Fast Estimation of Kendall's Tau and Conditional Kendall's Tau Matrices under Structural Assumptions

R.A.J. Van der Spek

TU Delft

Fast Estimation of Kendall's Tau and Conditional Kendall's Tau Matrices under Structural Assumptions

by

R.A.J. van der Spek

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

Student number:	4452259	
Project duration:	September 6, 2021 – March 29, 2022	
Thesis committee:	Prof. dr. ir. G. Jongbloed,	TU Delft
	Dr. D. Kurowicka,	TU Delft
	Dr. A. F. F. Derumigny,	TU Delft, supervisor

This thesis is to be defended publicly on Tuesday March 29, 2022 at 10:00 AM.

TU Delft

Abstract

Kendall's tau and conditional Kendall's tau matrices are multivariate (conditional) dependence measures between the components of a random vector. For large dimensions, available estimators are computationally expensive and can be improved by averaging. Under structural assumptions on the underlying Kendall's tau and conditional Kendall's tau matrices, we introduce new estimators that have a significantly reduced computational cost while keeping a similar error level. In the unconditional setting we assume that, up to reordering, the underlying Kendall's tau matrix is block-structured with constant values in each of the off-diagonal blocks. The estimators take advantage of this block structure by averaging over (part of) the pairwise estimates in each of the off-diagonal blocks. Derived explicit variance expressions show their improved efficiency. In the conditional setting, the conditional Kendall's tau matrix is assumed to have a constant block structure, independently of the conditioning variable. Conditional Kendall's tau matrix estimators are constructed similarly as in the unconditional case by averaging over (part of) the pairwise conditional Kendall's tau estimators. We establish their joint asymptotic normality, and show that the asymptotic variance is reduced compared to the naive estimators. Then, we perform a simulation study which displays the improved performance of both the unconditional and conditional estimators. Finally, the estimators are used for estimating the value at risk of a large stock portfolio; backtesting illustrates the obtained improvements compared to the previous estimators.

Acknowledgements

This thesis concludes my time as a master student Applied Mathematics at the Delft University of Technology. This also marks the end of my time as a university student, which has been a very fruitful and a very enjoyable period in my life. I embarked in Delft in 2015, first to complete the bachelors in Applied Mathematics and Applied Physics. Although I find physics incredibly interesting to hear and learn about, it soon became clear that my passion lies above all with mathematics and especially with its applications in finance. I thoroughly enjoyed completing the courses of the Financial Engineering specialisation, as well as the two internships I did in financial modelling. Therefore, when Alexis first brought up this project in the spring of 2021, I immediately became very enthusiastic.

I would like to take this opportunity to thank Alexis tremendously for his guidance on the project form start to finish. Your engagement in the project and the amount of time you spent with me each week, I really appreciate. You also helped me a lot in realising my fixed deadline, regarding the start of my future job shortly after. I really enjoyed unraveling the problem together, and have found our collaboration to be very pleasant. Furthermore, I would also like to thank to the two other members of the thesis committee, Prof. dr. ir. G. Jongbloed and Dr. D. Kurowicka for taking the time to read my report and attend the presentation.

Finally, I would like to thank my family and friends. In particular, Koen whom I have come to call a colleague these days through our many late-night sessions in the university library in which he kept me good company, and Rik who was willing to read through the thesis one last time.

R.A.J. van der Spek

Rotterdam, March 2022

Contents

1 Introduction

The problem of analysing interrelationships between different elements of a multivariate random vector $\mathbf{X} = (X_1, \ldots, X_p)$ arises in numerous applications in many fields [14]. This often includes the adoption of a certain statistical model involving the correlation matrix of all pairwise combinations of the random vector $\mathbf{X}$. It is therefore highly relevant to be able to estimate correlation matrices in all kinds of different types and settings [15].

For multivariate Gaussian data, the classical choice is Pearson's linear correlation coefficient, which is the ratio of the covariance of two variables and the product of their standard deviations. It is well known that Pearson's sample correlation matrix is generally not suitable for capturing correlations of heavy-tailed data. It is based on sub-Gaussian theory and is therefore very sensitive to outliers [28]. Consequently, robust measures of dependence should be used when heavy tails are involved [4].

Rank correlation coefficients measure the relationship between the ranking of two ordinal random variables and are hereby nonparametric and robust measures of correlation [3]. The most used ones are Spearman's rho and Kendall's tau, with the latter being of particular interest. This is due to a direct transformation between Kendall's tau and Pearson's correlation coefficient for the general class of elliptical distributions, a common model for financial data [32, 36]. This allows for robust covariance estimation after combining Pearson's correlation with the marginal standard deviations.

Estimation of the $p \times p$ Kendall's tau matrix $\mathbf{T}$ becomes particularly challenging in the high-dimensional setting when p is large. Simple use of the naive Kendall's tau matrix estimator of all pairwise sample Kendall's taus will result in noisy estimates with estimation errors piling up due to the estimates' individual imprecision [13]. Over the past two decades, various regularisation strategies have been proposed to reduce the aggregation of estimation errors. Ultimately, these methods all make certain assumptions on the underlying dependence structure and hereby reduce the number of free parameters to estimate.

In many instances, sparsity of the target matrix is assumed. For such settings, various (combinations of) thresholding and shrinkage methods have been proposed, see for example [2, 24, 44]. However, such assumptions are certainly not appropriate for the modelling of most financial data, e.g. market risk is reflected in all share prices and therefore their returns are certainly correlated. To this end, factor models are usually imposed, where the correlations depend on a number of common factors, which may or may not be latent, see [13, 16].

In 2019, Perreault et al. [40] have introduced an alternative approach to estimating large Kendall's tau matrices. They studied a model in which it is assumed that the set of variables could be partitioned into smaller clusters with exchangeable dependence. As such, after reorder-

ing of the variables by cluster, the corresponding Kendall's tau matrix is block-structured with constant values within each block. Following naturally is an improved estimation by averaging all pairwise Kendall's taus within each of the blocks. Additionally, they have proposed a robust algorithm identifying such structures (see also [41] for testing for the presence of such a structure).

In this thesis, we study a more flexible framework by assuming that the set of variables can be grouped in a way such that the pairwise Kendall's tau between variables of different groups is only dependent on group numbers. As such, after reordering of the variables by group the corresponding Kendall's tau matrix is again block-structured, but with arbitrary (non-constant) values within the diagonal blocks. The interest in this relaxation originates from the application to modelling financial stock returns; we could, for example, assume that the returns of stocks from a certain sector and economy have constant correlations with the returns of stocks from a certain other group, but that they have varying correlations with each other since they operate within the same sector and economy.

We propose an estimator similar to the one studied in [40], but that applies averaging to only the off-diagonal blocks, and study its efficiency. One of the drawbacks of the estimator studied in [40] is its computational cost, which is close to the one of the naive Kendall's tau matrix estimator: the number of pairwise sample Kendall's taus that are to be computed scales quadratically with the dimension p. Naturally, the idea of averaging among several Kendall's taus can be applied to part of the blocks, which allows for faster computations. As such, we propose several estimators that average among part of the Kendall's tau per off-diagonal block, and study their efficiencies and computational costs. For every off-diagonal block, we will consider averaging over elements in the same row, averaging over elements on the diagonal and averaging over a number of randomly selected elements. We will be referring to these estimators as the *row*, *diagonal* and *random* estimators; the estimator that averages over all elements is referred to as the *block* estimator.

Further, we investigate the extension of this model to the conditional setup when a d-dimensional covariate $\mathbf{Z} \in \mathcal{Z}$ is available. In this setting, Kendall's taus are depending on $\mathbf{Z}$ and we assume that the set of variables can be clustered such that for all $\mathbf{z} \in \mathcal{Z}$ the Kendall's tau conditional on $\mathbf{Z} = \mathbf{z}$ between variables of different groups is only depending on group numbers and on the value of $\mathbf{z}$. In view of applications to finance, the conditional version of our structural assumption could, for example, be seen as assuming that the correlations between European stocks of two different groups are equal and react equally in changes of some other American stock or portfolio. Furthermore, in [1, 12, 33], it was shown that stock returns actually exhibit higher correlations during market declines than during market upturns, and moreover that the same applies to exchange rates in [37]. This illustrates that analysing a model in which correlations depend on some conditioning variable is certainly relevant.

In this framework, we adopt nonparametric estimates of the conditional Kendall's tau based on kernel smoothing. Based on these nonparametric estimates we introduce conditional versions

of the averaging estimators and study their asymptotic behaviour as the sample size n tends to infinity. It is worth noting that conditional estimates of Kendall's tau using kernel smoothing carry significantly more computational cost than their unconditional counterparts, especially when the covariate's dimension d is large. Therefore, faster computations of conditional Kendall's tau matrices will be of particular use in the conditional, nonparametric setup.

This thesis is structured as follows. Chapter 2 provides an overview of the theory underlying the topic. In order to understand the basis of multivariate dependence structures we touch upon the concept of (conditional) copulas. Furthermore, we discuss the idea of rank-based correlation and examine the generic class of elliptical distributions. Next, we consider the theory behind U-statistics, which serves as the foundation for the proofs in subsequent chapters. Moreover, we discuss the concept of kernel smoothing for the construction of nonparametric conditional estimators. Chapter 3 sets the stage for improving the estimation of the unconditional Kendall's tau matrix. We formalise the proposed model, and follow up with a construction of estimators. Then, we derive their explicit variance expressions to compare them to the usual sample Kendall's tau matrix. Similarly, Chapter 4 is devoted to the improved estimation of the conditional Kendall's tau matrix. We discuss several nonparametric estimators of the conditional Kendall's tau in order to define the conditional versions of the averaging estimators. After formalising the conditional setup, we derive the estimators' joint asymptotic normality at different points of the covariate. This allows us to theoretically compare the performance of the different estimators. In Chapter 5 we perform a simulations study in order to support the theoretical findings. Here we examine the (conditional) estimators' performances for varying block dimensions and sample sizes. Additionally, we study the bandwidth choice that is associated with kernel regression. Finally, in Chapter 6, we examine a possible application to study the behaviour of the estimators in real data conditions. The estimators are used for the robust inference of the covariance matrix to calculate the value at risk of a large stock portfolio. We perform backtesting in order to asses the results. The conclusions of this thesis are summarised in Chapter 7.

2 Preliminaries

This chapter aims at providing the necessary fundamentals for the material discussed in subsequent chapters. We start by introducing the concept of (conditional) copulas in Section 2.1 in order to understand the basics of dependence in multivariate random variables. We then follow up with a short background on rank-based correlation in Section 2.2, where we formally define Spearman's rho and Kendall's tau. Furthermore, in Section 2.3 we give the definition of elliptical distributions together with some of their main properties, in particular that of the relation between Pearson's correlation coefficient and Kendall's tau. We then follow up with a brief background on U-statistics in Section 2.4, which serves as the basis for the proofs in subsequent chapters. Lastly, in Section 2.5 we discuss the concept of kernel regression estimation for understanding the basics of nonparametric conditional estimation.

2.1 Copulas

The concept of copulas, first introduced by Sklar [47], links the multivariate distribution to its one-dimensional marginal distributions. As such, copulas effectively capture the dependence structure between the components of a random vector. They have therefore become a popular modelling tool for when multivariate dependence has interest. Let us start by stating the formal definition.

Definition 1. *Let $p \geq 2$ be integer. A copula is a function $C : [0,1]^p \to [0,1]$ with the following properties:*

1. *For any $j = 1, \ldots, p$ and all $u_j \in [0,1]$, $C(u_1, \ldots, u_{j-1}, 0, u_{j+1}, \ldots, u_p) = 0$.*
2. *For any $j = 1, \ldots, p$ and all $u_j \in [0,1]$, $C(1, \ldots, 1, u_j, 1, \ldots, 1) = u_j$.*
3. *C is p-increasing, i.e. for each hyperrectangle $A = \Pi_{j=1}^{p} [a_j, b_j] \subseteq [0,1]^p$ the C-volume of A is non-negative:*

$$\int_A \mathrm{d}C(\mathbf{u}) \geq 0.$$

In probabilistic terms, copulas are defined as the joint cumulative distribution function (CDF) of a multivariate random vector on the unit cube with uniform marginal distributions. When given this definition, the concept of copulas may seem rather abstract. However, the following theorem shows how it effectively connects the the joint CDF of any multivariate random variable with its one-dimensional marginals. For the proof we refer to [48].

Theorem 2 (Sklar's Theorem). *Let $\mathbf{X} = (X_1, \dots, X_p)$ be a random vector with joint CDF F and univariate marginal CDFs $F_1, \dots, F_p$. Then, there exists a copula C such that for all $\mathbf{x} \in \mathbb{R}^p$,*

$$F(\mathbf{x}) = C(F_1(x_1), \dots, F_p(x_p)).$$

In addition, C is given for all $\mathbf{u} \in [0,1]^p$ by

$$C(u_1, \dots, u_p) = F(F_1^-(u_1), \dots, F_p^-(u_p)),$$

where F_j^- denotes the generalised inverse of F_j for $j = 1, \dots, p$. Therefore, if $\mathbf{X}$ is continuous, then C is unique.

It follows from Theorem 2 that any joint CDF can be written in terms of its marginal CDFs and a copula. Conversely, if $F_1, \dots, F_p$ are continuous CFDs on $\mathbb{R}$, and C is a copula on $[0,1]^p$, then the function $F : \mathbb{R}^p \to [0,1]$ given by

$$F(x) = C(F_1(x_1), \dots, F_p(x_p)),$$

is the joint CDF of a random variable on $\mathbb{R}^d$ with marginals $F_1, \dots, F_p$. As a consequence, there exists a bijection between the joint CDF F and the decomposition $(F_1, \dots, F_p, C)$. Hereby, we can decouple any random vector's joint CDF in a "dependent" and "marginal" part, given by the copula and all of its marginals respectively.

Given the definition of copulas, the concept of copula densities follows naturally. If it exists, then the copula's probability density function (PDF) is obtained in the usual way as

$$c(u_1, \dots, u_p) = \frac{\partial^p C\,(u_1, \dots, u_p)}{\partial u_1 \cdots \partial u_p},$$

and it follows that the joint PDF f of $\mathbf{X}$ is given by

$$f(\mathbf{x}) = f_1(x_1) \cdot \ldots \cdot f_p(x_p) \cdot c(F_1(x_1), \dots, F_p(x_p)),$$

where $f_1, \dots, f_p$ denote the marginal densities. Copulas and copula densities are studied in a wide variety of different parametric families. For an overview of the most well-known ones we refer to [11].

One of the key characteristics of copulas is that they are invariant under monotonic transformations of the marginal distributions. This will be useful later on when dealing with rank-based correlations. For a proof on Theorem 3 and for a thorough review on copulas, see [11, 29].

Theorem 3. *Let $\mathbf{X} = (X_1, \dots, X_p)$ be a random vector with continuous marginal CDFs $F_1, \dots, F_p$ and a continuous copula C. Let $T_i : \mathbb{R} \to \mathbb{R}$, for $i = 1, \dots, d$ be strictly increasing functions. Then the*

dependence structure of the random vector

$$(T_1(X_1), \dots, T_p(X_p))$$

is also given by the copula C.

When a d-dimensional conditioning variable $\mathbf{Z}$ is available, the concept of copulas can be generalized. Such a generalization was first introduced for time series in [38, 39]. Conditional copulas effectively allow for the modelling of (time-varying) multivariate processes that depend on a certain (time-varying) covariate. This was further generalised in [17]. In a similar way as in the unconditional case, conditional copulas form the link between a conditional joint CDF and the corresponding conditional marginal CDFs. Let us start with a formal definition.

Definition 4. *Let* $p \geq 2$ *and* $\mathbf{Z}$ *be a conditioning vector taking values in* $\mathcal{Z} \subset \mathbb{R}^d$. *A conditional copula is a measurable function* $C : [0,1]^p \times \mathcal{Z} \to [0,1]$ *such that for* $\mathbb{P}_{\mathbf{Z}}$*-almost every* $\mathbf{z} \in \mathcal{Z}$ *the following properties are satisfied:*

1. *For any* $j = 1, \dots, p$ *and all* $u_j \in [0,1]$, $C(u_1, \dots, u_{j-1}, 0, u_{j+1}, \dots, u_p | \mathbf{Z} = \mathbf{z}) = 0$.
2. *For any* $j = 1, \dots, p$ *and all* $u_j \in [0,1]$, $C(1, \dots, 1, u_j, 1, \dots, 1 | \mathbf{Z} = \mathbf{z}) = u_j$.
3. *For each hyperrectangle* $A = \Pi_{j=1}^{p} [a_j, b_j] \subseteq [0,1]^p$ *the* C*-volume of* A *is non-negative:*

$$\int_A \mathrm{d}C(\mathbf{u} | \mathbf{Z} = \mathbf{z}) \geq 0.$$

As such, conditional copulas are defined as the conditional joint CDF of a multivariate random vector on the unit cube with uniform marginals. Next, let us simply state the conditional version of Sklar's theorem. For a proof we refer to [38].

Theorem 5 (Sklar's Theorem for Conditional Copulas)**.** *Let* $\mathbf{X}$ *and* $\mathbf{Z}$ *be random vectors taking values in respectively* $\mathbb{R}^p$ *and* $\mathcal{Z} \subset \mathbb{R}^d$. *Let the conditional joint CDF of* $\mathbf{X}$ *given* $\mathbf{Z} = \mathbf{z}$, *denoted by* $F_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$, *have conditional marginals* $F_{1|\mathbf{Z}=\mathbf{z}}, \dots, F_{p|\mathbf{Z}=\mathbf{z}}$. *Then, there exists a conditional copula, denoted by* $C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$, *such that for all* $\mathbf{x} \in \mathbb{R}^p$ *and all* $\mathbf{z} \in \mathcal{Z}$,

$$C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(u_1, \dots, u_p) = F_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(F^{-}_{1|\mathbf{Z}=\mathbf{z}}(u_1), \dots, F^{-}_{p|\mathbf{Z}=\mathbf{z}}(u_p)),$$

where $F^{-}_{j|\mathbf{Z}=\mathbf{z}}$ *denotes the generalised inverse of* $F_{j|\mathbf{Z}=\mathbf{z}}$ *for* $j = 1, \dots, p$. *Therefore, if* $\mathbf{X}|\mathbf{Z} = \mathbf{z}$ *is continuous, then* $C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$ *is unique.*

Similarly to the unconditional case, we can combine any set of continuous marginal CDFs with any conditional copula to form a well-defined conditional CDF. It then follows that there

exists a bijection between $F_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$ and the decomposition $(F_{1|\mathbf{Z}=\mathbf{z}}, \dots, F_{p|\mathbf{Z}=\mathbf{z}}, C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}})$. As such, the concept of conditional copulas effectively connects the conditional multivariate CDF with its conditional marginals.

Again, the concept of conditional copula densities follows naturally. If it exists, then the conditional copula's PDF is obtained as

$$c(u_1, \dots, u_p|\mathbf{Z}=\mathbf{z}) = \frac{\partial^p C\left(u_1, \dots, u_p|\mathbf{Z}=\mathbf{z}\right)}{\partial u_1 \cdots \partial u_p},$$

and it follows that the conditional joint PDF of $\mathbf{X}$ given $\mathbf{Z}=\mathbf{z}$, denoted by $f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$, is given by

$$f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(\mathbf{x}) = f_{1|\mathbf{Z}=\mathbf{z}}(x_1) \cdot \ldots \cdot f_{p|\mathbf{Z}=\mathbf{z}}(x_p) \cdot c_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}(F_{1|\mathbf{Z}=\mathbf{z}}(x_1), \dots, F_{p|\mathbf{Z}=\mathbf{z}}(x_p)),$$

where $c_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$ denotes the conditional copula density and $f_{1|\mathbf{Z}=\mathbf{z}}, \dots, f_{1|\mathbf{Z}=\mathbf{z}}$ denote the conditional marginal densities.

Lastly, let us point out that the dependency of the conditional copula $C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$ with respect to the conditioning point $\mathbf{z}$ introduces significant complexities in terms of model specification and inference. As such, it is sometimes assumed that the so-called simplifying assumption holds, which assumes that $C_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$ is independent of $\mathbf{z}$. Note, however, that this does not mean that the conditional copula is equal to the unconditional copula, but merely that the conditional dependence structure is independent of the conditioning point. For a detailed discussion on the proper use of the simplifying assumption, as well as several methods of testing it, we refer to [8].

2.2 Rank correlation

To assess the dependence between two random variables, various types of dependence measures can be considered [20, 46]. Let us start by considering the Pearson correlation coefficient, which is essentially a normalised version of the covariance with values in $[-1, 1]$. Thus, Pearson's correlation coefficient only reflects a linear type of correlation, and ignores other types of correlation. Its population and sample versions are given in the following definition.

Definition 6. *Let X_1 and X_2 be real-valued random variables. The population Pearson correlation coefficient of X_1 and X_2 is defined by*

$$\varrho_{X_1,X_2} := \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}},$$

where σ_1, σ_2 denote the standard deviations of X_1, X_2. Further, let $\{(X_{1,1}, X_{2,1}), \dots, (X_{1,n}, Y_{2,n})\}$ be paired observations, then the sample Pearson correlation coefficient is defined by

$$r_{X_1,X_2} = \frac{\sum_{i=1}^n \left(X_{1,i} - \bar{X}_1\right)\left(X_{2,i} - \bar{X}_2\right)}{\sqrt{\sum_{i=1}^n \left(X_{1,i} - \bar{X}_1\right)^2}\sqrt{\sum_{i=1}^n \left(X_{2,i} - \bar{X}_2\right)^2}},$$

where $\bar{X}_1$ and $\bar{X}_2$ denote the sample means of $\{X_{1,1}, \ldots, X_{1,n}\}$ and $\{X_{2,1}, \ldots, Y_{2,n}\}$.

Note that the sample statistic r will be largely affected by an outlier in either X_1 or X_2 as it will have different impacts on the numerator and denominator. Especially for heavy-tailed data, this will lead to misleading results. For such settings, robust measures of correlation should be used. Accordingly, let us turn to measures of rank correlation that indicate the similarity of the orderings of the data when ranked by each of the quantiles. We recollect the definitions of two such measures, Spearman's rho and Kendall's tau. For continuous random variables X_1, X_2 with marginals F_1, F_2, set $U_i := F_i(X_i), i = 1, 2$ and note that U_1, U_2 have uniform marginals on $[0, 1]$ by Fisher's theorem [19]. Let us define Spearman's rho between variables X_1 and X_2 as the linear Pearson's correlation coefficient ϱ between variables U_1, U_2.

Definition 7. *Let X_1 and X_2 be real-valued continuous random variables. The population Spearman's rho of X_1 and X_2 is defined as*

$$\rho_{X_1,X_2} := \varrho_{U_1,U_2} = \frac{\mathrm{Cov}(U_1, U_2)}{\sigma_{U_1}\sigma_{U_2}} = \frac{\mathbb{E}[U_1U_2 - 1/4]}{1/12} = 12 \int_{[0,1]^2} u_1 u_2 \mathrm{d}C(u_1, u_2) - 3,$$

where σ_1, σ_2 denote the standard deviations of U_1, U_2 and C is the copula corresponding to (X_1, X_2). Further, let $\{(X_{1,1}, X_{2,1}), \ldots, (X_{1,n}, Y_{2,n})\}$ be paired observations with distinct integer ranks, then we define the sample Spearman's rho by

$$\widehat{\rho}_{X_1,X_2} = 1 - \frac{6\sum_{i=1}^{n}(R(X_{1,i}) - R(X_{2,i}))^2}{n\,(n^2 - 1)},$$

where $R(X_{1,i})$ and $R(X_{2,i})$ denote the ranks of observations $X_{1,i}$ and $X_{2,i}$, for $i = 1, \ldots, n$.

Note that Spearman's rho is a distribution-free measure of dependence, i.e. independent of marginals and thus invariant under monotonic transformations. As a consequence, it can be expressed in terms only of the copula. Further note that the corresponding sample statistic is robust to outliers as it only depends on ranked data. The same also applies to other rank correlation coefficients, e.g. Kendall's rank correlation coefficient. Let us define Kendall's tau as the difference between the probability of concordance and the probability of discordance of two independent versions of (X_1, X_2). A concordant pair is a pair of bivariate observations such that both elements of one observation are either greater than or less than the corresponding elements of the other observation. We speak of a discordant pair when only one of the elements of one observation is greater than the corresponding element of the other observation.

Definition 8. *Let X_1 and X_2 be real-valued random variables. The population Kendall's tau of X_1 and X_2 is defined as*

$$\tau_{X_1,X_2} := \mathbb{P}\left((X_{1,1} - X_{1,2})\,(X_{2,1} - X_{2,2}) > 0\right) - \mathbb{P}\left((X_{1,1} - X_{1,2})\,(X_{2,1} - X_{2,2}) < 0\right) \tag{1}$$

where $(X_{1,i}, X_{2,i})_{i=1,2}$ *are two independent versions of* (X_1, X_2)*. Further, we define the sample Kendall's tau for paired observations* $\{(X_{1,1}, X_{2,1}), \ldots, (X_{1,n}, Y_{2,n})\}$ *by*

$$\widehat{\tau}_{X_1,X_2} := \frac{2}{n(n-1)} \sum_{i_1<i_2} \operatorname{sign}\left((X_{1,i_1} - X_{1,i_2})(X_{2,i_1} - X_{2,i_2})\right).$$

It follows from the definition that, like Spearman's rho, Kendall's tau takes values in $[-1, 1]$, where a value of -1 corresponds to the variables being negatively dependent and a value of 1 corresponds to the variables being positively dependent, independent random variables a value of 0 is taken. Note that the Kendall's tau of independent variables is equal to 0, but conversely that a Kendall's tau of 0 does not necessarily mean that the variables are independent.

Furthermore, whenever the random variables are continuous, then the following expressions are equivalent to that of Definition 8

$$\begin{aligned} \tau_{X_1,X_2} &= 4\mathbb{P}(X_{1,1} < X_{1,2}, X_{2,1} < X_{2,2}) - 1 \qquad (2)\\ &= 1 - 4\mathbb{P}(X_{1,1} < X_{1,2}, X_{2,1} > X_{2,2}) \qquad (3)\\ &= 4\int C(u_1, u_2)\mathrm{d}C(u_1, u_2) - 1. \end{aligned}$$

As both Spearman's rho and Kendall's tau can be represented in terms of the copula, it follows that for most one-dimensional families of copulas, estimation of the parameter is equivalent to either estimating Kendall's tau or Spearman's rho. Kendall's taus thus provide much information on the underlying dependence structure, while being much easier to manipulate than the copulas themselves. Indeed, it should be noted that the Kendall's tau matrix (of all pairwise Kendall's tau) of a p-dimensional multivariate vector $\mathbf{X}$ lives in a space of dimension $\frac{1}{2}p(p-1)$, whereas the underlying copula is a multivariate function defined on a p-dimensional space. For notational convenience, we will write subscripts $\tau_{1,2}$ for τ_{X_1,X_2} when the appropriate variables are obvious.

Let us turn to the conditional setup when a multivariate covariate $\mathbf{Z}$ is available. Conditional Kendall's tau has been studied before, see [7, 9, 10, 23, 50]. Conditional Kendall's tau is a conditional dependence measure used to predict whether a pair of random variables is concordant or discordant conditionally on $\mathbf{Z}$. We define the conditional Kendall's tau in a similar manner as its unconditional counterpart.

Definition 9. *Let* X_1 *and* X_2 *be real-valued random variables and* $\mathbf{Z}$ *be a random vector taking values in* $\mathcal{Z} \subset \mathbb{R}^d$*. For any point* $\mathbf{z} \in \mathcal{Z}$*, we define Kendall's tau of* X_1 *and* X_2 *conditional on* $\mathbf{Z} = \mathbf{z}$ *by*

$$\begin{aligned} \tau_{1,2|\mathbf{Z}=\mathbf{z}} = {} & \mathbb{P}\left((X_{1,1} - X_{1,2})(X_{2,1} - X_{2,2}) > 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\right)\\ & - \mathbb{P}\left((X_{1,1} - X_{1,2})(X_{2,1} - X_{2,2}) < 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\right), \qquad (4) \end{aligned}$$

where $(X_{1,i}, X_{2,i}, \mathbf{Z}_i)_{i=1,2}$ *are two independent versions of* $(X_1, X_2, \mathbf{Z})$.

For every point in $\mathbf{z} \in \mathcal{Z}$ the conditional Kendall's tau takes values $[-1, 1]$, while the underlying conditional copula is a bivariate function for each $\mathbf{z} \in \mathcal{Z}$. As such, conditional Kendall's tau, and more generally conditional dependence measures, allow us to summarise the dependency between two variables under a changing covariate.

Furthermore, when the conditional marginal distributions of X_1 and X_2 given $\mathbf{Z} = \mathbf{z}$ are continuous, Definition 9 is equivalent to any of the following expressions

$$\begin{aligned}
\tau_{1,2|\mathbf{Z}=\mathbf{z}} &= 4\mathbb{P}\left(X_{1,1} < X_{1,2}, X_{2,1} < X_{2,2} | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\right) - 1 && (5)\\
&= 1 - 4\mathbb{P}\left(X_{1,1} < X_{1,2}, X_{2,1} > X_{2,2} | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\right), && (6)\\
&= 4\int_{[0,1]^2} C_{1,2|\mathbf{Z}=\mathbf{z}}\left(u_1, u_2\right) \mathrm{d}C_{1,2|\mathbf{Z}=\mathbf{z}}\left(u_1, u_2\right) - 1.
\end{aligned}$$

Since it can be expressed in terms of the underlying conditional copula, it follows that the conditional Kendall's tau is also invariant under increasing transformations. Clearly, the conditional Kendall's tau is equal to its unconditional counterpart if the variables X_1, X_2 are independent of $\mathbf{Z}$. However, the definition of appropriate estimators is less straightforward as we need to deal with the dependency on $\mathbf{Z}$. To this end, we discuss kernel regression estimation in Section 2.5.

2.3 Elliptical Distributions

The class of elliptical distributions is a generalization of the family of multivariate normal distributions. It is more flexible, e.g. providing for heavy tails, while still keeping some of its useful Gaussian properties. It is therefore often used in the modelling of different types of heavy tailed data. See [36] for a discussion on the use of elliptical distributions in finance.

Definition 10. *Let* $\boldsymbol{X}$ *be a* p*-dimensional random vector. Further, let* $\boldsymbol{\mu}$ *denote a vector in* $\mathbb{R}^p$, $\boldsymbol{\Sigma}$ *a* $p \times p$ *nonnegative definite symmetric matrix and* $g : \mathbb{R}^+ \to \mathbb{R}^+ \cup \{\infty\}$ *a measurable function. We say that* $\boldsymbol{X}$ *has an elliptical distribution with parameters* $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ *and* g*, if its density function w.r.t. the Lebesgue measure in* $\mathbb{R}^p$ *is given by*

$$f_{\mathbf{X}}(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2} g\left((\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right),$$

for all $\mathbf{x} \in \mathbb{R}^p$*, and we write* $X \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$.

Here, the function g is called the density generator of $\mathbf{X}$. Note that not every function g can be used to construct an elliptical distribution. Examples of elliptical distributions include the Gaussian, Student's t and Laplace distributions. With Definition 10 it may still be difficult to visualise the general concept of elliptic distributions. An alternative form of approaching elliptic distributions is given in the following theorem.

Theorem 11. *We have that* $\boldsymbol{X} \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ *with* $\operatorname{rank}(\boldsymbol{\Sigma}) = p$ *if and only if there exists a continuous random variable* $\xi \geq 0$ *independent of a random variable* $\mathbf{U}$*, where* $\mathbf{U}$ *is uniformly distributed on the unit sphere in* $\mathbb{R}^p$*, and a* $p \times p$ *matrix* $\boldsymbol{A}$ *with* $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}$*, such that*

$$\mathbf{X} \sim \boldsymbol{\mu} + \xi \mathbf{A}\mathbf{U}.$$

Note that the representation given in Theorem 11 is not unique. By substituting $\mathbf{A}$ and $\mathbf{U}$ by $\mathbf{AO}$ and $\mathbf{O}^T\mathbf{U}$ for some orthogonal $p \times p$ matrix $\mathbf{O}$, we end up with the same distribution. The existence of the inverse of $\boldsymbol{\Sigma}$ is ensured by requiring $\operatorname{rank}(\boldsymbol{\Sigma}) = p$, and if we set $\mathbb{E}\left[\xi^2\right] = p$, then $\boldsymbol{\Sigma}$ is equal to the covariance matrix of $\mathbf{X}$.

Next, let us state a well-known result for the relation between Kendall's tau and the linear Pearson's correlation coefficient in the following theorem. For the proof and for more details on the claim, we refer to [32].

Theorem 12. *Let* $\boldsymbol{X} \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$*. Assume that for all* $j = 1, \ldots, p$*,* $\mathbb{P}(X_j = \mu_i) = 0$*. Then,*

$$\tau_{j_1, j_2} = \frac{2}{\pi} \arcsin \varrho_{j_1, j_2},$$

for every $j_1, j_2 = 1, \ldots, p$*.*

Consequently, estimating the correlation matrix of an elliptically distributed random variable is equivalent to obtaining its Kendall's tau matrix. As such, Kendall's tau can be used for the robust inference of the covariance matrix. This makes the use of Kendall's tau particularly interesting as inference of the covariance matrix often plays a central role when analysing multivariate data.

2.4 U-statistics

The generic class of U-statistics is effective in constructing optimal estimators and was first studied by Hoeffding [27]. Let us start with some definitions. Let $\mathcal{P}$ be some general (nonparametric) family of probability distributions. Let $\theta(F)$ denote a real-valued function defined for $F \in \mathcal{F}$.

Definition 13. *We call* $\theta(F)$ *a regular parameter within* $\mathcal{F}$*, if for some integer* m *there exists an unbiased estimator of* $\theta(F)$ *based on* m *i.i.d. random variables with distribution* F*, i.e. there exists a real-valued measurable function* $g(x_1, \ldots, x_m)$ *such that*

$$\mathbb{E}\left[g(X_1, \ldots, X_m)\right] = \theta(F) \text{ for all } F \in \mathcal{F},$$

where $X_1, \ldots, X_m$ *are all independent and distributed according to* F*. The smallest* m *for which this applies is referred to as the degree of* $\theta(F)$*.*

There is no loss of generality in assuming that g is a symmetric, because it can always be replaced by the symmetric function

$$\frac{1}{m}\sum_{\mathbf{P}_m} g(x_{j_1},\dots,x_{j_m}),$$

where the summation is over the set $\mathbf{P}_m$ of all $m!$ possible permutations $(j_1,\dots,j_m)$ of $(1,\dots,m)$.

Definition 14. *Let $g(x_1,\dots,x_m)$ be a real-valued symmetric function and let $X_1,\dots,X_n$ be an i.i.d. sample of size $n \geq m$ distributed according to a distribution F. Then, a U-statistic with kernel g is defined as*

$$U_n = \binom{n}{m}^{-1} \sum_{\mathbf{C}_{m,n}} g(X_{i_1},\dots,X_{i_m}),$$

where $\sum_{\mathbf{C}_{m,n}}$ denotes the summation over all $\binom{n}{m}$ possible permutations $(i_1,\dots,i_m)$ of size m out of $(1,\dots,m)$.

U-statistics are unbiased estimators for $\theta(F)$ and are in fact optimal in most cases, i.e. uniformly minimum variance unbiased (UMVU). Let $\mathcal{F}$ be the family of probability distributions such that $\theta(F)$ is finite for all $F \in \mathcal{F}$. Since U_n is a symmetric function we can write in terms of the ordered statistic of $X_1,\dots,X_n$, which is known to be a complete and sufficient statistic for $F \in \mathcal{F}$. It follows then by the Lehmann-Scheffé theorem that U_n is in fact UMVU.

Let U_n be a U-statistic with kernel $g(x_1,\dots,x_m)$ of a sample size n. For the variance of a U_n there exists an explicit expression. To introduce this result, we consider a symmetric kernel $g(x_1,\dots,x_m)$ of order m satisfying $\mathbb{E}\left[g^2(X_1,\dots,X_m)\right] < \infty$ and let us denote $\theta = \theta(F) = \mathbb{E}\left[g(X_1,\dots,X_m)\right]$. For $c = 1,\dots,m-1$, we define

$$g_c(x_1,\dots,x_c) = \mathbb{E}\left[g(x_1,\dots,x_c,X_{c+1},\dots,X_m)\right],$$

and set $g_m := g$. Note that $\mathbb{E}\left[g_c(X_1,\dots,X_c)\right] = \theta$ for $c = 1,\dots,m$. Additionally, we set $\zeta_0 = 0$ and define for $1 \leq c \leq m$,

$$\zeta_c = \mathbb{V}\mathrm{ar}\left[g_c(X_1,\dots,X_c)\right] = \mathbb{E}\left[(g_c(X_1,\dots,X_m) - \theta(F))^2\right].$$

The following theorem gives us then the U-statistic's finite sample variance. The proof can be found in the work of Hoeffding [27].

Theorem 15. *If $\mathbb{E}\left[g^2(X_1,\dots,X_m)\right] < \infty$, then the variance of U_n is given by*

$$\mathbb{V}\mathrm{ar}\left[U_n\right] = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c}\zeta_c.$$

Finally, let us state the asymptotic normality of U-statistics in the following theorem. For the proof, we again refer to [27].

Theorem 16. *If* $\mathbb{E}\left[g^2(X_1,\ldots,X_m)\right] < \infty$ *and* $\zeta_1 > 0$, *then as* $n \to \infty$

$$n^{1/2}\left(U_n - \theta(F)\right) \xrightarrow{D} \mathcal{N}\left(0, m^2\zeta_1\right).$$

2.5 Kernel Regression Estimation

In this section, we introduce the notion of kernel regression for the construction of nonparametric conditional estimators. The presented derivations follow [21]. Let us be interested in the estimation of the expectation of a random variable Y conditional on some covariate $\mathbf{Z} \in \mathcal{Z}$. Then, in the case of a continuously distributed covariate, we will almost surely never observe that $\mathbf{Z} = \mathbf{z}$, for any $\mathbf{z} \in \mathcal{Z}$. Therefore a method is needed with which it is still possible to compute reasonable estimates. Most natural is the concept of kernel regression, also known as kernel smoothing. That is, to consider adjacent observations in which the variate $\mathbf{Z}$ is close to the point $\mathbf{z}$ at which we want the estimate.

Let us start by considering the definition of conditional expectation

$$m(\mathbf{z}) = \mathbb{E}\left[Y|\mathbf{Z}=\mathbf{z}\right] = \int y f_{Y|\mathbf{Z}=\mathbf{z}}(y)\mathrm{d}y = \frac{\int y f_{\mathbf{Z},Y}(\mathbf{z},y)\mathrm{d}y}{f_{\mathbf{Z}}(\mathbf{z})}. \tag{7}$$

In kernel regression the estimates of $f_{\mathbf{Z},Y}(\mathbf{z},y)$ and $f_{\mathbf{Z}}(\mathbf{z})$ are computed by kernel density estimation. That is, the density functions are approximated by adjacent observations of Z in the following way

$$\widehat{f}_{\mathbf{Z},Y}(\mathbf{z},y;h) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right)\frac{1}{h}K_2\left(\frac{y-Y_i}{h}\right) \tag{8}$$

$$\widehat{f}_{\mathbf{Z}}(\mathbf{z};h) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right), \tag{9}$$

where $K_1 : \mathbb{R}^d \to \mathbb{R}$ and $K_2 : \mathbb{R} \to \mathbb{R}$ are kernel functions, i.e. assumed to be symmetric, unimodal at zero and satisfying $\int K = 1$, the parameter h is a tuning parameter known as the bandwidth and n is the sample size. The bandwidth controls the sensitivity of the density estimates towards observations further away from z, whereas the kernel function defines the form of the dependency. Commonly used kernels are the Gaussian and Epanechnikov kernels. Note that for U-statistics the term kernel has a different meaning.

The smoothed estimate of $m(x)$ is obtained by replacing $f_{\mathbf{Z},Y}(\mathbf{z},y)$ and $f_{\mathbf{Z}}(\mathbf{z})$ in (7) with their

respective kernel density estimates (8) and (9). This gives,

$$\begin{aligned}\widehat{m}(\mathbf{z};h) :&= \frac{\int y\widehat{f}_{\mathbf{Z},Y}(\mathbf{z},y;h)\mathrm{d}y}{\widehat{f}_{\mathbf{Z}}(\mathbf{z};h)} \\ &= \frac{\int y\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right)\frac{1}{h}K_2\left(\frac{y-Y_i}{h}\right)\mathrm{d}y}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right)} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right)\int y\frac{1}{h}K_2\left(\frac{y-Y_i}{h}\right)\mathrm{d}y}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right)}.\end{aligned} \tag{10}$$

Then, by a change of variable $u_i = \frac{y-Y_1}{h}$ we find

$$\begin{aligned}\int y\frac{1}{h}K_2\left(\frac{y-Y_i}{h}\right)\mathrm{d}y &= \int (hu+Y_i)K(u)\mathrm{d}u \\ &= h\int uK_2(u)\mathrm{d}u + Y_i\int K_2(u)\mathrm{d}u \\ &= Y_i,\end{aligned} \tag{11}$$

where we have used that K_2 is symmetric and that $\int K_2 = 1$. By combining (11) and (10) we obtain

$$\begin{aligned}\widehat{m}(\mathbf{z};h) :&= \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right)Y_i}{\frac{1}{n}\sum_{i=1}^{n}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right)} \\ &= \sum_{i=1}^{n}\frac{\frac{1}{h^d}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right)}{\sum_{i=1}^{n}\frac{1}{h^d}K_1\left(\frac{\mathbf{z}-\mathbf{Z}_i}{h}\right)}Y_i.\end{aligned}$$

The resulting estimator can be seen as a weighted average of $Y_1,\ldots,Y_n$ by means of the so-called Nadaraya-Watson weights

$$\widehat{m}(\mathbf{z};h) := \sum_{i=1}^{n} w_{i,n}(\mathbf{z})Y_i,$$

where

$$w_{i,n}(\mathbf{z}) := \frac{K_h(\mathbf{Z}_i-\mathbf{z})}{\sum_{k=1}^{n}K_h(\mathbf{Z}_k-\mathbf{z})}$$

with $K_h := h^{-d}K(\cdot/h)$. Again, note that the bandwidth h controls the estimate's sensitivity towards observations $\mathbf{Z}_i$ that are further away from point $\mathbf{z}$. It will have a strong influence on the characteristics of the estimator and is closely related to the sample size. Reducing the bandwidth will decrease the estimator's bias and increase its variance, which is known as the bias-variance tradeoff. Hereby, larger sample sizes will allow for a smaller choice of the bandwidth. Further note that the volume of the space $\mathcal{Z}$ grows exponentially fast when increasing the dimensionality of $\mathbf{Z}$. As such, the density of observations within that space decreases at the same rate, calling for

an exponentially increasing bandwidth. In practice, this so-called curse of dimensionality means that we can only consider covariates of a few dimensions at most.

We conclude with introducing Bochner's lemma. This will be useful for further proofs involving kernel based estimators. The lemma and proof are multivariate adaptions of those found in [49].

Lemma 17 (Bochner's Lemma)**.** *Let f be a bounded function on $\mathbb{R}^d$, continuous in a neighborhood of the point $\mathbf{z} \in \mathbb{R}^d$ and let K be a function on $\mathbb{R}^d$ such that*

$$\int |K(\mathbf{u})| \mathrm{d}\mathbf{u} < \infty.$$

Then,

$$\lim_{h \to 0} \int K(\mathbf{u}) f(\mathbf{z} + h\mathbf{u}) \mathrm{d}\mathbf{u} = f(\mathbf{z}) \int K(\mathbf{u}) \mathrm{d}\mathbf{u}.$$

Proof. For every $h > 0$, we have

$$\begin{aligned}
\left| \int K(\mathbf{u}) f(\mathbf{z} + h\mathbf{u}) \mathrm{d}\mathbf{u} - f(\mathbf{z}) \int K(\mathbf{u}) \mathrm{d}\mathbf{u} \right| &= \left| \int (f(\mathbf{z} + h\mathbf{u}) - f(\mathbf{z})) K(\mathbf{u}) \, \mathrm{d}\mathbf{u} \right| \\
&\leq \int |f(\mathbf{z} + h\mathbf{u}) - f(\mathbf{z})| \, |K|(\mathbf{u}) \, \mathrm{d}\mathbf{u} \\
&\leq \sup_{|\mathbf{u}| \leq h^{-1/2}} |f(\mathbf{z} + h\mathbf{u}) - f(\mathbf{z})| \int |K|(\mathbf{u}) \mathrm{d}\mathbf{u} \\
&+ \int_{|\mathbf{u}| \geq h^{-1/2}} |f(\mathbf{z} + h\mathbf{u}) - f(\mathbf{z})| \, |K|(\mathbf{u}) \mathrm{d}\mathbf{u} \\
&\leq \sup_{|\mathbf{u}| \leq h^{-1/2}} |f(\mathbf{z} + h\mathbf{u}) - f(\mathbf{z})| \int |K|(\mathbf{u}) \mathrm{d}\mathbf{u} \\
&+ 2 \sup_{\mathbf{u}} |f|((\mathbf{u}) \int_{|\mathbf{u}| \geq h^{-1/2}} |K|(\mathbf{u}) \mathrm{d}\mathbf{u} \\
&\leq \sup_{|\mathbf{v}| \leq h^{1/2}} |f(\mathbf{z} + \mathbf{v}) - f(\mathbf{z})| \int |K|(\mathbf{u}) \mathrm{d}\mathbf{u} \\
&+ 2 \sup_{\mathbf{u}} |f|((\mathbf{u}) \int_{|\mathbf{u}| \geq h^{-1/2}} |K|(\mathbf{u}) \mathrm{d}\mathbf{u}.
\end{aligned}$$

and the result follows by letting h tend to 0. □

3 Improved Estimation of Kendall's Tau

This chapter is devoted to the improved estimation of the (unconditional) Kendall's tau matrix under our structural assumption. First, we formalise the proposed model in Section 3.1. Then in Section 3.2, we construct new estimators that take advantage of the underlying structural pattern of the population Kendall's tau matrix by averaging over pairwise Kendall's taus. In Section 3.3 we show how the variances of our new estimators are reduced compared to the usual sample Kendall's tau matrix.

3.1 The Structural Assumption

For the remainder of this work, we let $X_{j,i}$ denote the i^{th} observation $(i = 1, \dots, n)$ of the j^{th} variable $(j = 1, \dots, p)$. We assume that all $\mathbf{X}_i = (X_{1,i}, \dots, X_{p,i})$ are i.i.d. and we denote the marginal CDFs by $(F_1, \dots, F_p)$. As such, the $p \times p$ Kendall's tau matrix $\mathbf{T}$ of all pairwise Kendall's tau of $\mathbf{X}_i$ is independent of i. The task of estimating $\mathbf{T}$ becomes particularly challenging in the high-dimensional setting when p is large. The number of free parameters increases quadratically with p and therefore simple use of the sample Kendall's tau matrix can lead to noisy estimates, due to the accumulation of the sheer inaccuracies of all sample Kendall's taus [13]. As discussed in the Introduction, several regularisation techniques have been considered for reducing the number of free parameters to estimate. However, our interest lies in a particular model that has not been studied before: we assume that the set of variables can be grouped in such a way that the Kendall's taus between elements of different groups is only depending on group numbers. Let us formalise this in the following assumption. Here, $\mathbf{1}$ denotes the matrix with all components equal to 1.

Assumption A1 (Structural Assumption)**.** *There exists a $K > 0$ and a partition $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_K\}$ of $\{1, \dots, p\}$, such that for all distinct $k_1, k_2 \in \{1, \dots, K\}$,*

$$[\mathbf{T}]_{\mathcal{B}_{k_1,k_2}} = \tau_{k_1,k_2}\mathbf{1},$$

for some constants $\tau_{k_1,k_2} \in [-1, 1]$. Here, $\mathcal{B}_{k_1,k_2} = \{(j_1, j_2) \in (\mathcal{G}_{k_1} \times \mathcal{G}_{k_2})\}$, for all $k_1, k_2 \in \{1, \dots, K\}$.

It should be noted that $\tau_{k_1,k_2} = \tau_{k_2,k_1}$ for all distinct k_1, k_2, making the Kendall's tau matrix symmetric. Further, note that after reordering of the variables by group, the corresponding Kendall's tau matrix is block-structured having constant values in each of the off-diagonal blocks.

The interest in investigating this structural assumption originates from applications in stock return modelling. In this context, the clustering of the variables could for instance be considered as grouping companies by sector or economy. It then seems at least intuitive to assume that

companies from different groups have correlations that depend only on the groups they are in, without making any assumptions on the correlations between companies from the same group.

Obviously, the structural assumption is satisfied for any set of variables by using only groups of length 1. Therefore, assuming larger groups will make the assumption more constraining. Indeed, in this framework, Kendall's tau matrix depends on

$$\frac{1}{2}K(K-1)+\frac{1}{2}\sum_{k=1}^{K}|\mathcal{G}_k|(|\mathcal{G}_k|-1)$$

free parameters. For a dimension of 100, assuming we can split into $K = 10$ groups of equal size, this translates to a reduction by factor of 10 of the number of free parameters to estimate (from 4950 to 495). Such a reduction suggests that the use of appropriate estimators can lead to significant estimation improvements.

As mentioned in the Introduction, Perreault et al. (2019) [40] proposed a similar model with a more restrictive version of Assumption A1 by assuming that the variables could be partitioned into K clusters with exchangeable dependence. Their assumption is referred to as the Partial Exchangeability Assumption (PEA).

Assumption A2 (Partial Exchangeability Assumption)**.** *For $j \in \{1,\dots,p\}$, let $U_j = F_j(X_j)$. For any partition $\mathcal{G} = \{\mathcal{G}_1,\dots,G_K\}$ of $\{1,\dots,p\}$, let $\pi(G)$ be the set of permutations π of $\{1,\dots,p\}$ such that for all $j \in \{1,\dots,p\}$ and all $k \in \{1,\dots,K\}$, $\pi(j) \in \mathcal{G}_k$ if and only if $j \in \mathcal{G}_k$. A partition $\mathcal{G} = \{\mathcal{G}_1,\dots,G_K\}$ satisfies the PEA if for any $u_1,\dots,u_p \in [0,1]$ and any permutation $\pi \in \pi(G)$, one has*

$$C(u_1,\dots,u_p) = C\left(u_{\pi(1)},\dots,u_{\pi(p)}\right)$$

or equivalently, $(U_1,\dots,U_p) \overset{law}{=} \left(U_{\pi(1)},\dots,U_{\pi(p)}\right)$.

Note that the PEA imposes restrictions on the underlying copula, whereas Assumption A1 only does so on the underlying Kendall's tau matrix, making it a lot less restrictive. Additionally, under Assumption A2, the Kendall's tau matrix is fully block-structured including constant diagonal blocks as well, after reordering of the variables. This further restricts the number of free parameters to estimate. In contrast to Perreault's work [40], we are more interested in a model where we do not consider partial exchangeability nor constant interdependence of marginal variables within the same cluster. Particularly in view of the aforementioned application of stock returns, the PEA is seems too restrictive and a model without partial exchangeability in which companies from the same cluster have different mutual dependence is more plausible. For these reasons, we opt for a more flexible variant of Perreault's simpler model.

3.2 Construction of Estimators

In this section, we define estimators of the Kendall's tau matrix $\mathbf{T}$ under Assumption A1 for some known partitions $\mathcal{G}$ of $\{1,\ldots,p\}$. Note that such partition can also be inferred from the data, see [40, 41]. First, let us define the group membership function $\kappa : \{1,\ldots,p\} \to \{1,\ldots,K\}$, $\kappa(j) = k$, when $j \in \mathcal{G}_k$. Without loss of generality we assume that the ordering of variables is such that $j_1 < j_2 \Rightarrow \kappa(j_1) \leq \kappa(j_2)$, i.e. the variables are ordered by group and thus $\mathbf{T}$ has the proper block-structure. Naturally, we use the sample version of Kendall's tau as a first step statistic.

Following Definition 8, the sample Kendall's tau between X_1 and X_2 is given by

$$\widehat{\tau}_{1,2} := \frac{2}{n(n-1)} \sum_{i_1<i_2} \operatorname{sign}\left((X_{1,i_1} - X_{1,i_2})\,(X_{2,i_1} - X_{2,i_2})\right), \tag{12}$$

but following any of the equivalent expressions for Kendall's tau in (2) and (3), we might as well choose to define the sample statistic as

$$\frac{2}{n(n-1)} \sum_{i_1<i_2} 4\mathbb{1}\left\{X_{1,i_1} < X_{1,i_2}, X_{2,i_1} < X_{2,i_2}\right\} - 1, \tag{13}$$

or

$$\frac{2}{n(n-1)} \sum_{i_1<i_2} 1 - 4\mathbb{1}\left\{X_{1,i_1} < X_{1,i_2}, X_{2,i_1} > X_{2,i_2}\right\}, \tag{14}$$

where $\mathbb{1}$ denotes the indicator function. For continuous random variables, the properties of (12), (13) and (14) are all identical and we could choose any one of them. For convenience, we will use in the proofs the symmetrical version of expression (13), given by

$$\frac{2}{n(n-1)} \sum_{i_1<i_2} 2\left(\mathbb{1}\left\{X_{1,i_1} < X_{1,i_2}, X_{2,i_1} < X_{2,i_2}\right\} + \mathbb{1}\left\{X_{1,i_2} < X_{1,i_1}, X_{2,i_2} < X_{2,i_1}\right\}\right) - 1.$$

We denote the corresponding Kendall's tau matrix estimator by $\widehat{\mathbf{T}} = [\widehat{\tau}_{j_1,j_2}]_{p\times p}$, which serves as a first step estimator for obtaining a better estimator of the Kendall's tau matrix. The sample Kendall's tau matrix does not make any use of the underlying structure and will therefore be a rather naive tool in practice. Since we assume that the Kendall's taus in each of the off-diagonal blocks are equal, the idea of averaging the pairwise sample Kendall's tau follows naturally. Let us introduce the block estimator $\widehat{\mathbf{T}}^B$ that averages all sample Kendall's tau within each of the off-diagonal blocks. Formally, we have

$$\widehat{\mathbf{T}}^B := \left[\widehat{T}^B_{j_1,j_2}\right]_{p\times p} = \begin{cases} \widehat{\tau}_{j_1,j_2}, & \text{if } \kappa(j_1) = \kappa(j_2) \\ \widehat{\tau}^B_{k_1,k_2}, & \text{for distinct } \kappa(j_1) = k_1 \text{ and } \kappa(j_2) = k_2, \end{cases}$$

where

$$\widehat{\tau}^{B}_{k_1,k_2} := \frac{1}{|\mathcal{B}_{k_1,k_2}|} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \widehat{\tau}_{j_1,j_2}.$$

Under the PEA, [40] proposed a similar estimator that computes the average within each of the blocks, both off-diagonal and diagonal. It was shown that this estimator was asymptotically normal and optimal under the Mahalanobis distance. However, in terms of computational efficiency, the block estimator $\widehat{\mathbf{T}}^{B}$ does not show any improvement over the usual estimator $\widehat{\mathbf{T}}$, as both estimators require the computation of the usual Kendall's tau between all pairs of variables anyway.

To reduce the computation time, we propose not averaging over all Kendall's taus in the block but only over some of them. Naturally, the question arises over which elements then to average over. For this purpose, we introduce several estimators that average over different subsets of elements within each of the off-diagonal blocks.

We introduce two estimators that each average $N_{k_1,k_2} \in \{1,\dots,|\mathcal{G}_{k_1}| \vee |\mathcal{G}_{k_2}|\}$ pairs per off-diagonal block $\mathcal{B}_{k_1,k_2}$, so that we can compare estimators that either average pairs in the same row/column, or pairs on the diagonal. For averaging pairs on the diagonal, it is moreover required that $N_{k_1,k_2} \leq |\mathcal{G}_{k_1}| \wedge |\mathcal{G}_{k_2}|$. The number of Kendall's tau estimates is then reduced to scaling linearly with group size, which is a significant improvement over the previous quadratic growth. For distinct groups $k_1, k_2 \in \{1,\dots,K\}$, we set

$$\widehat{\tau}^{R}_{k_1,k_2} := \frac{1}{N_{k_1,k_2}} \sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_R} \widehat{\tau}_{j_1,j_2},$$
$$\widehat{\tau}^{D}_{k_1,k_2} := \frac{1}{N_{k_1,k_2}} \sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_D} \widehat{\tau}_{j_1,j_2},$$

where $(\mathcal{B}_{k_1,k_2})_R$ is the set of the N_{k_1,k_2} first pairs in the first line along the largest side of block $\mathcal{B}_{k_1,k_2}$ and $(\mathcal{B}_{k_1,k_2})_D$ is the set of the N_{k_1,k_2} first pairs on the first diagonal of the block. For example, if $N_{1,2} = 3$, $\mathcal{G}_1 = \{1,2,3,4\}$ and $\mathcal{G}_2 = \{5,6,7\}$, then $(\mathcal{B}_{1,2})_R = \{(1,5),(2,5),(3,5)\}$ and $(\mathcal{B}_{1,2})_D = \{(1,5),(2,6),(3,7)\}$. Then, analogously to the definition of $\mathbf{T}^{B}$, we define the row estimator $\widehat{\mathbf{T}}^{R}$ and the diagonal estimator $\widehat{\mathbf{T}}^{D}$ as

$$\widehat{\mathbf{T}}^{R} := \left[\widehat{T}^{R}_{j_1,j_2}\right]_{p\times p} = \begin{cases} \widehat{\tau}_{j_1,j_2}, & \text{if } \kappa(j_1) = \kappa(j_2) \\ \widehat{\tau}^{R}_{k_1,k_2}, & \text{for distinct } \kappa(j_1) = k_1 \text{ and } \kappa(j_2) = k_2, \end{cases}$$
$$\widehat{\mathbf{T}}^{D} := \left[\widehat{T}^{D}_{j_1,j_2}\right]_{p\times p} = \begin{cases} \widehat{\tau}_{j_1,j_2}, & \text{if } \kappa(j_1) = \kappa(j_2) \\ \widehat{\tau}^{D}_{k_1,k_2}, & \text{for distinct } \kappa(j_1) = k_1 \text{ and } \kappa(j_2) = k_2. \end{cases}$$

As such, for each of the off-diagonal blocks $\widehat{\mathbf{T}}^{R}$ averages only the pairs on the first line along the largest side, whereas $\widehat{\mathbf{T}}^{D}$ averages only the pairs along the first diagonal.

Lastly, we introduce the estimator that randomly selects pairs to average over per block. We denote the (deterministic) number of averaged pairs per block $\mathcal{B}_{k_1,k_2}$ by $N_{k_1,k_2} \in \{1,\dots,|\mathcal{G}_{k_1}| \times |\mathcal{G}_{k_2}|\}$ and the corresponding estimator by $\mathbf{T}^U$. The pairs are selected with uniform probability and without replacement. For distinct $k_1, k_2 \in \{1,\dots,K\}$ we define

$$\widehat{\tau}^U_{k_1,k_2} = \frac{1}{N_{k_1,k_2}} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} W_{j_1,j_2}\widehat{\tau}_{j_1,j_2},$$

where $\mathbf{W}$ is a $p \times p$ matrix of random weights that selects N_{k_1,k_2} pairs per off-diagonal block $\mathcal{B}_{k_1,k_2}$ with uniform probability and without replacement. $W_{j_1,j_2} = 1$ corresponds to selecting pair (X_{j_1}, X_{j_2}) and $W_{j_1,j_2} = 0$ corresponds to passing over it. The corresponding matrix estimator is then given by

$$\widehat{\mathbf{T}}^U = \left[\widehat{T}^U_{j_1,j_2}\right] = \begin{cases} \widehat{\tau}_{j_1,j_2}, & \text{if } \kappa(j_1) = \kappa(j_2) \\ \widehat{\tau}^U_{k_1,k_2}, & \text{for distinct } \kappa(j_1) = k_1 \text{ and } \kappa(j_2) = k_2. \end{cases}$$

3.3 Comparison of their Variances

Before we proceed with the main theoretical results on the estimators' variances, let us introduce some auxiliary notations. For every distinct $k_1, k_2 \in \{1,\dots,K\}$ we set

$$P_{k_1,k_2} := \mathbb{P}\left((X_{j_1,1} - X_{j_1,2})\,(X_{j_2,1} - X_{j_2,2}) > 0\right),$$

where the pair (j_1, j_2) lies within block $\mathcal{B}_{k_1,k_2}$. The quantity P_{k_1,k_2} is equal to the probability of concordance of the variables X_{j_1} and X_{j_2} and thus $\tau_{k_1,k_2} = 2P_{k_1,k_2} - 1$. As such, the structural assumption ensures that P_{k_1,k_2} is independent of the choice of pair $(j_1, j_2) \in \mathcal{B}_{k_1,k_2}$. Alternatively we can write P_{k_1,k_2} in terms of the copula C_{j_1,j_2} of (X_{j_1}, X_{j_2}) by

$$P_{k_1,k_2} = 2\int_{[0,1]^2} C_{j_1,j_2}(u_1,u_2)\mathrm{d}C_{j_1,j_2}(u_1,u_2).$$

Further, for every pair $(j_1, j_2) \in \{1,\dots,p\}^2$ we define

$$Q_{j_1,j_2} := \mathbb{P}\left((X_{j_1,1} - X_{j_1,2})\,(X_{j_2,1} - X_{j_2,2}) > 0, (X_{j_1,1} - X_{j_1,3})\,(X_{j_2,1} - X_{j_2,3}) > 0\right).$$

To derive the copula expression, we write

$$\begin{aligned} Q_{j_1,j_2} = {} & \mathbb{P}\big((X_{j_1},X_{j_2})_1 < (X_{j_1},X_{j_2})_2, (X_{j_1},X_{j_2})_1 < (X_{j_1},X_{j_2})_3\big) \\ & + \mathbb{P}\big((X_{j_1},X_{j_2})_1 < (X_{j_1},X_{j_2})_2, (X_{j_1},X_{j_2})_1 > (X_{j_1},X_{j_2})_3\big) \end{aligned}$$

$$+\mathbb{P}\big((X_{j_1},X_{j_2})_1>(X_{j_1},X_{j_2})_2,(X_{j_1},X_{j_2})_1>(X_{j_1},X_{j_2})_3\big)$$
$$+\mathbb{P}\big((X_{j_1},X_{j_2})_1>(X_{j_1},X_{j_2})_2,(X_{j_1},X_{j_2})_1<(X_{j_1},X_{j_2})_3\big)$$

where $(X_{j_1},X_{j_2})_i$ denotes the pair $(X_{j_1,i},X_{j_2,i})$. Let us write these probabilities in terms of the copula C_{j_1,j_2} of (X_{j_1},X_{j_2}). This gives

$$\begin{aligned}
Q_{j_1,j_2} &= \int_{[0,1]^2}\int_{(u_1,u_2)}^{(1,1)}\int_{(u_1,u_2)}^{(1,1)} \mathrm{d}C_{j_1,j_2}(u_5,u_6)\mathrm{d}C_{j_1,j_2}(u_3,u_4)\mathrm{d}C_{j_1,j_2}(u_1,u_2)\\
&+\int_{[0,1]^2}\int_{(0,0)}^{(u_1,u_2)}\int_{(u_1,u_2)}^{(1,1)} \mathrm{d}C_{j_1,j_2}(u_5,u_6)\mathrm{d}C_{j_1,j_2}(u_3,u_4)\mathrm{d}C_{j_1,j_2}(u_1,u_2)\\
&+\int_{[0,1]^2}\int_{(u_1,u_2)}^{(1,1)}\int_{(0,0)}^{(u_1,u_2)} \mathrm{d}C_{j_1,j_2}(u_5,u_6)\mathrm{d}C_{j_1,j_2}(u_3,u_4)\mathrm{d}C_{j_1,j_2}(u_1,u_2)\\
&+\int_{[0,1]^2}\int_{(0,0)}^{(u_1,u_2)}\int_{(0,0)}^{(u_1,u_2)} \mathrm{d}C_{j_1,j_2}(u_5,u_6)\mathrm{d}C_{j_1,j_2}(u_3,u_4)\mathrm{d}C_{j_1,j_2}(u_1,u_2)\\
&=\int_{[0,1]^2}\bar{C}^2_{j_1,j_2}(u_1,u_2)\mathrm{d}C_{j_1,j_2}(u_1,u_2)+\int_{[0,1]^2}\bar{C}_{j_1,j_2}(u_1,u_2)C_{j_1,j_2}(u_1,u_2)\mathrm{d}C_{j_1,j_2}(u_1,u_2)\\
&+\int_{[0,1]^2}C_{j_1,j_2}(u_1,u_2)\bar{C}_{j_1,j_2}(u_1,u_2)\mathrm{d}C_{j_1,j_2}(u_1,u_2)+\int_{[0,1]^2}C^2_{j_1,j_2}(u_1,u_2)\mathrm{d}C_{j_1,j_2}(u_1,u_2)\\
&=\int_{[0,1]^2}(C_{j_1,j_2}(u_1,u_2)+\bar{C}_{j_1,j_2}(u_1,u_2))^2\mathrm{d}C_{j_1,j_2}(u_1,u_2),
\end{aligned}$$

where $\bar{C}$ denotes the survival copula of variables (X_{j_1},X_{j_2}). For symmetric copulas, $\bar{C}(u_1,u_2)=C(1-u_1,1-u_2)$. Additionally, for all distinct $k_1,k_2\in\{1,\dots,K\}$ and for every $(j_1,j_2),(j_3,j_4)\in\mathcal{B}_{k_1,k_2}$ such that $|\{j_1,j_2\}\cap\{j_3,j_4\}|=1$ we define

$$\begin{aligned}
R_{j_1,j_2,j_3,j_4} &:= \mathbb{P}\left((X_{j_1,1}-X_{j_1,2})(X_{j_2,1}-X_{j_2,2})>0,(X_{j_3,1}-X_{j_3,2})(X_{j_4,1}-X_{j_4,2})>0\right)\\
S_{j_1,j_2,j_3,j_4} &:= \mathbb{P}\left((X_{j_1,1}-X_{j_1,2})(X_{j_2,1}-X_{j_2,2})>0,(X_{j_3,1}-X_{j_3,3})(X_{j_4,1}-X_{j_4,3})>0\right).
\end{aligned}$$

and for every $(j_1,j_2),(j_3,j_4)\in\mathcal{B}_{k_1,k_2}$ such that $\{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset$, we set

$$\begin{aligned}
T_{j_1,j_2,j_3,j_4} &:= \mathbb{P}\left((X_{j_1,1}-X_{j_1,2})(X_{j_2,1}-X_{j_2,2})>0,(X_{j_3,1}-X_{j_3,2})(X_{j_4,1}-X_{j_4,2})>0\right),\\
U_{j_1,j_2,j_3,j_4} &:= \mathbb{P}\left((X_{j_1,1}-X_{j_1,2})(X_{j_2,1}-X_{j_2,2})>0,(X_{j_3,1}-X_{j_3,3})(X_{j_4,1}-X_{j_4,3})>0\right).
\end{aligned}$$

Note that these quantities R,S,T,U can all be written as 8-dimensional integrals involving the copula of $X_{(j_1,j_2,j_3,j_4)}$. As a consequence, they are all functions of the joint law $\mathbb{P}_{(j_1,j_2,j_3,j_4)}$ of the random vector $\mathbf{X}_{(j_1,j_2,j_3,j_4)}$, as well as Q. Note further that under Assumption A1, the quantities Q_{j_1,j_2}, R_{j_1,j_2,j_3,j_4}, S_{j_1,j_2,j_3,j_4}, T_{j_1,j_2,j_3,j_4} and U_{j_1,j_2,j_3,j_4} are all depending on the choice of pairs within the off-diagonal block $\mathcal{B}_{k_1,k_2}$.

Let $Q^B_{k_1,k_2}$, $Q^R_{k_1,k_2}$, $Q^D_{k_1,k_2}$ denote the average among all Q_{j_1,j_2} of combinations of pairs within

$\mathcal{B}_{k_1,k_2}$, $(\mathcal{B}_{k_1,k_2})_R$ and $(\mathcal{B}_{k_1,k_2})_D$ respectively. Accordingly, we will denote the average among all R_{j_1,j_2,j_3,j_4} and all S_{j_1,j_2,j_3,j_4} of pairs within $\mathcal{B}_{k_1,k_2}$ and $(\mathcal{B}_{k_1,k_2})_R$ by respectively $R^B_{k_1,k_2}$, $R^R_{k_1,k_2}$ and $S^B_{k_1,k_2}$, $S^R_{k_1,k_2}$. Likewise, we set $T^B_{k_1,k_2}$, $T^D_{k_1,k_2}$ and $U^B_{k_1,k_2}$, $U^D_{k_1,k_2}$ for the averages among all T_{j_1,j_2,j_3,j_4} and all U_{j_1,j_2,j_3,j_4} of pairs within $\mathcal{B}_{k_1,k_2}$ and $(\mathcal{B}_{k_1,k_2})_D$.

Now that we have all auxiliary notations in place, let us start by showing that each of the estimators is in fact a U-statistic.

Lemma 18. *For distinct $k_1, k_2 \in \{1, \dots, K\}$ and all $(j_1, j_2) \in \mathcal{B}_{k_1,k_2}$, the estimators $\widehat{\tau}_{j_1,j_2}$, $\widehat{\tau}^B_{k_1,k_2}$, $\widehat{\tau}^R_{k_1,k_2}$, $\widehat{\tau}^D_{k_1,k_2}$ and $\widehat{\tau}^U_{k_1,k_2}$ are all second order U-statistics.*

Proof. First check that, trivially, the sample Kendall's tau $\widehat{\tau}_{j_1,j_2}$ is a U-statistic of order 2 with (symmetric) kernel

$$\begin{aligned} g^*\left((X_{j_1}, X_{j_2})_{i_1}, (X_{j_1}, X_{j_2})_{i_2}\right) =& 2\mathbb{1}\left\{X_{j_1,i_1} < X_{j_1,i_2}, X_{j_2,i_1} < X_{j_2,i_2}\right\} \\ &+ 2\mathbb{1}\left\{X_{j_1,i_2} < X_{j_1,i_1}, X_{j_2,i_2} < X_{j_2,i_1}\right\} - 1. \end{aligned}$$

Consequently,

$$\begin{aligned} \widehat{\tau}^B_{k_1,k_2} &= \frac{1}{|\mathcal{B}_{k_1,k_2}|} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \widehat{\tau}_{j_1,j_2} \\ &= \frac{1}{|\mathcal{B}_{k_1,k_2}|} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \frac{2}{n(n-1)} \sum_{i_1<i_2} g^*\left((X_{j_1}, X_{j_2})_{i_1}, (X_{j_1}, X_{j_2})_{i_2}\right) \\ &= \frac{2}{n(n-1)} \sum_{i_1<i_2} \frac{1}{|\mathcal{B}_{k_1,k_2}|} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} g^*\left((X_{j_1}, X_{j_2})_{i_1}, (X_{j_1}, X_{j_2})_{i_2}\right), \end{aligned}$$

and it follows that $\widehat{\tau}^B_{k_1,k_2}$ is a U-statistic with kernel

$$g^B_{k_1,k_2}\left(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}\right) = \frac{1}{|\mathcal{B}_{k_1,k_2}|} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} g^*\left((X_{j_1}, X_{j_2})_{i_1}, (X_{j_1}, X_{j_2})_{i_2}\right).$$

In a similar manner, it is easily seen that $\widehat{\tau}^R_{k_1,k_2}$, $\widehat{\tau}^D_{k_1,k_2}$ and $\widehat{\tau}^U_{k_1,k_2}$ are all U-statistics as well with respective kernels

$$\begin{aligned} g^R_{k_1,k_2}\left(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}\right) &= \frac{1}{N_{k_1,k_2}} \sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_R} g^*\left((X_{j_1}, X_{j_2})_{i_1}, (X_{j_1}, X_{j_2})_{i_2}\right), \\ g^D_{k_1,k_2}\left(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}\right) &= \frac{1}{N_{k_1,k_2}} \sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_D} g^*\left((X_{j_1}, X_{j_2})_{i_1}, (X_{j_1}, X_{j_2})_{i_2}\right), \\ g^U_{k_1,k_2}\left(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}\right) &= \frac{1}{N_{k_1,k_2}} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} W_{j_1,j_2} g^*\left((X_{j_1}, X_{j_2})_{i_1}, (X_{j_1}, X_{j_2})_{i_2}\right). \end{aligned}$$

Note that the kernel $g^U_{k_1,k_2}$ is random by depending on the weights $\mathbf{W}$. □

From Lemma 18 and the fact that $\mathbb{E}\left[g^*\left((X_{j_1}, X_{j_2})_{i_1}, (X_{j_1}, X_{j_2})_{i_2}\right)\right] = \tau_{k_1,k_2}$ when $(j_1, j_2) \in \mathcal{B}_{k_1,k_2}$, it follows that $\widehat{\mathbf{T}}$, $\widehat{\mathbf{T}}^B$, $\widehat{\mathbf{T}}^R$, $\widehat{\mathbf{T}}^D$ and $\widehat{\mathbf{T}}^U$ are all unbiased estimators of the Kendall's tau matrix. The finite sample variances are given in the following theorem.

Theorem 19. *Under Assumption A1, for distinct $k_1, k_2 \in \{1, \ldots, K\}$ and all $(j_1, j_2) \in \mathcal{B}_{k_1,k_2}$, the variances of $\widehat{\tau}_{j_1,j_2}$, $\widehat{\tau}^B_{k_1,k_2}$, $\widehat{\tau}^R_{k_1,k_2}$, $\widehat{\tau}^D_{k_1,k_2}$ and $\widehat{\tau}^U_{k_1,k_2}$ are given by*

(i)

$$\mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{j_1,j_2}\right] = \frac{8}{n(n-1)}\Big(P_{k_1,k_2} - P^2_{k_1,k_2} + 2(n-2)\big(Q_{j_1,j_2} - P^2_{k_1,k_2}\big)\Big),$$

(ii)

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^B_{k_1,k_2}\right] = \frac{8}{|\mathcal{B}_{k_1,k_2}|n(n-1)}\Bigg(&P_{k_1,k_2} - P^2_{k_1,k_2} + \big(|\mathcal{G}_{k_1}|-1\big)\big(|\mathcal{G}_{k_2}|-1\big)\big(T^B_{k_1,k_2} - P^2_{k_1,k_2}\big)\\ &+ \big(|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2\big)\big(R^B_{k_1,k_2} - P^2_{k_1,k_2}\big)\\ &+ 2(n-2)\Big(Q^B_{k_1,k_2} - P^2_{k_1,k_2} + \big(|\mathcal{G}_{k_1}|-1\big)\big(|\mathcal{G}_{k_2}|-1\big)\big(U^B_{k_1,k_2} - P^2_{k_1,k_2}\big)\\ &+ \big(|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2\big)\big(S^B_{k_1,k_2} - P^2_{k_1,k_2}\big)\Big)\Bigg),\end{aligned}$$

(iii)

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^R_{k_1,k_2}\right] = \frac{8}{N_{k_1,k_2}n(n-1)}\Bigg(&P_{k_1,k_2} - P^2_{k_1,k_2} + \big(N_{k_1,k_2}-1\big)\big(R^R_{k_1,k_2} - P^2_{k_1,k_2}\big)\\ &+ 2(n-2)\Big(Q^R_{k_1,k_2} - P^2_{k_1,k_2} + \big(N_{k_1,k_2}-1\big)\big(S^R_{k_1,k_2} - P^2_{k_1,k_2}\big)\Big)\Bigg),\end{aligned}$$

(iv)

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^D_{k_1,k_2}\right] = \frac{8}{N_{k_1,k_2}n(n-1)}\Bigg(&P_{k_1,k_2} - P^2_{k_1,k_2} + \big(N_{k_1,k_2}-1\big)\big(T^D_{k_1,k_2} - P^2_{k_1,k_2}\big)\\ &+ 2(n-2)\Big(Q^D_{k_1,k_2} - P^2_{k_1,k_2} + \big(N_{k_1,k_2}-1\big)\big(U^D_{k_1,k_2} - P^2_{k_1,k_2}\big)\Big)\Bigg),\end{aligned}$$

(v)

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}\right] = \frac{8}{N_{k_1,k_2}n(n-1)}\Bigg(&P_{k_1,k_2} - P^2_{k_1,k_2} + \frac{N_{k_1,k_2}-1}{|\mathcal{B}_{k_1,k_2}|-1}\\ &\times\Big(\big(|\mathcal{G}_{k_1}|-1\big)\big(|\mathcal{G}_{k_2}|-1\big)\big(T^B_{k_1,k_2} - P^2_{k_1,k_2}\big) + \big(|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2\big)\big(R^B_{k_1,k_2} - P^2_{k_1,k_2}\big)\Big)\\ &+ 2(n-2)\Big(Q^B_{k_1,k_2} - P^2_{k_1,k_2} + \frac{N_{k_1,k_2}-1}{|\mathcal{B}_{k_1,k_2}|-1}\Big(\big(|\mathcal{G}_{k_1}|-1\big)\big(|\mathcal{G}_{k_2}|-1\big)\big(U^B_{k_1,k_2} - P^2_{k_1,k_2}\big)\\ &+ \big(|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2\big)\big(S^B_{k_1,k_2} - P^2_{k_1,k_2}\big)\Big)\Big)\Bigg).\end{aligned}$$

The variance of the sample Kendall's tau estimator is in fact already known, see [22]. We will still prove it for self-completeness and to compare it with the other estimators. Let us proceed with the proof of Theorem 19.

Proof. Recall from Lemma 18 that $\widehat{\tau}_{j_1,j_2}, \widehat{\tau}^B_{k_1,k_2}, \widehat{\tau}^R_{k_1,k_2}, \widehat{\tau}^D_{k_1,k_2}, \widehat{\tau}^U_{k_1,k_2}$ can all be written as U-statistics of order 2 with the following symmetric kernels respectively

$$
\begin{aligned}
g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right) &= 2\mathbb{1}\left\{X_{j_1,i_1} < X_{j_1,i_2}, X_{j_2,i_1} < X_{j_2,i_2}\right\} \\
&\quad + 2\mathbb{1}\left\{X_{j_1,i_2} < X_{j_1,i_1}, X_{j_2,i_2} < X_{j_2,i_1}\right\} - 1 \\
g^B_{k_1,k_2}\left(\mathbf{X}_{i_1},\mathbf{X}_{i_2}\right) &= \frac{1}{|\mathcal{B}_{k_1,k_2}|} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right) \\
g^R_{k_1,k_2}\left(\mathbf{X}_{i_1},\mathbf{X}_{i_2}\right) &= \frac{1}{N_{k_1,k_2}} \sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_R} g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right), \\
g^D_{k_1,k_2}\left(\mathbf{X}_{i_1},\mathbf{X}_{i_2}\right) &= \frac{1}{N_{k_1,k_2}} \sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_D} g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right), \\
g^U_{k_1,k_2}\left(\mathbf{X}_{i_1},\mathbf{X}_{i_2}\right) &= \frac{1}{N_{k_1,k_2}} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} W_{j_1,j_2} g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right).
\end{aligned}
$$

We will use Theorem 15 which gives an expression for the variance of U-statistics. For a second order U-statistic U_n with symmetric kernel $g(\mathbf{X}_1,\mathbf{X}_2)$ satisfying $\mathbb{E}|g(\mathbf{X}_1,\mathbf{X}_2)| < \infty$, the variance is given by

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}[U_n] &= \binom{n}{2}^{-1} \sum_{c=1}^{2} \binom{2}{c}\binom{n-2}{2-c}\zeta_c \\
&= \frac{2}{n(n-1)}\left(2(n-2)\zeta_1 + \zeta_2\right), \qquad (15)
\end{aligned}
$$

where $\zeta_1 = \mathbb{V}\mathrm{ar}\left[g_1(\mathbf{Y})\right]$ and $\zeta_2 = \mathbb{V}\mathrm{ar}\left[g_2(\mathbf{Y}_1,\mathbf{Y}_2)\right]$. As before, we have set $g_1(\mathbf{x}) = \mathbb{E}[g(\mathbf{X}_1,\mathbf{x})]$ and $g_2 := g$. Further, note that $\mathbb{E}\left[g_1(\mathbf{X})\right] = \mathbb{E}\left[g_1\left(\mathbf{X}_1,\mathbf{X}_2\right)\right] = \mathbb{E}\left[U_T\right]$.

We proceed to evaluate ζ_1 and ζ_2 for the different kernels, and then substitute them into (15). Since the kernels g^*, g^B, g^R and g^D are all deterministic, this leaves us with the variances of the corresponding estimators. The variance of $\tau^U_{k_1,k_2}$, however, cannot be calculated directly because of its random kernel g^U. First we proof items (i)-(iv) and then proceed with the proof of (v), where we deal with the randomness of g^U.

(i) Under Assumption A1, we have for every distinct $k_1,k_2 \in \{1,\ldots,K\}$ and all $(j_1,j_2) \in \mathcal{B}_{k_1,k_2}$,

$$\mathbb{E}[g^*\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_2\right)] = \tau_{k_1,k_2} = 2P_{k_1,k_2} - 1.$$

Also,

$$g_1^*(x_{j_1}, x_{j_2}) = \mathbb{E}\left[2\left(\mathbb{1}\left\{x_{j_1} < X_{j_1,1}, x_{j_2} < X_{j_2,1}\right\} + \mathbb{1}\left\{X_{j_1,1} < x_{j_1}, X_{j_2,1} < x_{j_2}\right\}\right) - 1\right]$$

$$= 2P^c(x_{j_1}, x_{j_2}) - 1,$$

where $P^c(x_{j_1}, x_{j_2})$ denotes the probability of concordance of two versions of (X_{j_1}, X_{j_2}) given that one pair equals (x_{j_1}, x_{j_2}). Then,

$$\zeta_1 = \mathbb{V}\text{ar}\left[g_1^*(X_{j_1}, X_{j_2})\right]$$

$$= \mathbb{E}\left[(2P^c(X_{j_1}, X_{j_2}) - 1 - \tau_{k_1,k_2})^2\right]$$

$$= 4\mathbb{E}\left[P^c(X_{j_1}, X_{j_2})^2\right] - 4(1 + \tau_{k_1,k_2})\mathbb{E}\left[P^c(X_{j_1}, X_{j_2})\right] + (1 + \tau_{k_1,k_2})^2.$$

Note that $\mathbb{E}\left[P^c(X_{j_1}, X_{j_2})\right] = P_{k_1,k_2}$ and $\mathbb{E}\left[P^c(X_{j_1}, X_{j_2})^2\right] = Q_{j_1,j_2}$. Furthermore, substitution of $\tau_{k_1,k_2} = 2P_{k_1,k_2} - 1$ gives us

$$\zeta_1 = 4(Q_{j_1,j_2} - P_{k_1,k_2}^2). \tag{16}$$

For ζ_2 we find

$$\zeta_2 = \mathbb{V}\text{ar}\left[g^*\left((X_{j_1}, X_{j_2})_1, (X_{j_1}, X_{j_2})_2\right)\right]$$

$$= \mathbb{E}\Big[\big(2\left(\mathbb{1}\left\{X_{j_1,1} < X_{j_1,2}, X_{j_2,1} < X_{j_2,2}\right\} + \mathbb{1}\left\{X_{j_1,2} < X_{j_1,1}, X_{j_2,2} < X_{j_2,1}\right\}\right) - 1 - \tau_{k_1,k_2}\big)^2\Big]$$

$$= 4\mathbb{E}\left[\left(\mathbb{1}\left\{X_{j_1,1} < X_{j_1,2}, X_{j_2,1} < X_{j_2,2}\right\} + \mathbb{1}\left\{X_{j_1,2} < X_{j_1,1}, X_{j_2,2} < X_{j_2,1}\right\}\right)^2\right]$$
$$- 4(1 + \tau_{k_1,k_2})\mathbb{E}\left[\mathbb{1}\left\{X_{j_1,1} < X_{j_1,2}, X_{j_2,1} < X_{j_2,2}\right\} + \mathbb{1}\left\{X_{j_1,2} < X_{j_1,1}, X_{j_2,2} < X_{j_2,1}\right\}\right]$$
$$+ (1 + \tau_{k_1,k_2})^2.$$

Furthermore, note that

$$\mathbb{1}\left\{X_{j_1,1} < X_{j_1,2}, X_{j_2,1} < X_{j_2,2}\right\} + \mathbb{1}\left\{X_{j_1,2} < X_{j_1,1}, X_{j_2,2} < X_{j_2,1}\right\} \in \{0, 1\}$$

and that therefore the expression is equal to its square. We obtain

$$\zeta_2 = 4\mathbb{E}\left[\left(\mathbb{1}\left\{X_{j_1,1} < X_{j_1,2}, X_{j_2,1} < X_{j_2,2}\right\} + \mathbb{1}\left\{X_{j_1,2} < X_{j_1,1}, X_{j_2,2} < X_{j_2,1}\right\}\right)^2\right]$$

$$
\begin{aligned}
&- 4(1+\tau_{k_1,k_2})\mathbb{E}\left[\mathbb{1}\left\{X_{j_1,1} < X_{j_1,2}, X_{j_2,1} < X_{j_2,2}\right\} + \mathbb{1}\left\{X_{j_1,2} < X_{j_1,1}, X_{j_2,2} < X_{j_2,1}\right\}\right] \\
&+ (1+\tau_{k_1,k_2})^2 \\
&= -4\tau_{k_1,k_2}\mathbb{E}\left[\mathbb{1}\left\{X_{j_1,1} < X_{j_1,2}, X_{j_2,1} < X_{j_2,2}\right\} + \mathbb{1}\left\{X_{j_1,2} < X_{j_1,1}, X_{j_2,2} < X_{j_2,1}\right\}\right] \\
&+ (1+\tau_{k_1,k_2})^2 \\
&= -4(2P_{k_1,k_2}-1)P_{k_1,k_2} + (1+2P_{k_1,k_2}-1)^2 \\
&= 4(P_{k_1,k_2} - P_{k_1,k_2}^2), \qquad (17)
\end{aligned}
$$

where in the second step we have used that

$$
\mathbb{E}\left[\mathbb{1}\left\{X_{j_1,1} < X_{j_1,2}, X_{j_2,1} < X_{j_2,2}\right\} + \mathbb{1}\left\{X_{j_1,2} < X_{j_1,1}, X_{j_2,2} < X_{j_2,1}\right\}\right] = P_{k_1,k_2}.
$$

Substitution of (16) and (17) into (15) gives us the final expression

$$
\mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{j_1,j_2}\right] = \frac{8}{n(n-1)}\left(2(n-2)\left(Q_{j_1,j_2} - P_{k_1,k_2}^2\right) + P_{k_1,k_2} - P_{k_1,k_2}^2\right).
$$

(ii) Here, the following applies

$$
\mathbb{E}[g_{k_1,k_2}^B(\mathbf{X}_1,\mathbf{X}_2)] = \tau_{k_1,k_2} = 2P_{k_1,k_2} - 1,
$$

and

$$
\begin{aligned}
g_{k_1,k_2,1}^B(\mathbf{x}) &= \mathbb{E}\left[\frac{1}{|\mathcal{B}_{k_1,k_2}|}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} h\left((X_{j_1},X_{j_2})_1,(x_{j_1},x_{j_2})\right) - 1\right] \\
&= \frac{1}{|\mathcal{B}_{k_1,k_2}|}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} 2P^c(x_{j_1},x_{j_2}) - 1.
\end{aligned}
$$

For ζ_1 we then obtain

$$
\begin{aligned}
\zeta_1 = \mathbb{V}\mathrm{ar}\left[h_{k_1,k_2,1}^B(\mathbf{X})\right] &= \mathbb{E}\left[\left(\frac{1}{|\mathcal{B}_{k_1,k_2}|}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} 2P^c(X_{j_1},X_{j_2}) - 1 - \tau_{k_1,k_2}\right)^2\right] \\
&= \frac{1}{|\mathcal{B}_{k_1,k_2}|^2}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}}\Big(\mathbb{E}\left[(2P^c(X_{j_1},X_{j_2}) - 1 - \tau_{k_1,k_2})^2\right] \\
&+ \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2} \\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} \mathbb{E}\Big[\left(2P^c(x_{j_1},x_{j_2}) - 1 - \tau_{k_1,k_2}\right)\left(2P^c(X_{j_3},X_{j_4}) - 1 - \tau_{k_1,k_2}\right)\Big]
\end{aligned}
$$

$$+ \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} \mathbb{E}\Big[\left(2P^c(x_{j_1},x_{j_2}) - 1 - \tau_{k_1,k_2}\right)\left(2P^c(X_{j_3},X_{j_4}) - 1 - \tau_{k_1,k_2}\right)\Big]\Big).$$

Now check that when $(j_1,j_2),(j_3,j_4) \in \mathcal{B}_{k_1,k_2}$ such that $|\{j_1,j_2\}\cap\{j_3,j_4\}| = 1$, then

$$\mathbb{E}\left[P^c(X_{j_1},X_{j_2})P^c(X_{j_3},X_{j_4})\right] = S_{j_1,j_2,j_3,j_4},$$

and if $\{j_1,j_2\}\cap\{j_3,j_4\} = \emptyset$, then

$$\mathbb{E}\left[P^c(X_{j_1},X_{j_2})P^c(X_{j_3},X_{j_4})\right] = U_{j_1,j_2,j_3,j_4}.$$

This gives

$$\zeta_1 = \frac{1}{|\mathcal{B}_{k_1,k_2}|^2} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \Big(4(Q_{j_1,j_2} - P_{k_1,k_2}^2) + \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} 4(U_{j_1,j_2,j_3,j_4} - P_{k_1,k_2})^2$$
$$+ \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4(S_{j_1,j_2,j_3,j_4} - P_{k_1,k_2}^2)\Big). \tag{18}$$

Recall the definitions of $Q^B_{k_1,k_2}$, $S^B_{k_1,k_2}$ and $U^B_{k_1,k_2}$ and note that

$$Q^B_{k_1,k_2} = \frac{1}{|\mathcal{B}_{k_1,k_2}|} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} Q_{j_1,j_2},$$
$$S^B_{k_1,k_2} = \frac{1}{|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} S_{j_1,j_2,j_3,j_4},$$
$$U^B_{k_1,k_2} = \frac{1}{(|\mathcal{G}_{k_1}|-1)(|\mathcal{G}_{k_2}|-1)} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} U_{j_1,j_2,j_3,j_4}.$$

We combine them with (18) and obtain

$$\zeta_1 = \frac{4}{|\mathcal{B}_{k_1,k_2}|}\big(Q^B_{k_1,k_2} - P^2_{k_1,k_2} + (|\mathcal{G}_{k_1}|-1)(|\mathcal{G}_{k_2}|-1)(U^B_{k_1,k_2} - P^2_{k_1,k_2})$$
$$+ (|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2)(S^B_{k_1,k_2} - P^2_{k_1,k_2})\big). \tag{19}$$

For ζ_2 we have

$$\zeta_2 = \mathbb{V}\mathrm{ar}\left[g^B_{k_1,k_2}(\mathbf{X}_1,\mathbf{X}_2))\right]$$

$$= \mathbb{E}\left[\left(\frac{1}{|\mathcal{B}_{k_1,k_2}|}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} g^*\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_2\right)-\tau_{k_1,k_2}\right)^2\right]$$

$$= \frac{1}{|\mathcal{B}_{k_1,k_2}|^2}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}}\Big(\mathbb{E}\left[\left(g^*\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_2\right)-\tau_{k_1,k_2}\right)^2\right]$$

$$+ \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} \mathbb{E}\big[\left(g^*\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_2\right)-\tau_{k_1,k_2}\right)$$
$$\times\left(g^*\left((X_{j_3},X_{j_4})_1,(X_{j_3},X_{j_4})_2\right)-\tau_{k_1,k_2}\right)\big]$$

$$+ \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} \mathbb{E}\big[\left(g^*\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_2\right)-\tau_{k_1,k_2}\right)$$
$$\times\left(g^*\left((X_{j_3},X_{j_4})_1,(X_{j_3},X_{j_4})_1\right)-\tau_{k_1,k_2}\right)\big]\Big).$$

Note that when $(j_1,j_2),(j_3,j_4)\in\mathcal{B}_{k_1,k_2}$ such that $|\{j_1,j_2\}\cap\{j_3,j_4\}|=1$, then

$$\mathbb{E}\left[h\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_2\right)h\left((X_{j_3},X_{j_4})_1,(X_{j_3},X_{j_4})_2\right)\right]=4R_{j_1,j_2,j_3,j_4},$$

and if $\{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset$, then

$$\mathbb{E}\left[h\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_2\right)h\left((X_{j_3},X_{j_4})_1,(X_{j_3},X_{j_4})_2\right)\right]=4T_{j_1,j_2,j_3,j_4}.$$

This yields

$$\zeta_2=\frac{1}{|\mathcal{B}_{k_1,k_2}|^2}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}}\Big(4(P_{k_1,k_2}-P_{k_1,k_2}^2)+\sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ \{j_3,j_4\}\cap\{j_1,j_2\}=\emptyset}}4(T_{k_1,k_2}-P_{k_1,k_2}^2)$$
$$+\sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}}4(R_{k_1,k_2}-P_{k_1,k_2}^2)\Big)$$

Further, note that

$$R_{k_1,k_2}^B=\frac{1}{|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}}\sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}}R_{j_1,j_2,j_3,j_4},$$

$$T_{k_1,k_2}^B=\frac{1}{(|\mathcal{G}_{k_1}|-1)(|\mathcal{G}_{k_2}|-1)}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}}\sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}}T_{j_1,j_2,j_3,j_4}.$$

Therefore,

$$\zeta_2 = \frac{4}{|\mathcal{B}_{k_1,k_2}|}\Big(P_{k_1,k_2} - P_{k_1,k_2}^2 + (|\mathcal{G}_{k_1}|-1)\,(|\mathcal{G}_{k_2}|-1)\,(T_{k_1,k_2} - P_{k_1,k_2}^2) \\ + (|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2)\,(R_{k_1,k_2} - P_{k_1,k_2}^2)\Big) \tag{20}$$

Finally, by substituting (19) and (20) into (15) we find

$$\mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{k_1,k_2}^B\right] = \frac{8}{|\mathcal{B}_{k_1,k_2}|n(n-1)}\Bigg(2(n-2)\Big(Q_{k_1,k_2}^B - P_{k_1,k_2}^2 \\ + (|\mathcal{G}_{k_1}|-1)\,(|\mathcal{G}_{k_2}|-1)\,(U_{k_1,k_2}^B - P_{k_1,k_2}^2) + (|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2)\,(S_{k_1,k_2}^B - P_{k_1,k_2}^2)\Big) \\ + P_{k_1,k_2} - P_{k_1,k_2}^2 + (|\mathcal{G}_{k_1}|-1)\,(|\mathcal{G}_{k_2}|-1)\,(T_{k_1,k_2}^B - P_{k_1,k_2}^2) \\ + (|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2)\,(R_{k_1,k_2}^B - P_{k_1,k_2}^2)\Bigg).$$

(iii) In a similar manner to the proof of (ii) we obtain

$$\zeta_1 = \frac{1}{N_{k_1,k_2}^2}\sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_R}\Big(4(Q_{j_1,j_2} - P_{k_1,k_2}^2) + \sum_{\substack{(j_3,j_4)\in(\mathcal{B}_{k_1,k_2})_R \\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} 4(U_{j_1,j_2,j_3,j_4} - P_{k_1,k_2}^2) \\ + \sum_{\substack{(j_3,j_4)\in(\mathcal{B}_{k_1,k_2})_R \\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4(S_{j_1,j_2,j_3,j_4} - P_{k_1,k_2}^2)\Big),$$

Note that for all $(j_1,j_2) \in \mathcal{B}_{k_1,k_2}$, the set $\{(j_3,j_4) \in (\mathcal{B}_{k_1,k_2})_R : \{j_1,j_2\} \cap \{j_3,j_4\} = \emptyset\}$ is in fact empty and that

$$S_{k_1,k_2}^R = \frac{1}{N_{k_1,k_2}(N_{k_1,k_2}-1)}\sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_R}\sum_{\substack{(j_3,j_4)\in(\mathcal{B}_{k_1,k_2})_R \\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} S_{j_1,j_2,j_3,j_4}$$

Therefore,

$$\zeta_1 = \frac{4}{N_{k_1,k_2}}\left(Q_{k_1,k_2}^R - P_{k_1,k_2}^2 + (N_{k_1,k_2}-1)\,(S_{k_1,k_2}^R - P_{k_1,k_2}^2)\right).$$

Similarly, we find that

$$\zeta_2 = \frac{4}{N_{k_1,k_2}}\left(P_{k_1,k_2} - P_{k_1,k_2}^2 + (N_{k_1,k_2}-1)\,(R_{k_1,k_2}^R - P_{k_1,k_2}^2)\right).$$

Hence,

$$\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^R_{k_1,k_2}\right] = \frac{8}{N_{k_1,k_2} n(n-1)}\Bigg(2(n-2)\left(Q^R_{k_1,k_2} - P^2_{k_1,k_2} + (N_{k_1,k_2}-1)\left(S^R_{k_1,k_2} - P^2_{k_1,k_2}\right)\right)$$
$$+ P_{k_1,k_2} - P^2_{k_1,k_2} + (N_{k_1,k_2}-1)\left(R^R_{k_1,k_2} - P^2_{k_1,k_2}\right)\Bigg).$$

(iv) Again, analogous to the proof of (ii) we obtain

$$\zeta_1 = \frac{1}{N^2_{k_1,k_2}} \sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_D} \Bigg(4(Q_{j_1,j_2} - P^2_{k_1,k_2}) + \sum_{\substack{(j_3,j_4)\in(\mathcal{B}_{k_1,k_2})_D \\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} 4(U_{j_1,j_2,j_3,j_4} - P^2_{k_1,k_2})$$
$$+ \sum_{\substack{(j_3,j_4)\in(\mathcal{B}_{k_1,k_2})_D \\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4(S_{j_1,j_2,j_3,j_4} - P^2_{k_1,k_2})\Bigg)$$

We note that the set $\{(j_3,j_4)\in(\mathcal{B}_{k_1,k_2})_D : |\{j_1,j_2\}\cap\{j_3,j_4\}|=1\}$ is empty for all $(j_1,j_2)\in\mathcal{B}_{k_1,k_2}$ and that

$$U^D_{k_1,k_2} = \frac{1}{N_{k_1,k_2}(N_{k_1,k_2}-1)} \sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_D} \sum_{\substack{(j_3,j_4)\in(\mathcal{B}_{k_1,k_2})_D \\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} U_{j_1,j_2,j_3,j_4}.$$

Consequently,

$$\zeta_1 = \frac{4}{N_{k_1,k_2}}\left(Q^D_{k_1,k_2} - P^2_{k_1,k_2} + (N_{k_1,k_2}-1)\left(U^D_{k_1,k_2} - P^2_{k_1,k_2}\right)\right).$$

In a similar manner we obtain ζ_2

$$\zeta_2 = \frac{4}{N_{k_1,k_2}}\left(P_{k_1,k_2} - P^2_{k_1,k_2} + (N_{k_1,k_2}-1)\left(T^D_{k_1,k_2} - P^2_{k_1,k_2}\right)\right).$$

Hence,

$$\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^D_{k_1,k_2}\right] = \frac{8}{N_{k_1,k_2} n(n-1)}\Bigg(2(n-2)\left(Q^D_{k_1,k_2} - P^2_{k_1,k_2} + (N_{k_1,k_2}-1)\left(U^D_{k_1,k_2} - P^2_{k_1,k_2}\right)\right)$$
$$+ P_{k_1,k_2} - P^2_{k_1,k_2} + (N_{k_1,k_2}-1)\left(T^D_{k_1,k_2} - P^2_{k_1,k_2}\right)\Bigg).$$

(v) Lastly, for obtaining the variance of $\widehat{\tau}^U_{k_1,k_2}$ we need to deal with the random kernel $g^U_{k_1,k_2}$. To this end, we use the law of total variance and find

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}\right] &= \mathbb{V}\mathrm{ar}\left[\mathbb{E}\left[\widehat{\tau}^U_{k_1,k_2}|\mathbf{W}\right]\right] + \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}|\mathbf{W}\right]\right] \\ &= \mathbb{V}\mathrm{ar}\left[\tau_{k_1,k_2}\right] + \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}|\mathbf{W}\right]\right] \\ &= \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}|\mathbf{W}\right]\right].\end{aligned}$$

We can thus obtain the desired variance by first evaluating the variance of $\tau^U_{k_1,k_2}$ under a given $\mathbf{W}$ and then by taking the expectation with respect to $\mathbf{W}$. Note that $\widehat{\tau}^U_{k_1,k_2}|\mathbf{W}$ is a U-statistic with (deterministic) kernel $g^U_{k_1,k_2}|\mathbf{W}$. Therefore,

$$\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}\right] = \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}|\mathbf{W}\right]\right] = \frac{2}{n(n-1)}\left(2(n-2)\mathbb{E}\left[\zeta_1\right] + \mathbb{E}\left[\zeta_2\right]\right). \tag{21}$$

With similar steps as before we obtain the corresponding ζ_1 and ζ_2

$$\begin{aligned}\zeta_1 = \frac{1}{N^2_{k_1,k_2}} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \Big(& 4W^2_{j_1,j_2}(Q_{j_1,j_2} - P^2_{k_1,k_2}) \\ & + \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2} \\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} 4W_{j_1,j_2}W_{j_3,j_4}(U_{j_1,j_2,j_3,j_4} - P^2_{k_1,k_2}) \\ & + \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2} \\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4W_{j_1,j_2}W_{j_3,j_4}(S_{j_1,j_2,j_3,j_4} - P^2_{k_1,k_2})\Big),\end{aligned}$$

$$\begin{aligned}\zeta_2 = \frac{1}{N^2_{k_1,k_2}} \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \Big(& 4W^2_{j_1,j_2}(P_{k_1,k_2} - P^2_{k_1,k_2}) + \\ & \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2} \\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} 4W_{j_1,j_2}W_{j_3,j_4}(T_{j_1,j_2,j_3,j_4} - P_{k_1,k_2})^2 \\ & + \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2} \\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4W_{j_1,j_2}W_{j_3,j_4}(R_{j_1,j_2,j_3,j_4} - P^2_{k_1,k_2})\Big).\end{aligned}$$

Inserting them into (21) gives us

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}\right] = \frac{8}{N^2_{k_1,k_2}n(n-1)}\Big(& 2(n-2)\Big(\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \mathbb{E}\left[W^2_{j_1,j_2}\right](Q_{j_1,j_2} - P^2_{k_1,k_2}) \\ & + \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2} \\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} \mathbb{E}\left[W_{j_1,j_2}W_{j_3,j_4}\right](U_{j_1,j_2,j_3,j_4} - P^2_{k_1,k_2})\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2} \\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} \mathbb{E}\left[W_{j_1,j_2}W_{j_3,j_4}\right]\left(S_{j_1,j_2,j_3,j_4} - P_{k_1,k_2}^2\right)\Big) \\
&+ \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \mathbb{E}\left[W_{j_1,j_2}^2\right]\left(P_{k_1,k_2} - P_{k_1,k_2}^2\right) \\
&+ \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2} \\ \{j_1,j_2\}\cap\{j_3,j_4\}=\emptyset}} \mathbb{E}\left[W_{j_1,j_2}W_{j_3,j_4}\right]\left(T_{j_1,j_2,j_3,j_4} - P_{k_1,k_2}^2\right) \\
&+ \sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} \sum_{\substack{(j_3,j_4)\in\mathcal{B}_{k_1,k_2} \\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} \mathbb{E}\left[W_{j_1,j_2}W_{j_3,j_4}\right]\left(R_{j_1,j_2,j_3,j_4} - P_{k_1,k_2}^2\right)\Big) \qquad (22)
\end{aligned}
$$

Recall that we select N_{k_1,k_2} pairs out of the $|\mathcal{B}_{k_1,k_2}|$ possible pairs with uniform probability and without replacement. Therefore, for every distinct $(j_1,j_2),(j_3,j_4)\in\mathcal{B}_{k_1,k_2}$ we have the following expectations

$$
\mathbb{E}\left[W_{j_1,j_2}^2\right] = \mathbb{E}\left[W_{j_1,j_2}\right] = \frac{N_{k_1,k_2}}{|\mathcal{B}_{k_1,k_2}|},
$$

$$
\mathbb{E}\left[W_{j_1,j_2}W_{j_3,j_4}\right] = \frac{N_{k_1,k_2}(N_{k_1,k_2}-1)}{|\mathcal{B}_{k_1,k_2}|(|\mathcal{B}_{k_1,k_2}|-1)}.
$$

Lastly, by inserting these and the expressions of Q_{k_1,k_2}^B, R_{k_1,k_2}^B, S_{k_1,k_2}^B, T_{k_1,k_2}^B and U_{k_1,k_2}^B from the proof of item (ii) into (22), we establish the desired formula

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{k_1,k_2}^U\right] = \frac{8}{N_{k_1,k_2}n(n-1)}\Big(&2(n-2)\Big(Q_{k_1,k_2}^B - P_{k_1,k_2}^2 \\
&+ \frac{N_{k_1,k_2}-1}{|\mathcal{B}_{k_1,k_2}|-1}\big(\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)\left(U_{k_1,k_2}^B - P_{k_1,k_2}^2\right) \\
&+ \left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)\left(S_{k_1,k_2}^B - P_{k_1,k_2}^2\right)\big)\Big) \\
&+ P_{k_1,k_2} - P_{k_1,k_2}^2 + \frac{N_{k_1,k_2}-1}{|\mathcal{B}_{k_1,k_2}|-1}\big(\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)\left(T_{k_1,k_2}^B - P_{k_1,k_2}^2\right) \\
&+ \left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)\left(R_{k_1,k_2}^B - P_{k_1,k_2}^2\right)\big)\Big).
\end{aligned}
$$

□

Remark 20. *If the stronger Assumption A2 holds, all quantities Q_{j_1,j_2}, R_{j_1,j_2,j_3,j_4}, S_{j_1,j_2,j_3,j_4}, T_{j_1,j_2,j_3,j_4} and U_{j_1,j_2,j_3,j_4} are independent of the choice of pairs $(j_1,j_2),(j_3,j_4)\in\mathcal{B}_{k_1,k_2}$ for every distinct $k_1,k_2\in\{1,\ldots,K\}$. It then follows that $Q_{k_1,k_2}^B = Q_{k_1,k_2}^R = Q_{k_1,k_2}^D = Q_{i,j}$ and the same equality holds for quantities R,S,T and U.*

Remark 21. *Under Assumption A1, for distinct k_1, k_2 and all $(j_1, j_2) \in \mathcal{B}_{k_1,k_2}$, it holds that as $n \to \infty$*

$$n^{1/2}\left(\widehat{\tau}_{k_1,k_2} - \tau_{k_1,k_2}\right) \xrightarrow{\text{law}} \mathcal{N}\left(0, V_{k_1,k_2}\right),$$

where $\widehat{\tau}_{k_1,k_2}$ denotes any of the estimators $\widehat{\tau}_{j_1,j_2}$, $\widehat{\tau}^B_{k_1,k_2}$, $\widehat{\tau}^R_{k_1,k_2}$, $\widehat{\tau}^D_{k_1,k_2}$, $\widehat{\tau}^U_{k_1,k_2}$, and the corresponding asymptotic variances V_{k_1,k_2} are respectively given by

$$V_{j_1,j_2} = V_{j_1,j_2}(\mathbb{P}_{\mathbf{X}}) := 16(Q_{j_1,j_2} - P^2_{k_1,k_2}), \tag{23}$$

$$V^B_{k_1,k_2} = V^B_{k_1,k_2}(\mathbb{P}_{\mathbf{X}}) := \frac{16}{|\mathcal{B}_{k_1,k_2}|}\Big(Q^B_{k_1,k_2} - P^2_{k_1,k_2} + \big(|\mathcal{G}_{k_1}| - 1\big)\big(|\mathcal{G}_{k_2}| - 1\big)(U^B_{k_1,k_2} - P^2_{k_1,k_2}) + \big(|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2\big)\big(S^B_{k_1,k_2} - P^2_{k_1,k_2}\big)\Big), \tag{24}$$

$$V^R_{k_1,k_2} = V^R_{k_1,k_2}(\mathbb{P}_{\mathbf{X}}) := \frac{16}{N_{k_1,k_2}}\Big(Q^R_{k_1,k_2} - P^2_{k_1,k_2} + (N_{k_1,k_2} - 1)\left(S^R_{k_1,k_2} - P^2_{k_1,k_2}\right)\Big), \tag{25}$$

$$V^D_{k_1,k_2} = V^D_{k_1,k_2}(\mathbb{P}_{\mathbf{X}}) := \frac{16}{N_{k_1,k_2}}\Big(Q^D_{k_1,k_2} - P^2_{k_1,k_2} + \big(N_{k_1,k_2} - 1\big)\big(U^D_{k_1,k_2} - P^2_{k_1,k_2}\big)\Big), \tag{26}$$

$$V^U_{k_1,k_2} = V^U_{k_1,k_2}(\mathbb{P}_{\mathbf{X}}) := \frac{16}{N_{k_1,k_2}}\Bigg(Q^B_{k_1,k_2} - P^2_{k_1,k_2} + \frac{N_{k_1,k_2} - 1}{|\mathcal{B}_{k_1,k_2}| - 1} \times \Big(\big(|\mathcal{G}_{k_1}| - 1\big)\big(|\mathcal{G}_{k_2}| - 1\big)\big(U^B_{k_1,k_2} - P^2_{k_1,k_2}\big) + \big(|\mathcal{G}_{k_1}| + |\mathcal{G}_{k_2}| - 2\big)\big(S^B_{k_1,k_2} - P^2_{k_1,k_2}\big)\Big)\Bigg), \tag{27}$$

where $\mathbb{P}_{\mathbf{X}}$ denotes the law of the random vector $\mathbf{X}$. This can straightforwardly be derived by combining Theorem 16 and the computations of the corresponding ζ_1's in the proof of Theorem 19.

From the lengthy expressions in Theorem 19 we can derive the asymptotic variances in the setting where $n, |\mathcal{G}_{k_1}|, |\mathcal{G}_{k_2}|, N_{k_1,k_2} \to +\infty$.

Remark 22. *Under Assumption A1, as $n, |\mathcal{G}_{k_1}|, |\mathcal{G}_{k_2}|, N_{k_1,k_2} \to +\infty$, we have the following equivalents*

- $\mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{j_1,j_2}\right] \sim \dfrac{16}{n}\left(Q_{j_1,j_2} - P^2_{k_1,k_2}\right) = \dfrac{1}{n} \times V_{j_1,j_2}(\mathbb{P}_{\mathbf{X}})$,
- $\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^B_{k_1,k_2}\right] \sim \dfrac{16}{n}\left(U^B_{k_1,k_2} - P^2_{k_1,k_2}\right) = \dfrac{1}{n} \times \lim\limits_{|\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}| \to +\infty} V^B_{k_1,k_2}(\mathbb{P}_{\mathbf{X}})$,
- $\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^R_{k_1,k_2}\right] \sim \dfrac{16}{n}\left(S^R_{k_1,k_2} - P^2_{k_1,k_2}\right) = \dfrac{1}{n} \times \lim\limits_{N_{k_1,k_2} \to +\infty} V^R_{k_1,k_2}(\mathbb{P}_{\mathbf{X}})$,
- $\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^D_{k_1,k_2}\right] \sim \dfrac{16}{n}\left(U^D_{k_1,k_2} - P^2_{k_1,k_2}\right) = \dfrac{1}{n} \times \lim\limits_{N_{k_1,k_2} \to +\infty} V^D_{k_1,k_2}(\mathbb{P}_{\mathbf{X}})$,
- $\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}\right] \sim \dfrac{16}{n}\left(U^B_{k_1,k_2} - P^2_{k_1,k_2}\right) = \dfrac{1}{n} \times \lim\limits_{|\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|,N_{k_1,k_2} \to +\infty} V^U_{k_1,k_2}(\mathbb{P}_{\mathbf{X}})$,

Surprisingly, note that the variances do not depend on the block dimensions as soon as they are large enough. In the limit, the quality of the estimator will therefore not improve by averaging over additional elements. Moreover, for large sample sizes, only the quantities S, U and Q determine the levels of variance. Whether the asymptotic variance contains either S or U is determined by which terms (corresponding to the overlapping and non-overlapping combinations of pairs) get the overhand in the summing of the proof of Theorem 19 when dimensions tend to infinity. For the diagonal, random and block estimators the number of terms corresponding to non-overlapping combinations grow faster than the number of overlapping combinations. However, for the row estimator only pairs within the same row are averaged, and thus the limiting variance contains the quantity S (instead of U).

Furthermore, it should be noted that for every distinct k_1, k_2 we have $U^B_{k_1,k_2} \leq Q^B_{k_1,k_2}$ and $S^R_{k_1,k_2} \leq Q^R_{k_1,k_2}$ with equality if and only if the Kendall's taus within the diagonal blocks corresponding to k_1 and k_2 are equal to 1. This follows from the fact that averaging among several unbiased estimates reduces the total variance when compared to no averaging. Moreover, it would seem natural from the definitions of Q_{j_1,j_2}, S_{j_1,j_2,j_1,j_3} and U_{j_1,j_2,j_3,j_4} that

$$U_{j_1,j_2,j_3,j_4} \leq S_{j_1,j_2,j_1,j_3} \leq Q_{j_1,j_2}, \tag{28}$$

for all pairs (j_1, j_2), (j_1, j_3), (j_3, j_4) such these quantities are all well defined, i.e. (j_1, j_2), (j_1, j_3), $(j_3, j_4) \in \mathcal{B}_{k_1,k_2}$ such that k_1, k_2 are distinct. The intuition behind this is the following: the probability that both pairs of observations $(1, 2)$ and $(1, 3)$ are concordant should be naturally greater if the corresponding pairs of variables $\{j_1, j_2\}$ and $\{j_3, j_4\}$ overlap (assuming some kind of positive dependence). If so, then the row estimator performs worse than the block, diagonal and random estimator. Empirically, we indeed observe that this is the case. However, a proof of this inequality seems very difficult. For instance, even in the case of Gaussian copulas with constant correlation r, this would require the computation of involved integrals such as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-rx_1x_2-\frac{1}{2}x_1^2-\frac{1}{2}x_2^2} \times \operatorname{erf}(x_1+x_2+a)\operatorname{erf}(x_1+x_2+b)\operatorname{erf}(x_1+x_2+c)\operatorname{erf}(x_1+x_2+d)\,\mathrm{d}x_1\mathrm{d}x_2,$$

for some constant a, b, c, d, r, where erf is the Gauss error function. Using integration by part, a necessary condition would be the computation of the indefinite integral $\int \exp(-x^2)\operatorname{erf}(\alpha+\beta x)\mathrm{d}x$ for $\alpha, \beta \in \mathbb{R}$. Except when $\alpha = \beta = 0$, this integral does not appear in usual tables of integrals; the Computer algebra system Maple was also unable to compute this integral and we conjecture that it cannot be expressed using standard functions.

Interestingly, the block estimator, the diagonal estimator and the random estimator perform

(almost) equally well in the limit, assuming that the averages U^B and U^D are not far apart. Hence, we can greatly reduce computation time by using either one of the diagonal or random estimator instead of the block estimator, while still maintaining a low asymptotic variance. This is summarised in the following corollary. Furthermore, we set additional conditions to derive an ordering of the three estimators.

Corollary 23. *Under Assumption A1 we have the following:*

(i) Assume that the quantity U does not depend on the choice of pairs within the off-diagonal block $\mathcal{B}_{k_1,k_2}$. Then, the asymptotic variances of the block, diagonal and random estimators are equal:

$$\begin{aligned}\lim_{\substack{n\to+\infty \\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|\to+\infty}} n\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^B_{k_1,k_2}\right] &= \lim_{\substack{n\to+\infty \\ N_{k_1,k_2}\to+\infty}} n\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^D_{k_1,k_2}\right] \\ &= \lim_{\substack{n\to+\infty \\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|,N_{k_1,k_2}\to+\infty}} n\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}\right] = U_{k_1,k_2} - P^2_{k_1,k_2} := \sigma^2_A.\end{aligned}$$

(ii) Further, assume that quantities Q and S do not depend on the choice of pairs within $\mathcal{B}_{k_1,k_2}$ and that $U < S < \frac{1}{2}Q + \frac{1}{2}U$. Then, there is a natural ordering between the estimators, given by:

$$\begin{aligned}0 &< \lim_{\substack{n\to+\infty \\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|\to+\infty}} n\left(|\mathcal{G}_{k_1}|\wedge|\mathcal{G}_{k_2}|\right)\left(\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^B_{k_1,k_2}\right] - \sigma^2_A\right) \\ &< \lim_{\substack{n\to+\infty \\ N_{k_1,k_2}\to+\infty}} n\,N_{k_1,k_2}\left(\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^D_{k_1,k_2}\right] - \sigma^2_A\right) < \lim_{\substack{n\to+\infty \\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|,N_{k_1,k_2}\to+\infty}} n\,N_{k_1,k_2}\left(\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}\right] - \sigma^2_A\right),\end{aligned}$$

(iii) Assume, moreover, that quantities R and T do not depend on the choice of pairs within $\mathcal{B}_{k_1,k_2}$ and that $T < R < \frac{1}{2}P + \frac{1}{2}T$. Then, there is a natural ordering for finite sample sizes and block dimensions, given by:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^B_{k_1,k_2}\right] < \mathbb{V}\mathrm{ar}\left[\widehat{\tau}^D_{k_1,k_2}\right] < \mathbb{V}\mathrm{ar}\left[\widehat{\tau}^U_{k_1,k_2}\right].$$

Proof. (i) This follows from Remark 22 and the assumption that $U^B_{k_1,k_2} = U^D_{k_1,k_2} := U_{k_1,k_2}$ are independent of block dimensions.

(ii) We further have that $Q^B_{k_1,k_2} = Q^D_{k_1,k_2} := Q_{k_1,k_2}$ and $S^B_{k_1,k_2} := S_{k_1,k_2}$ are all independent of block dimensions. From the the expressions in Theorem 19 we then obtain

$$\begin{aligned}0 &< \lim_{\substack{n\to+\infty \\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|,N_{k_1,k_2}\to+\infty}} n\left(|\mathcal{G}_{k_1}|\wedge|\mathcal{G}_{k_2}|\right)\left(\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^B_{k_1,k_2}\right] - \sigma^2_A\right) \\ &= 8\lim_{|\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|,N_{k_1,k_2}\to+\infty} \frac{\left(|\mathcal{G}_{k_1}|\wedge|\mathcal{G}_{k_2}|\right)\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)}{|\mathcal{B}_{k_1,k_2}|}\left(\left(S_{k_1,k_2}-U_{k_1,k_2}\right) - P^2_{k_1,k_2}\right)\end{aligned}$$

$$\leq 16\left((S_{k_1,k_2}-U_{k_1,k_2})-P_{k_1,k_2}^2\right)$$

$$< 16\left((\frac{1}{2}Q_{k_1,k_2}+\frac{1}{2}U_{k_1,k_2}-U_{k_1,k_2})-P_{k_1,k_2}^2\right)$$

$$= 8\left((Q_{k_1,k_2}-U_{k_1,k_2})-P_{k_1,k_2}^2\right)$$

$$= \lim_{\substack{n\to+\infty \\ N_{k_1,k_2}\to+\infty}} n\,N_{k_1,k_2}\left(\mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{k_1,k_2}^D\right]-\sigma_A^2\right)$$

$$< 8\left((Q_{k_1,k_2}-U_{k_1,k_2})-P_{k_1,k_2}^2\right)$$

$$+\lim_{|\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|,N_{k_1,k_2}\to+\infty}\frac{N_{k_1,k_2}(N_{k_1,k_2}-1)}{|\mathcal{B}_{k_1,k_2}|}\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)\left((S_{k_1,k_2}-U_{k_1,k_2})-P_{k_1,k_2}^2\right)$$

$$= \lim_{\substack{n\to+\infty \\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|,N_{k_1,k_2}\to+\infty}} n\,N_{k_1,k_2}\left(\mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{k_1,k_2}^U\right]-\sigma_A^2\right),$$

where we have used that $U_{k_1,k_2} < S_{k_1,k_2} < \frac{1}{2}Q_{k_1,k_2}+\frac{1}{2}U_{k_1,k_2}$.

(iii) Additionally, we have that $T_{k_1,k_2}^B = T_{k_1,k_2}^D := T_{k_1,k_2}$ and $R_{k_1,k_2}^B = R_{k_1,k_2}$ are independent of block dimensions. Further, let us use the notations $C_{1,n} := 16(n-2)/(n(n-1))$ and $C_{2,n} := 8/(n(n-1))$. Following the expressions of Theorem 19 we then obtain

$$\mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{k_1,k_2}^B\right] = \frac{C_{1,n}}{|\mathcal{B}_{k_1,k_2}|}\Big(Q_{k_1,k_2}-P_{k_1,k_2}^2+\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)\left(U_{k_1,k_2}-P_{k_1,k_2}^2\right)$$

$$+\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)\left(S_{k_1,k_2}-P_{k_1,k_2}^2\right)\Big)$$

$$+\frac{C_{2,n}}{|\mathcal{B}_{k_1,k_2}|}\Big(P_{k_1,k_2}-P_{k_1,k_2}^2+\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)\left(T_{k_1,k_2}-P_{k_1,k_2}^2\right)$$

$$+\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)\left(R_{k_1,k_2}-P_{k_1,k_2}^2\right)\Big),$$

$$< \frac{C_{1,n}}{|\mathcal{B}_{k_1,k_2}|}\Big(Q_{k_1,k_2}-P_{k_1,k_2}^2+\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)\left(U_{k_1,k_2}-P_{k_1,k_2}^2\right)$$

$$+\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)\left(\frac{1}{2}Q_{k_1,k_2}+\frac{1}{2}U_{k_1,k_2}-P_{k_1,k_2}^2\right)\Big)$$

$$+\frac{C_{2,n}}{|\mathcal{B}_{k_1,k_2}|}\Big(P_{k_1,k_2}-P_{k_1,k_2}^2+\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)\left(T_{k_1,k_2}-P_{k_1,k_2}^2\right)$$

$$+\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)\left(\frac{1}{2}P_{k_1,k_2}+\frac{1}{2}T_{k_1,k_2}-P_{k_1,k_2}^2\right)\Big),$$

$$= \frac{C_{1,n}}{|\mathcal{B}_{k_1,k_2}|}\left(\frac{1}{2}\big(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|\big)(Q_{k_1,k_2}-P_{k_1,k_2}^2)+\Big(|\mathcal{B}_{k_1,k_2}|-\frac{1}{2}\big(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|\big)\Big)(U_{k_1,k_2}-P_{k_1,k_2}^2)\right)$$

$$+ \frac{C_{2,n}}{|\mathcal{B}_{k_1,k_2}|}\left(\frac{1}{2}\big(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|\big)(P_{k_1,k_2}-P_{k_1,k_2}^2)+\Big(|\mathcal{B}_{k_1,k_2}|-\frac{1}{2}\big(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|\big)\Big)(T_{k_1,k_2}-P_{k_1,k_2}^2)\right)$$

$$\leq \frac{C_{1,n}}{N_{k_1,k_2}}\Big(Q_{k_1,k_2}-P_{k_1,k_2}^2+(N_{k_1,k_2}-1)(U_{k_1,k_2}-P_{k_1,k_2}^2)\Big)$$

$$+ \frac{C_{2,n}}{N_{k_1,k_2}}\Big(P_{k_1,k_2}-P_{k_1,k_2}^2+(N_{k_1,k_2}-1)(T_{k_1,k_2}-P_{k_1,k_2}^2)\Big)$$

$$= \mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{k_1,k_2}^{D}\right]$$

$$= \frac{C_{1,n}}{N_{k_1,k_2}}\Big(Q_{k_1,k_2}-P_{k_1,k_2}^2+\frac{N_{k_1,k_2}-1}{|\mathcal{B}_{k_1,k_2}|-1}\big(\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)(U_{k_1,k_2}-P_{k_1,k_2}^2)$$

$$+\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)(U_{k_1,k_2}-P_{k_1,k_2}^2)\big)\Big)$$

$$+ \frac{C_{2,n}}{|\mathcal{B}_{k_1,k_2}|}\Big(P_{k_1,k_2}-P_{k_1,k_2}^2+\frac{N_{k_1,k_2}-1}{|\mathcal{B}_{k_1,k_2}|-1}\big(\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)(T_{k_1,k_2}-P_{k_1,k_2}^2)$$

$$+\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)(T_{k_1,k_2}-P_{k_1,k_2}^2)\big)\Big)$$

$$< \frac{C_{1,n}}{N_{k_1,k_2}}\Big(Q_{k_1,k_2}-P_{k_1,k_2}^2+\frac{N_{k_1,k_2}-1}{|\mathcal{B}_{k_1,k_2}|-1}\big(\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)(U_{k_1,k_2}-P_{k_1,k_2}^2)$$

$$+\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)(S_{k_1,k_2}-P_{k_1,k_2}^2)\big)\Big)$$

$$+ \frac{C_{2,n}}{|\mathcal{B}_{k_1,k_2}|}\Big(P_{k_1,k_2}-P_{k_1,k_2}^2+\frac{N_{k_1,k_2}-1}{|\mathcal{B}_{k_1,k_2}|-1}\big(\left(|\mathcal{G}_{k_1}|-1\right)\left(|\mathcal{G}_{k_2}|-1\right)(T_{k_1,k_2}-P_{k_1,k_2}^2)$$

$$+\left(|\mathcal{G}_{k_1}|+|\mathcal{G}_{k_2}|-2\right)(R_{k_1,k_2}-P_{k_1,k_2}^2)\big)\Big)$$

$$= \mathbb{V}\mathrm{ar}\left[\widehat{\tau}_{k_1,k_2}^{U}\right],$$

where we have used that $U_{k_1,k_2} < S_{k_1,k_2} < \frac{1}{2}Q_{k_1,k_2} + \frac{1}{2}U_{k_1,k_2}$ and that $T_{k_1,k_2} < R_{k_1,k_2} < \frac{1}{2}P_{k_1,k_2} + \frac{1}{2}T_{k_1,k_2}$.

□

As a consequence of this result, under the above conditions the variance of the block estimator converges the fastest to the asymptotic variance. This is coherent with Theorem 1 of [40]

which shows that under Assumption A2 the block averaging estimator is optimal with respect to the Mahalanobis distance. Furthermore, since the diagonal estimator averages solely over non-overlapping combinations, note that it should converge faster than that of the random estimator. Therefore, if computation costs are to be reduced, the diagonal estimator is preferable to both the random and the row estimator.

Lastly, we note that if only part of the row or diagonal is averaged, the asymptotic variances of the resulting estimators do not change. By doing so we can further lower computation times, but at the cost of attaining the limiting variances at slower rates. It therefore makes sense to choose N_{k_1,k_2} large enough to attain the asymptotic rates, but not too large to keep a low computation time.

4 Improved Estimation of Conditional Kendall's Tau

We extend the aforementioned setting to the conditional setup, when a d-dimensional covariate $\mathbf{Z}$ is available taking values in $\mathcal{Z} \subset \mathbb{R}^d$. The objective is now to estimate the $p \times p$ conditional Kendall's tau matrix $\mathbf{T}_{|\mathbf{Z}=\mathbf{z}} = \left[\tau_{j_1,j_2|\mathbf{Z}=\mathbf{z}}\right]$. To this end, we study several kernel based estimators of the conditional pairwise Kendall's tau in Section 4.1. Then, in Section 4.2 we formalise the assumptions and construct the estimators adjusted for the conditional setup. Lastly, the main theoretical results on the estimators' asymptotic variances are presented in Section 4.3.

4.1 Estimation of Conditional Kendall's Tau

For construction of nonparametric estimates of the conditional Kendall's tau, let us start by recalling the equivalent expressions (4), (5) and (6) of the conditional Kendall's tau

$$\begin{aligned}
\tau_{1,2|\mathbf{Z}=\mathbf{z}} &= \mathbb{P}\left((X_{1,1} - X_{1,2})\,(X_{2,1} - X_{2,2}) > 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\right) \\
&\quad - \mathbb{P}\left((X_{1,1} - X_{1,2})\,(X_{2,1} - X_{2,2}) < 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\right) \\
&= 4\mathbb{P}\left(X_{1,1} < X_{1,2}, X_{2,1} < X_{2,2} | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\right) - 1 \\
&= 1 - 4\mathbb{P}\left(X_{1,1} < X_{1,2}, X_{2,1} > X_{2,2} | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\right),
\end{aligned}$$

Following the approach of [10], we introduce several kernel estimators of $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ motivated by each of the equivalent definitions

$$\begin{aligned}
\widehat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} :=& \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1,n}(\mathbf{z}) w_{i_2,n}(\mathbf{z}) \Big(\mathbb{1}\left\{(X_{1,i_1} - X_{1,i_2})\,(X_{2,i_1} - X_{2,i_2}) > 0\right\} \\
& - \mathbb{1}\left\{(X_{1,i_1} - X_{1,i_2})\,(X_{2,i_1} - X_{2,i_2}) < 0\right\} \Big), \qquad (29) \\
\widehat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}} :=& 4\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1,n}(\mathbf{z}) w_{i_2,n}(\mathbf{z}) \mathbb{1}\left\{X_{1,i_1} < X_{1,i_2}, X_{2,i_1} < X_{2,i_2}\right\} - 1, \qquad (30) \\
\widehat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}} :=& 1 - 4\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1,n}(\mathbf{z}) w_{i_2,n}(\mathbf{z}) \mathbb{1}\left\{X_{1,i_1} < X_{1,i_2}, X_{2,i_1} > X_{2,i_2}\right\}, \qquad (31)
\end{aligned}$$

with Nadaraya-Watson weights $w_{i,n}$ as in (12) given by

$$w_{i,n}(\mathbf{z}) := \frac{K_h(\mathbf{Z}_i - \mathbf{z})}{\sum_{k=1}^{n} K_h(\mathbf{Z}_k - \mathbf{z})}, \qquad (32)$$

for some kernel K on $\mathbb{R}^d$ and bandwidth sequence $h = h(n)$ converging to zero as $n \to \infty$.

Further, let us define

$$\begin{aligned} g_1((X_1,X_2)_{i_1},(X_1,X_2)_{i_2}) :=& \mathbb{1}\left((X_{1,i_1}-X_{1,i_2})\left(X_{2,i_1}-X_{2,i_2}\right)>0\right) \\ & -\mathbb{1}\left((X_{1,i_1}-X_{1,i_2})\left(X_{2,i_1}-X_{2,i_2}\right)<0\right), \\ g_2((X_1,X_2)_{i_1},(X_1,X_2)_{i_2}) :=& 4\mathbb{1}\left(X_{1,i_1}<X_{1,i_1},X_{1,i_2}<X_{1,i_2}\right)-1, \\ g_3((X_1,X_2)_{i_1},(X_1,X_2)_{i_2}) :=& 1-4\mathbb{1}\left(X_{1,i_1}<X_{1,i_2},X_{2,i_1}>X_{2,i_2}\right). \end{aligned}$$

Then $\widehat{\tau}^{(s)}_{1,2|\mathbf{Z}=\mathbf{z}}$ is a smoothed estimator of $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{E}\left[g_s(\mathbf{X}_i,\mathbf{X}_j)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}\right]$, for any choice of $s=1,2,3$. However, note that the terms for which $i_1=i_2$ are treated differently by each of the estimators. Consequently, the estimators $\widehat{\tau}^{(s)}_{1,2|\mathbf{Z}=\mathbf{z}}$, $s=1,2,3$ all take values in different subsets of $[-1,1]$ given by

$$\begin{aligned} 1+s_n \leq \widehat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} \leq 1-s_n, \\ -1 \leq \widehat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}} \leq 1-2s_n, \\ -1+2s_n \leq \widehat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}} \leq 1, \end{aligned}$$

where we denote by s_n the sum of squared weights

$$s_n := \sum_{i=1}^{n} w^2_{i,n}(\mathbf{z}).$$

Furthermore, it was proved in [10] that in fact almost surely

$$\widehat{\tau}^{(1)}_{1,2|\mathbf{Z}=\mathbf{z}} = \widehat{\tau}^{(2)}_{1,2|\mathbf{Z}=\mathbf{z}} + s_n = \widehat{\tau}^{(3)}_{1,2|\mathbf{Z}=\mathbf{z}} - s_n.$$

Consequently, we can define a rescaled version taking values in $[-1,1]$ by

$$\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}} := \frac{\widehat{\tau}^{(1)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}}{1-s_n} = \frac{\widehat{\tau}^{(2)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}}{1-s_n} + \frac{s_n}{1-s_n} = \frac{\widehat{\tau}^{(3)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}}{1-s_n} - \frac{s_n}{1-s_n}. \tag{33}$$

Naturally, we prefer the rescaled version over the other estimators since it takes values in the whole interval $[-1,1]$.

4.2 The Conditional Setup

In the conditional setup we need adapted versions of the structural assumption as well as adapted versions of each of the estimators. We adapt the structural assumption by assuming that the underlying structural pattern applies to the conditional Kendall's tau matrix for any value of the covariate $\mathbf{Z}$. We formalise this in the following assumption.

Assumption A3 (Conditional Structural Assumption). *There exists a* $K > 0$ *and a partition* $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_K\}$ *of* $\{1, \dots, p\}$*, such that for all distinct* $k_1, k_2 \in \{1, \dots, K\}$ *and all* $\mathbf{z} \in \mathcal{Z}$*,*

$$\left[\mathbf{T}_{|\mathbf{Z}=\mathbf{z}}\right]_{\mathcal{B}_{k_1,k_2}} = \tau_{k_1,k_2}\mathbf{1}$$

for some constants $\tau_{k_1,k_2} \in [-1, 1]$*. Here,* $\mathcal{B}_{k_1,k_2} = \{(j_1, j_2) \in (\mathcal{G}_{k_1} \times \mathcal{G}_{k_2})\}$*, for all* $k_1, k_2 \in \{1, \dots, K\}$*.*

In terms of stock return modeling, Assumption A3 has the following interpretation: conditionally on a given market state or portfolio movement, the stocks of companies from different sectors/countries have equal rank correlations with every other pair from the respective groups. This could for instance be used for the computation of conditional risk measures. Note that we then also assume that the blocks do not change as a function of the conditioning variables, which seems reasonable in our perspective of groups chosen from sectors or countries. For any fixed $z \in \mathcal{Z}$, the number of free parameters is again reduced from $\frac{1}{2}p(p-1)$ to

$$\frac{1}{2}K(K-1) + \frac{1}{2}\sum_{k=1}^{K}|\mathcal{G}_k|(|\mathcal{G}_k| - 1),$$

allowing for an improved estimation based on the averaging principle. Note that here the free parameters are a function of z and therefore live in an infinite-dimensional space, contrary to the previous section where each parameter was a real number.

Let us denote the (unaveraged) conditional Kendall's tau matrix estimator as $\widehat{\mathbf{T}}_{|\mathbf{Z}=\mathbf{z}} = \left[\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}\right]_{p\times p}$ with $\widetilde{\tau}_{j_1,j_2}$ as in (33). As in the unconditional framework that was studied previously, we define conditional versions of the averaging estimators by

$$\widehat{\tau}^{B}_{k_1,k_2|\mathbf{Z}=\mathbf{z}} = \frac{1}{|\mathcal{B}_{k_1,k_2}|}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}}\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$$

$$\widehat{\tau}^{R}_{k_1,k_2|\mathbf{Z}=\mathbf{z}} = \frac{1}{N_{k_1,k_2}}\sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_R}\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}},$$

$$\widehat{\tau}^{D}_{k_1,k_2|\mathbf{Z}=\mathbf{z}} = \frac{1}{N_{k_1,k_2}}\sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_D}\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}},$$

$$\widehat{\tau}^{U}_{k_1,k_2|\mathbf{Z}=\mathbf{z}} = \frac{1}{N_{k_1,k_2}}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}}W_{j_1,j_2}\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}},$$

using the same notations $\mathcal{B}_{k_1,k_2}, (\mathcal{B}_{k_1,k_2})_R, (\mathcal{B}_{k_1,k_2})_D, N_{k_1,k_2}$ and $\mathbf{W}$ as before. The corresponding conditional Kendall's tau matrix estimators are denoted by $\widehat{\mathbf{T}}^{B}_{|\mathbf{Z}=\mathbf{z}}$, $\widehat{\mathbf{T}}^{R}_{|\mathbf{Z}=\mathbf{z}}$, $\widehat{\mathbf{T}}^{D}_{|\mathbf{Z}=\mathbf{z}}$, $\widehat{\mathbf{T}}^{U}_{|\mathbf{Z}=\mathbf{z}}$.

Next, let us define auxiliary notations P, Q, R, S, T and U for the conditional setup. Under

Assumption A3, we define the conditional version of P_{k_1,k_2} for each distinct $k_1, k_2 \in \{1, \dots, K\}$ and all $\mathbf{z} \in \mathcal{Z}$ as

$$P_{k_1,k_2|\mathbf{Z}=\mathbf{z}} := \mathbb{P}\left(\left(X_{j_1,1} - X_{j_1,2}\right)\left(X_{j_2,1} - X_{j_2,2}\right) > 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\right),$$

where (j_1, j_2) is a pair in block $\mathcal{B}_{k_1,k_2}$. Note that Assumption A3 ensures that $P_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$ is independent of the choice of pair $(j_1, j_2) \in \mathcal{B}_{k_1,k_2}$. Alternatively we can write $P_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$ in terms of the conditional copula $C_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$ of (X_{j_1}, X_{j_2}) by

$$P_{k_1,k_2|\mathbf{Z}=\mathbf{z}} = 2\int_{[0,1]^2} C_{j_1,j_2|\mathbf{Z}=\mathbf{z}}(u_1, u_2)\mathrm{d}C_{j_1,j_2|\mathbf{Z}=\mathbf{z}}(u_1, u_2).$$

Let the conditional version of Q_{j_1,j_2} for every $(j_1, j_2) \in \{1, \dots, p\}^2$ and all $\mathbf{z} \in \mathcal{Z}$ be given by

$$\begin{aligned} Q_{j_1,j_2|\mathbf{Z}=\mathbf{z}} := \mathbb{P}(&\left(X_{j_1,1} - X_{j_1,2}\right)\left(X_{j_2,1} - X_{j_2,2}\right) > 0, \\ &\left(X_{j_1,1} - X_{j_1,3}\right)\left(X_{j_2,1} - X_{j_2,3}\right) > 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}). \end{aligned}$$

or alternatively in terms of copulas

$$Q_{j_1,j_2|\mathbf{Z}=\mathbf{z}} = \int_{[0,1]^2} \left(C_{j_1,j_2|\mathbf{Z}=\mathbf{z}}(u_1, u_2) + \bar{C}_{\mathbf{Z}=\mathbf{z}}(u_1, u_2)\right) \mathrm{d}C_{j_1,j_2\mathbf{Z}=\mathbf{z}}(u_1, u_2),$$

where $\bar{C}_{j_1,j_2\mathbf{Z}=\mathbf{z}}$ denotes the survival copula of variables (X_{j_1}, X_{j_2}) conditional on $\mathbf{Z} = \mathbf{z}$. The derivation is analogous to the unconditional case. For all $\mathbf{z} \in \mathcal{Z}$ all distinct $k_1, k_2 \in \{1, \dots, K\}$ and for every $(j_1, j_2), (j_3, j_4) \in \mathcal{B}_{k_1,k_2}$ such that $|\{j_1, j_2\} \cap \{j_3, j_4\}| = 1$, we define

$$\begin{aligned} R_{j_1,j_2,j_3,j_4|\mathbf{Z}=\mathbf{z}} := \mathbb{P}(&\left(X_{j_1,1} - X_{j_1,2}\right)\left(X_{j_2,1} - X_{j_2,2}\right) > 0, \\ &\left(X_{j_3,1} - X_{j_3,2}\right)\left(X_{j_4,1} - X_{j_4,2}\right) > 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}) \\ S_{j_1,j_2,j_3,j_4|\mathbf{Z}=\mathbf{z}} := \mathbb{P}(&\left(X_{j_1,1} - X_{j_1,2}\right)\left(X_{j_2,1} - X_{j_2,2}\right) > 0, \\ &\left(X_{j_3,1} - X_{j_3,3}\right)\left(X_{j_4,1} - X_{j_4,3}\right) > 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}), \end{aligned}$$

and for every $(j_1, j_2), (j_3, j_4) \in \mathcal{B}_{k_1,k_2}$ such that $\{j_1, j_2\} \cap \{j_3, j_4\} = \emptyset$, we define

$$\begin{aligned} T_{j_1,j_2,j_3,j_4|\mathbf{Z}=\mathbf{z}} := \mathbb{P}(&\left(X_{j_1,1} - X_{j_1,2}\right)\left(X_{j_2,1} - X_{j_2,2}\right) > 0, \\ &\left(X_{j_3,1} - X_{j_3,2}\right)\left(X_{j_4,1} - X_{j_4,2}\right) > 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}), \\ U_{j_1,j_2,j_3,j_4|\mathbf{Z}=\mathbf{z}} := \mathbb{P}(&\left(X_{j_1,1} - X_{j_1,2}\right)\left(X_{j_2,1} - X_{j_2,2}\right) > 0, \\ &\left(X_{j_3,1} - X_{j_3,3}\right)\left(X_{j_4,1} - X_{j_4,3}\right) > 0 | \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}). \end{aligned}$$

As before, the superscripts Q^B,Q^R and Q^D are used to denote the auxiliary quantities averaged

over the whole block, row or diagonal respectively.

We will examine estimators' performances when the sample size $n \to \infty$, as finite sample results depend heavily on the bandwidth and the covariate's underlying distribution. Hence, the conditional versions of quantities R and T will turn out to be superfluous. Similar to the unconditional setting, these are only useful for presenting finite sample results.

4.3 Comparison of their Asymptotic Variances

Before proceeding with the asymptotic results, we need to formalise some regularity assumptions on the kernel K, the covariate $\mathbf{Z}$ and the bandwidth sequence $h = h(n)$. Since we follow the proof strategy of Derumigny et al. in [10], Proposition 9, we use an adapted version of their assumptions.

Assumption A4. *(a) The kernel K is bounded, compactly supported, symmetrical in the sense that $K(\mathbf{u}) = K(-\mathbf{u})$ for every $\mathbf{u} \in \mathbb{R}^d$ and satisfies $\int K = 1, \int |K| < \infty, \int K^2 < \infty$.*

(b) The kernel is of order α for some integer $\alpha > 1$, i.e. for all $k = 1, \dots, \alpha - 1$ and every indices $j_1, \dots, j_k$ in $\{1, \dots, d\}$,

$$\int K(\mathbf{u}) u_{j_1} \dots u_{j_k} d\mathbf{u} = 0.$$

(c) In addition, $\mathbb{E}\left[K_h(\mathbf{Z} - \mathbf{z})\right] > 0$ for every $\mathbf{z} \in \mathcal{Z}$ and $h > 0$.

Assumption A5. *For every $\mathbf{x} \in \mathbb{R}^p, \mathbf{z} \mapsto f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z})$ is continuous and almost everywhere differentiable on $\mathcal{Z}$ up to the order α. For every $0 \leq k \leq \alpha$ and every $1 \leq j_1, \dots, j_\alpha \leq d$, let*

$$\mathcal{H}_{k,\vec{j}}(\mathbf{u}, \mathbf{v}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) := \sup_{t \in [0,1]} \left| \frac{\partial^k f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_1} \dots \partial z_{j_k}} (\mathbf{x}_1, \mathbf{z} + th\mathbf{u}) \frac{\partial^{\alpha-k} f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_{k+1}} \dots \partial z_{j_\alpha}} (\mathbf{x}_2, \mathbf{z} + th\mathbf{v}) \right|$$

denoting $\vec{j} = (j_1, \dots, j_\alpha)$. Assume that $\mathcal{H}_{k,\vec{j}}(\mathbf{u}, \mathbf{v}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z})$ is integrable and there exists a finite constant $C_{\mathbf{XZ},\alpha} > 0$ such that, for every $\mathbf{z} \in \mathcal{Z}$ and every $h < 1$,

$$\int |K|(\mathbf{u}) |K|(\mathbf{v}) \sum_{k=0}^{\alpha} \binom{\alpha}{k} \sum_{j_1, \dots, j_\alpha = 1}^{p} \mathcal{H}_{k,\vec{j}}(\mathbf{u}, \mathbf{v}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) \left| u_{j_1} \dots u_{j_k} v_{j_{k+1}} \dots v_{j_\alpha} \right| d\mathbf{u} d\mathbf{v} d\mathbf{x}_1 d\mathbf{x}_2$$

is less than $C_{\mathbf{XZ},\alpha}$.

Assumption A6. *$nh^d \to \infty$ and $nh^{d+2\alpha} \to 0$ as $n \to \infty$.*

Assumption A4 is satisfied for the commonly used Epanechnikov kernel. It should however be noted that the Gaussian kernel does not meet the requirements of Assumption $A4$ due to its infinite support. Furthermore, under the above assumptions, the asymptotic distributions of the averaging estimators are in fact independent of the choice of the pairwise conditional Kendall's

tau estimator $\widehat{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$ or $\widetilde{\tau}^{(s)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$, $s=1,2,3$ on which the averaging is based. We formalise this in the following lemma.

Lemma 24. *Under Assumptions A4-A6, for any $\mathbf{z}\in\mathcal{Z}$ and any $j_1,j_2\in\{1,\dots,p\}$, the asymptotic laws of the pairwise estimators $\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$ and $\widehat{\tau}^{(s)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$, $s=1,2,3$ are all equal as $n\to\infty$. As such, the asymptotic laws of the averaging estimators $\widehat{\tau}^{B}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'}$, $\widehat{\tau}^{R}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'}$, $\widehat{\tau}^{D}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$ and $\widehat{\tau}^{U}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$ are all invariant under interchanging $\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$ for any of the $\widehat{\tau}^{(s)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$, $s=1,2,3$ in their respective definitions.*

Proof. Let us study the behaviour of

$$s_n=\sum_{i=1}^n w_{i,n}^2(\mathbf{z})=\frac{\frac{1}{n^2}\sum_{i=1}^n K_h^2\left(\mathbf{Z}_i-\mathbf{z}\right)}{\left(\frac{1}{n}\sum_{i_2=1}^n K_h\left(\mathbf{Z}_{i_2}-\mathbf{z}\right)\right)^2},$$

as $n\to\infty$. Under Assumption A4(a) and by Bochner's lemma, we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n K_h^2\left(\mathbf{Z}_i-\mathbf{z}\right)\right]&=\frac{1}{n}\int\frac{1}{h^{2d}}K^2\left(\frac{\mathbf{z}_1-\mathbf{z}}{h}\right)f_{\mathbf{Z}}(\mathbf{z_1})\mathrm{d}\mathbf{z}_1\\&=\frac{1}{nh^d}\int K^2(\mathbf{u})f_{\mathbf{Z}}(\mathbf{z}+h\mathbf{u})\mathrm{d}\mathbf{u}\\&\xrightarrow[h\to 0]{}\frac{1}{nh^d}f_{\mathbf{Z}}(\mathbf{z})\int K^2=O((nh^d)^{-1}).\end{aligned}\tag{34}$$

Similarly, we find by Assumption A4(a) that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n K_h\left(\mathbf{Z}_i-\mathbf{z}\right)\right]=O\left(1\right).$$

As a consequence, by dominated convergence $\mathbb{E}\left[s_n\right]=O((nh^d)^{-1})$ for any $\mathbf{z}\in\mathcal{Z}$, or equivalently using Markov's inequality, $s_n=O_P((nh^d)^{-1})$. By Assumption A6, this converges to zero. Recall from (33) that almost surely

$$\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}=\frac{\widehat{\tau}^{(1)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}}{1-s_n}=\frac{\widehat{\tau}^{(2)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}}{1-s_n}+\frac{s_n}{1-s_n}=\frac{\widehat{\tau}^{(3)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}}{1-s_n}-\frac{s_n}{1-s_n}.\tag{35}$$

It then follows by Slutsky's theorem that the limiting laws of $\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$ and $\widehat{\tau}^{(s)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$, $s=1,2,3$ are all equivalent as $n\to\infty$. As such, the asymptotic laws of the averaging estimators are therefore invariant under interchanging $\widehat{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$ for any of the $\widetilde{\tau}^{(s)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$, $s=1,2,3$. □

We now present our main theoretical result on the joint asymptotic normality at different points of the conditioning variable $\mathbf{Z}$.

Theorem 25. *(Joint asymptotic normality at different points). Let $\mathbf{z}'_1, \ldots, \mathbf{z}'_{n'}$ be fixed, distinct points in the set $\mathcal{Z} \subset \mathbf{R}^d$ and assume Assumptions A3-A6. For all distinct $k_1, k_2 \in \{1, \ldots, K\}$ and all $(j_1, j_2) \in \mathcal{B}_{k_1,k_2}$, as $n \to \infty$*

$$\left(nh^d\right)^{1/2} \left(\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} - \tau_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}\right)_{j'=1,\ldots,n'} \xrightarrow{\text{law}} \mathcal{N}\left(0, \mathbf{H}_{k_1,k_2}\right),$$

where $\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'}$ denotes any of the estimators $\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}'}$, $\widehat{\tau}^B_{k_1,k_2|\mathbf{Z}=\mathbf{z}'}$, $\widehat{\tau}^R_{k_1,k_2|\mathbf{Z}=\mathbf{z}'}$, $\widehat{\tau}^D_{k_1,k_2|\mathbf{Z}=\mathbf{z}'}$, $\widehat{\tau}^U_{k_1,k_2|\mathbf{Z}=\mathbf{z}'}$ and the corresponding $n' \times n'$ diagonal matrices $\mathbf{H}_{k_1,k_2}$ are respectively given by

(i)

$$\left[\mathbf{H}_{j_1,j_2}\right]_{j'_1,j'_2} = \frac{\int K^2 \mathbb{1}_{\{j'_1=j'_2\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)} V_{j_1,j_2}\left(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j'_1}}\right),$$

(ii)

$$\left[\mathbf{H}^B_{k_1,k_2}\right]_{j'_1,j'_2} = \frac{\int K^2 \mathbb{1}_{\{j'_1=j'_2\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)} V^B_{k_1,k_2}\left(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j'_1}}\right),$$

(iii)

$$\left[\mathbf{H}^R_{k_1,k_2}\right]_{j'_1,j'_2} = \frac{\int K^2 \mathbb{1}_{\{j'_1=j'_2\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)} V^R_{k_1,k_2}\left(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j'_1}}\right),$$

(iv)

$$\left[\mathbf{H}^D_{k_1,k_2}\right]_{j'_1,j'_2} = \frac{\int K^2 \mathbb{1}_{\{j'_1=j'_2\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)} V^D_{k_1,k_2}\left(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j'_1}}\right),$$

(v)

$$\left[\mathbf{H}^U_{k_1,k_2}\right]_{j'_1,j'_2} = \frac{\int K^2 \mathbb{1}_{\{j'_1=j'_2\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)} V^U_{k_1,k_2}\left(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j'_1}}\right),$$

for every $1 \leq j'_1, j'_2 \leq n'$, where the asymptotic variances V are as defined in Equations (23)-(27).

Proof. In [10] (see p. 299) it was already shown that the conditional Kendall's tau estimator defined in (33) is asymptotically normal at different points of the conditioning variable. We can use this result directly for the proof of (i). However, the asymptotic normalities of the averaging estimators remain to be proven. To this end, we follow their approach of studying the joint distribution of U-statistics at several conditioning points. The asymptotic covariance matrices are then

obtained by combining the results with the appropriate kernels under Assumption A3.

(i) Let us start with restating the result from [10] on the joint asymptotic normality of the conditional Kendall's tau estimator at different points. That is, as $n \to \infty$

$$\left(nh^d\right)^{1/2}\left(\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}'_{j'}} - \tau_{j_1,j_2|\mathbf{Z}=\mathbf{z}'_{j'}}\right)_{j'=1,\ldots,n'} \xrightarrow{\text{law}} \mathcal{N}\left(0, \mathbf{H}_{j_1,j_2}\right),$$

where $\mathbf{H}_{j_1,j_2}$ is a $n' \times n'$ diagonal matrix given by

$$\begin{aligned}\left[\mathbf{H}_{j_1,j_2}\right]_{j'_1,j'_2} = &\frac{4\int K^2 \mathbb{1}_{\{j'_1=j'_2\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)} \\ &\Big(\mathbb{E}\big[g^*\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_2\right) g^*\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_3\right) \mid \mathbf{Z_1}=\mathbf{Z}_2=\mathbf{Z}_3=\mathbf{z}'_{j'_1}\big] \\ &- \tau^2_{j_1,j_2|\mathbf{Z}=\mathbf{z}'_{j'}}\Big),\end{aligned}$$

where $1 \le j'_1, j'_2 \le n'$, and

$$\begin{aligned}g^*\left((X_{j_1},X_{j_2})_1,(X_{j_1},X_{j_2})_2\right) = &2\mathbb{1}\left\{X_{j_1,1} < X_{j_1,2}, X_{j_2,1} < X_{j_2,2}\right\} \\ &+2\mathbb{1}\left\{X_{j_1,2} < X_{j_1,1}, X_{j_2,2} < X_{j_2,1}\right\} - 1.\end{aligned}$$

Then, by similar algebraic steps as in the derivation of ζ_1 from the proof of Theorem 19(i), we obtain

$$\left[\mathbf{H}_{j_1,j_2}\right]_{j'_1,j'_2} = \frac{16\int K^2 \mathbb{1}_{\{j'_1=j'_2\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)}\left(Q_{j_1,j_2|\mathbf{Z}=\mathbf{z}'_{j'_1}} - P^2_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'_1}}\right).$$

(ii)-(v) From Lemma 24 we known that the asymptotic distributions of the averaging estimators are invariant under interchanging $\widehat{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$ for any of the $\widetilde{\tau}^{(s)}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$, $s=1,2,3$ in their definitions. We continue the proof using the symmetric version of the estimator $\widetilde{\tau}^{(2)}_{j_1,j_2}$, so that the proof stays similar to the one of Theorem 19. Again, we set the following notations

$$\begin{aligned}g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right) = &2\mathbb{1}\left\{X_{j_1,i_1} < X_{j_1,i_2}, X_{j_2,i_1} < X_{j_2,i_2}\right\} \\ &+ 2\mathbb{1}\left\{X_{j_1,i_2} < X_{j_1,i_1}, X_{j_2,i_2} < X_{j_2,i_1}\right\} - 1 \\ g^B_{k_1,k_2}\left(\mathbf{X}_{i_1},\mathbf{X}_{i_2}\right) = &\frac{1}{|\mathcal{B}_{k_1,k_2}|}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right) \\ g^R_{k_1,k_2}\left(\mathbf{X}_{i_1},\mathbf{X}_{i_2}\right) = &\frac{1}{N_{k_1,k_2}}\sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_R} g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right), \\ g^D_{k_1,k_2}\left(\mathbf{X}_{i_1},\mathbf{X}_{i_2}\right) = &\frac{1}{N_{k_1,k_2}}\sum_{(j_1,j_2)\in(\mathcal{B}_{k_1,k_2})_D} g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right), \\ g^U_{k_1,k_2}\left(\mathbf{X}_{i_1},\mathbf{X}_{i_2}\right) = &\frac{1}{N_{k_1,k_2}}\sum_{(j_1,j_2)\in\mathcal{B}_{k_1,k_2}} W_{j_1,j_2} g^*\left((X_{j_1},X_{j_2})_{i_1},(X_{j_1},X_{j_2})_{i_2}\right).\end{aligned}$$

Furthermore, for any measurable function $g : \mathbb{R}^{2d} \to \mathbb{R}$, let us define the second order U-statistic

$$U_{n,j'}(g) := \frac{1}{n(n-1)} \sum_{1 \le i_1 \ne i_2 \le n} \frac{g\left(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}\right) K_h\left(\mathbf{z}'_{j'} - \mathbf{Z}_{i_1}\right) K_h\left(\mathbf{z}'_{j'} - \mathbf{Z}_{i_2}\right)}{\mathbb{E}\left[K_h\left(\mathbf{z}'_{j'} - \mathbf{Z}\right)\right]^2}$$
$$=: \frac{1}{n(n-1)} \sum_{1 \le i_1 \ne i_2 \le n} g_{i_1,i_2}. \tag{36}$$

It follows easily that the averaging estimators can be written in terms of $U_{n,j'}$ by

$$\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} = \frac{U_{n,j'}(g_{k_1,k_2})}{U_{n,j'}(1) + \epsilon_{n,j'}}, \tag{37}$$

where we write $\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}$ for any of the estimators $\widehat{\tau}^B_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}, \widehat{\tau}^R_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}, \widehat{\tau}^D_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}, \widehat{\tau}^U_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}$, with g_{k_1,k_2} given by, respectively, $g^B_{k_1,k_2}, g^R_{k_1,k_2}, g^D_{k_1,k_2}, g^U_{k_1,k_2}$. The residual term $\epsilon_{n,j'}$ is given by

$$\epsilon_{n,j'} := \frac{\sum_{i=1}^n K_h^2\left(\mathbf{z}'_{j'} - \mathbf{Z}_i\right)}{n(n-1)\mathbb{E}\left[K_h(\mathbf{z}'_{j'} - \mathbf{Z})\right]^2}.$$

Further, we set

$$\widetilde{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} := \frac{U_{n,j'}(g_{k_1,k_2})}{U_{n,j'}(1)} \tag{38}$$

and find that under Assumption A4

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\epsilon_{n,j'}}(\widetilde{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} - \widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}})\right] &= \mathbb{E}\left[\frac{1}{\epsilon_{n,j'}}\left(\frac{U_{n,j'}(h_{k_1,k_2})}{U_{n,j'}(1)} - \frac{U_{n,j'}(h_{k_1,k_2})}{U_{n,j'}(1)+\epsilon_{n,j'}}\right)\right] \\
&= \mathbb{E}\left[\frac{1}{U_{n,j'}(1)}\frac{U_{n,j'}(h_{k_1,k_2})}{U_{n,j'}(1)+\epsilon_{n,j'}}\right] \\
&= \mathbb{E}\left[\frac{1}{U_{n,j'}(1)}\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}\right] \\
&= O(1),
\end{aligned}$$

or equivalently $\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} - \widetilde{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} = O_P(\epsilon_{n,j'})$ using Markov's inequality. By Assumption A4(c) and by the same argument as in (34) we see that $\epsilon_{n,j'} = O_P((nh^d)^{-1})$. It then follows by Assumption A6 that

$$(nh^d)^{1/2}(\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} - \widetilde{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}) = O_P((nh^d)^{1/2}\epsilon_{n,j'}) = o_P(1).$$

It therefore suffices to obtain the limiting law of $(nh^d)^{1/2}(\widetilde{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} - \tau_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}})$ as $n \to \infty$.

Now let us apply the result from [10] (see p. 318) on the joint asymptotic law of U-statistics of the form $U_{n,j'}$. That is, under Assumptions A4-A6 and for any two bounded measurable functions

g_1 and g_2

$$\left(nh^d\right)^{1/2}\left(\left(U_{n,j'}\left(g_1\right)-\mathbb{E}\left[U_{n,j'}\left(g_1\right)\right]\right)_{j'=1,\ldots,n'},\left(U_{n,j'}\left(g_2\right)-\mathbb{E}\left[U_{n,j'}\left(g_2\right)\right]\right)_{j'=1,\ldots,n'}\right) \xrightarrow{\text{law}} \mathcal{N}\left(0,\begin{bmatrix} M_\infty\left(g_1\right) & M_\infty\left(g_1,g_2\right) \\ M_\infty\left(g_1,g_2\right) & M_\infty\left(g_2\right)\end{bmatrix}\right), \tag{39}$$

as $n\to\infty$, where

$$\left[M_\infty\left(g_1,g_2\right)\right]_{j_1',j_2'} := \frac{4\int K^2\mathbb{1}_{\{\mathbf{z}'_{j_1'}=\mathbf{z}'_{j_2'}\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j_1'}\right)}\int g_1\left(\mathbf{x}_1,\mathbf{x}\right)g_2\left(\mathbf{x}_2,\mathbf{x}\right) \times f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j_1'}}(\mathbf{x})f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j_1'}}\left(\mathbf{x}_1\right)f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j_1'}}\left(\mathbf{x}_2\right)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2. \tag{40}$$

Let us investigate the expectation of $U_{n,j'}(g)$. We write by (36)

$$\mathbb{E}\left[U_{n,j'}(g)\right]=\frac{1}{\mathbb{E}\left[K_h(\mathbf{z}'_{j'}-\mathbf{Z})\right]^2}\mathbb{E}\Big[g\left(\mathbf{X}_1,\mathbf{X}_2\right)K_h\left(\mathbf{z}'_{j'}-\mathbf{Z}_1\right)K_h\left(\mathbf{z}'_{j'}-\mathbf{Z}_2\right)\Big].$$

Further, by a change of variable we find

$$\begin{aligned}&\mathbb{E}\Big[g\left(\mathbf{X}_1,\mathbf{X}_2\right)K_h\left(\mathbf{z}'_{j'}-\mathbf{Z}_1\right)K_h\left(\mathbf{z}'_{j'}-\mathbf{Z}_2\right)\Big]\\ &=\int g\left(\mathbf{x}_1,\mathbf{x}_2\right)K_h\left(\mathbf{z}'_{j'}-\mathbf{z}_1\right)K_h\left(\mathbf{z}'_{j'}-\mathbf{z}_2\right)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x_1},\mathbf{z}_1)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x_2},\mathbf{z}_2)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{z}_1\mathrm{d}\mathbf{z}_2\\ &=\int g\left(\mathbf{x}_1,\mathbf{x}_2\right)K\left(u_1\right)K\left(u_2\right)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x_1},\mathbf{z}'_{j'}+hu_1)f_{\mathbf{X},\mathbf{Z}}(\mathbf{x_2},\mathbf{z}'_{j'}+hu_2)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{u}_1\mathrm{d}\mathbf{u}_2.\end{aligned} \tag{41}$$

Let us define the function $\phi_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t) := f_{\mathbf{X},\mathbf{Z}}\left(\mathbf{x}_1,\mathbf{z}'_{j'}+th\mathbf{u}_1\right)f_{\mathbf{X},\mathbf{Z}}\left(\mathbf{x}_2,\mathbf{z}'_{j'}+th\mathbf{u}_2\right)$ for $t\in[0,1]$. By Assumption A5, $\phi_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t)$ is α times differentiable, allowing us to apply the Taylor-Lagrange formula. This gives

$$\phi_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t)=\sum_{k=0}^{\alpha-1}\frac{1}{k!}\phi^{(k)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(0)+\frac{1}{\alpha!}\phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t^*), \tag{42}$$

for some $t^*\in[0,1]$ and where $\phi^k_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t)$ is equal to

$$\sum_{l=0}^{k}\binom{k}{l}\sum_{j_1,\ldots,j_k=1}^{d}h^k u_{j_1,1}\ldots u_{j_l,1}u_{j_{l+1},2}\ldots u_{j_k,2} \frac{\partial^k f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_1}\ldots\partial z_{j_l}}\left(\mathbf{x}_1,\mathbf{z}'_{j'}+th\mathbf{u}_1\right)\frac{\partial^{k-l}f_{\mathbf{X},\mathbf{Z}}}{\partial z_{j_{l+1}}\ldots\partial z_{j_k}}\left(\mathbf{x}_2,\mathbf{z}'_{j'}+th\mathbf{u}_2\right).$$

After substituting (42) into (41) we obtain

$$
\begin{aligned}
&\int g\left(\mathbf{x}_1, \mathbf{x}_2\right) K\left(\mathbf{u}_1\right) K\left(\mathbf{u}_2\right) \left(\sum_{k=0}^{\alpha-1} \frac{1}{k!} \phi^{(k)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(0) + \frac{1}{\alpha!} \phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t^*)\right) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \\
&= \int g\left(\mathbf{x}_1, \mathbf{x}_2\right) K\left(\mathbf{u}_1\right) K\left(\mathbf{u}_2\right) \left(\phi_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(0) + \frac{1}{\alpha!} \phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t^*)\right) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \\
&= \int g\left(\mathbf{x}_1, \mathbf{x}_2\right) K\left(\mathbf{u}_1\right) K\left(\mathbf{u}_2\right) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x_1}, \mathbf{z}'_{j'}) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x_2}, \mathbf{z}'_{j'}) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \\
&+ \frac{1}{\alpha!} \int g\left(\mathbf{x}_1, \mathbf{x}_2\right) K\left(\mathbf{u}_1\right) K\left(\mathbf{u}_2\right) \phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t^*) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \\
&= f^2_{\mathbf{Z}}(\mathbf{z}'_{j'}) \mathbb{E}\Big[g\left(\mathbf{X}_1, \mathbf{X}_2\right) \mid \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_{j'}\Big] \\
&+ \frac{1}{\alpha!} \int g\left(\mathbf{x}_1, \mathbf{x}_2\right) K\left(\mathbf{u}_1\right) K\left(\mathbf{u}_2\right) \phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t^*) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2,
\end{aligned}
$$

where we have used the fact that $\int K(\mathbf{u}) u_{j_1} \ldots u_{j_k} d\mathbf{u} = 0$ for all $k = 1, \ldots, \alpha - 1$ as stated in Assumption A4(b).

Furthermore, by Assumption A5 we have

$$
\begin{aligned}
&\frac{1}{\alpha!} \left| \int g\left(\mathbf{x}_1, \mathbf{x}_2\right) K\left(\mathbf{u}_1\right) K\left(\mathbf{u}_2\right) \phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t^*) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \right| \\
&\leq \int |K|\left(\mathbf{u}_1\right) |K|\left(\mathbf{u}_2\right) |\phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}|(t^*) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{u}_1 \mathrm{d}\mathbf{u}_2 \\
&\leq C_{\mathbf{X},\mathbf{Z}} h^{\alpha}.
\end{aligned}
$$

Therefore,

$$
\mathbb{E}\Big[g\left(\mathbf{X}_1, \mathbf{X}_2\right) K_h\left(\mathbf{z}'_{j'} - \mathbf{Z}_1\right) K_h\left(\mathbf{z}'_{j'} - \mathbf{Z}_2\right)\Big] = f^2_{\mathbf{Z}}(\mathbf{z}'_j) \mathbb{E}\Big[g\left(\mathbf{X}_1, \mathbf{X}_2\right) \mid \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_{j'}\Big] + O(h^{\alpha}),
$$

and by the same reasoning we obtain

$$
\mathbb{E}\left[K_h(\mathbf{z}'_{j'} - \mathbf{Z})\right] = f^2_{\mathbf{Z}}(\mathbf{z}'_j) + O(h^{\alpha}).
$$

Consequently, we find that

$$
\begin{aligned}
\mathbb{E}\left[U_{n,j'}(g)\right] &= \frac{1}{\mathbb{E}\left[K_h(\mathbf{z}'_{j'} - \mathbf{Z})\right]^2} \mathbb{E}\Big[g\left(\mathbf{X}_1, \mathbf{X}_2\right) K_h\left(\mathbf{z}'_{j'} - \mathbf{Z}_1\right) K_h\left(\mathbf{z}'_{j'} - \mathbf{Z}_2\right)\Big] \\
&= \mathbb{E}\Big[g\left(\mathbf{X}_1, \mathbf{X}_2\right) \mid \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_{j'}\Big] + r_{n,j'},
\end{aligned} \tag{43}
$$

where $|r_{n,j'}| \leq C_0 h^{\alpha}$ for some constant C_0 independent of j'. Then, by Assumption A6

$$
(nh^d)^{1/2} \left(\mathbb{E}[U_{n,j'}(g)] - \mathbb{E}[g(\mathbf{X}_1, \mathbf{X}_2) \mid \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z'}_{j'}]\right) = O((nh^d)^{1/2} h^{\alpha}) = o(1).
$$

Therefore, the asymptotic law of (39) still holds after replacing $\mathbb{E}[U_{n,j'}(g)]$ with $\mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2) \mid \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_{j'}]$. As such,

$$\left(nh^d\right)^{1/2}\left(\left(U_{n,j'}\left(g_{k_1,k_2}\right)-\tau_{k_1,k_2|\mathbf{Z}=\mathbf{z}_{j'}}\right)_{j'=1,\ldots,n'},\left(U_{n,j'}\left(1\right)-1\right)_{j'=1,\ldots,n'}\right) \xrightarrow{\text{law}} \mathcal{N}\left(0,\left[\begin{array}{cc} M_\infty\left(g_{k_1,k_2},g_{k_1,k_2}\right) & M_\infty\left(g_{k_1,k_2},1\right) \\ M_\infty\left(g_{k_1,k_2},1\right) & M_\infty\left(1,1\right)\end{array}\right]\right), \tag{44}$$

as $n \to \infty$, where we have used that $\mathbb{E}[g_{k_1,k_2}(\mathbf{X}_1,\mathbf{X}_2) \mid \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}'_{j'}] = \tau_{k_1,k_2|\mathbf{Z}=\mathbf{z}_{j'}}$ under Assumption A3.

In order to derive the asymptotic law of $(nh^d)^{1/2}(\widetilde{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} - \tau_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}})$ we apply the Delta-method on (44) with the function $\gamma(\mathbf{x}), := \mathbf{x}/\mathbf{y}$, that divides two real vectors $\mathbf{x},\mathbf{y}$ of size n' component-wise. The corresponding Jacobian is given by the $n' \times 2n'$ matrix

$$J_\gamma(\mathbf{x},\mathbf{y}) = \left[\text{Diag}\left(y_1^{-1},\ldots y_{n'}^{-1}\right), \text{Diag}\left(-x_1y_1^{-2},\cdots - x_{n'}y_{n'}^{-2}\right)\right].$$

Hence, as $n \to \infty$

$$\left(nh^d\right)^{1/2}\left(\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}} - \tau_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}\right)_{j'=1,\ldots,n'} \xrightarrow{\text{law}} \mathcal{N}\left(0,\mathbf{H}_{k_1,k_2}\right),$$

setting

$$\mathbf{H}_{k_1,k_2} := J_\gamma(\vec{\tau}_{k_1,k_2},\mathbf{e})\left[\begin{array}{cc} M_\infty\left(g_{k_1,k_2}\right) & M_\infty\left(g_{k_1,k_2},1\right) \\ M_\infty\left(g_{k_1,k_2},1\right) & M_\infty(1)\end{array}\right] J_\rho(\vec{\tau}_{k_1,k_2},\mathbf{e})^T,$$

where $\vec{\tau}_{k_1,k_2}$ and $\mathbf{e}$ denote n'-dimensional vectors filled with respectively τ_{k_1,k_2} and 1. This gives

$$\begin{aligned}\mathbf{H}_{k_1,k_2} &= M_\infty\left(g_{k_1,k_2},g_{k_1,k_2}\right) - \text{Diag}(\vec{\tau}_{k_1,k_2})M_\infty\left(g_{k_1,k_2},1\right) - M_\infty\left(g_{k_1,k_2},1\right)\text{Diag}(\vec{\tau}_{k_1,k_2}) \\ &\quad + \text{Diag}(\vec{\tau}_{k_1,k_2})M_\infty(1,1)\,\text{Diag}(\vec{\tau}_{k_1,k_2})\end{aligned}$$

and for $1 \le j'_1, j'_2 \le n'$, we find

$$\begin{aligned}\left[M_\infty\left(g_{k_1,k_2},g_{k_1,k_2}\right)\right]_{j'_1,j'_2} &= \frac{4\int K^2\mathbb{1}_{\{\mathbf{z}'_{j'_1}=\mathbf{z}'_{j'_2}\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)} \\ &\quad \mathbb{E}\left[g_{k_1,k_2}\left(\mathbf{X}_1,\mathbf{X}\right)g_{k_1,k_2}\left(\mathbf{X}_2,\mathbf{X}\right) \mid \mathbf{Z}=\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_{j'_1}\right],\end{aligned}$$

$$\left[\text{Diag}(\vec{\tau}_{k_1,k_2})M_\infty\left(g_{k_1,k_2},1\right)\right]_{j'_1,j'_2} = \frac{4\int K^2\mathbb{1}_{\{\mathbf{z}'_{j'_1}=\mathbf{z}'_{j'_2}\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)}\tau_{k_1,k_2}\mathbb{E}\left[g_{k_1,k_2}\left(\mathbf{X}_1,\mathbf{X}\right) \mid \mathbf{Z}=\mathbf{Z}_1=\mathbf{z}'_{j'_1}\right]$$

$$
\begin{aligned}
&= \frac{4\int K^2 \mathbb{1}_{\{\mathbf{z}'_{j'_1}=\mathbf{z}'_{j'_2}\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)} \tau^2_{k_1,k_2} \\
&= \left[M_\infty\left(g_{k_1,k_2},1\right)\mathrm{Diag}(\vec{\tau}_{k_1,k_2})\right]_{j'_1,j'_2} \\
&= \left[\mathrm{Diag}(\vec{\tau}_{k_1,k_2} M_\infty\left(1,1\right)\mathrm{Diag}(\vec{\tau}_{k_1,k_2})\right]_{j'_1,j'_2},
\end{aligned}
$$

and thus,

$$
\begin{aligned}
[\mathbf{H}_{k_1,k_2}]_{j'_1,j'_2} = &\frac{4\int K^2 \mathbb{1}_{\left\{\mathbf{z}'_{j'_1}=\mathbf{z}'_{j'_2}\right\}}}{f_{\mathbf{Z}}\left(\mathbf{z}'_{j'_1}\right)} \\
&\left(\mathbb{E}\left[g_{k_1,k_2}\left(\mathbf{X}_1,\mathbf{X}_2\right) g_{k_1,k_2}\left(\mathbf{X}_1,\mathbf{X}_3\right) \mid \mathbf{Z}=\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_{j'_1}\right] - \tau^2_{k_1,k_2|\mathbf{Z}=\mathbf{z}'_{j'}}\right).
\end{aligned}
$$

Lastly, by substituting the appropriate kernels and by similar steps as in the derivation of the corresponding ζ_1's from the proof of Theorem 19, it is easily seen that under Assumption A3 we obtain the desired asymptotic covariance matrices.

□

From Theorem 25 we see that asymptotically the estimators are all unbiased. This is not surprising, as the bandwidth then goes to zero. For finite sample sizes there is a tradeoff between bias and variance, which we can control by the choice of bandwidth. For more theory on bandwidth selection methods see [30]. On the variance, we remark the following.

Remark 26. *Under the assumptions of Theorem 25, we have for $\mathbb{P}_{\mathbf{Z}}$-almost all $\mathbf{z} \in \mathcal{Z}$,*

$$
\lim_{n\to+\infty}\left(nh^d\right)^{1/2}\left(\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}} - \tau_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right) \overset{\text{law}}{=} \frac{\int K^2}{f_{\mathbf{Z}}(\mathbf{z})} \times \lim_{n\to+\infty} n^{1/2}\big(\widehat{\tau}_{k_1,k_2}(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}) - \tau_{k_1,k_2}(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}})\big),
$$

where on the left-hand side $\widehat{\tau}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$ denotes any of the estimators $\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$, $\widehat{\tau}^B_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$, $\widehat{\tau}^R_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$, $\widehat{\tau}^D_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$, $\widehat{\tau}^U_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$, and on the right-hand side $\widehat{\tau}_{k_1,k_2}(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}})$ denotes the estimated Kendall's tau if we had observed a sample of size n from the distribution $\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$ and $\tau_{k_1,k_2}(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}) = \tau_{k_1,k_2|\mathbf{Z}=\mathbf{z}}$ denotes Kendall's tau of the distribution $\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$.

Remark 27. *Under the same assumptions as in Theorem 25 and by letting the sample size and dimensions tend to infinity, the following holds for $\mathbb{P}_{\mathbf{Z}}$-almost all $\mathbf{z} \in \mathcal{Z}$,*

- $\mathbb{V}\mathrm{ar}\left[\widetilde{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}\right] \sim \frac{16\int K^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}\left(Q_{j_1,j_2|\mathbf{Z}=\mathbf{z}} - P^2_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right)$,
- $\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^B_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right] \sim \frac{16\int K^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}\left(U^B_{k_1,k_2|\mathbf{Z}=\mathbf{z}} - P^2_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right)$,
- $\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^R_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right] \sim \frac{16\int K^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}\left(S^R_{k_1,k_2|\mathbf{Z}=\mathbf{z}} - P^2_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right)$,

- $\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^{D}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right] \sim \frac{16\int K^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}\left(U^{D}_{k_1,k_2|\mathbf{Z}=\mathbf{z}} - P^2_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right),$
- $\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^{U}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right] \sim \frac{16\int K^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}\left(U^{B}_{k_1,k_2|\mathbf{Z}=\mathbf{z}} - P^2_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right),$

As seen before, we remark that the asymptotic variances have analogous expressions to that of their unconditional counterparts. Note that therefore the same ordering as in Corollary 23 holds under adapted conditions, where the (unconditional) distribution of $\mathbf{X}$ is replaced by $\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$. Since the proof of Corollary 28 is analogous to that of Corollary 23, it is omitted.

Corollary 28. *Under the assumptions of Theorem 25 we have the following:*

(i) Assume that the quantity $U_{|\mathbf{Z}=\mathbf{z}}$ does not depend on the choice of pairs within the off-diagonal block $\mathcal{B}_{k_1,k_2}$. Then, the asymptotic variances of the block, diagonal and random estimators are equal:

$$\begin{aligned}\lim_{\substack{n\to+\infty\\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|\to+\infty}} nh^d\,\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^{B}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right] &= \lim_{\substack{n\to+\infty\\ N_{k_1,k_2}\to+\infty}} nh^d\,\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^{D}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right]\\ &= \lim_{\substack{n\to+\infty\\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|,N_{k_1,k_2}\to+\infty}} nh^d\,\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^{U}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right]\\ &= U_{k_1,k_2|\mathbf{Z}=\mathbf{z}} - P^2_{k_1,k_2|\mathbf{Z}=\mathbf{z}} := \sigma^2_{A|\mathbf{Z}=\mathbf{z}}.\end{aligned}$$

(ii) Further, assume that quantities $Q_{|\mathbf{Z}=\mathbf{z}}$ and $S_{|\mathbf{Z}=\mathbf{z}}$ do not depend on the choice of pairs within $\mathcal{B}_{k_1,k_2}$ and that $U_{|\mathbf{Z}=\mathbf{z}} < S_{|\mathbf{Z}=\mathbf{z}} < \frac{1}{2}Q_{|\mathbf{Z}=\mathbf{z}} + \frac{1}{2}U_{|\mathbf{Z}=\mathbf{z}}$. Then, there is a natural ordering between the estimators, given by:

$$\begin{aligned}0 &< \lim_{\substack{n\to+\infty\\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|\to+\infty}} nh^d\,(|\mathcal{G}_{k_1}|\wedge|\mathcal{G}_{k_2}|)\left(\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^{B}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right] - \sigma^2_{A|\mathbf{Z}=\mathbf{z}}\right)\\ &< \lim_{\substack{n\to+\infty\\ N_{k_1,k_2}\to+\infty}} nh^d\,N_{k_1,k_2}\left(\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^{D}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right] - \sigma^2_{A|\mathbf{Z}=\mathbf{z}}\right)\\ &< \lim_{\substack{n\to+\infty\\ |\mathcal{G}_{k_1}|,|\mathcal{G}_{k_2}|,N_{k_1,k_2}\to+\infty}} nh^d\,N_{k_1,k_2}\left(\mathbb{V}\mathrm{ar}\left[\widehat{\tau}^{U}_{k_1,k_2|\mathbf{Z}=\mathbf{z}}\right] - \sigma^2_{A|\mathbf{Z}=\mathbf{z}}\right),\end{aligned}$$

It again follows that all averaging estimators exhibit a lower asymptotic variance than the naive conditional Kendall's tau estimator. Also, the row averaging estimator intuitively performs less than the block, diagonal and random estimators and for growing dimensions the block, diagonal and random estimator perform (almost) equally in the limit, assuming that the averages of U are not far apart. Again, we can greatly reduce computation time by using either one of the diagonal or random averaging estimator instead of the block averaging estimator, since then only part of all conditional Kendall's taus have to be computed. Under mild assumptions, the block estimator estimator converges fastest to the U-asymptotic variance, followed by the diagonal and

then the random estimator. Lastly, it also holds that if only part of the row or diagonal is averaged, the asymptotic variances of the resulting estimators do not change. By doing so we can further decrease computation time, but at the cost of attaining the limiting variances at slower rates.

5 Simulation Study

We perform a simulation study to assess the finite sample properties of our estimators. Firstly, in Section 5.1, we compare the unconditional estimators by studying their variances and computation times for varying block and sample sizes. In the unconditional setting, we consider the row, diagonal, block and sample Kendall's tau estimators. In Section 5.2, we focus on the conditional versions of the diagonal and block estimators and we let the Kendall's taus depend on a one-dimensional covariate. Similarly, we compare their accuracy and computational efficiency for varying sample size and block dimensions. In addition, we examine the estimators' optimal bandwidths under varying conditional dependencies of the Kendall's tau matrix. The simulations are all executed with the help of the statistical environment R [43]. For simplicity, we choose $N_{k_1,k_2} = |\mathcal{G}_{k_1}| \wedge |\mathcal{G}_{k_2}|$, so that diagonal and row estimators average over the same number of terms.

5.1 Unconditional Kendall's Tau

In the unconditional framework, we compare the block, row, diagonal and sample Kendall's tau matrix estimators. The random estimator is left out as it is mainly of interest from a theoretical point of view. We will examine how the estimators' variance changes as a function of the block dimensions and the sample size. For this purpose, we consider mean squared errors (MSEs), which is a measure of variance as all unconditional estimators are unbiased. Furthermore, we measure computation times for comparing the computational efficiency. For computing the pairwise sample Kendall's taus, we use the function `cor.fk()` in the R package `pcaPP` [18], which can efficiently calculate the sample Kendall's tau with runtime $O(n \log(n))$.

In each of the simulations, data is generated using either the Gaussian distribution or the Student's t distribution with one degree of freedom, with zero mean and unit marginal variances. As both distributions are elliptical, correlation matrices can be directly obtained from the Kendall's tau matrix using Theorem 12. We let the underlying Kendall's tau matrix be block-structured corresponding to two groups of equal size, which we will refer to as the block size. This results in two diagonal blocks and a single distinct off-diagonal block due to symmetry.

As obtained in Theorem 19, the variances depend on the averages of the auxiliary quantities P, Q, R, S, T and U of pairs either along the row, the diagonal or over the entire block. For a fair comparison of the different estimators, we need all different averages of these auxiliary quantities to be equal. As such, in addition to having identical Kendall's tau values in the off-diagonal block, we take the values within the diagonal blocks to be identical as well. In that case, all auxiliary quantities are independent of the choice of pairs within the off-diagonal block, and moreover the Partial Exchangeability Assumption holds.

In Section 5.1.1 we study MSEs and computation times for different sample sizes. For this,

we set the diagonal block Kendall's taus to 0.3 and the off-diagonal block Kendall's taus to 0.1, which can be considered realistic for the modelling of stock returns. Similarly, using these values, we study the estimators' performances for different block dimensions in Section 5.1.2. To confirm these results, we are interested in the estimators' behavior for a different Kendall's tau matrix: we now set the diagonal block Kendall's taus to 0.5 and the off-diagonal block Kendall's taus to -0.5. Note that the diagonal blocks need to have a positive value to ensure positive definiteness. For this data-generating process, we will study the influence of the block dimension on the MSEs in Section 5.1.2.

For performance analysis, we will focus on estimates of the single off-diagonal block, as all estimators treat the diagonal blocks equally. As such, computation times and MSEs result from only estimating the single off-diagonal block.

5.1.1 Effect of the Sample Size

In the first experiment, we study the dependency of the MSE on the sample size. To this end, the sample size is varied and the block size is fixed to 10, resulting in a block size of 10×10. We examine data generated from both the Gaussian distribution and the Student's t-distribution. For each estimator, the MSEs and 95% confidence intervals are calculated using 4000 replications. The results can be found on a log-log scale in Figure 1.

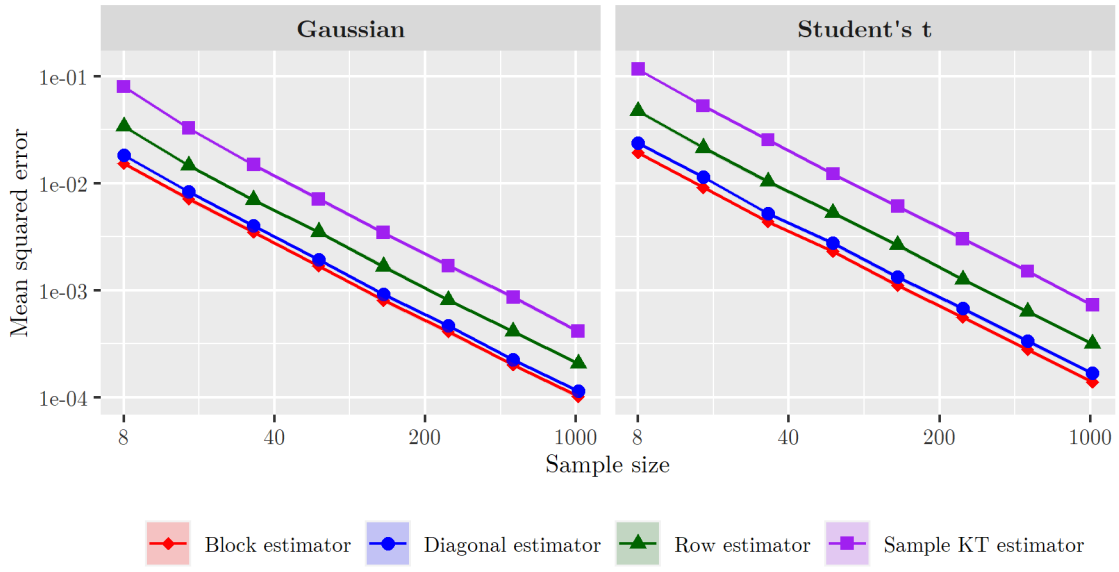


Figure 1: Log-log plots of the estimators' mean squared error as a function of the sample size including 95% confidence intervals, calculated using a block size of 10. The diagonal block Kendall's taus are set at 0.3 and the off-diagonal block values at 0.1.

Figure 1 shows that for each estimator the MSEs generated from the Gaussian distribution are slightly lower than those generated from the Student's t distribution, but apart from that there is little difference. In the figure we clearly observe straight lines for all of the estimators, with slopes indicating an inverse relationship. This not only confirms that the limiting variances are inversely proportional to the sample size, but that this also applies accurately for small sample sizes. In addition, we see that averaging the sample Kendall's taus does indeed lead to better estimates. This applies to any given sample size, as all estimates depend equally on it. As expected, the block estimator behaves best, only closely followed by the diagonal estimator.

Next, we study the dependency of the computation times on sample size. For this experiment, we set the block size to 40 and for each estimator we calculate the average of the computation time by performing 400 replications. The results are shown in Figure 2 on log-log scale. It shows that as the sample size increases, the computation times gradually increase to a point where they appear to scale linearly with each other. These observations are in line with the runtime of the pairwise sample Kendall's tau estimator used of $O(n \log(n))$. As expected, the computation times of the block and sample Kendall's tau matrix estimator are very similar, as are the computation times of the row and diagonal estimators, with the latter two being significantly more efficient for any given sample size.

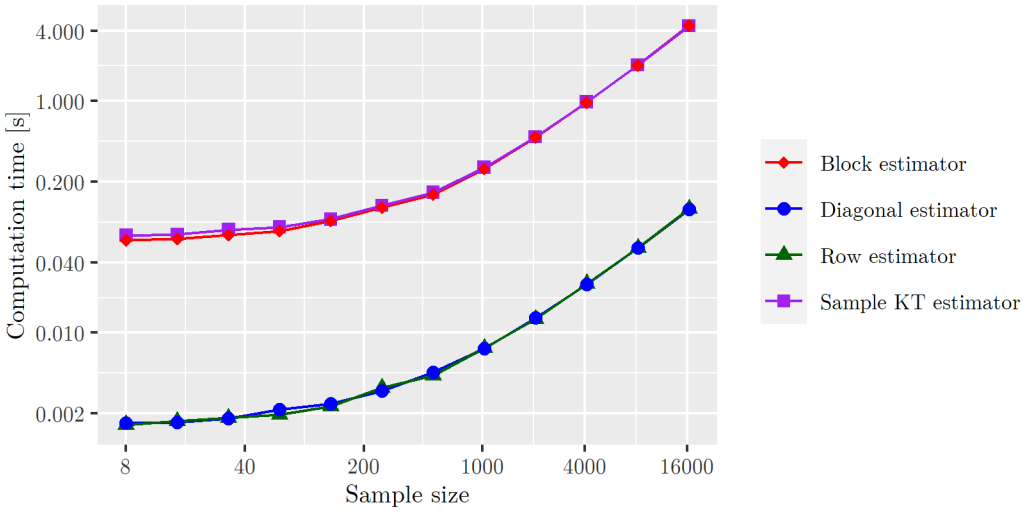


Figure 2: Log-log plot of the estimators' mean computation time [s] as a function of the sample size, calculated using a block size of 40.

5.1.2 Effect of the Block Size

We first study the behaviour of the MSE with respect to varying block sizes with off-diagonal block Kendall's taus of 0.1 and diagonal block Kendall's taus of 0.3. In this experiment, we set the sample size to 4 to reduce the computational cost of running a sufficient number of replications. Again, we examine data generated from the Gaussian and the Student's t distributions. The MSEs and 95% pointwise confidence intervals are calculated using 4000 replications. See Figure 3 for a log-log plot of the MSEs as a function block size.

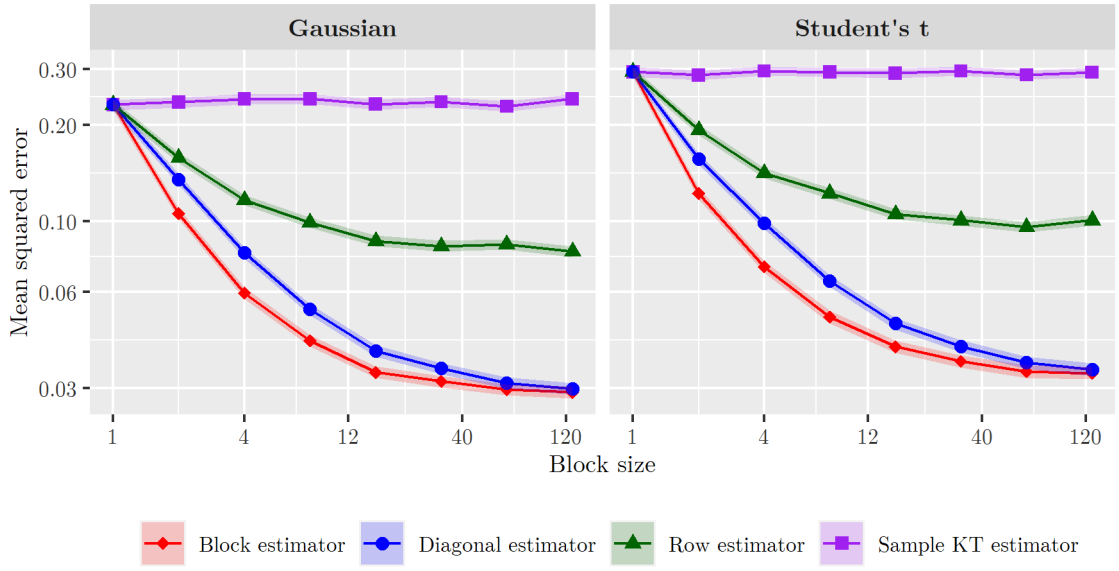


Figure 3: Log-log plots of the mean squared error as a function of the block size including 95% confidence intervals, calculated using a sample size of 4. The diagonal block Kendall's taus are set at 0.3 and the off-diagonal block values at 0.1.

Figure 3 shows that all of the averaging estimators perform increasingly better than the sample Kendall's tau estimator for growing block dimensions. For large block dimensions the MSEs seem to reach constant values, confirming that the asymptotic variances do not depend on block dimensions. As expected, the block and diagonal averaging estimators both converge to the lowest limiting variance, approached fastest by the block averaging estimator. The row averaging estimator performs considerably less. The only difference we observe between the Gaussian and the Student's t distribution is again that that the MSEs of the latter are slightly higher. This is not surprising as all of the variances depend on the underlying copula.

To better visualise the difference in order of magnitude between the MSEs of the estimators $\widehat{\tau}$, we plot the ratio $\mathrm{MSE}(\widehat{\tau})/\mathrm{MSE}(\widehat{\tau}^B)$ as a function of the block size, where $\widehat{\tau}$ denotes any of the row, diagonal and naive estimators. The results are depicted in Figure 4 including 95% confidence intervals.

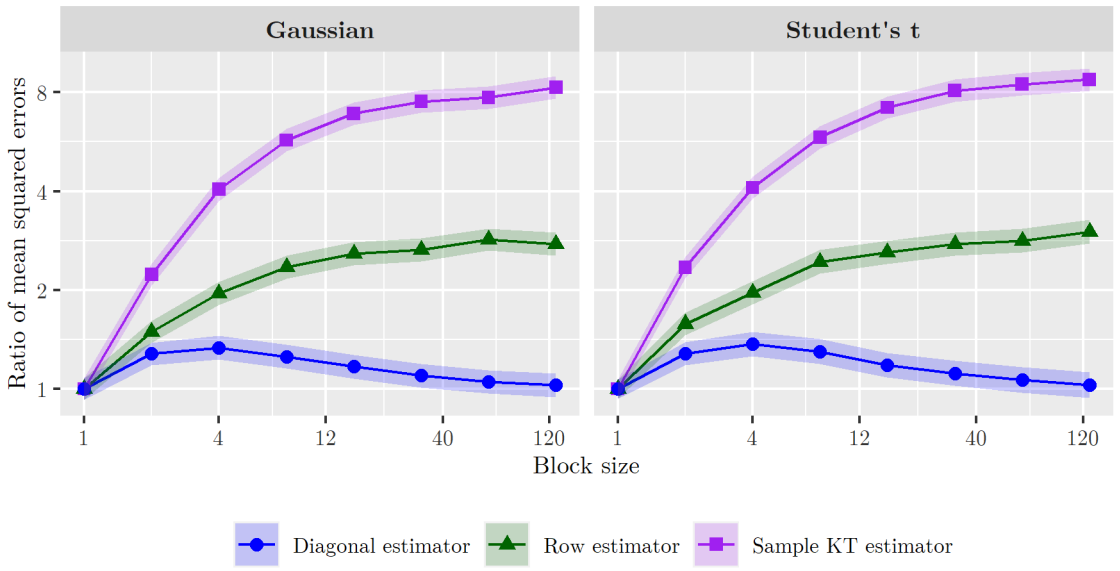


Figure 4: Log-log plots of the ratio $\mathrm{MSE}(\widehat{\tau})/\mathrm{MSE}(\widehat{\tau}^B)$ as a function of the block size including 95% confidence intervals, calculated using a sample size of 4. The diagonal block Kendall's taus are set at 0.3 and the off-diagonal block values at 0.1.

On Figure 4, we see that for both the Gaussian and Student's t distributions the averaging estimators seem to improve on the sample Kendall's tau estimator by roughly the same order of magnitude. Note, however, that these improvements do theoretically depend on the underlying copula. This result thus indicates that the exact choice of copula makes little difference to the improvement on the sample Kendall's tau matrix estimator when it resembles to some extent the Gaussian and Student's t copulas. Furthermore, we find that the relative difference between the diagonal and block estimator is largest for small dimensions, but here there they are still well within a factor of 1.5 of each other. As the dimensions increase, the MSEs of the diagonal estimator converge rapidly to that of the block estimator, again confirming that the block and diagonal estimators have similar variances for large block dimensions.

Next, let us investigate the estimators' MSEs under the less realistic target values of -0.5 in the off-diagonal block and 0.5 in the diagonal blocks. Similarly, we set the sample size to 4 and examine data generated from both the Gaussian and Student's t distributions. The MSEs and 95% confidence intervals are calculated using 4000 replications. See Figure 5 for a plot of the MSEs as a function of the block size and see Figure 6 for the corresponding ratio plot.

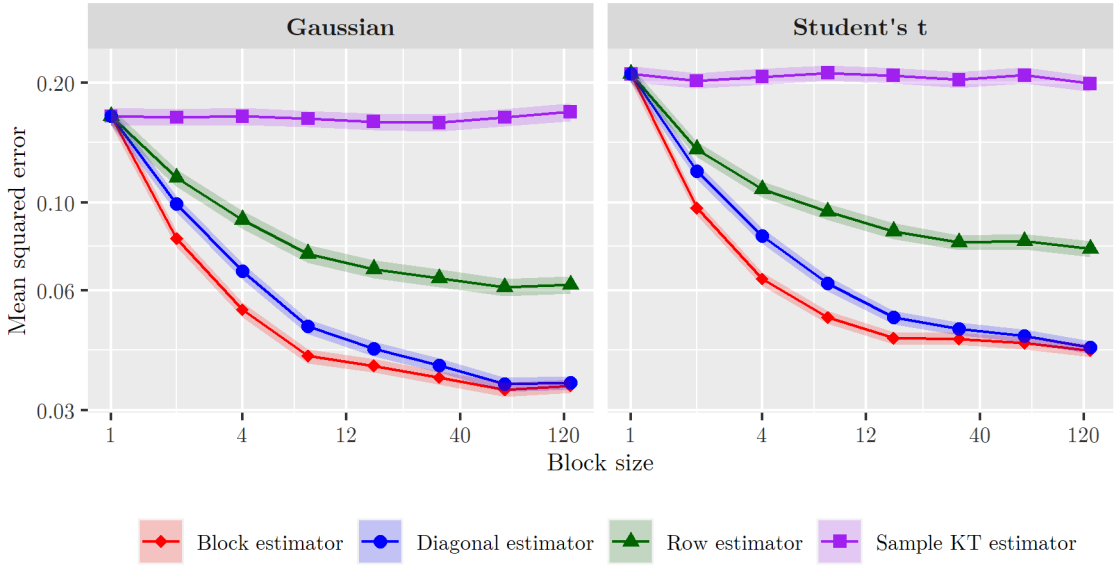


Figure 5: Log-log plots of the estimators' mean squared error as a function of the block size including 95% confidence intervals, calculated using a sample size of 4. The diagonal block target Kendall's taus are set at 0.5 and the off-diagonal block values at -0.5.

On Figure 5, we observe analogous results to the more realistic target values: all estimators are improvements over the naive estimator, the block and diagonal estimators have the best limiting variance, with the block estimator having the fastest convergence. Also, the MSEs generated from the student's t distribution are slightly higher than for the Gaussian distribution, but the order of magnitudes with which the averaging estimators improve are comparable again. It is however noticeable that these order of magnitudes are considerably lower when compared to the setting with the more realistic target Kendall's tau values, i.e. roughly 4 versus 8 for the block and diagonal estimators. This comes as no surprise; the diagonal block Kendall's tau are higher, reducing the degree of independence of the pairwise estimates of the off-diagonal block's, and thus reducing the effect of averaging.

For comparison of the computation efficiency, we perform 1000 replications with the same parameters as before. In Figure 7 the averages of the computation times can be found on log-log scale. As expected, we observe that both the sample Kendall's tau matrix estimator and the block estimator scale quadratically with block size and that the row and diagonal estimators both scale linearly with block size. Consequently, for larger block dimensions one may prefer the diagonal estimator over the block estimator to gain substantial computational efficiency and lose only little precision.

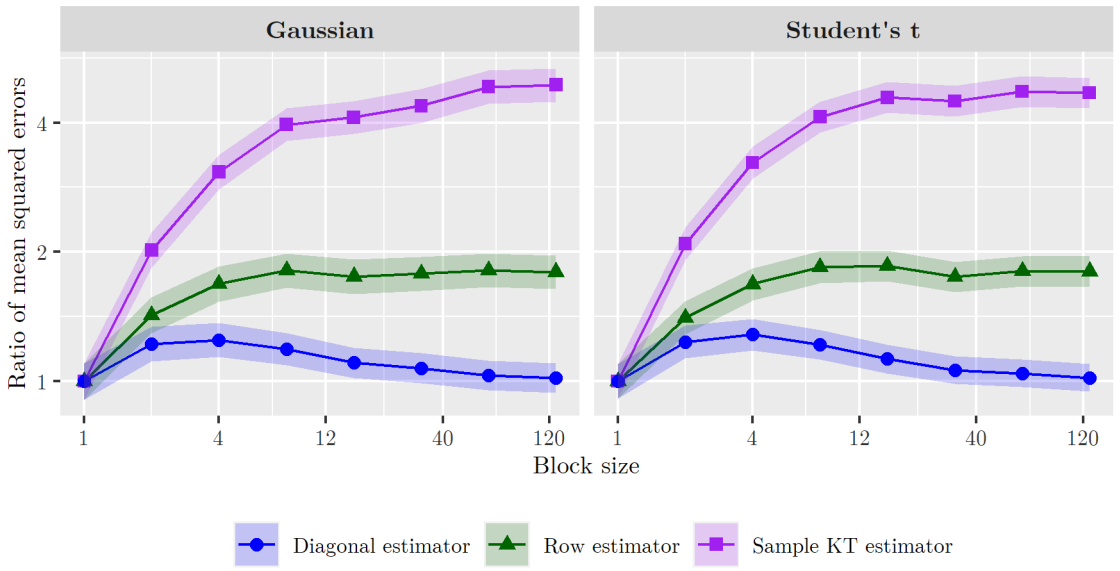


Figure 6: Log-log plots of the ratio $\mathrm{MSE}(\widehat{\tau})/\mathrm{MSE}(\widehat{\tau}^B)$ as a function of the block size including 95% confidence intervals, calculated a sample size of 4. The diagonal block Kendall's taus are set at 0.5 and the off-diagonal block values at -0.5.

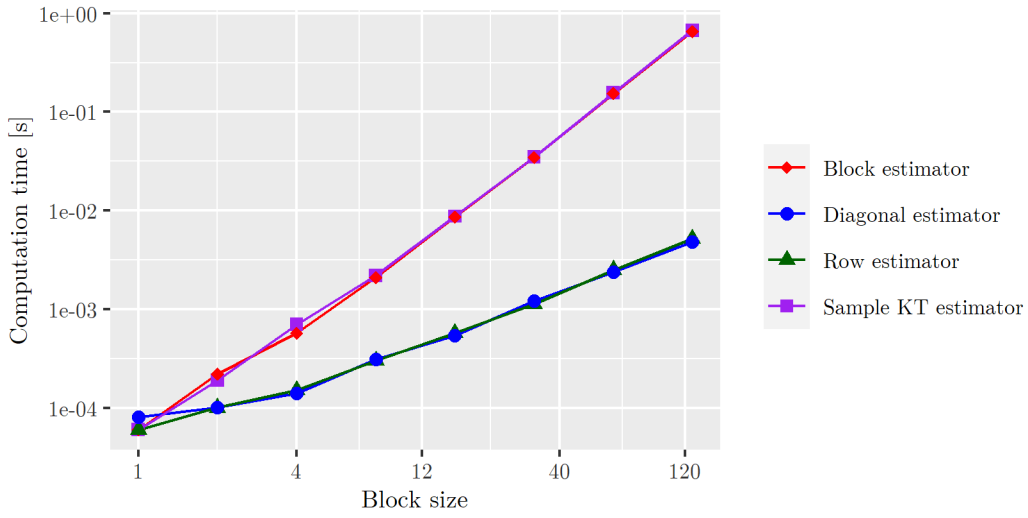


Figure 7: Log-log plot of the estimators' mean computation time [s] as a function of the block size, calculated using a sample size of 4.

5.2 Conditional Kendall's Tau

In this section, we study the conditional versions of the block and diagonal estimators. Since the estimators make use of kernel regression, a larger sample size is needed for obtaining stable results. We therefore consider only a one-dimensional covariate Z, so that we do not need to increase the sample size even further and can run a sufficient number of replications. Kernel estimation is carried out on the basis of the Epanechnikov kernel and simulations are performed with the help of the R package `CondCopulas` [5].

In each of the experiments, we let the covariate Z be uniformly distributed on the interval $[0, 1]$. We will estimate conditional Kendall's taus for points z ranging from 0 to 1 in steps of 0.1. We generate data with the Gaussian distribution, as the student's t distribution yields similar results. All variables will have a mean of Z and variance of $1 + Z^2$. The Kendall's tau matrix is again block-structured corresponding to two groups of equal size. Similarly to the unconditional case, we focus on the estimates of the single off-diagonal block. We set all Kendall's taus within the diagonal blocks to a constant value of 0.3, which is independent of Z. Finally, we let the Kendall's taus within the off-diagonal block depend on the covariate Z.

In Section 5.2.1 and Section 5.2.2, we examine the accuracy and computational efficiency of the estimates under varying sample size and block dimensions. To this end, we set Kendall's tau in the off-diagonal blocks to $0.1\,Z$. As Z is distributed on $[0, 1]$, the conditional Kendall's taus range from 0 to 0.1. As such, the underlying variables are again partially exchangeable conditional on any $z \in [0, 1]$. It follows that the biases of the pairwise estimates in the off-diagonal block are all equal and thus that averaging over them does not change the total bias. Furthermore, note that this property is in fact guaranteed by the structural assumption. Since therefore all estimators have equal biases, we focus on the sample variances instead of the MSEs for a comparison of accuracies.

Then, in Section 5.2.3 we study optimal bandwidths where we vary the way in which the off-diagonal block Kendall's taus depend on Z. We consider a model in which we let the off-diagonal block conditional Kendall's taus be given by

$$\left[\mathbf{T}_{|Z=z}\right]_{\mathcal{B}_{1,2}} = 0.1(\cos(0.5\pi\omega z) + 1)\mathbf{1},$$

with frequencies ω in $\{1, 2, 3, 4\}$. As such, these conditional Kendall's taus range from 0 until 0.2. For comparing the accuracies under varying bandwidths, we study mean integrated squared errors (MISE) calculated by averaging the MSEs of conditional estimates computed at conditioning points ranging from 0 to 1 in steps of 0.1.

5.2.1 Effect of the Sample Size

In this experiment, we study the dependency of the variances on the sample size. To this end, we vary the sample size under a fixed block size of 4 and a bandwidth of 0.5. We use this relatively large bandwidth to ensure stable results even at lower sample sizes. The sample variances and 95% confidence intervals are calculated using 8000 replications. For each grid point z we plot the resulting sample variances on log-log scale in Figure 10.

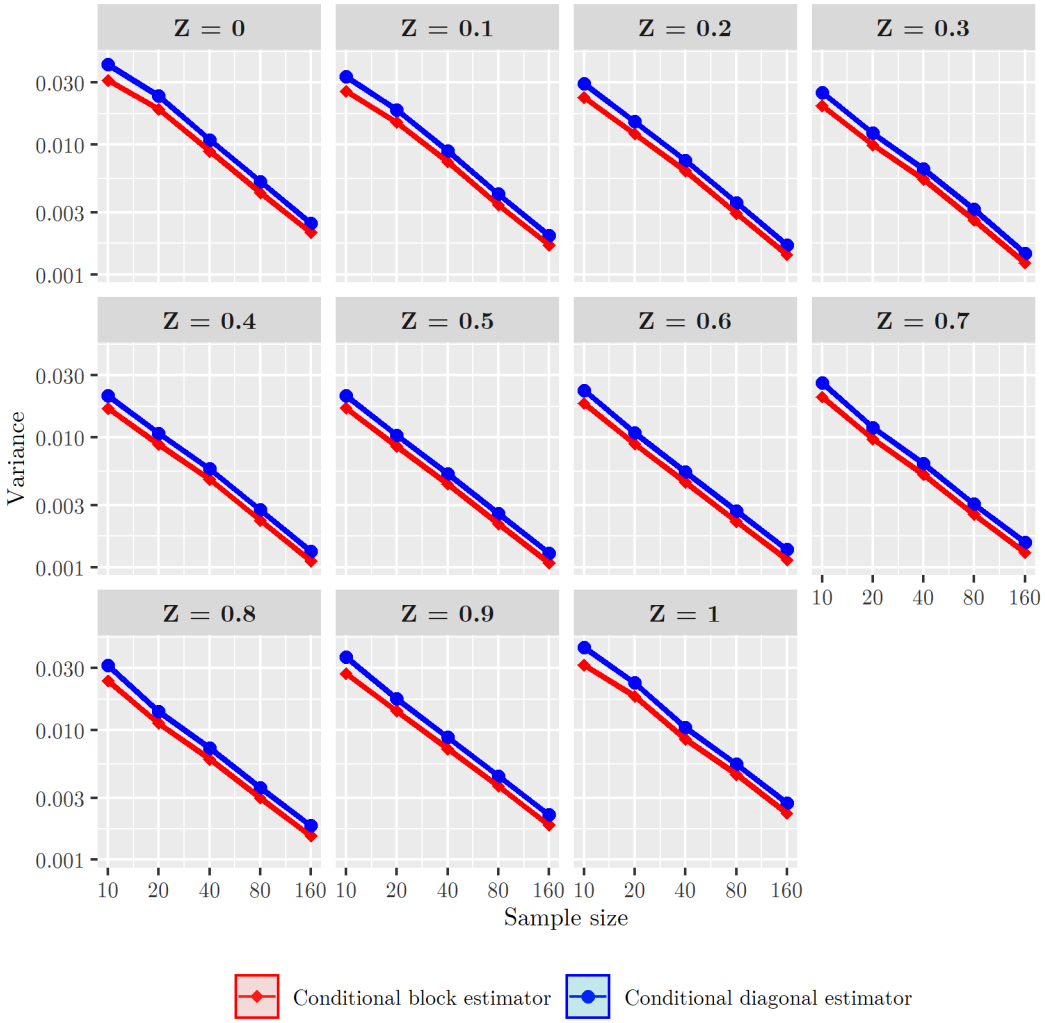


Figure 8: Log-log plots of the estimators' variances as a function of the sample size on several conditioning points including 95% confidence intervals, calculated using a block size of 4 and a bandwidth of 0.5.

Unsurprisingly, the conditional variances are also inversely related to the sample size. It follows that if bandwidths are kept constant, MSEs converge to the bias. As such, appropriate band-

widths are naturally smaller for increasing sample sizes. Furthermore, it is seen that the estimates near the edges of the interval $[0, 1]$ are less accurate than those in the middle. This can be attributed to the fact that there are fewer observations of Z near grid points close to the edges than near grid points in the middle, since the observations can be found there on both sides. Evidently, a change in the distribution of Z also changes the level of the variances.

Next, let us study the dependency of the computation time on the sample size. We leave the setting unchanged, though the results correspond to the calculation of the conditional block estimates on a single grid point. The results are calculated using 500 replications and are represented on log-log scale in Figure 9. Here it is seen that the computation times gradually increase with the sample sample to a point where they appear to scale quadratically with each other. This behaviour follows from the fact that the conditional estimates require the calculation of a double sum of n terms. Note that the computation times of the diagonal and block estimators are relatively close since only a block size of 4 is used here.

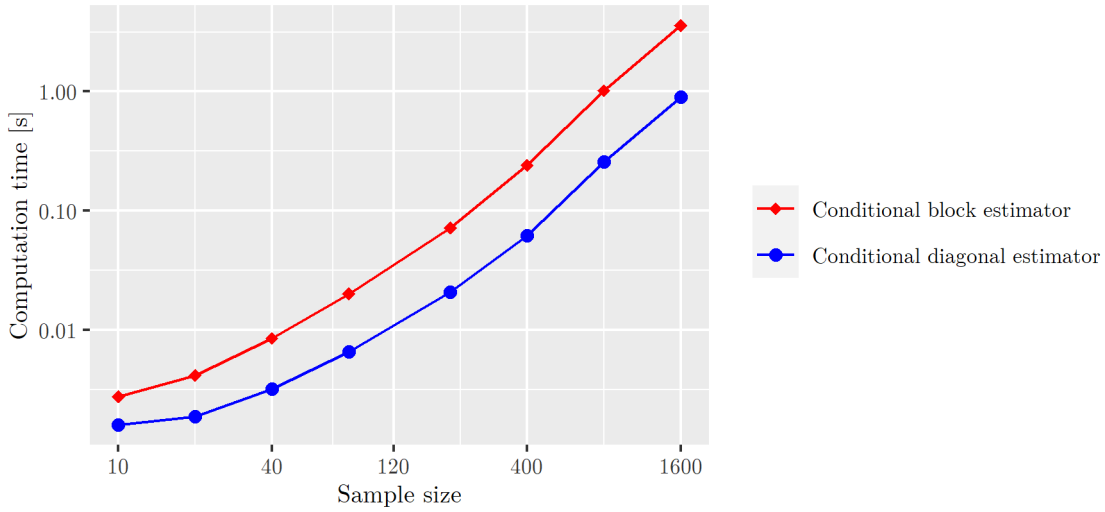


Figure 9: Log-log plot of the mean computation time [s] versus the sample size, calculated using a block size of 4.

5.2.2 Effect of the Block Size

We first study the estimators' variance under varying block dimensions. In order to run a sufficient number of replications we set the sample size to 20 and consequently the bandwidth to 0.5. The sample variances and 95% confidence intervals are calculated using 30000 replications. For each grid point z, the resulting sample variances are displayed on log-log scale in Figure 10.

From the figure we observe that the estimators' variances behave similarly to the unconditional setting under varying block dimensions, for each of the grid points. That is, both estimators

are improvements over the naive estimator, both limiting variances are identical, and the block estimator converges slightly faster than the diagonal estimator. It further follows that since averaging reduces variance, it also reduces the optimal bandwidth. This will be studied in more detail in Section 5.2.3. Again, as there are fewer observations of Z near grid points close to the edges of $[0, 1]$, the variance levels vary slightly over the different grid points.

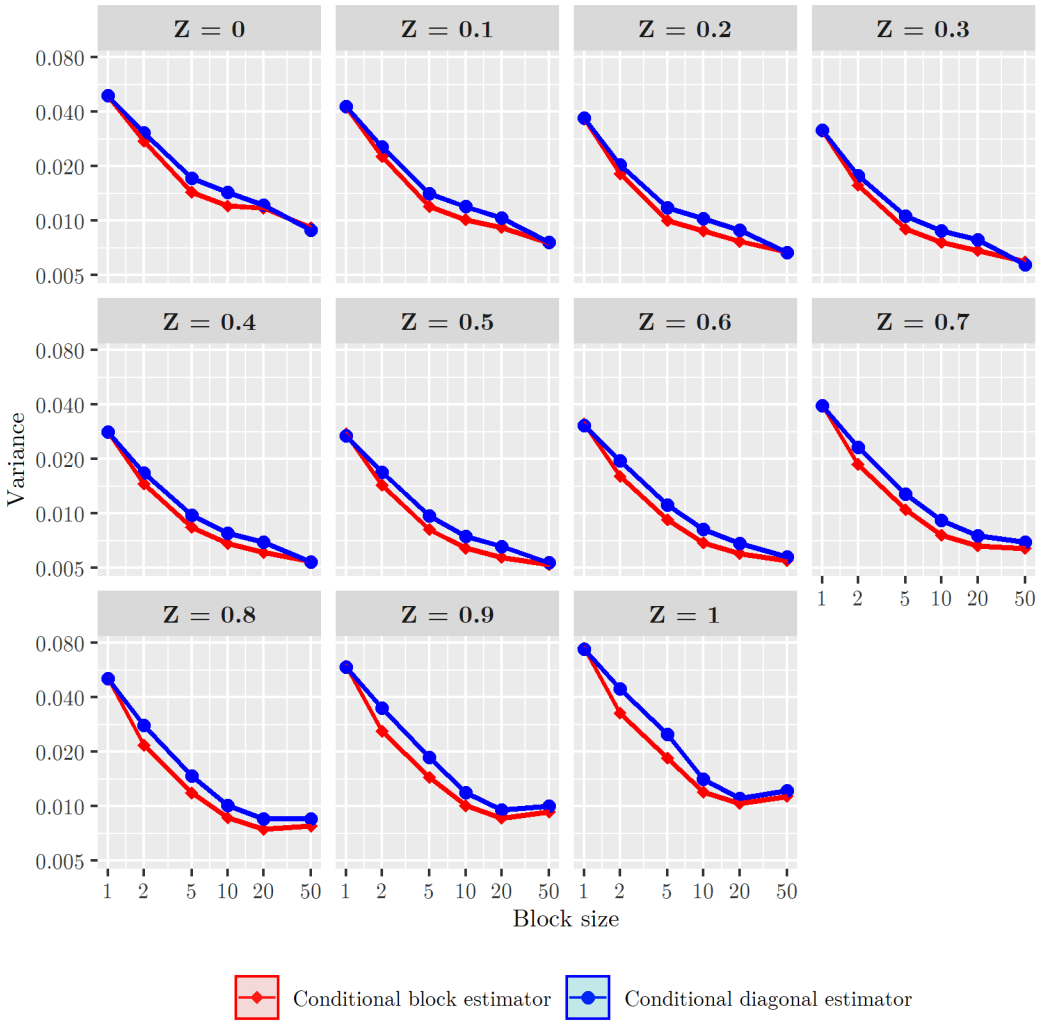


Figure 10: Log-log plots of the estimators' variances as a function of the sample size on several conditioning points including 95% confidence intervals, calculated using a sample size of 20 and a bandwidth of 0.5.

As for the computation times, there is clearly no fundamental change in how these depend on the block size when compared to the unconditional setting. However, since the conditional estimators are kernel-based, it should be noted that they generally require more time than their unconditional counterparts, as was also seen in Figure 9. For the sake of completeness, we still

include a plot of the average computation time against the block size, see Figure 11. The results correspond to estimating the off-diagonal block conditional Kendall's taus simultaneously on the 11 grid points, and follow from 10000 replications with a sample size of 150. As expected, the block estimator scales quadratically with block size, while the diagonal estimator scales linearly with block size. Therefore, as in the unconditional case, one may prefer the diagonal estimator over the block estimator to gain substantial computational efficiency and lose only little precision.

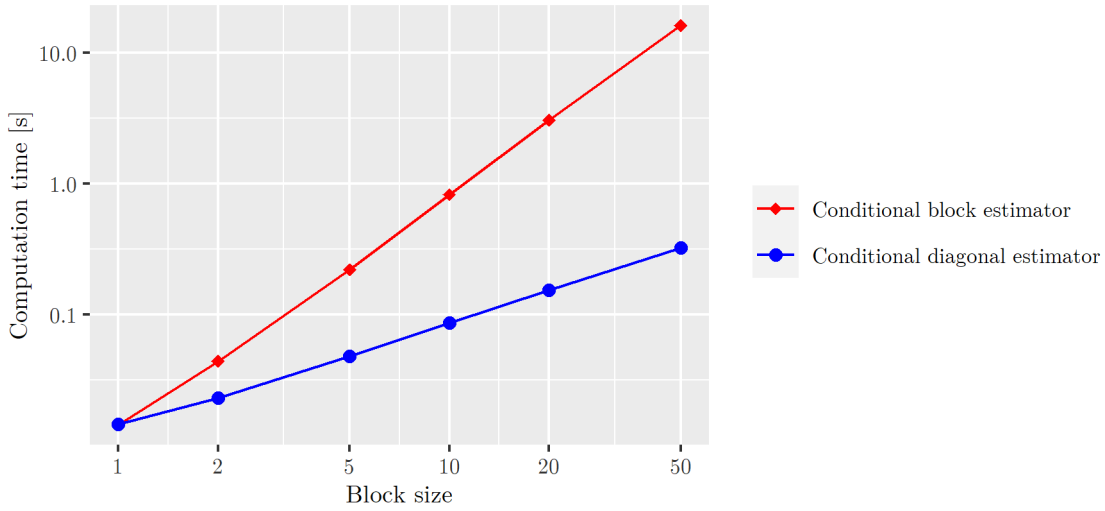


Figure 11: Log-log plot of the estimators' mean computation time [s] as a function of the block size, calculated using a sample size of 150.

5.2.3 Bandwidth Selection

Let us compare the estimators' MISEs for different bandwidths. In this experiment, we set the diagonal block Kendall's taus to 0.3 and the off-diagonal block Kendall's taus conditionally at $Z = z$ to

$$0.1(\cos(0.5\,\pi\,\omega\,z) + 1),$$

with frequencies $\omega \in \{1, 2, 3, 4\}$. The block size is fixed at 8 and the sample size at 200. The MISEs and 95% confidence intervals are calculated using 100 replications, see Figure 12.

The figure confirms that indeed the averaging estimators have smaller optimal bandwidths than the naive estimator. It should be noted that only a block size of 8 is used here, and that the optimal bandwidth decreases with block size until the limit values are reached. Furthermore, the figure shows that as the frequency increases, the optimal bandwidth is reduced. This is fully consistent with kernel regression theory: increasing the frequency increases the difference in Kendall's tau values conditionally on adjacent points of z, and hereby we generally need to pick a smaller bandwidth. Lastly, it should be noted that as the bandwidth increases the effect of

averaging is less and less visible. This can be attributed to the fact that by increasing the bandwidth, the variance term within the MISE becomes less and less prominent, while the bias term generally increases.

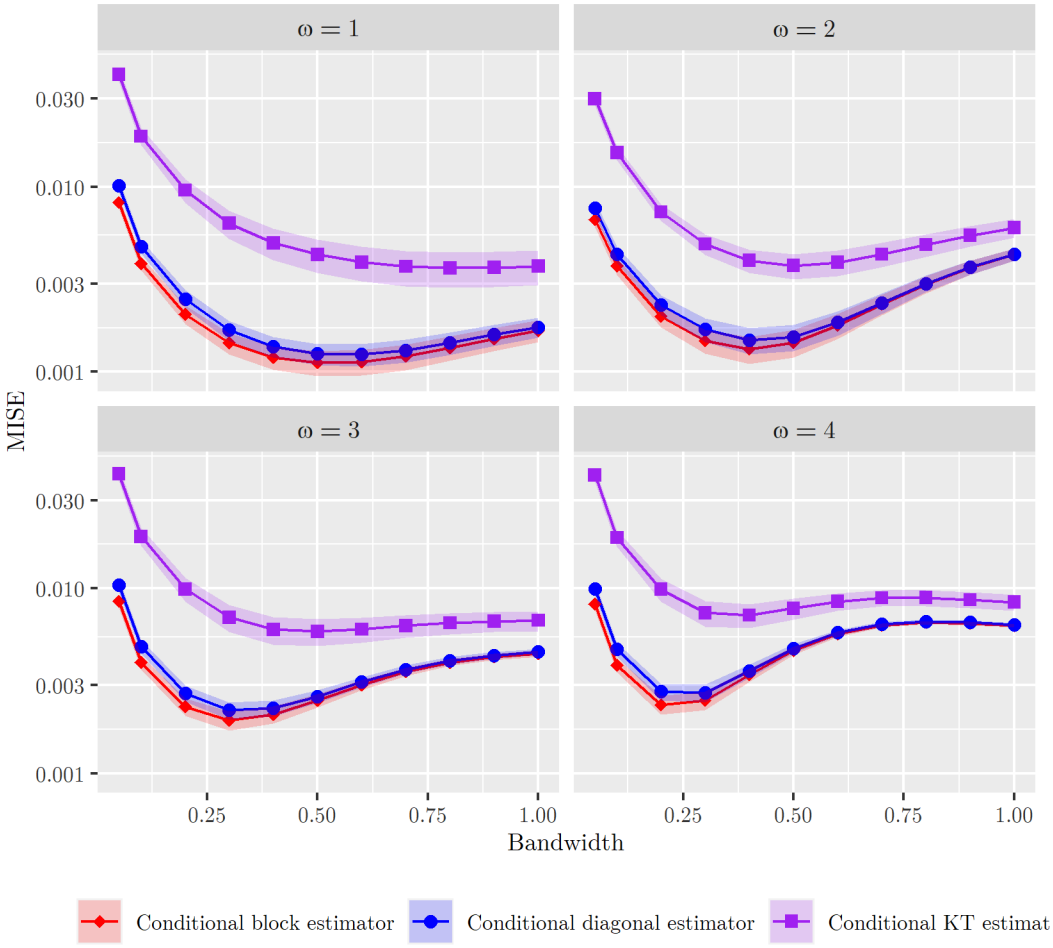


Figure 12: Log-plots of the estimators' MISEs as a function of the bandwidth for different frequencies ω including 95% confidence intervals, calculated using a sample size of 200 and a block size of 8.

6 Application to Real Data

In this section, we study the behaviour of the estimators under real data conditions and provide value at risk (VaR) computations of a large stock portfolio as an example of possible applications. In Section 6.1, we discuss the necessary background on VaR calculations for elliptical distributions Then in Section 6.2, we describe the methods used to estimate the VaR input parameters. The results are presented in Section 6.3, where backtesting is applied to asses the viability. All computations have been done using the `R` statistical environment [43].

6.1 VaR for Elliptical Distributions

The value at risk is a widely used risk measure in a variety of financial fields, ranging from auditing and financial reporting to risk management and the calculation of regulatory capital [35]. It is used to quantify potential losses over a specific time frame of some financial entity or portfolio of assets. We only discuss the theory necessary for calculating the value at risk for elliptical distrbutions. For a more comprehensive theory on the value at risk, we refer to [45]. We will follow the approach of [42, 45], in which explicit expressions for the value at risk of elliptical distributions was derived.

Let X be a loss function (with negative losses and positive profits), then we define the VaR at level $\alpha \in (0,1)$ as the smallest number x such that the probability that X does not exceed x is at least $1-\alpha$. More formally,

$$\mathrm{VaR}_\alpha(X) := -\inf\left\{x \in \mathbb{R} : F_X(x) > \alpha\right\},$$

or equivalently by setting $Y := -X$,

$$\mathrm{VaR}_\alpha(X) = F_Y^{-1}(1-\alpha). \tag{45}$$

To calculate the VaR of a given portfolio of assets, it is often assumed that the portfolio's profits and losses are a linear function of the returns of the individual constituents. More formally, a portfolio with value $\Pi(t)$ at time t is called linear if its profit and loss $\Delta\Pi(t) = \Pi(t) - \Pi(0)$ over a time window $[0, \mathrm{t}]$ is a linear function of the returns $X_1(t), \ldots, X_p(t)$:

$$\Delta\Pi(t) = \delta_1 X_1(t) + \delta_2 X_2(t) + \ldots + \delta_p X_p(t).$$

This clearly applies to any common stock portfolio by using the ordinary returns of the individual shares and when considering the log returns, this holds to a good approximation provided that the time window $[0,t]$ is small, e.g. for daily log returns. The time window t will be kept constant

and will therefore be omitted from future notations.

Furthermore, we will assume that the X_j are elliptically distributed with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$ with Cholesky decomposition $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$ and density generator g:

$$(X_1 \dots, X_p) \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g).$$

Thus, the probability density $f_{\mathbf{X}}$ of $\mathbf{X} = (X_1, \dots, X_p)$ is given by

$$f_{\mathbf{X}}(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2} g\left((\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right).$$

When considering elliptically distributed risk factors, we cannot simply use the well-known Delta-Normal approach to calculate the VaR, as it relies on the stronger assumption of Gaussianity. A generalisation of the Delta-Normal method was derived for the class of elliptical distributions in [45].

Let us start by noting that the VaR of the portfolio profits and losses $\Delta\Pi(t)$ as given in (45) can be rewritten as

$$\mathbb{P}\left(\Delta\Pi < -\operatorname{VaR}_\alpha\right) = \alpha.$$

Then, given the linearity of the portfolio and the fact that $\mathbf{X}$ follows an elliptical distribution, the VaR is obtained by solving the following equation:

$$\alpha = |\boldsymbol{\Sigma}|^{-1/2} \int_{\{\boldsymbol{\delta}\mathbf{x}^T \leq -\operatorname{VaR}_\alpha\}} g\left((\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})^T\right) \mathrm{d}\mathbf{x},$$

where $\boldsymbol{\delta}$ denotes the vector of weights $(\delta_1, \dots, \delta_p)$. We change variables to $\mathbf{y} = (\mathbf{x} - \mu)\mathbf{A}^{-1}$, $\mathrm{d}\mathbf{y} = |\mathbf{A}|\mathrm{d}\mathbf{x}$ and obtain

$$\alpha = \int_{\{\boldsymbol{\delta}\mathbf{A}\mathbf{y}^T \leq -\boldsymbol{\delta}\boldsymbol{\mu}^T - \operatorname{VaR}_\alpha\}} g\left(|\mathbf{y}|^2\right) \mathrm{d}\mathbf{y}.$$

Now let $\mathbf{R}$ be the rotation matrix such that $(|\boldsymbol{\delta}\mathbf{A}|, 0, \dots, 0)\mathbf{R} = \boldsymbol{\delta}\mathbf{A}$, then by changing to $\mathbf{y} = \mathbf{z}\mathbf{R}$ we get

$$\alpha = \int_{\{|\boldsymbol{\delta}\mathbf{A}| z_1 \leq -\boldsymbol{\delta}\boldsymbol{\mu}^T - \operatorname{VaR}_\alpha\}} g\left(|\mathbf{z}|^2\right) \mathrm{d}\mathbf{z}.$$

By writing $|\mathbf{z}|^2 = z_1^2 + |\mathbf{z}'|^2$ with $\mathbf{z}' \in \mathbb{R}^{p-1}$ we have

$$\alpha = \int_{\mathbb{R}^{p-1}} \int_{-\infty}^{\frac{-\boldsymbol{\delta}\boldsymbol{\mu}^T - \operatorname{VaR}_\alpha}{|\boldsymbol{\delta}\mathbf{A}|}} g\left(z_1^2 + |\mathbf{z}'|^2\right) \mathrm{d}z_1 \mathrm{d}\mathbf{z}'.$$

Lastly, we change to a hyperspherical coordinate system $\mathbf{z}' = r\boldsymbol{\xi}$ with $\boldsymbol{\xi} \in S_{p-2}$, where S_{p-2} is the

unit sphere in $\mathbb{R}^{p-1}$. This gives

$$\alpha = |S_{p-2}| \int_0^\infty r^{p-2} \int_{-\infty}^{\frac{-\boldsymbol{\delta\mu}^T - \mathrm{VaR}_\alpha}{|\boldsymbol{\delta}\mathbf{A}|}} g\left(z_1^2 + r^2\right) \mathrm{d}z_1 \mathrm{d}r, \tag{46}$$

where

$$|S_{p-2}| = \frac{2\pi^{\frac{p-1}{2}}}{\Gamma(\frac{p-1}{2})}.$$

Let us now introduce the function

$$\begin{aligned} G(s) &= \frac{2\pi^{\frac{p-1}{2}}}{\Gamma(\frac{p-1}{2})} \int_{-\infty}^{-s} \int_0^\infty r^{n-2} g(z_1^2 + r^2) \mathrm{d}r \mathrm{d}z_1 \\ &= \frac{\pi^{\frac{p-1}{2}}}{\Gamma(\frac{p-1}{2})} \int_s^\infty \int_z^\infty (u - z^2)^{\frac{p-3}{2}} g(u) \mathrm{d}u \mathrm{d}z, \end{aligned} \tag{47}$$

where we have changed variables to $u = r^2 + z_1^2$ and $z = -z_1$. Let us denote by $q^g_{\alpha,p}$ the unique solution of the so-called transcendental equation

$$\alpha = G(q^g_{\alpha,p}). \tag{48}$$

It then finally follows from expressions (46) and (47) that the Delta-Elliptic VaR is given by

$$\begin{aligned} \mathrm{VaR}_\alpha &= -\boldsymbol{\delta\mu}^T + q^g_{\alpha,p} |\boldsymbol{\delta}\mathbf{A}| \\ &= -\boldsymbol{\delta\mu}^T + q^g_{\alpha,p} \sqrt{\boldsymbol{\delta\Sigma\delta}^T}. \end{aligned} \tag{49}$$

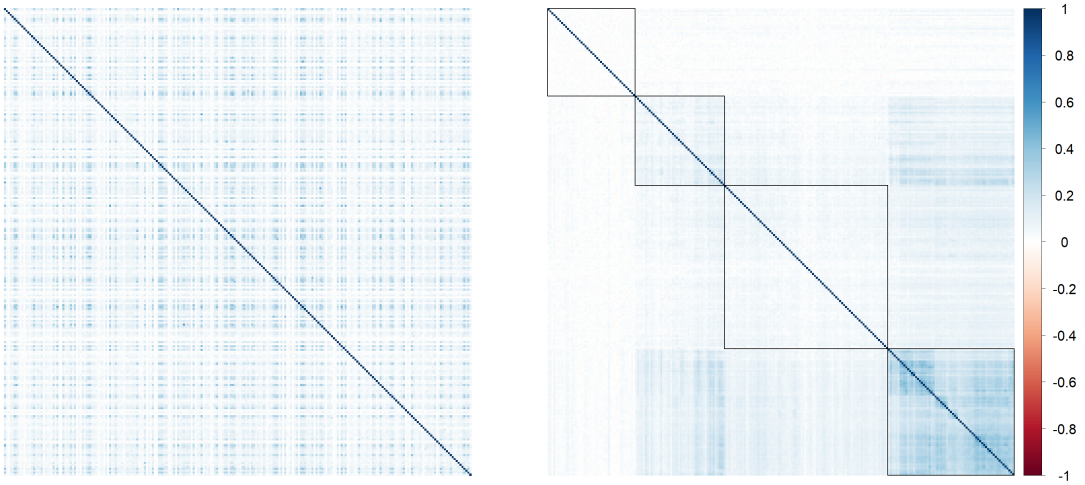
Note that this equation has a clear financial interpretation: the portfolio's average return is given by $\boldsymbol{\delta\mu}^T$ and the portfolio's standard deviation by $\sqrt{\boldsymbol{\delta\Sigma\delta}^T}$. Further note that the result is analogous to that of the Delta-Normal VaR, in which we simply replace $q^g_{\alpha,p}$ with the $1-\alpha$ quantile of the standard-normal distribution.

6.2 Estimation Procedure

In order to test the estimators in real data conditions, we consider a portfolio consisting of 240 different stocks. All stocks are listed on the Euronext markets and data is being sourced from Yahoo Finance. The complete list of all shares involved is available in Appendix A. We will estimate the portolio's daily VaR assuming that the price is set at a level of 100 and that all stocks in the portfolio are equally weighted. To this end, we model the daily log returns of the individual stocks, assuming they follow an elliptical distribution.

In order to achieve a proper clustering, we compute the pairwise Kendall's tau matrix over a long time period from 01 January 2007 to 14 January 2022, after which we reorder the variables in

order to obtain the intended block structure. Since we have not proposed a clustering method, we consider several methods for rearranging matrices that are built into the R-package `seriation` [25]. On the basis of visual inspection, we have applied the so-called `GW_Ward` method along with a few manual adjustments. The resulting reordering corresponds to four large groups, which are specified further in Appendix A. See Figure 13 for a heatmap of the pairwise Kendall's tau matrix before and after reordering the variables by group. To indicate the groups, lines have been drawn around the diagonal blocks. It should be noted that, if studied carefully, the large groups can be broken down into smaller and more accurate groups. Nevertheless, these large groups already seem to be quite useful and therefore we will simply use them for our further analysis.



(a) unclustered (b) clustered

Figure 13: Heatmap plots of the sample Kendall's tau matrix computed on the daily log returns from 01 January 2007 until 14 January 2022 of all portfolio stocks.

Based on the groups displayed in Figure 13, the objective is to calculate the VaR at 30 June 2017, leaving sufficient future data for backtesting the results. To this end, we estimate the Kendall's tau matrix of the log returns using the block, row, diagonal and sample Kendall's tau matrix estimators using data points over the period 01 August 2015 to 30 June 2017. The row, diagonal and block estimators are now available as part of the package `ElliptCopulas` [6], and can be computed using the function `KTMatrixEst`. See Appendix B for the written code. It also contains the code for the conditional estimators. To estimate the standard deviations and averages over the same period, we make use of the sample mean and sample standard deviation statistics.

Following the elliptical assumption, we can now obtain covariance matrix estimates from each of the Kendall's tau matrix estimates. Subsequently, we can compute nonparametric estimates of the density generator for each of these inputs. To this end, we make use of the function `EllDistrEst` within the `ElliptCopulas` package [6] which implements Liebscher's proce-

dure [31]. We improved this function so that it also works for high dimensions, invoking the package `Rmpfr` [34] to use multiple precision floating point computations, as a higher degree of accuracy is needed. This is now also contained in the `ElliptCopulas` package (see also the code in Appendix B). For the density generator estimation we require a complete data set with no missing values. As such, the interval on which we estimate the density generator will be chosen as shorter (01 June 2016 to 30 June 2017). The kernel function will be chosen as the Epanechnikov kernel. Further we use Silverman's rule of thumb for bandwidth selection to estimate elliptical density generators [42], which for a sample size of n is given by

$$h = 1.06\sqrt{\mathrm{Var}\left[\widehat{\boldsymbol{\xi}}\right]}\, n^{1/5}, \tag{50}$$

where

$$\widehat{\xi}_i = -1 + \left(1 + ((\mathbf{x}_i - \boldsymbol{\mu})\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}))^{p/2}\right)^{2/p},$$

for $i = 1, \ldots, n$ and $p = 240$. Here, $\mathbf{x}_i$ stands for the vector of log returns at the i'th date, $\widehat{\boldsymbol{\Sigma}}$ stands for one of the covariance matrix estimates and $\boldsymbol{\mu}$ stands for the log returns' sample mean. It should be noted that this is merely a rule of thumb and can be further optimised, see [26] for further details. Clearly, by using this bandwidth selection method, the use of different Kendall's tau matrix estimators yields different values of the bandwidth. In order to get a better idea of the effects of the bandwidth choice, we also consider several deterministic bandwidth choices, and compare the performance of the estimators for each of them.

Finally, we can numerically solve the transcendental equation as given in (46) to arrive at the corresponding quantiles. As such, we have discussed all ingredients for calculating the VaR as in (49). In order to test the results, we perform backtesting on two intervals, one in the future from 01 July 2017 to 14 January 2022 and one during the period on which the estimations are based, from 01 August 2015 to 30 June 2017.

6.3 Results

We compute the portfolio's 5% and 10% VaR values by following the estimation procedure described in Section 6.2. Table 1 shows the quantile estimates obtained by solving the transcendental equation for each of the different density generator estimates. The density generators were calculated using each of the block, row, diagonal and naive Kendall's tau matrix estimators and using varying values of the bandwidth.

The table shows that the averaging estimators yield very similar quantiles which are all relatively constant for different choices of the bandwidth. In contrast, the quantiles of the naive estimator lie substantially higher and vary significantly for the different bandwidths. In that sense, the estimates obtained with the averaging estimators seem to be much more stable. Moreover,

the Silverman's bandwidths of the averaging estimators are also all very similar, while that of the naive estimator is again considerably larger.

Table 1: Estimated quantiles corresponding to the 5% and 10% VaRs calculated by estimating an elliptical distribution for the daily log returns using each of the different Kendall's tau matrix estimates and several values of the bandwidth.

Quantiles $q_{\alpha,p}^{\widehat{g_h}}$		Estimated			
α	Estimator	$h = 20$	$h = 40$	$h = 100$	Silverman's h
5%	Naive	2.11	1.94	1.98	2.12 ($h = 586.8$)
	Block	1.60	1.60	1.60	1.60 ($h = 40.8$)
	Row	1.60	1.60	1.60	1.60 ($h = 41.1$)
	Diagonal	1.59	1.59	1.60	1.59 ($h = 40.5$)
10%	Naive	1.48	1.38	1.40	1.53 ($h = 586.8$)
	Block	1.23	1.23	1.23	1.23 ($h = 40.8$)
	Row	1.24	1.24	1.23	1.24 ($h = 41.1$)
	Diagonal	1.23	1.23	1.23	1.23 ($h = 40.5$)

Table 2 shows the VaR estimates for each of the different estimators and bandwidths, and also the backtested VaR values. As discussed in Section 6.2, backtests were conducted at two intervals, interval 1 refers to the upcoming interval from 01 July 2017 until 14 January 2022, and interval 2 refers to the interval on which the estimation is based, from 01 August 2015 until 30 June 2017.

Table 2: Estimated 5% and 10% VaRs including the corresponding backtesting results on two intervals. Interval 1 corresponds to 01 July 2017 until 14 January 2022 and interval 2 to 01 August 2015 until 30 June 2017.

VaR		Estimated				Backtested	
α	Estimator	$h = 20$	$h = 40$	$h = 100$	Silverman's h	Interval 1	Interval 2
5%	Naive	1.647	1.512	1.544	1.655 ($h = 586.8$)	1.392	1.262
	Block	1.320	1.320	1.320	1.320 ($h = 40.8$)		
	Row	1.332	1.332	1.332	1.332 ($h = 41.1$)		
	Diagonal	1.284	1.284	1.292	1.284 ($h = 40.5$)		
10%	Naive	1.147	1.083	1.068	1.187 ($h = 586.8$)	0.861	0.839
	Block	1.008	1.008	1.008	1.008 ($h = 40.8$)		
	Row	1.026	1.017	1.026	1.026 ($h = 41.1$)		
	Diagonal	0.987	0.987	0.987	0.987 ($h = 40.5$)		

This clearly shows that the averaging estimators have performed significantly better than the naive estimator when compared to both backtesting intervals. For both α-levels, it can be seen that the VaRs generated using the naive estimator are considerably larger than those using the averaging estimators, which themselves produce relatively similar values. Furthermore, it can be seen that the 5% VaRs of the averaging estimators agree fairly well with the results of the

backtesting, unlike those of the naive estimator. However, the 10% VaR estimates are not as accurate and all estimators yield considerably higher VaRs than those obtained by backtesting. This could indicate that the log returns are not elliptically distributed, or that the interval at which we generate the density generator is too short. Recall that the interval on which we estimate the density generator is merely from 01 June 2016 until 30 June 2017.

To get a better understanding of how well the VaR estimates correspond with the backtesting results, we examine how often the estimates are exceeded by the portfolio's losses in each of the backtesting periods. Table 3 and 4 show the number of exceedances in interval 1 and interval 2 respectively.

Table 3: The number of exceedances of the estimated 5% and 10% VaRs during backtesting interval 1, from 01 July 2017 until 14 January 2022.

# Exceedances		Estimated				Backtested
α	Estimator	$h = 20$	$h = 40$	$h = 100$	Silverman's h	Interval 1
5%	Naive	47	53	53	46 ($h = 586.8$)	58
	Block	58	58	58	58 ($h = 40.8$)	
	Row	58	58	58	58 ($h = 41.1$)	
	Diagonal	61	61	59	61 ($h = 40.5$)	
10%	Naive	76	85	83	72 ($h = 586.8$)	116
	Block	94	94	94	94 ($h = 40.8$)	
	Row	91	91	92	91 ($h = 41.1$)	
	Diagonal	100	100	100	100 ($h = 40.5$)	

Table 4: The number of exceedances of the estimated 5% and 10% VaRs during backtesting interval 2, from 01 August 2015 until 30 June 2017.

# Exceedances		Estimated				Backtested
α	Estimator	$h = 20$	$h = 40$	$h = 100$	Silverman's h	Interval 2
5%	Naive	9	12	12	9 ($h = 586.8$)	25
	Block	20	20	20	20 ($h = 40.8$)	
	Row	20	20	20	20 ($h = 41.1$)	
	Diagonal	22	22	21	22 ($h = 40.5$)	
10%	Naive	30	32	32	30 ($h = 586.8$)	49
	Block	35	35	35	35 ($h = 40.8$)	
	Row	35	35	35	35 ($h = 41.1$)	
	Diagonal	35	35	35	35 ($h = 40.5$)	

Both tables show that the difference between the theoretical and the observed number of exceedances is much larger when using the naive sample Kendall's tau matrix estimator than when using any of the averaging estimators and this applies to both α-levels as well as to all band-

widths. As such, the averaging estimators are overall significantly better performers than the naive estimator. In addition, although there are subtle differences in the performance of the block, row and diagonal estimators, there is no clear winner in this example. This shows that computing all Kendall's tau using the block estimators incur no additional benefits compared to using only the row or diagonal estimators, that are computationally much cheaper.

7 Conclusion

In this thesis, we have provided an alternative approach to the generally challenging task of estimating Kendall's tau and conditional Kendall's tau matrices in high-dimensional settings. By imposing structural assumptions on the underlying (conditional) Kendall's tau matrix, we have introduced new estimators that have significantly reduced computational costs without much loss in performance.

For the unconditional case, a model was studied in which the set of variables could be grouped in such a way that the Kendall's taus of variables from different groups depends only on the group numbers. After reordering the variables by group, the underlying Kendall's tau matrix is then block-structured with constant values in the off-diagonal blocks. We have proposed several (unbiased) estimators that take advantage of this block-structure by averaging over the usual pairwise Kendall's tau estimates in each of the off-diagonal blocks: the block estimator averages over all pairwise estimates, whereas the row, the diagonal and the random estimators only average over part of the off-diagonal blocks (respectively, over the pairs on the first row, on the first diagonal and over a random selection of pairs). This makes them computationally cheaper than the usual sample Kendall's tau matrix estimator and the block estimator, for which all pairwise estimates have to be computed.

We have formally derived variance expressions, which show not only that all estimators are improvements over the usual sample Kendall's tau matrix estimator, but also, interestingly, that the asymptotic variances do not depend on the block dimensions. Furthermore, we have seen that the block, the diagonal and the random estimators have very similar asymptotic variances, whereas that of the row estimator was different. The former depend on the auxiliary quantity U, while the latter depends on S. In each example that has been studied, we saw that the U-asymptotic variances were lower than the S-asymptotic variances, but a formal characterisation of the set of copulas to which this applies is left for future work. Nevertheless, we recommend to avoid using the row estimator. Under light assumptions, we have shown that U-asymptotic variances are equal, and that it is approached fastest by the block estimator, followed by the diagonal estimator and then the random estimator. Hence, if the computational costs were to be reduced, the diagonal estimator is preferable to both the random and the row estimator.

Furthermore, a model was studied in which the Kendall's taus depend on a conditioning variable. Here it was assumed that the conditional Kendall's tau matrix has the above-mentioned block structure and, moreover, that it is preserved under fluctuations of the conditioning variable. We have adopted nonparametric, kernel-based estimates of the conditional Kendall's tau in order to construct the conditional versions of the block, row, diagonal and random estimators. Under some additional regularity assumptions, we have shown that the estimators are all asymptotically

normal conditionally to different values of the covariate. Following from these expressions, we have seen that the asymptotic variances have analogous expressions to their unconditional counterparts. As such, all estimators are again improvements over the naive estimator, with the block estimator having the best performance. Similarly, if computational costs were to be reduced, the diagonal estimator is preferable to both the random and the row estimator. Moreover, the reduction of computing costs becomes particularly relevant in the conditional setting, as the use of kernel smoothing introduces additional complexity.

We have performed a simulation study in order to support the theoretical findings. In the unconditional setting, simulations were performed with the Gaussian and Student's t distributions. It was indeed found that the averaging estimators yield significant improvements over the naive Kendall's tau matrix estimator. These improvements were similar for both the Gaussian and Student's t distributions. Such a result suggests that the exact form of the underlying data distribution has little influence. It was furthermore confirmed that the diagonal and the block estimator indeed have the lowest asymptotic variance, with the block estimator converging the fastest, though closely followed by the diagonal estimator. This emphasises the practical use of the diagonal estimator.

We remarked again that the conditional estimators' variances decrease in a similar fashion for growing block dimensions. The estimation bias associated with the use of kernel smoothing techniques is not affected by the averaging principle. As a consequence, the averaging estimators allow for a reduced optimal bandwidth; this was indeed confirmed in the simulations. This makes the averaging estimators perfectly suited for practical applications, as reducing the bandwidth goes hand in hand with reducing the estimation bias.

Lastly, we have demonstrated the use of the estimators in a real world application. The estimators were used to model the daily log returns of a large stock portfolio consisting of 240 Euronext listed stocks. After restructuring the sample Kendall's tau matrix, computed on a long time interval, the proposed block structure was clearly visible. Building on these groups, robust estimates of the correlation matrix were obtained by assuming that the log returns follow an elliptical distribution. Using each of the estimates, the portfolio's 5% and 10% VaR values were calculated. The results of the averaging estimators were much more stable under changes in the bandwidth used for the estimation of the density generator. Moreover, the averaging VaRs were significant improvements over the naive estimates when compared to the backtesting results. Although the 5% VaRs were very accurate, the 10% values turned out to be overestimates. A possible explanation could be that the log returns are either not elliptically distributed or that the interval on which the density generator is based is too short. Nevertheless, this example confirmed that the proposed block structures are well reflected in real data conditions and that the averaging estimators lead to significantly more stable and accurate results.

The problem of robustly estimating (conditional) covariance matrices arises in numerous applications for all types of different data. While the structural assumptions were designed to be applied to the modelling of (conditional) correlations of stock returns, the results give rise to further modelling questions. The idea of imposing a structural assumption on the underlying Kendall's tau matrix can be applied to any situation in which multiple pairwise Kendall's taus are assumed to be constant. The block structure is mere one example of a structural assumption, but there are of course many other structures that can be considered. In such research, the question remains of the formal comparison between U (non-overlapping pairs) and S (overlapping pairs), here related to the question of whether there are copulas for which the row estimator performs better than the block or diagonal estimators.

References

[1] A. Ang and G. Bekaert. International asset allocation with regime shifts. *Review of Financial Studies*, 15:1137–1187, 2002.

[2] P. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[3] M. Bin Abdullah. On a robust correlation coefficient. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 39(4):455–460, 1990.

[4] S. Delvin, R. Gnanadesikan, and J. Kettenring. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3):531–545, 1975.

[5] A. Derumigny. *CondCopulas: Estimation and Inference for Conditional Copulas Models*. R package version 0.0.4. 2021.

[6] A. Derumigny. *ElliptCopulas: Inference of Elliptical Copulas and Elliptical Distributions*. R package version 0.1.1. 2021.

[7] A. Derumigny and J.-D. Fermanian. A classification point-of-view about conditional kendall's tau. *Computational Statistics & Data Analysis*, 135(C):70–94, 2019.

[8] A. Derumigny and J.-D. Fermanian. About tests of the "simplifying" assumption for conditional copulas. *Dependence Modeling*, 5:154–197, 2017.

[9] A. Derumigny and J.-D. Fermanian. On kendall's regression. *Journal of Multivariate Analysis*, 178:104610, 2020.

[10] A. Derumigny and J.-D. Fermanian. On kernel-based estimation of conditional kendall's tau: finite-distance bounds and asymptotic behavior. *Dependence Modeling*, 7(1):292–321, 2019.

[11] F. Durante and C. Sempi. *Principles of Copula Theory*. Chapman and Hall/CRC, 2015.

[12] C. Erb, C. Harvey, and T. Viskanta. Forecasting international equity correlations. *Financial Analysts Journal*, 50:32–45, 1994.

[13] J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

[14] J. Fan, F. Han, and H. Liu. Challenges of Big Data analysis. *National Science Review*, 1(2):293–314, 2014.

[15] J. Fan, Y. Liao, and H. Liu. An overview on the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19, 2015.

[16] J. Fan, Y. Liao, and W. Wang. Projected principal component analysis in factor models. *The Annals of Statistics*, 44(1):219–254, 2016.

[17] J.-D. Fermanian and M. Wegkamp. Time-dependent copulas. *Journal of Multivariate Analysis*, 110:19–29, 2012.

[18] P. Filzmoser, H. Fritz, and K. Kalcher. *pcaPP: Robust PCA by Projection Pursuit*. R package version 1.9.74. 2021.

[19] R. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

[20] D. Freedman, R. Pisani, and R. Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[21] E. García-Portugués. *Notes for Predictive Modeling*. 2022. URL: `https://bookdown.org/egarpor/PM-UC3M/`. Version 5.9.8.

[22] C. Genest, J. Nešlehová, and N. Ghorbal. Estimators based on kendall's tau in multivariate copula models. *Australian & New Zealand Journal of Statistics*, 53:157–177. 2011.

[23] I. Gijbels, N. Veraverbeke, and M. Omelka. Conditional copulas, association measures and their applications. *Computational Statistics & Data Analysis*, 55:1919–1932, 2011.

[24] H. Gray, G. G. R. Leday, C. A. Vallejos, and S. Richardson. Shrinkage estimation of large covariance matrices using multiple shrinkage targets. *arXiv: Methodology*, 2018.

[25] M. Hahsler, C. Buchta, and K. Hornik. *seriation: Infrastructure for Ordering Objects Using Seriation*. R package version 1.3.2. 2022.

[26] W. K. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer-Verlag Berlin and Heidelberg, 2004.

[27] W. Hoeffding. A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics*, 19(4):546–557, 1948.

[28] P. Huber, J. Wiley, and W. InterScience. *Robust statistics*. Wiley New York, 1981.

[29] H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1997.

[30] M. Köhler, A. Schindler, and S. Sperlich. A review and comparison of bandwidth selection methods for kernel regression. *International Statistical Review / Revue Internationale de Statistique*, 82(2):243–274, 2014.

[31] E. Liebscher. A semiparametric density estimator based on elliptical distributions. *Journal of Multivariate Analysis*, 92(1):205–225, 2005.

[32] F. Lindskog, A. McNeil, and U. Schmock. Kendall's tau for elliptical distributions. *Contributions to Economics*:149–156, 2003.

[33] F. Longin and B. Solnik. Extreme value correlation of international equity markets. *The Journal of Finance*, 56:649–676, 2001.

[34] M. Maechler, R. Heiberger, J. Nash, and H. Borchers. *Rmpfr: Multiple Precision Floating-Point Reliable*. R package version 0.8.7. 2021.

[35] A. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton University Press, 2005.

[36] J. Owen and R. Rabinovitch. On the class of elliptical distributions and their applications to the theory of portfolio choice. *The Journal of Finance*, 38(3):745–752, 1983.

[37] A. Patton. Modeling asymmetric exchange rate dependence. *International Economic Review*, 47:527–556, 2006.

[38] A. Patton. Modelling time-varying exchange rate dependence using the conditional copula. *Econometrics eJournal*, 2001.

[39] A. J. Patton. Estimation of copula models for time series of possibly different lengths. *University of California at San Diego, Economics Working Paper Series*, 2001(17), 2001.

[40] S. Perreault, T. Duchesne, and J. G. Nešlehová. Detection of block-exchangeable structure in large-scale correlation matrices. *Journal of Multivariate Analysis*, 169(C):400–422, 2019.

[41] S. Perreault, J. Nešlehová, and T. Duchesne. Hypothesis tests for structured rank correlation matrices. *ArXiv preprint, arXiv:2007.09738*, 2020.

[42] I. Pimenova. *Semi-parametric estimation of elliptical distribution in case of high dimensionality*. Master's thesis, Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, 2012.

[43] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021.

[44] A. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

[45] J. Sadefo Kamdem. Value-at-risk and expected shortfall for linear portfolios with elliptically distributed risk factors. *International Journal of Theoretical and Applied Finance (IJTAF)*, 08:537–551, 2005.

[46] A. Sharma. *Text Book of Correlations and Regression*. DPH mathematics series. Discovery Publishing House, 2005.

[47] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

[48] A. Sklar. Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6):449–460, 1973.

[49] A. Tsybakov. *Introduction à l'estimation non paramétrique*, volume 41. Springer Science & Business Media, 2003.

[50] N. Veraverbeke, M. Omelka, and I. Gijbels. Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics*, 38:766–780, 2011.

A List of Stocks Used

Group 1

1. Cafom (CAFO)
2. Techstep (TECH)
3. Gold by Gold (ALGLD)
4. Fonciere Inea (INEA)
5. NSC Groupe (ALNSC)
6. Hofseth BioCare (HBC)
7. GC Rieber Shipping (RISH)
8. Aega (AEGA)
9. i2S (ALI2S)
10. Moury Construct (MOUR)
11. Gascogne (ALBI)
12. Thunderbird (TBIRD)
13. Hydratec Industries (HYDRA)
14. Sparebank 1 Ostfold Akershus (SOAG)
15. Altareit (AREIT)
16. Unibel (UNBL)
17. Cheops Technology France (MLCHE)
18. Zenobe Gramme Cert (ZEN)
19. Indel Fin (INFE)
20. Artois Nom (ARTO)
21. IDS (MLIDS)
22. Musee Grevin (GREV)
23. Robertet (CBE)
24. Aurskog Sparebank (AURG)
25. Alliance Developpement Capital (ALDV)
26. Fonciere Atland (FATL)
27. FREYR Battery (FREY)
28. Hotels de Paris (HDP)
29. Phone Web (MLPHW)
30. Maroc Telecom (IAM)
31. Sporting (SCP)
32. MG International (ALMGI)
33. Ucar (ALUCR)
34. Cumulex (CLEX)
35. Televerbier (TVRB)
36. Alan Allman Associates (AAA)
37. Serma Group (ALSER)
38. Planet Media (ALPLA)
39. Philly Shipyard (PHLY)
40. Augros Cosmetic Packaging (AUGR)
41. Sequa Petroleum (MLSEQ)
42. EMOVA Group (ALEMV)
43. Streamwide (ALSTW)
44. Accentis (ACCB)
45. Smalto (MLSML)
46. Signaux Girod (ALGIR)

Group 2

47. Interoil (IOX)
48. Ensurge Micropower (ENSU)
49. Idex Biometrics (IDEX)
50. SD Standard Drilling (SDSD)
51. SpareBank 1 Nord-Norge (NONG)
52. Bonheur (BONHR)
53. Eidesvik Offshore (EIOF)
54. DOF (DOF)
55. Solstad Offshore (SOFF)
56. Havila Shipping (HAVI)
57. Awilco LNG (ALNG)
58. FLEX LNG (FLNG)
59. Avance Gas Holding (AGAS)
60. Hunter Group (HUNT)
61. Itera (ITERA)
62. Q-Free (QFR)
63. Photocure (PHO)
64. PCI Biotech Holding (PCIB)
65. Hexagon Composites (HEX)
66. Nel (NEL)
67. McPhy Energy (MCPHY)
68. Vow (VOW)
69. Axactor (ACR)
70. Magseis Fair Fairfield (MSEIS)
71. NRC Group (NRC)
72. Petrolia (PSE)
73. Ctac (CTAC)
74. StrongPoint (STRO)
75. Crescent (OPTI)
76. Magnora (MGN)
77. Rec Silicon (RECSI)
78. Questerre Energy Corp (QEC)
79. ElectroMagnetic GeoServices (EMGS)
80. ABG Sundal Collier Holding (ABG)
81. Nekkar (NKR)
82. SeaBird Exploration (GEG)
83. Wilh. Wilhelmsen Holding (WWI)
84. Golden Ocean Group (GOGL)
85. Frontline (FRO)
86. Euronav (EURN)
87. Norwegian Air Shuttle (NAS)
88. Atea (ATEA)
89. Vopak (VPK)
90. Orkla (ORK)
91. Corbion (CRBN)
92. Otello Corporation (OTEC)

Group 3

93. Aures Technologies (AURS)
94. Keyrus (ALKEY)
95. Nextedia (ALNXT)
96. Cabasse Group (ALCG)
97. Groupe Guillin (ALGIL)
98. Guillemot (GUI)
99. Solutions 30 (S30)
100. Esker (ALESK)
101. Wavestone (WAVE)
102. Groupe Open (OPN)
103. Envea (ALTEV)
104. Stern Groep (STRN)
105. IT Link (ALITL)
106. Lectra (LSS)
107. Groupe CRIT (CEN)
108. Aubay (AUB)
109. Sword Group (SWP)
110. NRJ Group (NRG)
111. Van de Velde (VAN)
112. Hunter Douglas (HDG)
113. Oeneo (SBT)
114. Axway Software (AXW)
115. SES-imagotag (SESL)
116. Ateme (ATEME)
117. Infotel (INF)
118. Sergeferrari Group (SEFER)
119. Umanis (ALUMS)
120. Corticeira Amorim (COR)
121. Pharmagest Interactive (PHA)
122. Asetek (ASTK)
123. ID Logistics Group (IDL)
124. Scana (SCANA)
125. Acheter Louer fr (ALOLO)
126. Adomos (ALADO)
127. Glintt (GLINT)
128. Inapa (INA)
129. Cegedim (CGM)
130. Lavide Holding (LVIDE)
131. TIE Kinetix (TIE)
132. Alumexx (ALX)
133. Ober (ALOBR)
134. Cibox Interactive (CIB)
135. Evolis (ALTVO)
136. Proactis (PROAC)
137. Visiodent (SDT)
138. Fashion B Air (ALFBA)
139. Adthink (ALADM)
140. Innelec Multimedia (ALINN)
141. Herige (ALHRG)
142. Egide (GID)

143. U10 Corp (ALU10)
144. Mr. Bricolage (ALMRB)
145. Coheris (COH)
146. Pcas (PCA)
147. Rosier (ENGB)
148. Itesoft (ITE)
149. Gea Grenobl.Elect. (GEA)
150. Immobel (IMMO)
151. IGE+XAO Group (IGE)
152. Koninklijke Brill (BRILL)
153. Argan (ARG)
154. Fonciere Lyonnaise (FLY)
155. Covivio Hotels (COVH)
156. Electricite de Strasbourg (ELEC)
157. Robertet (RBT)
158. Norway Royal Salmon (NRS)
159. Olympique Lyonnais Groupe (OLG)
160. GeoJunxion (GOJXN)
161. Hybrid Software Group (HYSG)
162. Cast (CAS)
163. Acteos (EOS)
164. HF Company (ALHF)
165. Vranken-Pommery Monopole (VRAP)
166. Generix Group (GENX)
167. Union Technologies Infor. (FPG)
168. Diagnostic Medical Systems (DGM)
169. Capelli (CAPLI)
170. EXEL Industries (EXE)
171. Groupe LDLC (ALLDL)
172. genOway (ALGEN)
173. CBo Territoria (CBOT)
174. Aurea (AURE)
175. EO2 (ALEO2)
176. RAK Petroleum (RAKP)

Group 4

177. DNO (DNO)
178. Archer Limited (ARCH)
179. Odfjell Drilling (ODL)
180. BW Offshore (BWO)
181. Panoro Energy (PEN)
182. PGS (PGS)
183. TGS (TGS)
184. Subsea 7 (SUBC)
185. Equinor (EQNR)
186. Aker BP (AKRBP)
187. Aker (AKER)
188. Aker Solutions (AKSO)
189. Akastor (AKAST)
190. Etablissements Maurel et Prom (MAU)

191. Vallourec (VK)
192. CGG (CGG)
193. TechnipFMC (FTI)
194. Fugro (FUR)
195. SBM Offshore (SBMO)
196. Galp Energia (GALP)
197. TotalEnergies (TTE)
198. Royal Dutch Shell B (RDSB)
199. Schlumberger Limited (LSD)
200. Alten (ATE)
201. Capgemini (CAP)
202. Atos (ATO)
203. STMicroelectronics (STM)
204. ASML Holding (ASML)
205. ASM International (ASM)
206. BE Semiconductors (BESI)
207. Banco Comercial Portugues (BCP)
208. Mota-Engil (EGL)
209. Altri (ALTR)
210. The Navigator Company (NVG)
211. Semapa (SEM)
212. Trigano (TRI)
213. Ipsos (IPS)
214. Barco (BAR)
215. TomTom (TOM2)
216. PostNL (PNL)
217. TF1 (TFI)
218. Derichebourg (DBG)
219. Heijmans (HEIJM)
220. Koninklijke BAM Groep (BAMNB)
221. Aegon (AGN)
222. AXA (CS)
223. Agaas (AGS)
224. Bouygues (EN)
225. VINCI (DG)
226. Eiffage (FGR)
227. Faurecia (EO)
228. Valeo (FR)
229. Michelin (ML)
230. Ackermans & Van Haaren (ACKB)
231. Royal Boskalis Westminster (BOKA)
232. Imerys (NK)
233. Solvay (SOLB)
234. Umicore (UMI)
235. AkzoNobel (AKZA)
236. Air Liquide (AI)
237. Koninklijke Philips (PHIA)
238. Aperam (APAM)
239. Eramet (ERA)
240. Norsk Hydro (NHY)

B R Code

KTMatrixEst()

```
#' Fast estimation of Kendall's tau matrix
#'
#'
#' @param dataMatrix matrix of size \code{(n,d)} containing \code{n}
#' observations of a \code{d}-dimensional random vector.
#'
#' @param averaging type of averaging used for fast estimation.
#' Possible choices are \itemize{
#'   \item \code{no}: no averaging;
#'   \item \code{all}: averaging all Kendall's taus in each block;
#'   \item \code{diag}: averaging along diagonal blocks elements;
#'   \item \code{row}: averaging Kendall's tau along the smallest block side.
#' }
#'
#' @param blockStructure list of vectors.
#' Each vector corresponds to one group of variables
#' and contains the indexes of the variables that belongs to this group.
#' \code{blockStructure} must be a partition of \code{1:d},
#' where \code{d} is the number of columns in \code{dataMatrix}.
#'
#'
#' @return matrix with dimensions depending on \code{averaging}.
#' \itemize{
#'   \item If \code{averaging = no},
#'   the function returns a matrix of dimension \code{(n,n)}
#'   which estimates the Kendall's tau matrix.
#'
#'   \item Else, the function returns a matrix of dimension
#'   \code{(length(blockStructure) , length(blockStructure))}
#'   giving the estimates of the Kendall's tau for each block with ones
#'   on the diagonal.
#'
#' }
#'
#' @author Rutger van der Spek, Alexis Derumigny
#'
```

```
#' @examples
#' # Estimating off-diagonal block Kendall's taus
#' matrixCov = matrix(c(1   , 0.5, 0.3 ,0.3,
#'                      0.5,   1, 0.3, 0.3,
#'                      0.3, 0.3,   1, 0.5,
#'                      0.3, 0.3, 0.5,   1), ncol = 4 , nrow = 4)
#' dataMatrix = mvtnorm::rmvnorm(n = 100, mean = rep(0, times = 4),
#'                               sigma = matrixCov)
#' blockStructure = list(1:2, 3:4)
#' KTMatrixEst(dataMatrix = dataMatrix, blockStructure = blockStructure,
#'             averaging = "all")
#'
#'
#' @export
#'
KTMatrixEst <- function(dataMatrix, blockStructure = NULL, averaging = "no")
{
  d = ncol(dataMatrix)
  n = nrow(dataMatrix)

  if (averaging == "no"){
    estimate <- pcaPP::cor.fk(dataMatrix)
    return(estimate)
  }

  # We now assume that one of the averaging method is used.

  if(is.null(blockStructure))
  {
    stop("To use averaging estimators, a block structure must be specified.")
  }
  if (length(blockStructure) == 1){
    stop("To use averaging estimators, there must be more than one block.")
  }
  if( sum(1:d %in% unlist(blockStructure)) == d & length(blockStructure) == d)
  {
    stop(paste0("The block structure is not a partition of 1:", d ))
  }

  totalGroups = length(blockStructure)
```

```
estimate = matrix(data = 1 , nrow = totalGroups , ncol = totalGroups)

switch(
  averaging,

  "diag" = {

    for (g1 in 2:totalGroups) {
      for (g2 in 1:g1) {
        diagSize = min(length(blockStructure[[g1]]) ,
                       length(blockStructure[[g2]]) )

        vectorBlockKT = rep(NA, times = diagSize)
        for (j in 1:diagSize)
        {
          vectorBlockKT[j] = pcaPP::cor.fk(
            x = dataMatrix[ , blockStructure[[g1]][j] ] ,
            y = dataMatrix[ , blockStructure[[g2]][j] ])
        }
        blockKT = mean(vectorBlockKT)
        estimate[g1,g2] = blockKT
        estimate[g2,g1] = blockKT
      }
    }
  } ,

  "all" = {
    for (g1 in 2:totalGroups) {
      for (g2 in 1:g1) {
        matrixBlockKT = matrix(nrow = length(blockStructure[[g1]]) ,
                               ncol = length(blockStructure[[g2]]) )
        for (j1 in 1:length(blockStructure[[g1]]) )
        {
          for (j2 in 1:length(blockStructure[[g2]]) )
          {
            matrixBlockKT[j1,j2] = pcaPP::cor.fk(
              x = dataMatrix[ , blockStructure[[g1]][j1] ] ,
              y = dataMatrix[ , blockStructure[[g2]][j2] ] )
          }
        }
```

```
          blockKT = mean(matrixBlockKT)
          estimate[g1,g2] = blockKT
          estimate[g2,g1] = blockKT
        }
      }
    } ,

    "row" = {

      for (g1 in 2:totalGroups) {
        for (g2 in 1:g1) {
          g1Size = length(blockStructure[[g1]])
          g2Size = length(blockStructure[[g2]])

          rowSize = min(g1Size, g2Size)

          gSmall = ifelse(g1Size <= g2Size, g1, g2)
          gLarge = ifelse(g1Size <= g2Size, g2, g1)

          vectorBlockKT = rep(NA, times = rowSize)

          for (j in 1:diagSize)
          {
            vectorBlockKT[j] = pcaPP::cor.fk(
              x = dataMatrix[ , blockStructure[[gSmall]][j] ] ,
              y = dataMatrix[ , blockStructure[[gLarge]][1] ] )
          }
          blockKT = mean(vectorBlockKT)
          estimate[g1,g2] = blockKT
          estimate[g2,g1] = blockKT
        }
      }
    }

  )

  return(estimate)
}
```

CKTmatrix.kernel()

```
#' Estimate the conditional Kendall's tau matrix
#' at different conditioning points
#'
#' Assume that we are interested in a random vector \eqn{(X, Z)},
#' where \eqn{X} is of dimension \eqn{d > 2} and \eqn{Z} is of dimension
#' \eqn{1}. We want to estimate the dependence across the elements of the
#' conditioned vector \eqn{X} given \eqn{Z=z}.
#' This function takes in parameter observations of \eqn{(X,Z)}
#' and returns kernel-based estimators of \deqn{\tau_{i,j | Z=zk}}
#' which is the conditional Kendall's tau between \eqn{X_i} and \eqn{X_j}
#' given to \eqn{Z=zk}, for every conditioning point \eqn{zk} in \code{gridZ}.
#'
#' If the conditional Kendall's tau matrix has a block structure,
#' then improved estimation is possible by averaging over the kernel-based
#'  estimators of pairwise conditional Kendall's taus.
#' Groups of variables composing the same blocks can be defined
#' using the parameter \code{blockStructure}, and the averaging can be set
#' on using the parameter \code{averaging=all}, or \code{averaging=diag}
#' for faster estimation by averaging only over diagonal elements of each
#' block.
#'
#' @param dataMatrix a matrix of size \code{(n,d)} containing \code{n}
#'  observations of a \code{d}-dimensional random vector \eqn{X}.
#'
#' @param observedZ vector of observed points of a conditioning variable
#' \eqn{Z}. It must have the same length as the number of rows of
#' \code{dataMatrix}.
#'
#' @param h bandwidth. It can be a real, in this case the same \code{h}
#' will be used for every element of \code{gridZ}.
#' If \code{h} is a vector then its elements are recycled to match the length
#' of \code{gridZ}.
#'
#' @param gridZ points at which the conditional Kendall's tau is computed.
#' @param typeEstCKT type of estimation of the conditional Kendall's tau.
#' @param kernel.name name of the kernel used for smoothing.
#' Possible choices are: \code{"Gaussian"} (Gaussian kernel)
#' and \code{"Epa"} (Epanechnikov kernel).
#'
```

```
#' @param averaging type of averaging used for fast estimation.
#' Possible choices are \itemize{
#'   \item \code{no}: no averaging;
#'   \item \code{all}: averaging all Kendall's taus in each block;
#'   \item \code{diag}: averaging along diagonal blocks elements.
#' }
#'
#' @param blockStructure list of vectors.
#' Each vector corresponds to one group of variables
#' and contains the indexes of the variables that belongs to this group.
#' \code{blockStructure} must be a partition of \code{1:d},
#' where \code{d} is the number of columns in \code{dataMatrix}.
#'
#'
#' @return array with dimensions depending on \code{averaging}:
#' \itemize{
#'   \item If \code{averaging = "no"}:
#'   it returns an array of dimensions \code{(n, n, length(gridZ))},
#'   containing the estimated conditional Kendall's tau matrix given
#'   \eqn{Z = z}. Here, \code{n} is the number of rows in \code{dataMatrix}.
#'
#'   \item If \code{averaging = "all"} or \code{"diag"}:
#'   it returns an array of dimensions
#'   \code{(length(blockStructure), length(blockStructure), length(gridZ))},
#'   containing the block estimates of the conditional Kendall's tau given
#'   \eqn{Z = z} with ones on the diagonal.
#' }
#'
#' @seealso \code{\link{CKT.kernel}} for kernel-based estimation of
#' conditional Kendall's tau between two variables (i.e. the equivalent of
#' this function when \eqn{X} is bivariate and \code{d=2}).
#' \code{ElliptCopulas::\link[ElliptCopulas]{KTMatrixEst}()} for the fast
#' estimation of Kendall's tau matrix in the unconditional case (i.e., without
#'  Z and without smoothing).
#'
#' @examples
#'
#' # Data simulation
#' n = 100
#' Z = runif(n)
```

```
#' d = 5
#' CKT_11 = 0.8
#' CKT_22 = 0.9
#' CKT_12 = 0.1 + 0.5 * cos(pi * Z)
#' data_X = matrix(nrow = n, ncol = d)
#' for (i in 1:n){
#'   CKT_matrix = matrix(data =
#'     c(  1      , CKT_11   , CKT_11   , CKT_12[i], CKT_12[i] ,
#'       CKT_11   ,   1      , CKT_11   , CKT_12[i], CKT_12[i] ,
#'       CKT_11   , CKT_11   ,    1     , CKT_12[i], CKT_12[i] ,
#'       CKT_12[i], CKT_12[i], CKT_12[i],   1      , CKT_22    ,
#'       CKT_12[i], CKT_12[i], CKT_12[i], CKT_22   ,   1
#'       ) ,
#'      nrow = 5, ncol = 5)
#'   sigma = sin(pi * CKT_matrix/2)
#'   data_X[i, ] = mvtnorm::rmvnorm(n = 1, sigma = sigma)
#' }
#' plot(as.data.frame.matrix(data_X))
#'
#' # Estimation of CKT matrix
#' h = 1.06 * sd(Z) * n^{-1/5}
#' gridZ = c(0.2, 0.8)
#' estMatrixAll <- CKTmatrix.kernel(
#'   dataMatrix = data_X, observedZ = Z, gridZ = gridZ, h = h)
#' # Averaging estimator
#' estMatrixAve <- CKTmatrix.kernel(
#'   dataMatrix = data_X, observedZ = Z, gridZ = gridZ,
#'   averaging = "diag", blockStructure = list(1:3,4:5), h = h)
#'
#' # The estimated CKT matrix conditionally to Z=0.2 is:
#' estMatrixAll[ , , 1]
#' # Using the averaging estimator,
#' # the estimated CKT between the first group (variables 1 to 3)
#' # and the second group (variables 4 and 5) is
#' estMatrixAve[1, 2, 1]
#'
#' # True value (of CKT between variables in block 1 and 2 given Z = 0.2):
#' 0.1 + 0.5 * cos(pi * 0.2)
#'
#'
```

```
#' @author Rutger van der Spek, Alexis Derumigny
#'
#' @export
#'
CKTmatrix.kernel <- function(dataMatrix, observedZ, gridZ,
                             averaging = "no", blockStructure = NULL,
                             h, kernel.name = "Epa",
                             typeEstCKT = 4)
{
  d = ncol( dataMatrix )
  n = nrow( dataMatrix )
  nz = length( gridZ )

  if(length(observedZ) != n)
  {
    stop("The length of observedZ and the number of rows in dataMatrix must be
         equal.")
  }

  arrayWeights = array(data = NA , dim = c(n ,n , nz) )
  matrixWeights = matrix(data = NA, nrow = n, ncol = nz)
  for(i in 1:length(gridZ))
  {
    matrixWeights[,i] = computeWeights.univariate(vectorZ = observedZ,
                                                  h = h,
                                                  pointZ = gridZ[i],
                                                  kernel.name = kernel.name,
                                                  normalization = TRUE)
    arrayWeights[,,i] = matrixWeights[,i] %*% t(matrixWeights[,i])
  }


  if(averaging == "no")
  {
    estimate = array(data = 1 , dim = c(d , d , nz))
    for(j1 in 2:d)
    {
      for(j2 in 1:(j1-1))
      {
        vectorX1 = dataMatrix[ , j1]
```

```
      vectorX2 = dataMatrix[ , j2]
      matrixSigns = computeMatrixSignPairs(vectorX1 = vectorX1,
                                           vectorX2 = vectorX2,
                                           typeEstCKT = typeEstCKT)
      estimate[j1,j2,] = apply(X = arrayWeights , MARGIN = 3 ,
                               FUN = function(x){return(sum(x*matrixSigns))
                                   })
      estimate[j2,j1,] = estimate[j1,j2,]
    }
  }
} else if(averaging == "diag")
{
  if(is.null(blockStructure))
  {
    stop(paste("blockStructure not specified, when averaging = ", averaging)
        )
  }
  if( all.equal( sort(unname(unlist(blockStructure))) , 1:d ) != TRUE )
  {
    stop("blockStructure must be a partition.")
  }
  totalGroups = length(blockStructure)
  estimate = array(data = 1 , dim = c(totalGroups , totalGroups , nz))
  for (g1 in 2:totalGroups)
  {
    for (g2 in 1:(g1-1))
    {

      diagSize = min(length(blockStructure[[g1]]),
                     length(blockStructure[[g2]]) )
      matrixBlockCKT = matrix(NA, nrow = diagSize , ncol = nz)
      for (j in 1:diagSize)
      {
        vectorX1 = dataMatrix[ , blockStructure[[g1]][j] ]
        vectorX2 = dataMatrix[ , blockStructure[[g2]][j] ]
        matrixSigns = computeMatrixSignPairs(vectorX1 = vectorX1,
                                             vectorX2 = vectorX2,
                                             typeEstCKT = typeEstCKT)
        matrixBlockCKT[j,] = apply(X = arrayWeights , MARGIN = 3,
                                   FUN = function(x)
```

```
                                    {return(sum(x * matrixSigns))})
      }
      blockCKT = apply(X = matrixBlockCKT , MARGIN = 2 , mean)
      estimate[g1 , g2 ,] = blockCKT
      estimate[g2 , g1 ,] = blockCKT

    }
  }
} else if (averaging == "all")
{
  if(is.null(blockStructure))
  {
    stop(paste("blockStructure not specified, when averaging = ", averaging)
        )
  }
  if( all.equal( sort(unname(unlist(blockStructure))) , 1:d ) != TRUE )
  {
    stop("blockStructure must be a partition.")
  }
  totalGroups = length(blockStructure)
  estimate = array(data = 1 , dim = c(totalGroups , totalGroups , nz))
  for (g1 in 2:totalGroups)
  {
    for (g2 in 1:(g1-1))
    {
      arrayBlockCKT = array(NA, dim = c(length(blockStructure[[g1]]),
                                        length(blockStructure[[g2]]), nz) )
      for (j1 in 1:length(blockStructure[[g1]]) )
      {
        for (j2 in 1:length(blockStructure[[g2]]) )
        {
          vectorX1 = dataMatrix[ , blockStructure[[g1]][j1] ]
          vectorX2 = dataMatrix[ , blockStructure[[g2]][j2] ]
          matrixSigns = computeMatrixSignPairs(vectorX1 = vectorX1,
                                               vectorX2 = vectorX2,
                                               typeEstCKT = typeEstCKT)
          arrayBlockCKT[j1,j2,] = apply(X = arrayWeights , MARGIN = 3,
                                        FUN = function(x)
                                        {return(sum(x * matrixSigns))})
        }
```

```
        }
        blockCKT = apply(X = arrayBlockCKT , MARGIN = 3 , mean)
        estimate[g1,g2,] = blockCKT
        estimate[g2,g1,] = blockCKT

      }
    }
  } else
  {
    stop("'averaging' must be one of: 'no', 'all' or 'diag'.")
  }

  return(estimate)
}
```

EllDistrEst()

```
#' Nonparametric estimation of the density generator of an elliptical
#' distribution
#'
#' This function uses Liebscher's algorithm to estimate the density generator
#' of an elliptical distribution by kernel smoothing.
#'
#' @param X matrix of observations.
#' @param mu (estimated) mean of X.
#' @param Sigma_m1 (estimated) inverse of the covariance matrix of X.
#'
#' @param grid grid of values on which to estimate the density generator
#' @param h bandwidth of the kernel
#' @param Kernel kernel used for the smoothing
#' @param a tuning parameter to improve the performance at 0.
#' See Liebscher (2005), Example p.210.
#' @param mpfr if \code{mpfr = TRUE}, multiple precision floating point is
#' set. This allows for a higher accuracy, at the expense of computing times.
#' It is recommended to use this option for higher dimensions.
#' @param precBits number of precBits used for floating point precision
#' (only used if \code{mpfr = TRUE}).
```

```
#' @param dopb if \code{dopb = TRUE}, a progressbar is displayed.
#'
#' @return the values of the density generator of the elliptical copula,
#' estimated at each point of the `grid`.
#'
#' @references Liebscher, E. (2005).
#' A semiparametric density estimator based on elliptical distributions.
#' Journal of Multivariate Analysis, 92(1), 205.
#' \doi{10.1016/j.jmva.2003.09.007}
#'
#' @seealso \code{\link{EllDistrSim}} for the simulation of elliptical
#' distribution samples, \code{\link{EllCopEst}} for the estimation of
#' elliptical copulas.
#'
#' @examples
#' # Comparison between the estimated and true generator of the Gaussian
#'   distribution
#' X = matrix(rnorm(500*3), ncol = 3)
#' grid = seq(0,5,by=0.1)
#' g_3 = EllDistrEst(X = X, grid = grid, a = 0.7, h=0.05)
#' g_3mpfr = EllDistrEst(X = X, grid = grid, a = 0.7, h=0.05,
#'                       mpfr = TRUE, precBits = 20)
#' plot(grid, g_3, type = "l")
#' lines(grid, exp(-grid/2)/(2*pi)^{3/2}, col = "red")
#'
#' # In higher dimensions
#' \donttest{
#' d = 250
#' X = matrix(rnorm(500*d), ncol = d)
#' grid = seq(0, 400, by = 25)
#' true_g = exp(-grid/2) / (2*pi)^{d/2}
#'
#' g_d = EllDistrEst(X = X, grid = grid, a = 100, h=40)
#'
#' g_dmpfr = EllDistrEst(X = X, grid = grid, a = 100, h=40,
#'                       mpfr = TRUE, precBits = 10000)
#' ylim = c(min(c(true_g, as.numeric(g_dmpfr[which(g_dmpfr>0)]))),
#'          max(c(true_g, as.numeric(g_dmpfr)), na.rm=TRUE) )
#' plot(grid, g_dmpfr, type = "l", col = "red", ylim = ylim, log = "y")
#' lines(grid, g_d, type = "l")
```

```
#' lines(grid, true_g, col = "blue")
#' }
#'
#' @author Alexis Derumigny, Rutger van der Spek
#'
#' @export
#' @importClassesFrom Rmpfr mpfr mpfrMatrix
#' @importFrom Rmpfr mean
#'
EllDistrEst <- function(X, mu = 0, Sigma_m1 = diag(d),
                        grid, h, Kernel = "epanechnikov", a = 1,
                        mpfr = FALSE, precBits = 100, dopb = TRUE)
{
  kernelFun = getKernel(Kernel = Kernel)
  d = ncol(X)
  n = nrow(X)
  n1 = length(grid)

  if(mpfr) {
    # We don't need to convert this to higher precision
    # -> mpfr is needed only for the exponentiation
    # X = Rmpfr::mpfr(X, precBits = precBits)
    # mu = Rmpfr::mpfr(mu, precBits = precBits)
    # Sigma_m1 = Rmpfr::mpfr(Sigma_m1, precBits = precBits)
    # h = Rmpfr::mpfr(h, precBits = precBits)
    # s_d = Rmpfr::Const("pi")^(d/2) / Rmpfr::igamma(d/2,0)

    a = Rmpfr::mpfr(a, precBits = precBits)
    d = Rmpfr::mpfr(d, precBits = precBits)
    grid = Rmpfr::mpfr(grid, precBits = precBits)

  }
  s_d = pi^(d/2) / gamma(d/2)
  vector_Y = rep(NA , n)
  grid_g = rep(NA, n1)

  if (dopb){ pb = pbapply::startpb(max = n + n1) }

  for (i in 1:n) {
    # The matrix product is the expensive part (in high dimensions)
```

```
    # and should not use the mpfr library.
    # (mpfr is only used in the exponentiation, after)
    vector_Y[i] = as.numeric(
      -a + (a ^ (d/2) + ( (X[i,] - mu) %*% Sigma_m1 %*% (X[i,] - mu) )
            ^ (d/2) ) ^ (2/d) )
    if (dopb){ pbapply::setpb(pb, i) }
  }

  for (i1 in 1:n1){
    z = grid[i1]
    psiZ = as.numeric( -a + (a ^ (d/2) + z^(d/2)) ^ (2/d) )
    psiPZ = z^(d/2 - 1) * (a ^ (d/2) + z^(d/2)) ^ (2/d - 1)
    # This should use mean.default() (not the mpfr version) to save
        computation time.
    h_ny = (1/h) * mean( kernelFun((psiZ - vector_Y)/h) + kernelFun((psiZ +
        vector_Y)/h) )
    gn_z = 1/s_d * z^(-d/2 + 1) * psiPZ * h_ny
    grid_g[i1] = as.numeric(gn_z)

    if (dopb){ pbapply::setpb(pb, n + i1) }
  }

  if (dopb){ pbapply::closepb(pb) }


  return (grid_g)
}
```