(matrix(nrow=6, ncol=3))
for (s in 1:6){
```

```
  m <- mean(syntheticdata[[s+1]])
  v <- var(syntheticdata[[s+1]])
  p0 <- (v-m)/(m^2-m+v)
  lam <- m/(1-p0)
  if (p0<0){ #only poisson left, mean=lam & var=lam
    p0 <- 0
    lam <- (m+v)/2
  }
  params[s,] <- c(s, p0, lam)
}


# Calculate pearson cor matrix, considering used for data generation
cormat <- cor(syntheticdata[2:7], method = "pearson")
View(cormat)


# COMPUTE TAUS

# trivariate
joetau <- sample_joe(trivariatedata)
taub <- cor(trivariatedata, method="kendall")
avgtaub <- (taub[[1,2]] + taub[[1,3]] + taub[[2,3]])/3
tauphz <- c(calculate_phz_2022(trivariatedata$sizeA, trivariatedata$sizeB),
            calculate_phz_2022(trivariatedata$sizeA, trivariatedata$sizeC),
            calculate_phz_2022(trivariatedata$sizeB, trivariatedata$sizeC))
avgtauphz <- (tauphz[[1]] + tauphz[[2]] + tauphz[[3]])/3

# not trivariate
joetau <- sample_joe6(syntheticdata[2:7])
taub <- cor(syntheticdata[2:7], method="kendall")
tauphz <- c()
for (i in 2:6){
  for (j in (i+1):7){
    tauphz <- c(tauphz, calculate_phz_2022(syntheticdata[[i]], syntheticdata[[j]]))
  }
}



# FUNCTIONS
# calculate joe's sixvariate tau coording to his 1990 paper
# [this still has mistakes i believe]
sample_joe6 <- function(x){
  x <- data.frame(matrix(unlist(x), nrow=nrow(x)))
  w <- function (d, k){
    return((choose(k,2)+choose(d-k,2)-k*(d-k))/choose(d,2))
  }
  # necessary w's
  w3 <- w(6,3)
  w4 <- w(6,4)
  w5 <- w(6,5)
  w6 <- w(6,6)

  # intitialize counters
  Ik6 <- 0
  Ik5 <- 0
  Ik4 <- 0
  Ik3 <- 0

  n <- nrow(x)
  m <- c(1,2,3,4,5,6)
  for (i in 1:(n-1)){
    x1 = x[i,]
    #ignore ties
    temptie <- sum(x1[1]==x[(i+1):n,1] | x1[2]==x[(i+1):n,2] | x1[3]==x[(i+1):n,3]
        |
```

```
                    x1[4]==x[(i+1):n,4] | x1[5]==x[(i+1):n,5] | x1[6]==x[(i+1):n
                        ,6]) #ignore ties
    #concordance
    temIk6 <- sum( (x1[1]>x[(i+1):n,1] & x1[2]>x[(i+1):n,2] & x1[3]>x[(i+1):n,3] &
        x1[4]>x[(i+1):n,4] & x1[5]>x[(i+1):n,5] & x1[6]>x[(i+1):n,6]) |
                      (x1[1]<x[(i+1):n,1] & x1[2]<x[(i+1):n,2] & x1[3]<x[(i+1):n
                          ,3] & x1[4]<x[(i+1):n,4] & x1[5]<x[(i+1):n,5] & x1[6]<x
                          [(i+1):n,6]))
    temIk5 <- 0
    temIk4 <- 0
    for (k in 1:6){
      tm <- m[m!=k]
      temIk5 <- temIk5 + sum( (x1[tm[1]]>x[(i+1):n,tm[1]] & x1[tm[2]]>x[(i+1):n,tm
          [2]] & x1[tm[3]]>x[(i+1):n,tm[3]] &
                                x1[tm[4]]>x[(i+1):n,tm[4]] & x1[tm[5]]>x[(i+1):n,tm
                                    [5]] & x1[k]<x[(i+1):n,k]) |
                                (x1[tm[1]]<x[(i+1):n,tm[1]] & x1[tm[2]]<x[(i+1):n,tm
                                    [2]] & x1[tm[3]]<x[(i+1):n,tm[3]] &
                                x1[tm[4]]<x[(i+1):n,tm[4]] & x1[tm[5]]<x[(i+1):n,tm
                                    [5]] & x1[k]>x[(i+1):n,k]) )
      for (j in k:6){
        tm2 <- tm[tm!=j]
        temIk4 <- temIk4 + sum( (x1[tm2[1]]>x[(i+1):n,tm2[1]] & x1[tm2[2]]>x[(i+1):
            n,tm2[2]] & x1[tm2[3]]>x[(i+1):n,tm2[3]] &
                                  x1[tm2[4]]>x[(i+1):n,tm2[4]] & x1[j]<x[(i+1):n,j
                                      ] & x1[k]<x[(i+1):n,k]) |
                                 (x1[tm2[1]]<x[(i+1):n,tm2[1]] & x1[tm2[2]]<x[(i
                                     +1):n,tm2[2]] & x1[tm2[3]]<x[(i+1):n,tm2[3]]
                                     &
                                    x1[tm2[4]]<x[(i+1):n,tm2[4]] & x1[j]>x[(i+1):n
                                        ,j] & x1[k]>x[(i+1):n,k]) )
      }
    }

    Ik6 <- Ik6 + temIk6
    Ik5 <- Ik5 + temIk5
    Ik4 <- Ik4 + temIk4
    Ik3 <- Ik3 + (n-i-temIk6 - temIk5 - temIk4 - temptie)
  }
  joetau <- 2/(n*(n-1)) *(w6 * Ik6 + w5 * Ik5 + w4 * Ik4 + w3 *  Ik3)
  return(joetau)
}

sample_joeJ6 <- function(x){
  x <- data.frame(matrix(unlist(x), nrow=nrow(x)))
  w <- function (d, k){
    return((choose(k,2)+choose(d-k,2)-k*(d-k))/choose(d,2))
  }
  # necessary w's
  w3 <- w(6,3)
  w4 <- w(6,4)
  w5 <- w(6,5)
  w6 <- w(6,6)

  # intitialize counters
  Ik6 <- 0
  Ik5 <- 0
  Ik4 <- 0
  Ik3 <- 0

  n <- nrow(x)
  m <- c(1,2,3,4,5,6)
  for (i in 1:(n-1)){
    x1 = x[i,]
    #ignore ties
    #temptie <- sum(x1[1]==x[(i+1):n,1] | x1[2]==x[(i+1):n,2] | x1[3]==x[(i+1):n,3]
        |
```

```
    #                   x1[4]==x[(i+1):n,4] | x1[5]==x[(i+1):n,5] | x1[6]==x[(i+1):n
        ,6]) #ignore ties
    #concordance
    temIk6 <- sum( (x1[1]>x[(i+1):n,1] & x1[2]>x[(i+1):n,2] & x1[3]>x[(i+1):n,3] &
        x1[4]>x[(i+1):n,4] & x1[5]>x[(i+1):n,5] & x1[6]>x[(i+1):n,6]) |
                     (x1[1]<x[(i+1):n,1] & x1[2]<x[(i+1):n,2] & x1[3]<x[(i+1):n,3]
                         & x1[4]<x[(i+1):n,4] & x1[5]<x[(i+1):n,5] & x1[6]<x[(i+1):
                         n,6]))
    temIk5 <- 0
    temIk4 <- 0
    for (k in 1:6){
      tm <- m[m!=k]
      temIk5 <- temIk5 + sum( (x1[tm[1]]>x[(i+1):n,tm[1]] & x1[tm[2]]>x[(i+1):n,tm
          [2]] & x1[tm[3]]>x[(i+1):n,tm[3]] &
                                  x1[tm[4]]>x[(i+1):n,tm[4]] & x1[tm[5]]>x[(i+1):n,
                                      tm[5]] & x1[k]<x[(i+1):n,k]) |
                                 (x1[tm[1]]<x[(i+1):n,tm[1]] & x1[tm[2]]<x[(i+1):n,
                                     tm[2]] & x1[tm[3]]<x[(i+1):n,tm[3]] &
                                    x1[tm[4]]<x[(i+1):n,tm[4]] & x1[tm[5]]<x[(i+1):n
                                        ,tm[5]] & x1[k]>x[(i+1):n,k]) )
      for (j in k:6){
        tm2 <- tm[tm!=j]
        temIk4 <- temIk4 + sum( (x1[tm2[1]]>x[(i+1):n,tm2[1]] & x1[tm2[2]]>x[(i+1):
            n,tm2[2]] & x1[tm2[3]]>x[(i+1):n,tm2[3]] &
                                  x1[tm2[4]]>x[(i+1):n,tm2[4]] & x1[j]<x[(i+1):n,j
                                      ] & x1[k]<x[(i+1):n,k]) |
                                 (x1[tm2[1]]<x[(i+1):n,tm2[1]] & x1[tm2[2]]<x[(i
                                     +1):n,tm2[2]] & x1[tm2[3]]<x[(i+1):n,tm2[3]]
                                    &
                                    x1[tm2[4]]<x[(i+1):n,tm2[4]] & x1[j]>x[(i+1):n
                                        ,j] & x1[k]>x[(i+1):n,k]) )
      }
    }

    Ik6 <- Ik6 + temIk6
    Ik5 <- Ik5 + temIk5
    Ik4 <- Ik4 + temIk4
    Ik3 <- Ik3 + (n-i-temIk6 - temIk5 - temIk4)
  }
  joetau <- 2/(n*(n-1)) *(w6 * Ik6 + w5 * Ik5 + w4 * Ik4 + w3 *  Ik3)
  return(joetau)
}
```