Fast estimation of Kendall's Tau and conditional Kendall's Tau matrices under structural assumptions

van der Spek, Rutger; Derumigny, Alexis

DOI
10.1515/demo-2025-0012

Publication date
2025

Document Version
Final published version

Published in
Dependence Modeling

Research Article

Rutger van der Spek and Alexis Derumigny*

Fast estimation of Kendall's Tau and conditional Kendall's Tau matrices under structural assumptions

https://doi.org/10.1515/demo-2025-0012
received September 29, 2023; accepted February 4, 2025

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. Consequences on the underlying correlation matrix are then discussed. 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 block structure, for some value 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 that 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.

Keywords: Kendall's tau matrix, block structure, kernel smoothing, conditional dependence measure

MSC 2020: Primary: 62H20, Secondary: 62F30, 62G05

1 Introduction

In dependence modeling, the main object of interest is the copula, which is a cumulative distribution function on $[0, 1]^p$ with uniform margins, describing the links between elements of a p-dimensional random vector $\mathbf{X}$. However, the copula belongs to an infinite-dimensional space, and it is not easy to represent it as soon as p is larger than 3 or 4. In such cases, finite-dimensional statistics becomes more useful to understand the dependence, the most well known of them being Kendall's tau matrix.

Kendall's tau between two random variables X_i and X_j, denoted by $\tau_{i,j} = \tau(\mathbb{P}_{i,j})$, is defined as the probability of concordance between two independent replications from the distribution $\mathbb{P}_{i,j}$ of (X_i, X_j) minus the probability of discordance. The equality $\tau(\mathbb{P}_{i,j}) = 4\int C_{i,j} dC_{i,j} - 1$ relates Kendall's tau with the copula $C_{i,j}$ of X_i and X_j; we refer to the previous study [32] for an extensive introduction to Kendall's tau and copulas.

* **Corresponding author: Alexis Derumigny,** Department of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, Netherlands, e-mail: a.f.f.derumigny@tudelft.nl
Rutger van der Spek: Department of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, Netherlands

When a covariate $\mathbf{Z} \in \mathbb{R}^d$ is available, we can extend the definition of Kendall's tau to the conditional setting. Conditional Kendall's tau is then defined as $\tau_{i,j|\mathbf{Z}=\mathbf{z}} \coloneqq \tau(\mathbb{P}_{(i,j)|\mathbf{Z}=\mathbf{z}})$, where $\mathbb{P}_{(i,j)|\mathbf{Z}=\mathbf{z}}$ denotes the conditional law of (X_i, X_j) given $\mathbf{Z} = \mathbf{z}$, for some $\mathbf{z} \in \mathbb{R}^d$. In the previous studies [10,20,45], smoothing-based estimators of conditional Kendall's tau are studied. In [9], it is shown that the estimation of conditional Kendall's tau can be written as a classification task; they proposed to use classification algorithms to estimate conditional Kendall's tau. In [11], a regression-type model is used to estimate conditional Kendall's tau in a parametric conditional framework. [3] uses conditional Kendall's tau for hypothesis testing.

For a random vector $\mathbf{X}$, we define Kendall's tau matrix by $\mathbf{T} \coloneqq [\tau_{i,j}]_{1\le i,j\le p}$, which contains all pairwise Kendall's taus; respectively, in the conditional framework, a natural counterpart is conditional Kendall's tau matrix, denoted by $\mathbf{T}_{|\mathbf{Z}=\mathbf{z}} \coloneqq [\tau_{i,j|\mathbf{Z}=\mathbf{z}}]_{1\le i,j\le p}$. Kendall's tau matrix is especially useful for elliptical graphical models and their generalizations, see [4,26]. In the study by Lu et al. [28], a time-varying graphical model is studied using an estimate of conditional Kendall's tau matrix. Kendall's tau matrix plays an important role since it allows robust estimation of the dependence, and can be used to fit an appropriate copula [19]. In an elliptical distribution framework, it can also be used to estimate the Value at Risk of a portfolio, see [37,42].

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 [15]. Over the past two decades, various regularization strategies have been proposed to reduce the aggregation of estimation errors. Ultimately, these methods all make certain assumptions on the underlying dependence structure, hereby reducing 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 [5,21,39]. However, such assumptions are certainly not appropriate for the modeling 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 [15,16].

In studies by Perreault [34,35], an alternative approach to estimating large Kendall's tau matrices was introduced. 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 reordering 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 sample Kendall's taus within each of the blocks. In addition, they have proposed a robust algorithm identifying such structures (see also [36] for testing for the presence of such a structure).

In this article, we study a similar framework as in previous studies [35], where we relax the partial exchangeability assumption: we only assume that off-diagonal blocks of Kendall's tau matrix are constant. One of the drawbacks of the estimator studied in [35] 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.

We then extend this model to the conditional setup: conditional Kendall's taus are depending on $\mathbf{z}$ and are assumed to be clustered such that for all $\mathbf{z} \in \mathbb{R}^d$, the Kendall's tau conditionally 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 to changes of some other

American stock or portfolio. Furthermore, in previous studies [2,14,27], 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 [33]. In such a model, it is also important to limit computation times and study improved estimators that can take advantage of the block structure of the Kendall's tau matrix.

In this framework, we adopt nonparametric estimates of the conditional Kendall's tau based on kernel smoothing. On the basis of these nonparametric estimates, we introduce conditional versions of the averaging estimators and study their asymptotic behavior 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.

The rest of this article is structured as follows. In Section 2, we present the unconditional framework, and detail a few consequences on the correlation matrix. Then we construct the different estimators in this framework and derive variance expressions. Similarly, Section 3 is devoted to the improved estimation of the conditional Kendall's tau matrix, where we propose averaged conditional estimators and we derive the estimators' joint asymptotic normality. In Section 4, we perform a simulation study in order to support the theoretical findings. Finally, in Section 5, we examine a possible application to study the behavior of the estimators in real data conditions. The estimators are used for the robust inference of the covariance matrix to estimate the value at risk of a large stock portfolio. Proofs are postponed to the Appendix.

Notations. We denote by $\mathbf{1}$ be the vector and the matrix with all entries equal to 1, where the dimensions can be inferred from the context. For a matrix M of size $p \times p$, and a set of indices $J \subset \{1, \dots, p\}^2$, we denote by $[M]_J$ the submatrix $(M_j)_{j \in J}$.

2 Fast estimation of Kendall's tau matrix

2.1 The structural assumption

Let $n \geq 2$ and assume that we observe n i.i.d. replications $\mathbf{X}_i = (X_{i,1}, \dots, X_{i,p})$ of a random vector $\mathbf{X} = (X_1, \dots, X_p) \in \mathbb{R}^p$, for $i = 1, \dots, n$. Moreover, assume that the Kendall's tau matrix $\mathbf{T} := [\tau_{i,j}]_{1 \leq i,j \leq p}$ of $\mathbf{X}$ satisfies the following structural assumption.

Assumption 1. (Structural assumption) There exists $K > 0$, a partition $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_K\}$ of $\{1, \dots, p\}$, a set $J \subset \{1, \dots, K\}^2$ and some constants $(\tau_{k_1,k_2})_{(k_1,k_2) \in J} \in [-1, 1]^J$ such that for all $(k_1, k_2) \in J$,

$$[\mathbf{T}]_{\mathcal{G}_{k_1} \times \mathcal{G}_{k_2}} = \tau_{k_1,k_2} \cdot \mathbf{1}.$$

Note that after reordering of the variables by group, the corresponding Kendall's tau matrix is block structured with constant values in some of the off-diagonal blocks. The interest in investigating this structural assumption originates from applications in stock return modeling. In this context, the clustering of the variables could 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. We will therefore call τ_{k_1,k_2} **the intergroup Kendall's tau** between groups $\mathcal{G}_{k_1}$ and $\mathcal{G}_{k_2}$.

This can be clearly seen in Figure 1: in each of the off-diagonal blocks, the Kendall's tau is mostly homogeneous, but significant differences can be seen in the fourth diagonal block. Indeed, it gathers companies whose link to other groups is constant, but with different relationships inside the group. This may be explained by the presence of subgroups inside this fourth group, even if the relationship with variables from other groups do not seem to be related to these subgroup structures.

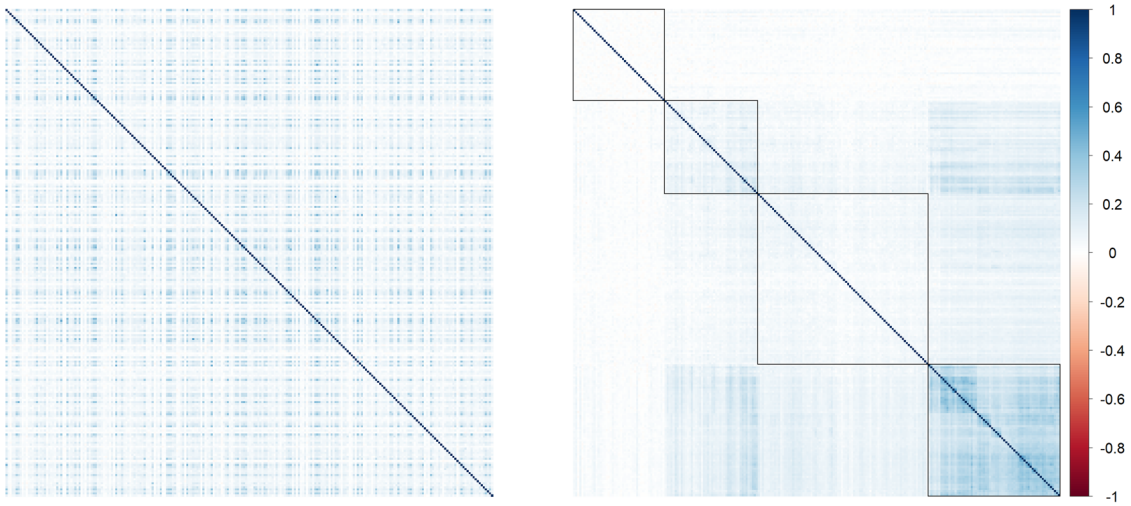

Figure 1: Heatmap plots of the sample Kendall's tau matrix computed on the daily log returns from 01 January 2007 until 14 January 2022 of all 240 portfolio stocks (whose list is available in Appendix C). (a) Unclustered and (b) clustered.

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. We will define estimators of the Kendall's tau matrix $\mathbf{T}$ under Assumption 1 for some known partitions $\mathcal{G}$ of $\{1, \ldots, p\}$. Note that such partition can also be inferred from the data, see [35,36] and the thesis [34]. Even if Assumption 1 is not satisfied, the estimators that we will propose can still be of interest, for example, to do linear shrinkage.

Note that although our results are only interesting when $k_1 \neq k_2$, they also hold if a pair of form (k_1, k_1) belongs to J. Indeed, in this case, we must have $\tau_{k_1,k_1} = 1$, and then all pairs of observations in the block will be concordant, so all our estimators will be equal to 1.

In this article, we are interested in the estimation of the intergroup Kendall's tau τ_{k_1,k_2} for some pair (k_1, k_2) belonging to J, i.e., such that the block $[\mathbf{T}]_{\mathcal{G}_{k_1}\times\mathcal{G}_{k_2}}$ is a constant block. This means that the random vector $(\mathbf{X}_{\mathcal{G}_{k_1}}^\top, \mathbf{X}_{\mathcal{G}_{k_2}}^\top)^\top$ satisfies the following assumption.

Assumption 2. (Simplified Structural assumption) Let $\mathbf{X}$ be a p-dimensional random vector of interest. Kendall's tau matrix of the random vector $\mathbf{X}$ can be written in the block form

$$\mathbf{T} = \begin{pmatrix} \cdot & \tau\mathbf{1} \\ \tau\mathbf{1} & \cdot \end{pmatrix},$$

where $\tau\mathbf{1}$ represents a block filled with the value $\tau \in [-1, 1]$ and the symbol $\cdot$ represents any matrices of respective sizes $b_1 \times b_1$ and $b_2 \times b_2$ for some $b_1 \in \{1, \ldots, p\}$ and $b_2 := p - b_1$.

Under Assumption 2, we call τ **the intergroup Kendall's tau**, as a particular case of the previous framework. As discussed earlier, Assumption 2 may seem more constraining than the previous Assumption 1, but both assumptions are actually equivalent insofar as we are only interested in the estimation of each interblock

Kendall's tau τ_{k_1,k_2}. Therefore, and in order to simplify the notations, we will choose to assume Assumption 2. Similarly as the more general Assumption 1, Assumption 2 can be tested using the framework developed in [36].

Note that [35] proposed a similar model with a more restrictive version of Assumption 1, the partial exchangeability assumption, by assuming that the variables could be partitioned into K clusters with exchangeable dependence.

Assumption 3. (Partial exchangeability assumption) For $j \in \{1, \ldots, p\}$, let $U_j = F_j(X_j)$, where F_j is the cumulative distribution function of X_j, and C be the copula of $\mathbf{X}$. For any partition $\mathcal{G} = \{\mathcal{G}_1, \ldots, G_K\}$ of $\{1, \ldots, p\}$, let $\pi(G)$ be the set of permutations π of $\{1, \ldots, p\}$ such that for all $j \in \{1, \ldots, p\}$ and all $k \in \{1, \ldots, K\}$, $\pi(j) \in \mathcal{G}_k$ if and only if $j \in \mathcal{G}_k$.

A partition $\mathcal{G} = \{\mathcal{G}_1, \ldots, G_K\}$ satisfies the partial exchangeability assumption if for any $u_1, \ldots, u_p \in [0, 1]$ and any permutation $\pi \in \pi(G)$, one has

$$C(u_1, \ldots, u_p) = C(u_{\pi(1)}, \ldots, u_{\pi(p)}),$$

or, equivalently, $(U_1, \ldots, U_p) \overset{\text{law}}{=} (U_{\pi(1)}, \ldots, U_{\pi(p)})$.

Note that the partial exchangeability assumption imposes restrictions on the underlying copula, whereas Assumption 1 only does so on the underlying Kendall's tau matrix, making it a lot less restrictive. Further, under Assumption 3, Kendall's tau matrix is fully block structured including constant diagonal blocks as well, after reordering of the variables. In contrast to [35], 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 partial exchangeability assumption seems quite restrictive and a model without partial exchangeability in which companies from the same cluster have different mutual dependence is more plausible (see Figure 1). For these reasons, we opt for a more flexible variant of this model.

2.2 Consequences of the block structure on the correlation matrix

As explained in [24, Chapter 3], if $\mathbf{X}$ follows a multivariate Gaussian distribution with exchangeable dependence and correlation $\rho \in (-1, 1)$, then $\mathrm{Corr}(\mathbf{X})$ is positive definite if and only if $\rho > -1/(p-1)$. In terms of Kendall's tau, this constraint translates as $\tau > -(2/\pi)\,\mathrm{Arcsin}(1/(p-1))$. Assumption 1 is weaker than exchangeable dependence, and even in the Gaussian setting with two groups of variables, it allows for *arbitrary* negative correlation in the off-diagonal block. The results presented in this section are related to the results of [6] who study the eigenstructure of such block structured correlation matrices, but without giving precise constraint on the allowed values of the correlation. McNeil et al. [30] discuss attainability of Kendall's tau matrices in a general framework, i.e., without discussing the block structure specifically.

Proposition 1. *Let $b_1, b_2 \geq 2$ be integers. Let $\rho_1, \rho_2, \rho_3 \in (-1, 1)$, and define the block matrix $M \in \mathbb{R}^{(b_1+b_2)^2}$ with blocks of size b_1 and b_2 by*

$$M := \begin{pmatrix} I + \rho_1 \tilde{I} & \rho_3 \mathbf{1} \\ \rho_3 \mathbf{1} & I + \rho_2 \tilde{I} \end{pmatrix},$$

where I is the identity matrix and $\tilde{I} := \mathbf{1} - I$ is the matrix with 1 at each off-diagonal entry and 0 on the diagonal. Then M is positive definite if and only if

$$(b_2 b_1 - b_1 - b_2 + 1)\rho_1\rho_2 - b_1 b_2 \rho_3^2 + (b_2 - 1)\rho_2 + (b_1 - 1)\rho_1 + 1 > 0. \tag{1}$$

Furthermore, this inequality is satisfied as soon as

$$\begin{cases} \rho_1 > \dfrac{b_1 b_2 \rho_3^2 - (b_2 - 1)\rho_2 - 1}{(b_1 b_2 - b_1 - b_2 + 1)\rho_2 + b_1 - 1}, \\ \rho_2 > -\dfrac{b_1 - 1}{b_1 b_2 - b_1 - b_2 + 1}. \end{cases}$$

This result is proved in Appendix B.1. We can remark that the constraint (1) is always satisfied as soon as

$$\rho_1 \rho_2 \geq \rho_3^2, \tag{2}$$

i.e., the absolute value of ρ_3 has to be smaller than the geometric mean of ρ_1 and ρ_2. Furthermore, in the high-dimensional setting, where $b_1, b_2 \to +\infty$ and ρ_1, ρ_2 are fixed and positive, (1) actually becomes equivalent to the simplified constraint (2). Note that (1) allows for situations, where ρ_3 is arbitrarily close to -1, for any choice of block sizes. This is possible, for example, by setting $\rho_1 = \rho_2 = |\rho_3|$. Concretely, if all variables of one group have a high correlation with all variables of the second group, then in each group, the intragroup correlation should be high. Such a result translates directly for Kendall's tau matrix by using the relationship $\tau = (2/\pi)\operatorname{Arcsin}(\rho)$, allowing for Kendall's tau matrix with arbitrary entries in the off-diagonal blocks.

Rather surprisingly, as soon as the group sizes b_1 and b_2 are large enough, they don't appear anymore in the constraint (2). This phenomenon is in fact typical of block-structured matrices and will appear again in the performance of our estimators in the next sections. We now give a lower bound for the intergroup Kendall's tau in the setting, where K groups are present, with equal intergroup Kendall's tau. It is proved in Appendix B.1.

Proposition 2. *Let $K \geq 2$ and let $b_1, \ldots, b_K$ be positive integers. Let $\rho \in (-1, 1)$, $\Sigma_1, \ldots, \Sigma_K$ be K correlation matrices of size $b_1, \ldots, b_K$, respectively, and let $M \in \mathbb{R}^{(b_1+b_2+\cdots+b_K)^2}$ be the block matrix defined by*

$$M := \begin{pmatrix} \Sigma_1 & \rho\mathbf{1} & \cdots & \rho\mathbf{1} \\ \rho\mathbf{1} & \Sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho\mathbf{1} \\ \rho\mathbf{1} & \cdots & \rho\mathbf{1} & \Sigma_K \end{pmatrix}.$$

Then

$$\inf\{\rho \in (-1, 1) : \exists \Sigma_1, \ldots, \Sigma_K \text{ such that } M \text{ is a correlation matrix}\} = \frac{-1}{K - 1}.$$

Interestingly, this bound does not depend on the sizes of the blocks $b_1, \ldots, b_K$. This shows that, for such a matrix, the structure of each block does not have a strong influence on the possible choice of intergroup Kendall's tau. The constraint $\rho \geq -1/(p-1)$ for exchangeable correlation matrices becomes in this framework $\rho \geq -1/(K-1)$ for exchangeable intergroup dependence, suggesting that the number of blocks becomes the relevant dimension for this problem (instead of the number of variables). Nevertheless, the knowledge of the intragroup correlations matrices $\Sigma_1, \ldots, \Sigma_K$ will still constrain the range of possible values of ρ, as this was the case in Proposition 1 for the particular case $K = 2$ and exchangeable dependence in each block.

2.3 Construction of estimators

First, note that we can naturally rely on the usual estimator of Kendall's tau between X_{j_1} and X_{j_2} defined by

$$\hat{\tau}_{j_1, j_2} := \frac{2}{n(n-1)} \sum_{i_1 < i_2} \operatorname{sign}((X_{i_1, j_1} - X_{i_2, j_1})(X_{i_1, j_2} - X_{i_2, j_2})), \tag{3}$$

for any $1 \leq j_1, j_2 \leq p$. We denote the corresponding Kendall's tau matrix estimator by $\hat{\mathbf{T}} = [\hat{\tau}_{j_1, j_2}]_{1 \leq j_1, j_2 \leq p}$, which serves as a first step estimator for obtaining a better estimator of the Kendall's tau matrix. This estimated Kendall's tau matrix $\hat{\mathbf{T}}$ does not make any use of the underlying structure and will therefore be a rather naive

tool in practice. As discussed in the previous section, we will introduce several estimators in the setting of Assumption 2, with straightforward generalizations under Assumption 1. More precisely, for every estimator $\hat{\tau}^{A2}$ of τ under Assumption 2, we define a corresponding matrix estimator $\widehat{\mathbf{T}}^{A1}$ under Assumption 1 by

$$\widehat{\mathbf{T}}^{A1} \coloneqq [\widehat{T}^{A1}_{j_1,j_2}]_{1\le j_1,j_2\le p} = \begin{cases} \hat{\tau}^{A2}(\mathcal{G}_{k_1},\mathcal{G}_{k_2}), & \text{if} \quad j_1 \in \mathcal{G}_{k_1}, \quad j_2 \in \mathcal{G}_{k_2} \quad \text{and} \quad (k_1,k_2) \in J, \\ \hat{\tau}_{j_1,j_2}, & \text{else.} \end{cases} \tag{4}$$

Remember that J is the set of pairs (k_1, k_2) of block indices such that Kendall's tau τ_{j_1,j_2} does not depend on $j_1 \in \mathcal{G}_{k_1}$, $j_2 \in \mathcal{G}_{k_2}$. Since we assume that the Kendall's taus in (some 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 $\hat{\tau}^{B}$ that averages all sample Kendall's taus within each of the off-diagonal blocks. Formally, we have

$$\hat{\tau}^{B} \coloneqq \frac{1}{b_1 b_2} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} \hat{\tau}_{j_1,j_2}.$$

We define $\widehat{\mathbf{T}}^{B}$ the corresponding estimator under Assumption 1 by Equation (4).

Under the partial exchangeability assumption, Perreault et al. [35] showed that the estimator $\widehat{\mathbf{T}}^{B}$ is 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. This would lead to computationally cheaper estimates. 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 $N \in \{1, \dots, b_1 \vee b_2\}$ pairs in the off-diagonal block under Assumption 2, so that we can compare estimators that either average pairs in the same row/column, or pairs on the diagonal. Without loss of generality, since we can switch both blocks of variables, we assume that $b_1 \le b_2$, and we will average over the row. For averaging pairs on the diagonal, it is moreover required that $N \le b_1 \wedge b_2$. 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. We set

$$\hat{\tau}^{R} \coloneqq \frac{1}{N} \sum_{j=1}^{N} \hat{\tau}_{1,b_1+j}, \quad \text{and} \quad \hat{\tau}^{D} \coloneqq \frac{1}{N} \sum_{j=1}^{N} \hat{\tau}_{j,b_1+j}.$$

Then, the "row-based" Kendall's tau matrix estimator $\widehat{\mathbf{T}}^{R}$ and the "diagonal-based" Kendall's tau matrix estimator $\widehat{\mathbf{T}}^{D}$ are defined by Equation (4) for the choices $\hat{\tau}^{A2} = \hat{\tau}^{R}$ and $\hat{\tau}^{A2} = \hat{\tau}^{D}$. 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.

Finally, we introduce the estimator that randomly selects pairs to average over per block. We denote the (deterministic) number of averaged pairs per block by $N \in \{1, \dots, b_1 \times b_2\}$ and the corresponding estimator by $\mathbf{T}^{U}$. The pairs are selected with uniform probability and without replacement. We define

$$\hat{\tau}^{U} \coloneqq \frac{1}{N} \sum_{j_1=1}^{b_1} \sum_{j_2=1}^{b_2} W_{j_1,j_2} \hat{\tau}_{j_1,j_2},$$

where $\mathbf{W}$ is a $b_1 \times b_2$ matrix of random weights that selects N pairs per off-diagonal block 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 denoted by $\widehat{\mathbf{T}}^{U}$ with $\boldsymbol{W}$ defined as selecting N pairs in each of the averaged blocks.

2.4 Comparison of their variances

Before we proceed with the main theoretical results on the estimators' variances, let us introduce some auxiliary notations. For every $j_1, j_2 \in \{1, \dots, p\}$, we set $P_{j_1,j_2} \coloneqq \mathbb{P}((X_{1,j_1} - X_{2,j_1})(X_{1,j_2} - X_{2,j_2}) > 0)$. The quantity P_{j_1,j_2} is equal to the probability of concordance of the variables X_{j_1} and X_{j_2}, and thus, $\tau_{j_1,j_2} = 2P_{j_1,j_2} - 1$. As such, the structural assumption 2 ensures that P_{j_1,j_2} is independent of the choice of pair whenever j_1 and j_2 are not in the same block. Alternatively we can write P_{j_1,j_2} in terms of the copula C_{j_1,j_2} of (X_{j_1}, X_{j_2}) by $P_{j_1,j_2} = 2\int_{[0,1]^2} C_{j_1,j_2}(u_1, u_2)\mathrm{d}C_{j_1,j_2}(u_1, u_2)$. By extension, we define

$$\begin{aligned}
P_{j_1,j_2,j_3,j_4} &\coloneqq \mathbb{P}((X_{1,j_1} - X_{2,j_1})(X_{1,j_2} - X_{2,j_2}) > 0, (X_{1,j_3} - X_{2,j_3})(X_{1,j_4} - X_{2,j_4}) > 0),\\
Q_{j_1,j_2,j_3,j_4} &\coloneqq \mathbb{P}((X_{1,j_1} - X_{2,j_1})(X_{1,j_2} - X_{2,j_2}) > 0, (X_{1,j_3} - X_{3,j_3})(X_{1,j_4} - X_{3,j_4}) > 0),\\
S_{j_1,j_2,j_3,j_4} &\coloneqq \mathbb{P}((X_{1,j_1} - X_{2,j_1})(X_{1,j_2} - X_{2,j_2}) > 0, (X_{4,j_3} - X_{3,j_3})(X_{4,j_4} - X_{3,j_4}) > 0) = P_{j_1,j_2}P_{j_3,j_4},
\end{aligned}$$

for every $j_1, j_2, j_3, j_4 \in \{1, \dots, p\}$. Note that both P and Q quantities can be understood as some kind of "cross-concordance measures," but there is an important difference between them: for the Q measure of cross-concordance, there is a third copy $\mathbf{X}_{3,1:p}$ needed. When $j_1 = j_3$ and $j_2 = j_4$, we obtain $P_{j_1,j_2,j_3,j_4} = P_{j_1,j_2}$. Q-type measures of cross-concordance naturally appears in the asymptotic variance of the usual estimator of Kendall's tau through the particular case

$$\begin{aligned}
Q_{j_1,j_2} \coloneqq Q_{j_1,j_2,j_1,j_2} &= \mathbb{P}((X_{1,j_1} - X_{2,j_1})(X_{1,j_2} - X_{2,j_2}) > 0, (X_{1,j_1} - X_{3,j_1})(X_{1,j_2} - X_{3,j_2}) > 0)\\
&= \int_{[0,1]^2} (C_{j_1,j_2}(u_1, u_2) + \bar{C}_{j_1,j_2}(u_1, u_2))^2\mathrm{d}C_{j_1,j_2}(u_1, u_2),
\end{aligned} \tag{5}$$

where $\bar{C}$ denotes the survival function of a copula C. For S, a fourth independent copy is needed, but by independence, it reduces to the product $P_{j_1,j_2}P_{j_3,j_4}$. For completeness, this expression is derived in Appendix B.2. Note that this equality was already given in [19, Equation (8)]. Note that both P and Q can be written as eight-dimensional integrals involving the copula of $(X_{j_1}, X_{j_2}, X_{j_3}, X_{j_4})$. As a consequence, they are functions of the joint law $\mathbb{P}_{(j_1,j_2,j_3,j_4)}$ of the random vector $\mathbf{X}_{(j_1,j_2,j_3,j_4)}$. Note further that even under Assumption 1, they both depend on the choice of pairs within the considered off-diagonal block. Obviously, this is not the case if we assume the stronger Assumption 3.

To give explicit expressions for our estimators, we need to average P, Q, and S quantities. We need to separate these averages depending on the number of common variables in the cross-concordance terms. We define

$$\begin{aligned}
P_{B,2} &\coloneqq \frac{1}{b_1 b_2} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} P_{j_1,j_2},\\
P_{B,1} &\coloneqq \frac{1}{b_1(b_1-1)b_2 + b_2(b_2-1)b_1}\left(\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}\sum_{j_3=1,j_3\neq j_1}^{b_1} P_{j_1,j_2,j_3,j_2} + \sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}\sum_{j_4=b_1+1,j_4\neq j_2}^{p} P_{j_1,j_2,j_1,j_4}\right),\\
P_{B,0} &\coloneqq \frac{1}{b_1(b_1-1)b_2(b_2-1)} \sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}\sum_{j_3=1,j_3\neq j_1}^{b_1}\sum_{j_4=b_1+1,j_4\neq j_2}^{p} P_{j_1,j_2,j_3,j_4},
\end{aligned}$$

and similarly, we define $P_{\alpha,\beta}$, $Q_{\alpha,\beta}$, $S_{\alpha,\beta}$ for $\alpha \in \{B, R, D\}$ and $\beta \in \{0, 1, 2\}$. In the latter expression, α denotes the type of averaging (respectively, over the block, over the row and over the diagonal) and β denotes the number of common variables, i.e., the size of $\{j_1, j_2\} \cap \{j_3, j_4\}$. This means that for $\beta = 2$, $P_{B,2}$ is the average of the P_{j_1,j_2,j_3,j_4} over the set of (j_1, j_2, j_3, j_4) such that $j_1 = j_3$ and $j_2 = j_4$ (2 common variables, since both variables are the same). $P_{B,1}$ is the average of the P_{j_1,j_2,j_3,j_4} over the set of (j_1, j_2, j_3, j_4) such that $j_1 = j_3$ or $j_2 = j_4$ (1 common variable, since one is the same and one is different). Finally, $P_{B,0}$ is the average of the P_{j_1,j_2,j_3,j_4} over the set of (j_1, j_2, j_3, j_4) such that $j_1 \neq j_3$ and $j_2 \neq j_4$ (0 common variables, all variables are different). To deal with the random estimator, we need to define the corresponding weighted quantities for a $p \times p$ matrix $\mathbf{w}$:

$$P_{\mathbf{w},2} \coloneqq \frac{1}{b_1 b_2} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} w_{j_1,j_2} P_{j_1,j_2},$$

$$P_{\mathbf{w},1} \coloneqq \frac{1}{b_1(b_1-1)b_2 + b_2(b_2-1)b_1} \left(\sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} \sum_{j_3=1, j_3 \neq j_1}^{b_1} w_{j_1,j_2} w_{j_3,j_2} P_{j_1,j_2,j_3,j_2} + \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} \sum_{j_4=b_1+1, j_4 \neq j_2}^{p} w_{j_1,j_2} w_{j_1,j_4} P_{j_1,j_2,j_1,j_4} \right),$$

$$P_{\mathbf{w},0} \coloneqq \frac{1}{b_1(b_1-1)b_2(b_2-1)} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} \sum_{j_3=1, j_3 \neq j_1}^{b_1} \sum_{j_4=b_1+1, j_4 \neq j_2}^{p} w_{j_1,j_2} w_{j_3,j_4} P_{j_1,j_2,j_3,j_4},$$

and similarly for $Q_{\mathbf{w},\beta}$ and $S_{\mathbf{w},\beta}$ for $\beta \in \{0, 1, 2\}$.

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 3. *Under* Assumption 2, *the estimators* $\hat{\tau}_{j_1,j_2}$, $\hat{\tau}^B$, $\hat{\tau}^R$, $\hat{\tau}^D$, *and* $\hat{\tau}^U$ *are all second order U-statistics, in the sense that they can be written as* $(n(n-1))^{-1} \sum_{1 \leq i_1 \neq i_2 \leq n} g(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})$, *for some real-valued function g. We denote these functions, respectively, by* g^*, g^B, g^R, g^D, *and* g^U.

This lemma is proved in Appendix B.3 where the expressions of the kernels g^*, g^B, g^R, g^D, and g^U are given. From Lemma 3 and the fact that $\mathbb{E}[g(\mathbf{X}_1, \mathbf{X}_2)] = \tau$ for each kernel g, it follows that $\hat{\mathbf{T}}$, $\hat{\mathbf{T}}^B$, $\hat{\mathbf{T}}^R$, $\hat{\mathbf{T}}^D$, and $\hat{\mathbf{T}}^U$ are all unbiased estimators of the Kendall's tau matrix under Assumption 1. The finite sample variances are given in the following theorem, proved in Appendix B.4.

Theorem 4. *Let* $1 \leq j_1, j_2 \leq p$, $1 \leq b_1 \leq p$ *and let* $b_2 \coloneqq p - b_1$. *The variances of* $\hat{\tau}_{j_1,j_2}$, $\hat{\tau}^B$, $\hat{\tau}^R$, $\hat{\tau}^D$, *and* $\hat{\tau}^U$, *and the conditional variance of* $\hat{\tau}^U$ *conditionally to* $\mathbf{W}$ *are given by*

(*i*)

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

(*ii*)

$$\mathbb{V}\mathrm{ar}[\hat{\tau}^B] = \frac{8}{b_1 b_2 n(n-1)} \Big(P_{B,2} - S_{B,2} + (b_1 + b_2 - 2)(P_{B,1} - S_{B,1}) + (b_1 - 1)(b_2 - 1)(P_{B,0} - S_{B,0}) + 2(n-2)(Q_{B,2} - S_{B,2} + (b_1 + b_2 - 2)(Q_{B,1} - S_{B,1}) + (b_1 - 1)(b_2 - 1)(Q_{B,0} - S_{B,0})) \Big).$$

(*iii*)

$$\mathbb{V}\mathrm{ar}[\hat{\tau}^R] = \frac{8}{Nn(n-1)} \Big(P_{R,2} - S_{R,2} + (N-1)(P_{R,1} - S_{R,1}) + 2(n-2)(Q_{R,2} - S_{R,2} + (N-1)(Q_{R,1} - S_{R,1})) \Big).$$

(*iv*)

$$\mathbb{V}\mathrm{ar}[\hat{\tau}^D] = \frac{8}{Nn(n-1)} \Big(P_{D,2} - S_{D,2} + (N-1)(P_{D,0} - S_{D,0}) + 2(n-2)(Q_{D,2} - S_{D,2} + (N-1)(Q_{D,0} - S_{D,0})) \Big).$$

(*v*)

$$\mathbb{V}\mathrm{ar}[\hat{\tau}^U | \mathbf{W}] = \frac{8}{b_1 b_2 n(n-1)} \Big(P_{\mathbf{W},2} - S_{\mathbf{W},2} + (b_1 + b_2 - 2)(P_{\mathbf{W},1} - S_{\mathbf{W},1}) + (b_1 - 1)(b_2 - 1)(P_{\mathbf{W},0} - S_{\mathbf{W},0}) + 2(n-2)(Q_{\mathbf{W},2} - S_{\mathbf{W},2} + (b_1 + b_2 - 2)(Q_{\mathbf{W},1} - S_{\mathbf{W},1}) + (b_1 - 1)(b_2 - 1)(Q_{\mathbf{W},0} - S_{\mathbf{W},0})) \Big).$$

(*vi*)

$$\mathbb{V}\mathrm{ar}[\hat{\tau}^U] = \mathbb{V}\mathrm{ar}[\tau_{J_1,J_2}] + \frac{8}{b_1 b_2 n(n-1)} \Big(2(n-2) \times (Q_{B,2} - S_{B,2} + \frac{N-1}{b_1 b_2 - 1}(b_1 + b_2 - 2)(Q_{B,1} - S_{B,1}) + \frac{N-1}{b_1 b_2 - 1}(b_1 - 1)(b_2 - 1)(Q_{B,0} - S_{B,0})) + (P_{B,2} - S_{B,2} + \frac{N-1}{b_1 b_2 - 1}(b_1 + b_2 - 2)(P_{B,1} - S_{B,1}) + \frac{N-1}{b_1 b_2 - 1}(b_1 - 1)(b_2 - 1)(P_{B,0} - S_{B,0})) \Big),$$

where J_1 and J_2 are independent random variables, uniformly distributed on $\{1, \dots, b_1\}$ and $\{b_1 + 1, \dots, p\}$, respectively.

Note that the variance of the usual Kendall's tau estimator $\hat{\tau}_{j_1,j_2}$ is already known [19]. We can also remark that this theorem always holds, even if Assumption 2 is not satisfied. However, in such a case, the estimators will have different expectations: this would be the average of the considered Kendall's tau over the corresponding pairs. Note that the conditional variance $\mathbb{V}\mathrm{ar}[\hat{\tau}^U|\mathbf{W}]$ is the variance of the estimator $\hat{\tau}^U$ for a fixed choice of pairs to average; it only measure the variability due to the randomness of the sample. On the contrary, the unconditional variance $\mathbb{V}\mathrm{ar}[\hat{\tau}^U]$ takes into account both the randomness of the sample and the randomness of the choice of the pairs.

Remark 5. If Assumption 2 holds, the first term in the variance of $\hat{\tau}^U$ vanishes, and $P_{B,2} = P_{R,2} = P_{D,2} = P_{j_1,j_2}$ and all S terms become equal to $P^2_{j_1,j_2}$. If the stronger Assumption 3 holds, all quantities P_{j_1,j_2,j_3,j_4} and Q_{j_1,j_2,j_3,j_4} becomes independent of the choice of indices (j_1, j_2, j_3, j_4). Therefore, all P-type averages becomes equal, as well as all Q-type averages.

Corollary 6. *Under* Assumption 2, $P^2 \coloneqq P^2_{j_1,j_2}$ *does not depend on the choice of* j_1, j_2, *and it holds that as* $n \to \infty$

$$n^{1/2}(\hat{\tau} - \tau) \xrightarrow{\text{law}} \mathcal{N}(0, V).$$

Here $\hat{\tau}$ denotes any of the estimators $\hat{\tau}_{j_1,j_2}$, $\hat{\tau}^B$, $\hat{\tau}^R$, $\hat{\tau}^D$, $\hat{\tau}^U$ given $\mathbf{W}$, $\hat{\tau}^U$, and the corresponding asymptotic variances V are, respectively, given by

$$V_{j_1,j_2} = V_{j_1,j_2}(\mathbb{P}_{\mathbf{X}}) \coloneqq 16(Q_{j_1,j_2} - P^2), \tag{6}$$

$$V^B = V^B(\mathbb{P}_{\mathbf{X}}) \coloneqq \frac{16}{b_1 b_2}(Q_{B,2} - P^2 + (b_1 + b_2 - 2)(Q_{B,1} - P^2) + (b_1 - 1)(b_2 - 1)(Q_{B,0} - P^2)), \tag{7}$$

$$V^R = V^R(\mathbb{P}_{\mathbf{X}}) \coloneqq \frac{16}{N}(Q_{R,2} - P^2 + (N - 1)(Q_{R,1} - P^2)), \tag{8}$$

$$V^D = V^D(\mathbb{P}_{\mathbf{X}}) \coloneqq \frac{16}{N}(Q_{D,2} - P^2 + (N - 1)(Q_{D,0} - P^2)), \tag{9}$$

$$V^{U|\mathbf{W}} = V^{U|\mathbf{W}}(\mathbb{P}_{\mathbf{X}}) \coloneqq \frac{16}{N}(Q_{\mathbf{W},2} - P^2 + (b_1 + b_2 - 2)(Q_{\mathbf{W},1} - P^2) + (b_1 - 1)(b_2 - 1)(Q_{\mathbf{W},0} - P^2)), \tag{10}$$

$$V^U = V^U(\mathbb{P}_{\mathbf{X}}) \coloneqq \frac{16}{N}\left(Q_{B,2} - P^2 + \frac{N - 1}{b_1 b_2 - 1}((b_1 + b_2 - 2)(Q_{B,1} - P^2) + (b_1 - 1)(b_2 - 1)(Q_{B,0} - P^2))\right), \tag{11}$$

where $\mathbb{P}_{\mathbf{X}}$ denotes the law of the random vector $\mathbf{X}$. Furthermore, these distributions are degenerate whenever the corresponding Q-type averages are equal to P^2. A sufficient condition for this to happen is when all Kendall's tau are equal to 1.

This corollary can straightforwardly be derived by combining Theorem A of [43, Section 5.5.1] and the computations of the corresponding ζ_1's in the proof of Theorem 4. The asymptotic normality of $\hat{\tau}^B$ was already known to hold under the stronger assumption of partial exchangeability [34, Theorem 1.2]. From the lengthy expressions in Theorem 4, we can derive the asymptotic variances in the setting where $n, b_1, b_2, N \to +\infty$.

Corollary 7. *Under* Assumption 2, *as* $n, b_1, b_2, N \to +\infty$, *we have the following equivalents*:

- $\mathbb{V}\mathrm{ar}[\hat{\tau}_{j_1,j_2}] \sim \frac{16}{n}(Q_{j_1,j_2} - P^2) = \frac{1}{n} \times V_{j_1,j_2}(\mathbb{P}_{\mathbf{X}})$,
- $\mathbb{V}\mathrm{ar}[\hat{\tau}^B] \sim \frac{16}{n}(Q_{B,0} - P^2) = \frac{1}{n} \times \lim_{b_1,b_2 \to +\infty} V^B(\mathbb{P}_{\mathbf{X}})$,
- $\mathbb{V}\mathrm{ar}[\hat{\tau}^R] \sim \frac{16}{n}(Q_{R,1} - P^2) = \frac{1}{n} \times \lim_{N \to +\infty} V^R(\mathbb{P}_{\mathbf{X}})$,
- $\mathbb{V}\mathrm{ar}[\hat{\tau}^D] \sim \frac{16}{n}(Q_{D,0} - P^2) = \frac{1}{n} \times \lim_{N \to +\infty} V^D(\mathbb{P}_{\mathbf{X}})$,

- $\mathbb{V}\mathrm{ar}[\hat{\tau}^U|\mathbf{W}] \sim \frac{16}{n}(Q_{\mathbf{W},0} - P^2) = \frac{1}{n} \times \lim_{b_1,b_2,N\to+\infty} V^{U|\mathbf{W}}(\mathbb{P}_{\mathbf{X}})$,
- $\mathbb{V}\mathrm{ar}[\hat{\tau}^U] \sim \frac{16}{n}(Q_{B,0} - P^2) = \frac{1}{n} \times \lim_{b_1,b_2,N\to+\infty} V^{U}(\mathbb{P}_{\mathbf{X}})$,

assuming, respectively, that Q_{j_1,j_2}, $Q_{B,0}$, $Q_{R,1}$, $Q_{D,0}$, $Q_{\mathbf{W},0}$, and $Q_{B,0}$ are strictly larger than P^2. If this is not the case, the corresponding variances converges to 0 at a rate faster than $O(1/n)$.

Surprisingly, note that the variances **do not depend on the block dimensions** as soon as they are large enough. This is also true if the block dimensions tends to the infinity at different rates. In the limit, the quality of the estimator will therefore not improve by averaging over additional elements in general. Note that this is coherent, since we do not assume the dependence to converge to 0, which would correspond to some mixing assumption. Therefore, averaging must have only a limited effect, as in the simpler statistical model $Y_i = \theta + \varepsilon_i$ where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ follows a centered exchangeable Normal distribution with correlation $\rho > 0$. In this case, the average $\overline{Y_n}$ is inconsistent for estimating θ since $\mathbb{E}[(\overline{Y_n} - \theta)^2] = n^{-2}\sum_{i,j=1}^{n}\mathbb{E}[\varepsilon_i\varepsilon_j] = 1/n + (n-1)\rho/n \to \rho$ as $n \to \infty$.

For large sample sizes, only the Q-type averages determine the levels of variance. 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 $Q_{R,1}$ (instead of a hypothetical $Q_{R,0}$).

Open problem 1. As all asymptotic variances are equal up to a constant, it seems logical to ask which is the best estimator. For this, we need to compare these constants. However, these constants are defined using eight-dimensional integrals, making explicit computations difficult.

Interestingly, the block estimator and the random estimator perform equally well in the limit. Hence, we can greatly reduce computation time by using the random estimator instead of the block estimator, while still maintaining a low asymptotic variance.

This is coherent with Theorem 1 of [35] which shows that under Assumption 3 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.

Finally, 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. Therefore, it makes sense to choose N large enough to attain the asymptotic regime, but not too large to keep a low computation time.

3 Fast estimation of conditional Kendall's tau matrix

We extend the aforementioned setting to the conditional setup, when a d-dimensional covariate $\mathbf{Z}$ is available taking values in $\mathcal{Z} \subset \mathbb{R}^d$. Formally, this means that we observe a sample $(\mathbf{X}_i, \mathbf{Z}_i)_{i=1,\ldots,n}$ of n independent and identically distributed replications of a random vector $(\mathbf{X}, \mathbf{Z}) \in \mathbb{R}^{p+d}$. The objective is now to estimate the $p \times p$ conditional Kendall's tau matrix $\mathbf{T}_{|\mathbf{Z}=\mathbf{z}} = [\tau_{j_1,j_2|\mathbf{Z}=\mathbf{z}}]_{1\le j_1,j_2\le p}$ for a given point $\mathbf{z} \in \mathcal{Z}$.

3.1 Estimation of conditional Kendall's tau

For construction of nonparametric estimates of the conditional Kendall's tau, let us start by recalling the expression of the conditional Kendall's tau, following [10]:

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{P}((X_{1,1} - X_{1,2})(X_{2,1} - X_{2,2}) > 0|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}) - \mathbb{P}((X_{1,1} - X_{1,2})(X_{2,1} - X_{2,2}) < 0|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}).$$

Following the approach of [10], we introduce a kernel-based estimator of $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ as follows:

$$\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := \frac{1}{1-s_n}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1,n}(\mathbf{z})w_{i_2,n}(\mathbf{z})g(\mathbf{X}_{i_1,(1,2)}, \mathbf{X}_{i_2,(1,2)}), \tag{12}$$

$$\text{where } g(\mathbf{X}_{i_1,(1,2)}, \mathbf{X}_{i_2,(1,2)}) := \operatorname{sign}((X_{i_1,1} - X_{i_2,1})(X_{i_1,2} - X_{i_2,2}))$$

with Nadaraya-Watson weights $w_{i,n}$ given by

$$w_{i,n}(\mathbf{z}) := \frac{\mathcal{K}_h(\mathbf{Z}_i - \mathbf{z})}{\sum_{k=1}^{n}\mathcal{K}_h(\mathbf{Z}_k - \mathbf{z})}, \tag{13}$$

and $s_n := \sum_{i=1}^{n} w_{i,n}^2(\mathbf{z})$, for some kernel $\mathcal{K}$ on $\mathbb{R}^d$ and a bandwidth sequence $h = h(n)$ converging to zero as $n \to \infty$. In this sense, $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$ is a smoothed estimator of $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \mathbb{E}[g(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}]$. The factor $1/(1 - s_n)$ tends to 1 as $n \to \infty$, and ensures that the estimated conditional Kendall's tau takes values in the whole interval $[-1, 1]$.

We adapt the simplified structural Assumption 2 by assuming that the underlying structural pattern applies to the conditional Kendall's tau matrix given $\mathbf{Z} = \mathbf{z}$.

Assumption 4. (Simplified structural assumption conditionally to $\mathbf{Z} = \mathbf{z} \in \mathcal{Z}$) Kendall's tau matrix of the random vector $\mathbf{X}$ can be written in the block form

$$\mathbf{T}_{|\mathbf{Z}=\mathbf{z}} = \begin{pmatrix} \cdot & \tau_{|\mathbf{Z}=\mathbf{z}}\mathbf{1} \\ \tau_{|\mathbf{Z}=\mathbf{z}}\mathbf{1} & \cdot \end{pmatrix}$$

for some value $\tau_{|\mathbf{Z}=\mathbf{z}} \in [-1, 1]$, where $\tau_{|\mathbf{Z}=\mathbf{z}}\mathbf{1}$ represent a block filled with the value $\tau_{|\mathbf{Z}=\mathbf{z}} \in [-1, 1]$ and the $\cdot$ represent any matrices of respective sizes $b_1 \times b_1$ and $b_2 \times b_2$ for some $b_1 \in \{1, \ldots, p\}$ and $b_2 := p - b_1$.

In terms of stock return modeling, Assumption 4 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 Assumption 4 only concern a fixed value of $\mathbf{z}$, and is quite general in the sense that it allows the existence of different block structures depending on the value of the conditional variable $\mathbf{z}$.

Let us denote the naive (unaveraged) conditional Kendall's tau matrix estimator as $\hat{\mathbf{T}}_{|\mathbf{Z}=\mathbf{z}} = [\hat{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}]_{1\le j_1,j_2\le p}$ with $\hat{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$ as in (12). As in the unconditional framework that was studied previously, we define averaged versions of the conditional estimators $\hat{\tau}^{B}_{|\mathbf{Z}=\mathbf{z}}$, $\hat{\tau}^{R}_{|\mathbf{Z}=\mathbf{z}}$, $\hat{\tau}^{D}_{|\mathbf{Z}=\mathbf{z}}$, $\hat{\tau}^{U}_{|\mathbf{Z}=\mathbf{z}}$, that, respectively, average over the whole block, the first row, the first diagonal, and uniformly chosen entries of the off-diagonal block.

3.2 Comparison of their asymptotic variances

Before proceeding with the asymptotic results, we need to formalize some regularity assumptions on the kernel $\mathcal{K}$, the covariate $\mathbf{Z}$ and the bandwidth sequence $h(n)$. Since we will give similar results as in [10, Proposition 9], where the case of the (bivariate) conditional Kendall's tau was treated, we give an adapted version of their assumptions.

Assumption 5.

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

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

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

(c) In addition, $\mathbb{E}[\mathcal{K}_h(\mathbf{Z} - \mathbf{z})] > 0$ for every $h > 0$.

These assumptions are classical in nonparametric statistics to obtain convergence rates of kernel-based estimators. The assumption of compactness of the support means that the estimators taken at different points in $\mathcal{Z}$ are independent if the sample size is large enough, since the bandwidth h_n tends to 0. The assumptions that the kernel is bounded and that $\int|\mathcal{K}|$ and $\int\mathcal{K}^2$ are finite rule out too much irregular kernels that have fat tails or irregular behavior. Higher-order kernels allows to obtain estimators that converge faster, under the assumption below that the joint density $f_{\mathbf{X},\mathbf{Z}}$ is smooth enough, see Section 1.2.1 of [44]. Assumption 5(c) ensures that the weights that appear in Equation (13) are well-defined asymptotically since the denominator will converge to a strictly positive value by the law of large numbers.

Assumption 6. 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 a neighborhood of $\mathbf{z}$ up to the order α. For every $0 \le k \le \alpha$ and every $1 \le j_1, \dots, j_\alpha \le d$, let

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

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 $h < 1$,

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

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

The regularity condition on $f_{\mathbf{X},\mathbf{Z}}$ can be interpreted in the classical way: smoother functions are easier to estimate than very irregular ones. Therefore, densities $f_{\mathbf{X},\mathbf{Z}}$ that are α-times differentiable allows to use a larger range of bandwidths for large α; this can be seen in the following Assumption 7.

Note that $C_{\mathbf{XZ},\alpha} \le C_{\alpha,p}\|K\|_\infty^2 \sup_{k\in\{0,\dots,\alpha\}} \sup_{\vec{j}\in\{1,\dots,d\}^k} \|\partial^k f_{\mathbf{X},\mathbf{Z}}/\partial z_{j_1} \dots \partial z_{j_k}\|_\infty^2$, where $C_{\alpha,p}$ is a constant that only depends on α and p. In this sense, the second part of Assumption 6 can be seen as a relaxed version of a uniform control on the higher-order derivatives of the density. Therefore, Assumption 6 is implied by Assumption 5 and by the (stronger) assumption that all partial derivatives of $f_{\mathbf{X},\mathbf{Z}}$ with respect to components of $\mathbf{z}$ exist up to the order α and are bounded.

Assumption 7. $nh_n^d \to \infty$ and $nh_n^{d+2\alpha} \to 0$ as $n \to \infty$.

This assumption controls the rate at which the sequence (h_n) tends to 0. It should tend to 0 fast enough but not too fast: the second condition controls the bias (see Equation (A16) in the proof), while the first condition ensures that the rate nh_n^d is meaningful and allows us to verify Lyapunov's condition (second part of [10, Section A.10]). We now present our main theoretical result on the joint asymptotic normality at different points of the conditioning variable $\mathbf{Z}$, at the nonparametric rate $(nh_n^d)^{1/2}$.

Theorem 8. (Joint asymptotic normality at different points) *Let $n' > 0$, and let $\mathbf{z}'_1, \dots, \mathbf{z}'_{n'}$ be a collection of n' points of $\mathcal{Z} \subset \mathbf{R}^d$ such that* Assumptions 4–7 *are satisfied for any choice $\mathbf{z} \in \{\mathbf{z}'_1, \dots, \mathbf{z}'_{n'}\}$. Then, as $n \to \infty$,*

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

where the diagonal matrix $\mathbf{H}$ is given by

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

and the asymptotic variance functions V are, respectively, defined in Equations (6)–(11).

The proof of this result is given in Appendix B.5.

Remark 9. Under the assumptions of Theorem 8, we have for a given $\mathbf{z} \in \mathcal{Z}$,

$$\lim_{n\to+\infty} (nh_n^d)^{1/2}(\hat{\tau}_{|\mathbf{Z}=\mathbf{z}} - \tau_{|\mathbf{Z}=\mathbf{z}}) \overset{\text{law}}{=} \frac{\int \mathcal{K}^2}{f_{\mathbf{Z}}(\mathbf{z})} \times \lim_{n\to+\infty} n^{1/2}(\hat{\tau}(\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}) - \tau_{|\mathbf{Z}=\mathbf{z}}),$$

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

Corollary 10. *Under the same assumptions as in* Theorem 8 *and by letting the sample size and dimensions tend to infinity, the following holds for* $\mathbb{P}_{\mathbf{Z}}$*-almost all* $\mathbf{z} \in \mathcal{Z}$,

- $\mathbb{V}\mathrm{ar}[\hat{\tau}_{j_1,j_2|\mathbf{Z}=\mathbf{z}}] \sim \frac{16\int \mathcal{K}^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}(Q_{j_1,j_2|\mathbf{Z}=\mathbf{z}} - P^2_{|\mathbf{Z}=\mathbf{z}})$,
- $\mathbb{V}\mathrm{ar}[\hat{\tau}^B_{|\mathbf{Z}=\mathbf{z}}] \sim \frac{16\int \mathcal{K}^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}(Q_{B,0|\mathbf{Z}=\mathbf{z}} - P^2_{|\mathbf{Z}=\mathbf{z}})$,
- $\mathbb{V}\mathrm{ar}[\hat{\tau}^R_{|\mathbf{Z}=\mathbf{z}}] \sim \frac{16\int \mathcal{K}^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}(Q_{R,1|\mathbf{Z}=\mathbf{z}} - P^2_{|\mathbf{Z}=\mathbf{z}})$,
- $\mathbb{V}\mathrm{ar}[\hat{\tau}^D_{|\mathbf{Z}=\mathbf{z}}] \sim \frac{16\int \mathcal{K}^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}(Q_{D,0|\mathbf{Z}=\mathbf{z}} - P^2_{|\mathbf{Z}=\mathbf{z}})$,
- $\mathbb{V}\mathrm{ar}[\hat{\tau}^U_{|\mathbf{Z}=\mathbf{z}}|\mathbf{W}] \sim \frac{16\int \mathcal{K}^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}(Q_{\mathbf{W},0|\mathbf{Z}=\mathbf{z}} - P^2_{|\mathbf{Z}=\mathbf{z}})$,
- $\mathbb{V}\mathrm{ar}[\hat{\tau}^U_{|\mathbf{Z}=\mathbf{z}}] \sim \frac{16\int \mathcal{K}^2}{nh^d f_{\mathbf{Z}}(\mathbf{z})}(Q_{B,0|\mathbf{Z}=\mathbf{z}} - P^2_{|\mathbf{Z}=\mathbf{z}})$,

assuming, respectively, that $Q_{j_1,j_2|\mathbf{Z}=\mathbf{z}}$, $Q_{B,0|\mathbf{Z}=\mathbf{z}}$, $Q_{R,1|\mathbf{Z}=\mathbf{z}}$, $Q_{D,0|\mathbf{Z}=\mathbf{z}}$, $Q_{\mathbf{W},0|\mathbf{Z}=\mathbf{z}}$, *and* $Q_{B,0|\mathbf{Z}=\mathbf{z}}$ *are strictly larger than* $P^2_{|\mathbf{Z}=\mathbf{z}}$, *where conditional versions of P, Q, and their averages are defined with the same expression as in the previous section, but applied to the conditional law* $\mathbb{P}_{\mathbf{X}|\mathbf{Z}=\mathbf{z}}$.

If these assumptions are not met, the corresponding variances converges to 0 at a rate faster than $O(1/(nh^d))$.

As seen earlier, we remark that the asymptotic variances have analogous expressions to that of their unconditional counterparts. Therefore, all averaging estimators exhibit a lower asymptotic variance than the naive conditional Kendall's tau estimator. Also, the row averaging estimator intuitively performs worse 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 Q 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. Finally, it again 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.

4 Simulation study

We perform a simulation study to assess the finite sample properties of our estimators. First, in Section 4.1, we compare the unconditional estimators by studying their variances and computation times for varying block and sample sizes. In Section 4.2, 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` [38] on the DelftBlue supercomputer [7]. For simplicity, we choose $N = \min(b_1, b_2)$, so that diagonal, row and random estimators average over the same number of terms.

4.1 Unconditional Kendall's tau

In the unconditional framework, we compare the block, row, diagonal, random, and naive Kendall's tau matrix estimators. 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 here 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 `wdm` in the `R` package `wdm` [31], which can efficiently calculate sample and weighted Kendall's tau with time complexity $O(n \log(n))$. The row, diagonal, random, and block estimators are now available as part of the package `ElliptCopulas` [13] and can be computed using the function `KTMatrixEst`.

In each simulation, data is generated using a meta-elliptical copula [1,12,17,18]. A copula is said to be meta-elliptical if it is the copula of a distribution with density $|\mathbf{\Sigma}|^{-1/2} g(\mathbf{x}^\top \Sigma \mathbf{x})$ for a covariance matrix Σ (here chosen to be a correlation matrix) and a function $g : \mathbb{R}_+ \to \mathbb{R}_+$, called the generator of the meta-elliptical copula. This means that we simulate data from the Gaussian copula, or from other meta-elliptical copulas with different generators. Note that for meta-elliptical copulas, the matrix Σ can be directly obtained from the Kendall's tau matrix. 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.

Open problem 2. Simulating from meta-elliptical copulas is easy, for example, using the `ElliptCopulas` package [13] as they rely on elliptical distributions, which are well-understood distributions. Constructing explicit nonmeta-elliptical models of dependence that satisfy Assumption 2 seems difficult (unless both groups are independent) and is left for future research. Indeed, even in the low-dimensional case, where $b_1 = b_2 = 2$ (so that the two blocks are {1,2} and {3,4}), a D-vine (for example) would give a decomposition of the copula $c_{1,2,3,4} = c_{1,2}c_{2,3}c_{3,4}c_{1,3|2}c_{2,4|3}c_{1,4|2,3}$ of (X_1, X_2, X_3, X_4). So Kendall's tau $\tau_{2,3}$ can be easily chosen through the specification of the copula $c_{2,3}$. However, this would not give an explicit expression for the other interblocks Kendall's taus $\tau_{1,3}$, $\tau_{1,4}$, $\tau_{2,4}$. Indeed, they depend in a complicated way of all the copulas' and conditional copulas' parameters through integration.

As obtained in Theorem 4, the variances depend on the averages of the auxiliary quantities P, Q, and S 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.

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.

4.1.1 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 32×32. The true Kendall's tau values are fixed in the following way: the intragroup Kendall's tau are $\tau_{1,1} = \tau_{2,2} = 0.5$ and the intergroup Kendall's tau is $\tau_{1,2} = 0.3$. We examine data generated from the Gaussian distribution. For each estimator, the MSEs are calculated using 3,000 replications. The results can be found on a log–log scale in Figure 2.

In Figure 2, we clearly observe almost 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 also that this 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 compute the average computation time. The results are shown in Figure 3 on a log–log scale. It shows that the computation times gradually increase with the sample size, to a point where they appear to scale almost linearly with each other. These observations are in line with the computation time $O(n \log(n))$ of the pairwise sample Kendall's tau estimator. 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.

4.1.2 Effect of the block size

We first study the behavior of the MSE with respect to varying block sizes with off-diagonal block Kendall's taus of 0.3 and diagonal block Kendall's taus of 0.5. 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

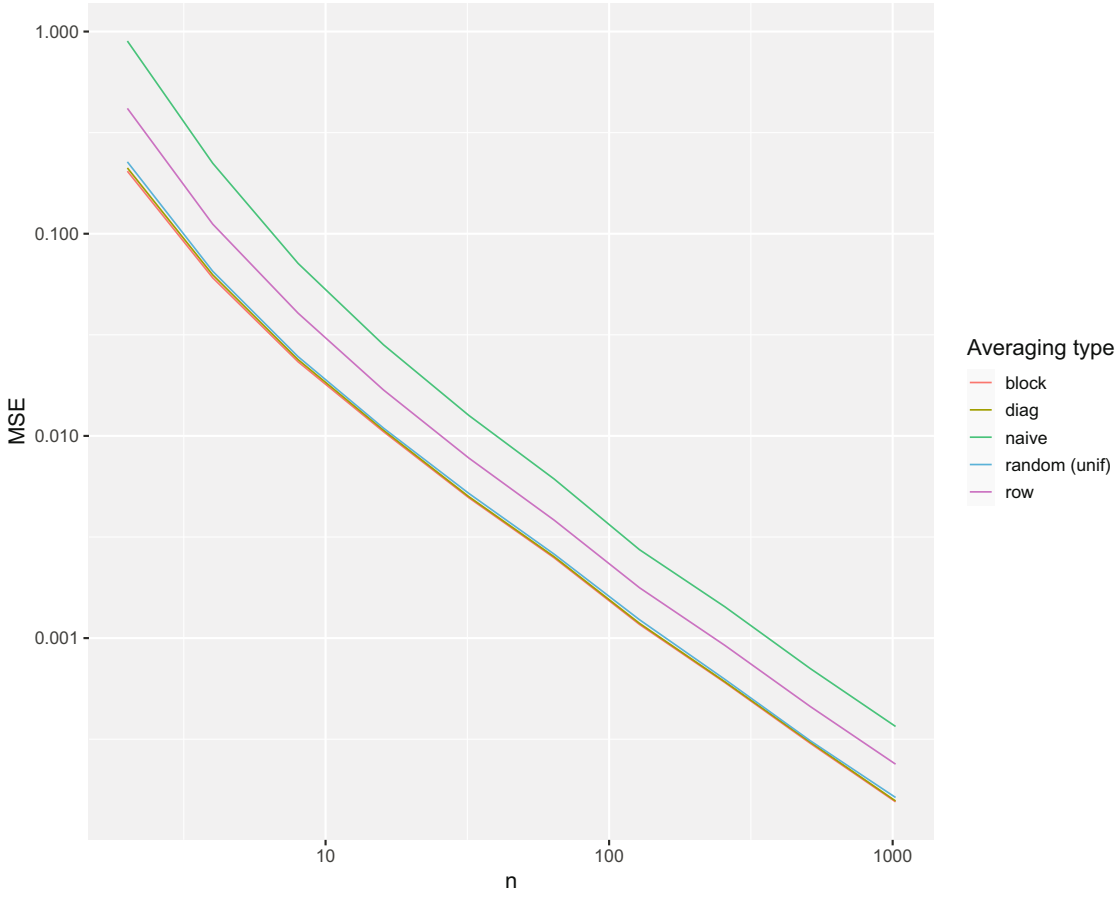

Figure 2: Log–log plots of the MSE of the estimators $\hat{\tau}_{j_1,j_2}$ ("naive"), $\hat{\tau}^B$ ("block"), $\hat{\tau}^R$ ("row"), $\hat{\tau}^D$ ("diag") and $\hat{\tau}^U$ ("random" with uniform selection of pairs) as a function of the sample size. The diagonal block Kendall's taus are set at 0.5 and the off-diagonal block values at 0.3.

Gaussian distributions. The MSEs are calculated using 3,000 replications. See Figure 4 for a log–log plot of the MSEs as a function of the block size.

Figure 4 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, 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 and the random averaging estimator perform considerably less.

Furthermore, we find that the relative difference between the diagonal and block estimator is largest for small dimensions, but they are still well within a factor of 1.5 of each other. As the dimension increases, the MSEs of the diagonal estimator converge rapidly to that of the block estimator, again confirming that the block and diagonal estimators have close variances for large block dimensions. A more detailed presentation of the interplay between the sample size n and the size b_1 and b_2 of the blocks is available in the Supplementary file "MSE_n,b1,b2.pdf."

4.1.3 Effect of the value of the true Kendall's taus

In this section, we fix the sample size $n = 16$ and the block sizes $b_1 = b_2 = 4$, and we change the value of Kendall's tau. The MSE is displayed in Figure 5 for different combinations of $\tau_{1,1}$, $\tau_{1,2}$, and $\tau_{2,2}$. Note that some combinations are not present due to the constraints presented in Section 2.2.

As expected, the relative order of the estimator is the same for all values of the Kendall's tau. A more detailed version of this figure is available as supplementary material (File "MSE_tau.pdf"), showing the same

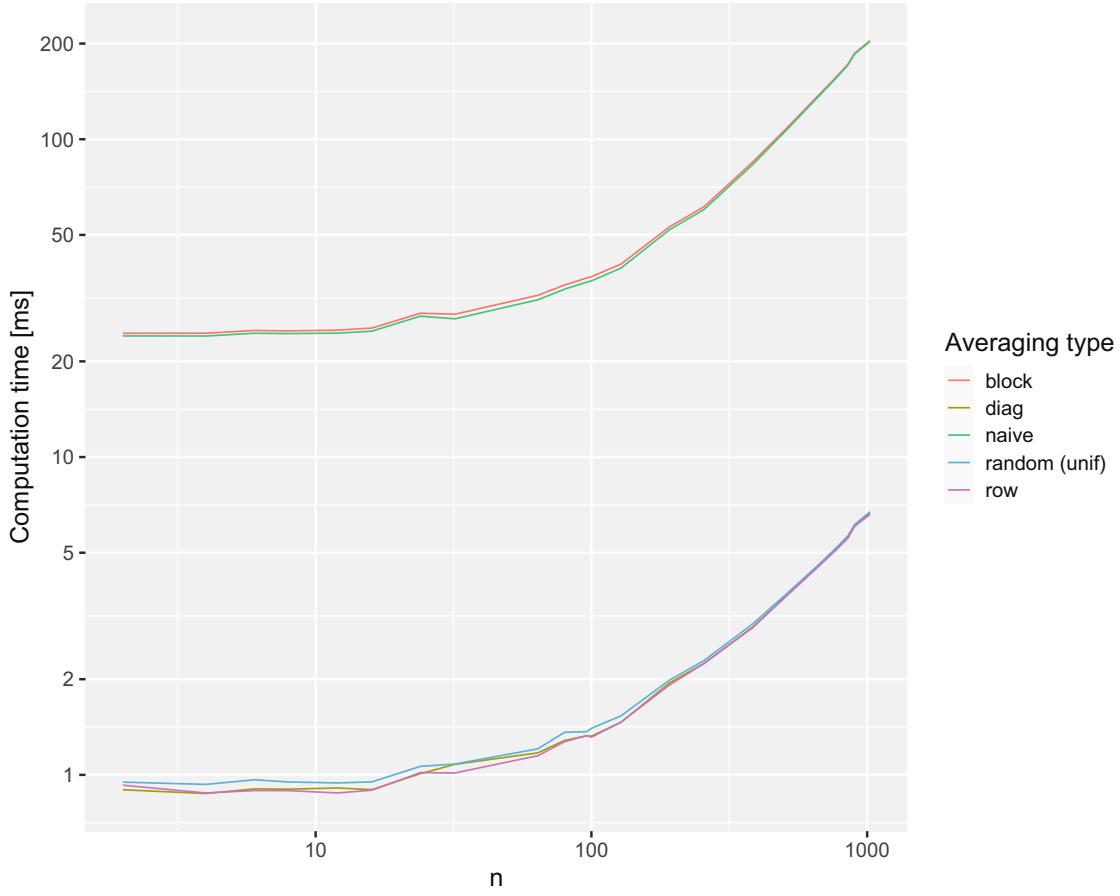

Figure 3: Log–log plot of the mean computation time [ms] of the estimators $\hat{\tau}_{j_1,j_2}$ ("naive"), $\hat{\tau}^B$ ("block"), $\hat{\tau}^R$ ("row"), $\hat{\tau}^D$ ("diag"), and $\hat{\tau}^U$ ("random" with uniform selection of pairs) as a function of the sample size, calculated using a block size of 32.

phenomena for a larger range of value for the Kendall's taus. Interestingly, the performance of all estimators is very similar when both intragroup Kendall's taus are equal to 0.9. Indeed, in this case, the variables in each blocks are mostly the same, and then averaging does not change the situation anymore.

4.1.4 Effect of the copula

In this section, we fix the sample size, block sizes, and Kendall's tau value. We vary instead the copula of the distribution. For this, we use different meta-elliptical copulas, because of their natural relationships between Kendall's tau and the underlying correlation matrix. These meta-elliptical copulas are defined as the copulas of the distributions with densities $|\mathbf{\Sigma}|^{-1/2} g(\mathbf{x}^\top \Sigma \mathbf{x})$ for a covariance matrix Σ (here chosen to be a correlation matrix) and a function $g : \mathbb{R}_+ \to \mathbb{R}_+$ respectively, chosen as

- $x/(1 + x^3)$,
- $1/(1 + x^2)$,
- $\exp(-x) \times |\cos(x)|$,
- $\exp(-x/2) + \exp(-x) \times |\cos(x)|$,
- $\exp(-x) + \exp(-x/3) \times \cos(x)^2$.

The results are displayed in Figure 6. We can observe that the relative order of the estimators is mostly the same as mentioned earlier, with the averaging estimator the best and the no-averaging estimator having the highest mean square error.

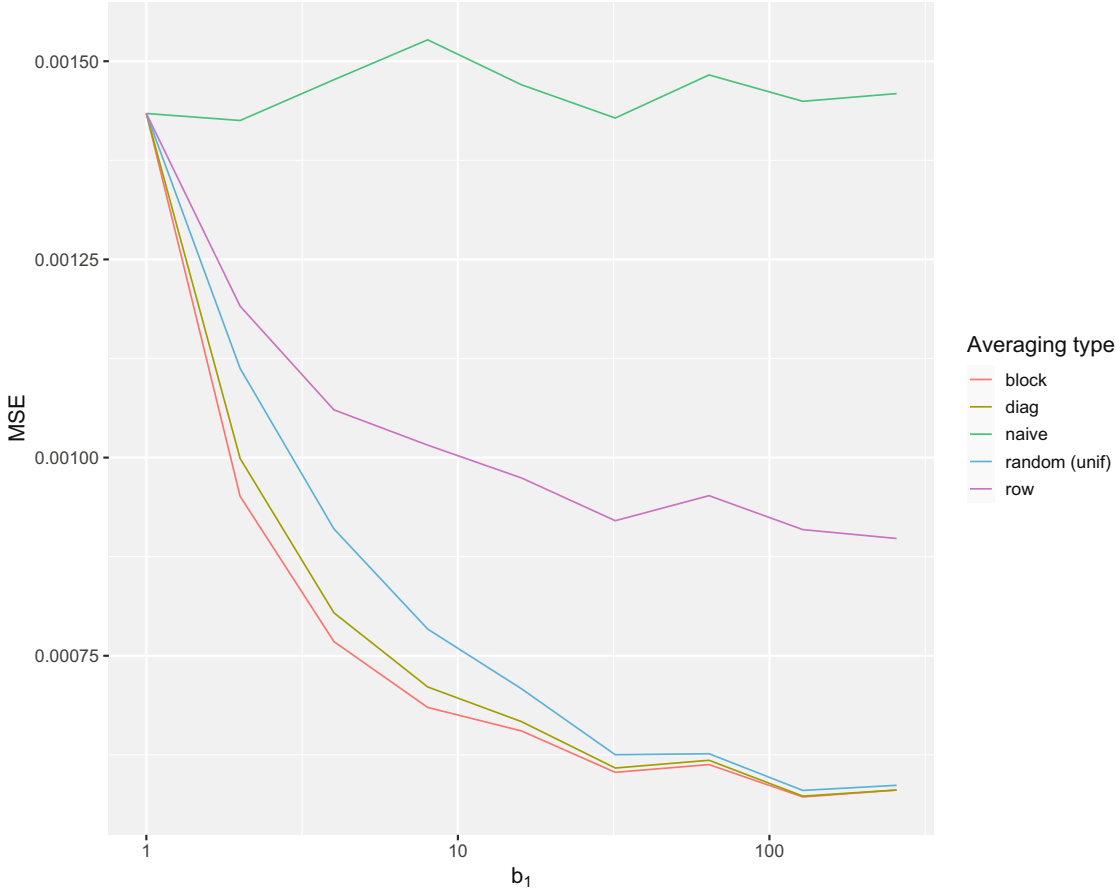

Figure 4: Log–log plots of the mean squared error of the estimators $\hat{\tau}_{j_1,j_2}$ ("naive"), $\hat{\tau}^B$ ("block"), $\hat{\tau}^R$ ("row"), $\hat{\tau}^D$ ("diag"), and $\hat{\tau}^U$ ("random" with uniform selection of pairs) as a function of the block size. The diagonal block Kendall's taus are set at 0.5 and the off-diagonal block values at 0.3.

4.2 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 with the Epanechnikov kernel [44], and the estimation procedures are now available in the function `CKTmatrix.kernel` of the `R` package `CondCopulas` [8].

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 other distribution yield 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 only 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 4.2.1, we examine the accuracy and computational efficiency of the estimates under varying sample size. A similar analysis for the effect of the block size is done in Appendix A.1. To this end, we set Kendall's tau in the off-diagonal blocks to $0.1Z$. 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 conditionally to $Z = z$ for any $z \in [0, 1]$. It follows that the biases of the pairwise estimates in the off-diagonal block are all equal and thus

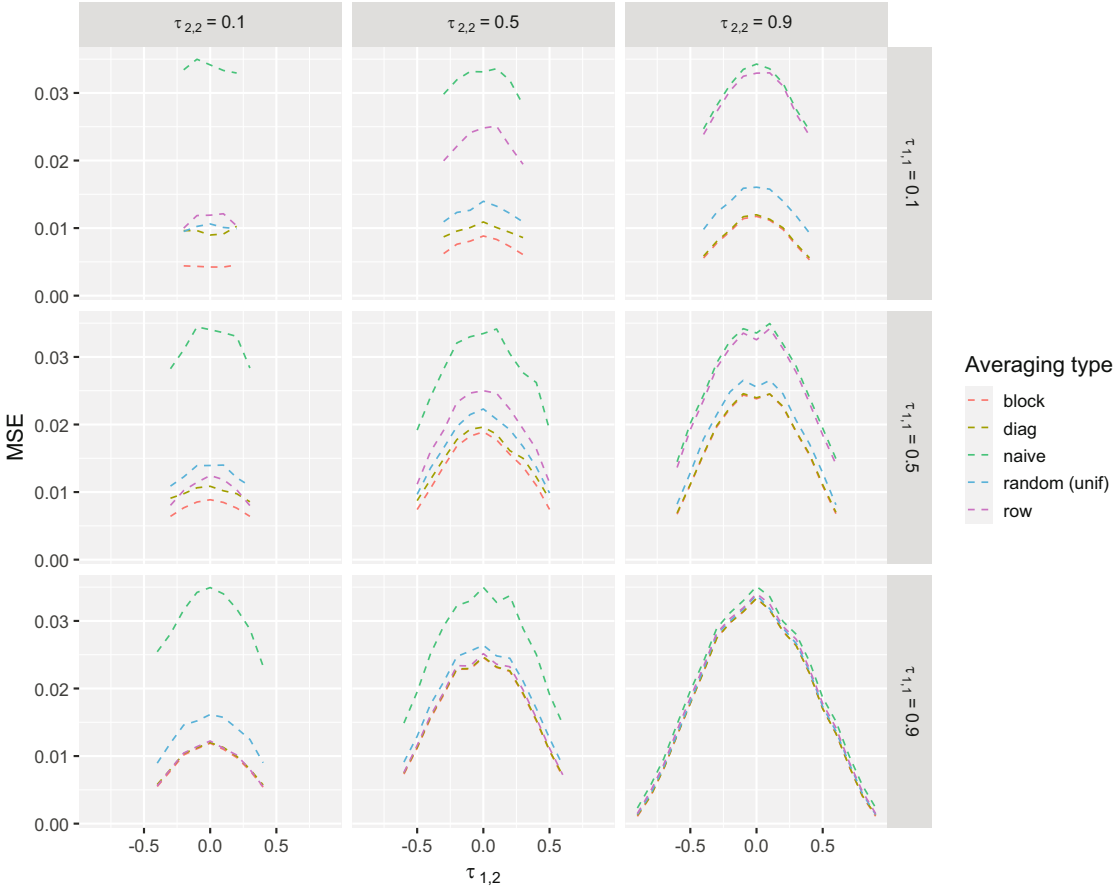

Figure 5: MSE as a function of the intergroup Kendall's tau $\tau_{1,2}$ for different combinations of $\tau_{1,1}$ and $\tau_{2,2}$. We use the sample size $n = 16$ and the block sizes $b_1 = b_2 = 4$.

that averaging over them does not change the total bias. 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 4.2.2, 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

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

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 (MISEs) computed by averaging the MSEs of conditional estimates computed at conditioning points ranging from 0 to 1 in steps of 0.1.

4.2.1 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 integrated variance IVAR $\coloneqq \int_{z=0}^{1} \mathbb{V}\mathrm{ar}[\hat{\tau}_{|Z=z}]\mathrm{d}z$ is estimated as the average of sample variances for the grid points $z_i = i/n$, $i = 0, \ldots, 10$, using 3,000 replications. Supplementary plots showing the influence of the sample size on each term $\mathbb{V}\mathrm{ar}[\hat{\tau}_{|\mathbf{Z}=z_i}]$ are available in the Appendix A, Figure A1 while the influence of the sample size on IVAR is shown in Figure 7.

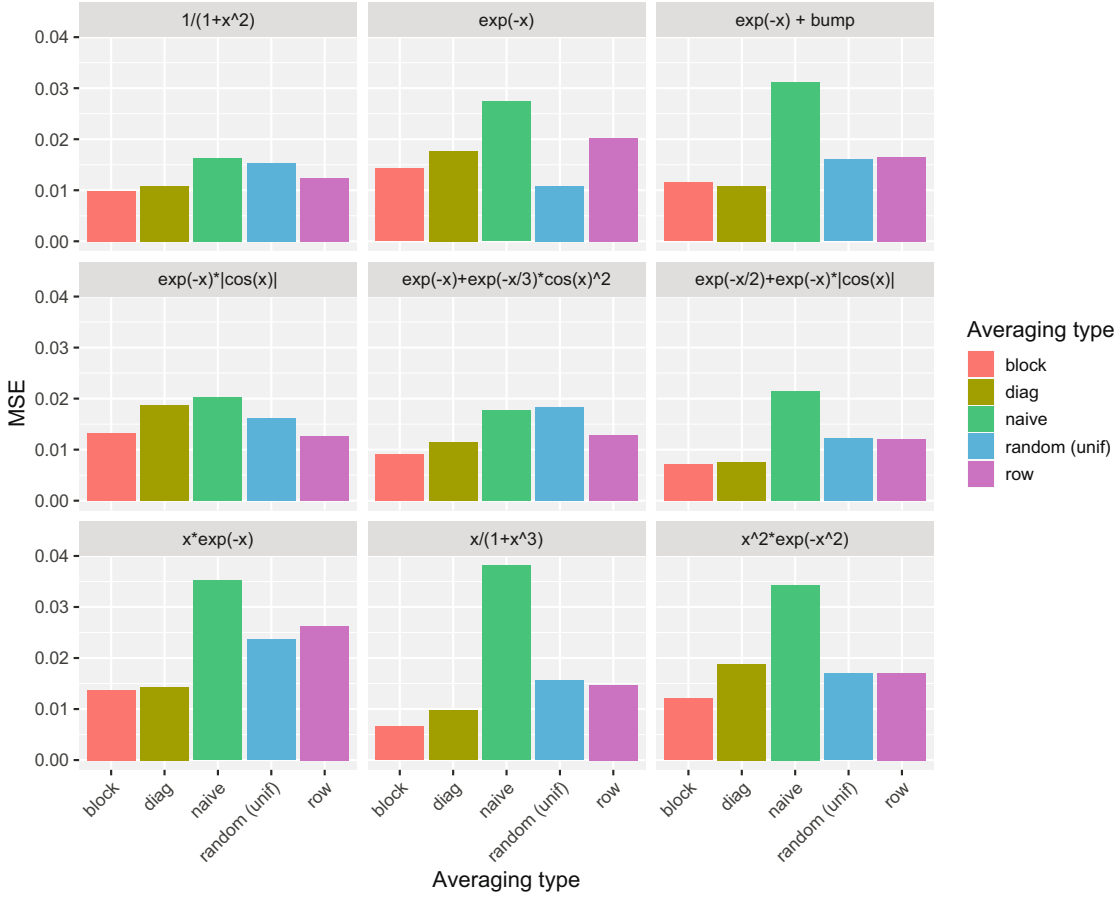

Figure 6: MSE for different meta-elliptical copulas and different estimators.

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 bandwidths are naturally smaller for larger 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 computed using 500 replications and are represented on log–log scale in Figure 8. Here, it is seen that the computation times gradually increase with the sample size to a point where they appear to scale quadratically with each other. This behavior 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.

4.2.2 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 estimated using 100 replications (Figure 9).

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 therefore, we need to pick a smaller bandwidth. Finally, it should be noted that as the bandwidth increases the effect of averaging is less and

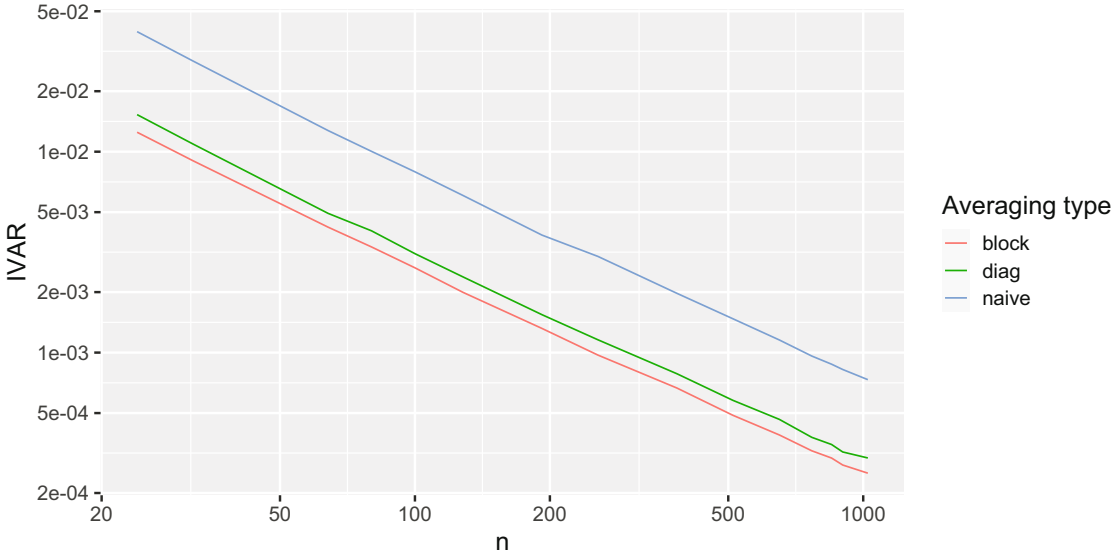

Figure 7: Log–log plots of the conditional estimators' integrated variance as a function of the sample size, using a block size of 4 and a bandwidth of 0.5.

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.

5 Application to real data

In this section, we study the behavior 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 5.2, we describe the methods used to estimate the VaR input parameters. The results are presented in Section 5.3, where backtesting is applied to asses the viability. All computations have been done using the R statistical environment [38].

5.1 Value at risk for elliptical distributions

The value at risk (VaR) 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 [29]. It is used to quantify potential losses over a specific time frame of some financial entity or portfolio of assets. We will follow the approach of [37,42], in which explicit expressions for the VaR of elliptical distributions was derived. For the reader's convenience, we recall these expressions in the present section.

Let X be the loss of a given portfolio, i.e., $X > 0$ means that the portfolio manager is loosing X euros. The VaR at level $\alpha \in (0, 1)$ is defined as the quantile of X at level $(1 - \alpha)$. To estimate 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 $[0, t]$ 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.

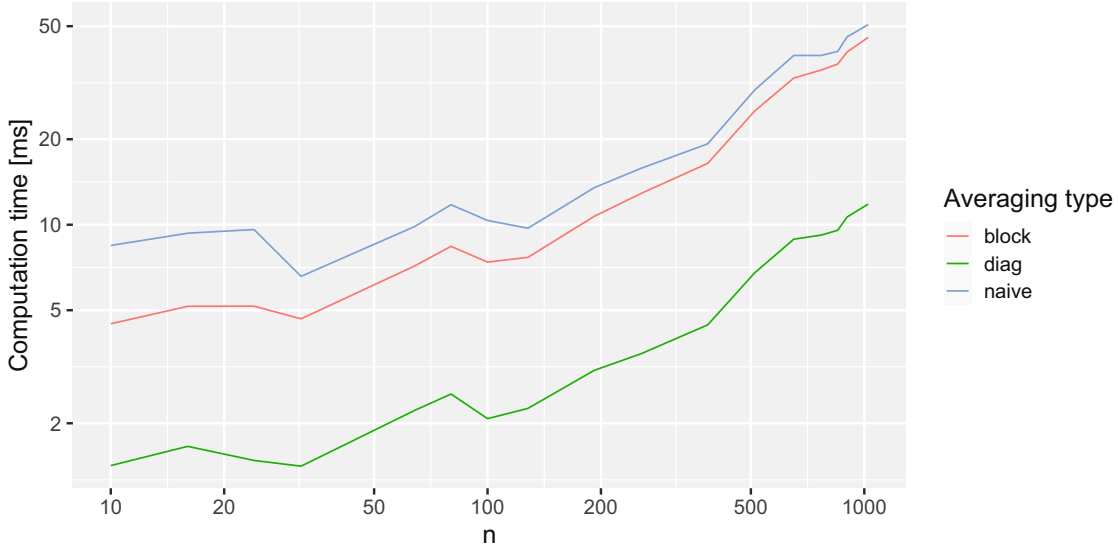

Figure 8: Log–log plot of the estimated mean computation time [ms] of the conditional estimator as a function of the sample size, for a block size of 4.

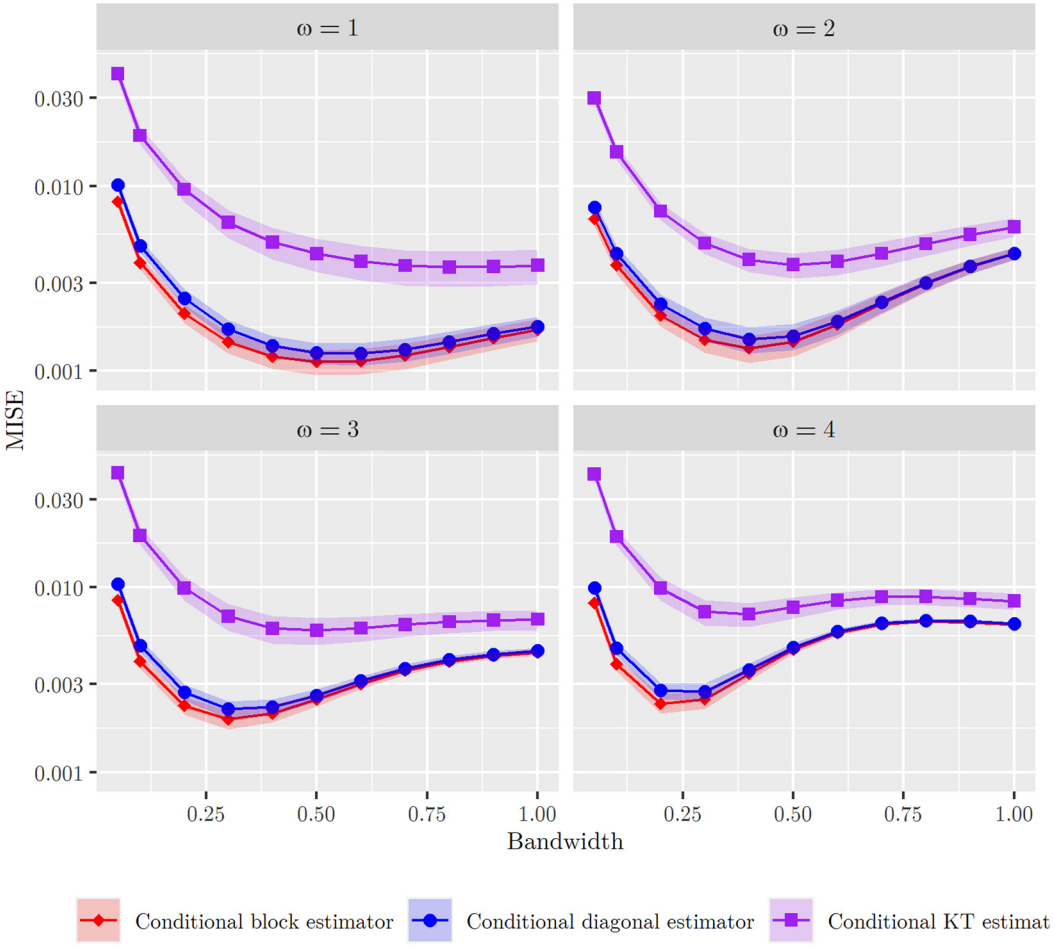

Figure 9: Log-plots of the conditional estimators' MISEs as a function of the bandwidth for different frequencies ω including 95% confidence intervals, for a sample size of 200 and a block size of 8. "Conditional KT estimate" refers to the naive estimator of conditional Kendall's tau $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}}$.

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

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

When considering elliptically distributed risk factors, we cannot simply use the delta-normal approach to calculate the VaR, as it relies on the stronger assumption of normality. A generalization of the delta-normal method was derived for the class of elliptical distributions in [42].

Let us start by noting that the VaR of the portfolio profits and losses $\Delta\Pi(t)$ can be rewritten as $\mathbb{P}(\Delta\Pi < -\text{VaR}_\alpha) = \alpha$. 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:

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

where $\boldsymbol{\delta}$ denotes the vector of weights $(\delta_1, \ldots, \delta_p)$. After several changes of variables, we obtain

$$\alpha = |S_{p-2}| \int_0^\infty r^{p-2} \int_{-\infty}^{\frac{-\boldsymbol{\delta}\boldsymbol{\mu}^T - \text{VaR}_\alpha}{|\boldsymbol{\delta}\mathbf{A}|}} g(z_1^2 + r^2) \mathrm{d}z_1 \mathrm{d}r, \tag{14}$$

where $|S_{p-2}| \coloneqq 2\pi^{\frac{p-1}{2}}/\Gamma(\frac{p-1}{2})$. Let us now introduce the function

$$G(s) = \frac{2\pi^{\frac{p-1}{2}}}{\Gamma(\frac{p-1}{2})}\int_{-\infty}^{-s}\int_{0}^{\infty} r^{n-2}g(z_1^2 + r^2)\mathrm{d}r\mathrm{d}z_1 = \frac{\pi^{\frac{p-1}{2}}}{\Gamma(\frac{p-1}{2})}\int_{s}^{\infty}\int_{z}^{\infty}(u - z^2)^{\frac{p-3}{2}}g(u)\mathrm{d}u\mathrm{d}z, \tag{15}$$

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

$$\alpha = G(q_{\alpha,p}^g). \tag{16}$$

It then finally follows from expressions (14) and (15) that the Delta-Elliptic VaR is given by

$$\mathrm{VaR}_\alpha = -\boldsymbol{\delta}\boldsymbol{\mu}^T + q_{\alpha,p}^g|\boldsymbol{\delta}\mathbf{A}| = -\boldsymbol{\delta}\boldsymbol{\mu}^T + q_{\alpha,p}^g\sqrt{\boldsymbol{\delta}\boldsymbol{\Sigma}\boldsymbol{\delta}^T}. \tag{17}$$

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

5.2 Estimation procedure

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 has been downloaded from Yahoo Finance. The complete list of all shares involved is available in Appendix C. We will estimate the portfolio'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.

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 simply use the method `GW_Ward` method from package `seriation` [22], along with a few manual adjustments. The resulting reordering corresponds to four large groups, which are specified further in Appendix C. See Figure 1 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.

Based on the groups displayed in Figure 1, the objective is to compute 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. To estimate the standard deviations and averages over the same period, we use the sample mean and sample standard deviation.

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` from the `EllipticCopulas` package [13], which implements Liebscher's procedure [25].

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. Choice of tuning parameters in this setting is discussed in [41]. For simplicity, we use Silverman's rule of thumb for bandwidth selection to estimate elliptical density generators [37], which for a sample size of n is given by

$$h = 1.06\sqrt{\mathrm{Var}[\widehat{\boldsymbol{\xi}}]}n^{1/5}, \tag{18}$$

where

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

for $i = 1, \ldots, n$ and $p = 240$. Here, $\mathbf{x}_i$ stands for the vector of log returns at the ith date, $\hat{\boldsymbol{\Sigma}}$ stands for one of the covariance matrix estimates and $\boldsymbol{\mu}$ stands for the log returns' sample mean. Clearly, by using this bandwidth selection method, the use of different Kendall's tau matrix estimators yields different values for the bandwidth. To get a better idea of the effects of the bandwidth choice, we also consider several deterministic bandwidths, and compare the performance of the estimators for each of them.

Finally, we can numerically solve the transcendental equation as given in (14) to arrive at the corresponding quantiles. As such, we have discussed all ingredients for calculating the VaR as in (17). To test the results, we perform backtesting on two intervals, one in the future from 1 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.

5.3 Results

We compute the portfolio's 5 and 10% VaR values by following the estimation procedure described in Section 5.2. Table 1 shows the quantile estimates obtained by solving the transcendental equation for each of the different density generator estimates. The density generators were estimated 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 2 shows the VaR estimates for each of the different estimators and bandwidths, and also the backtested VaR values. As discussed in Section 5.2, backtests were conducted at two intervals, interval 1 refers to the upcoming interval from 1 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.

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.

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 $q_{\alpha,p}^{\hat{g}_h}$		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: 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 ($h = 40.5$)		

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 estimate 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. This lack of performance is hard to relate directly to the block sizes. Indeed, the "naive" estimator of Kendall's tau (i.e., without any averaging) corresponds to the case where all the blocks have size 1, so all block sizes are as small as possible. Still the VaR estimates from this estimator are the worst.

To obtain 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. Tables 3 and 4 show the number of exceedances in interval 1 and interval 2, respectively.

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 bandwidths. 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 clear additional benefits compared to using only the row or diagonal estimators, which are computationally much cheaper.

Table 3: The number of exceedances of the estimated 5 and 10% VaRs during backtesting interval 1, from 1 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 ($h = 40.5$)	

6 Conclusion

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 depend 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).

We have formally derived variance expressions, which showed 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 similar asymptotic variances, whereas that of the row estimator was different. In most examples, the diagonal estimator performed the best, but a formal characterization of the set of such copulas is left for future work. Under light assumptions, we have shown that 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 aforementioned 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 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 different meta-elliptical copulas. It was furthermore confirmed that the diagonal and the block estimator indeed have the lowest asymptotic variance in most cases, with the block estimator converging the fastest, though closely followed by the diagonal estimator. This emphasizes the practical use of the diagonal estimator.

We remarked again that the conditional estimators' variances decrease in a similar fashion for growing block dimensions. 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.

Finally, 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 clustering the sample Kendall's tau matrix, 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 these estimates, the portfolio's 5 and 10% VaR values were estimated. 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. 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.

Acknowledgements: The authors thank Thomas Nagler for useful comments on a previous draft, and Dorota Kurowicka for a discussion and references that lead to Section 2.2. The authors also thank the Associate Editor and two anonymous reviewers for their useful comments which significantly improved the manuscript.

Funding information: The authors state that no specific funding was involved in this research.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results, and approved the final version of the manuscript. RVDS: conceptualization, formal analysis, software, investigation, writing. AD: conceptualization, methodology, formal analysis, software, writing, supervision.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: All data were obtained using the `getSymbols` function from the `quantmod` R package [40], fetching the data from Yahoo Finance.

Appendix

A Additional figures

Figure A1

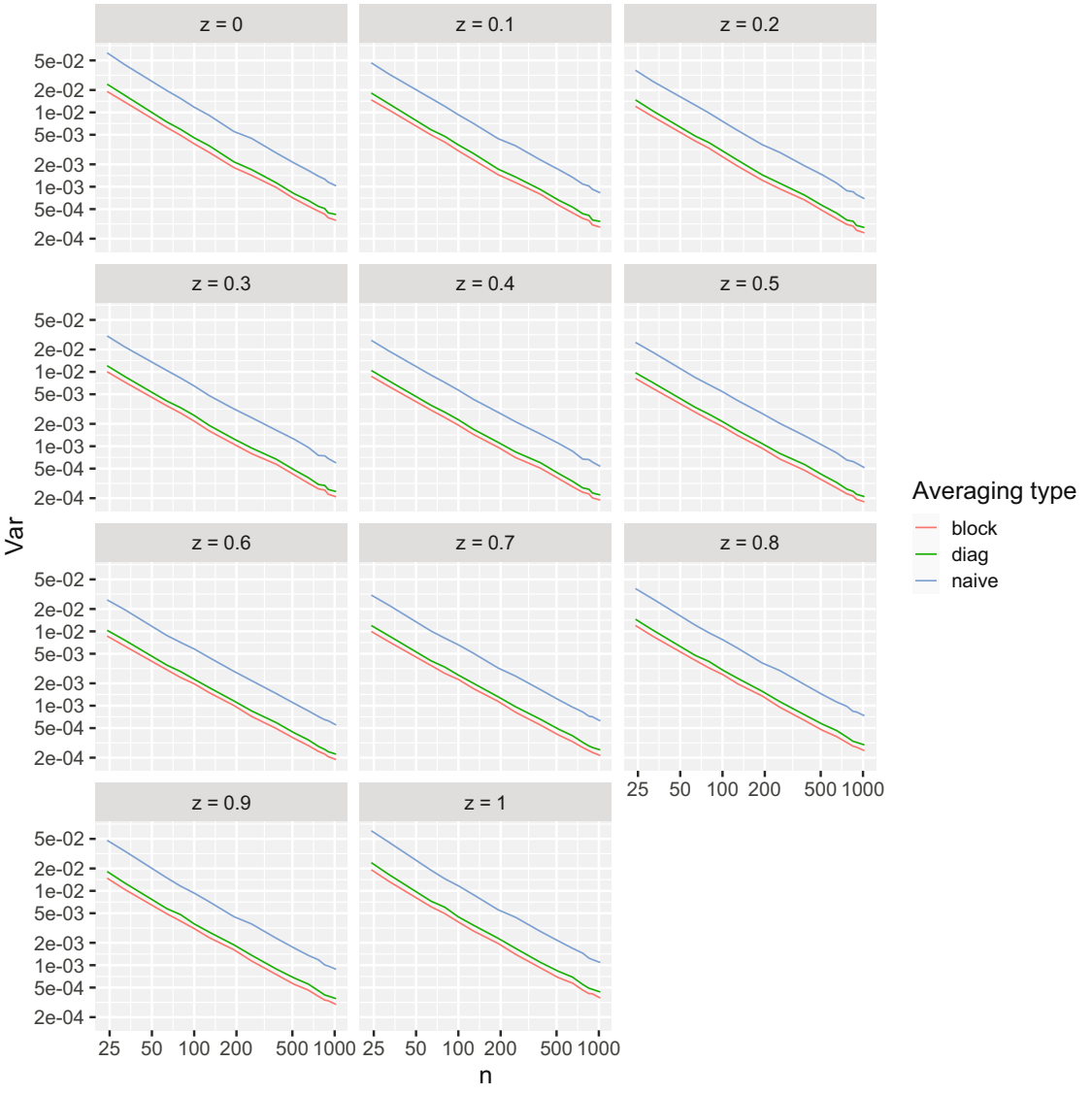

Figure A1: Log–log plots of the conditional estimators' variances as a function of the sample size on several conditioning points, using a block size of 4 and a bandwidth of 0.5.

A.1 Effect of the block size in the conditional framework (Section 4.2)

We first study the estimators' variance under varying block dimensions. To run a sufficient number of replications, we set the sample size to 20 and consequently the bandwidth to 0.5. The variances and 95% confidence intervals are estimated using 30,000 replications. For each grid point z, the resulting sample variances are displayed on log–log scale in Figure A2.

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

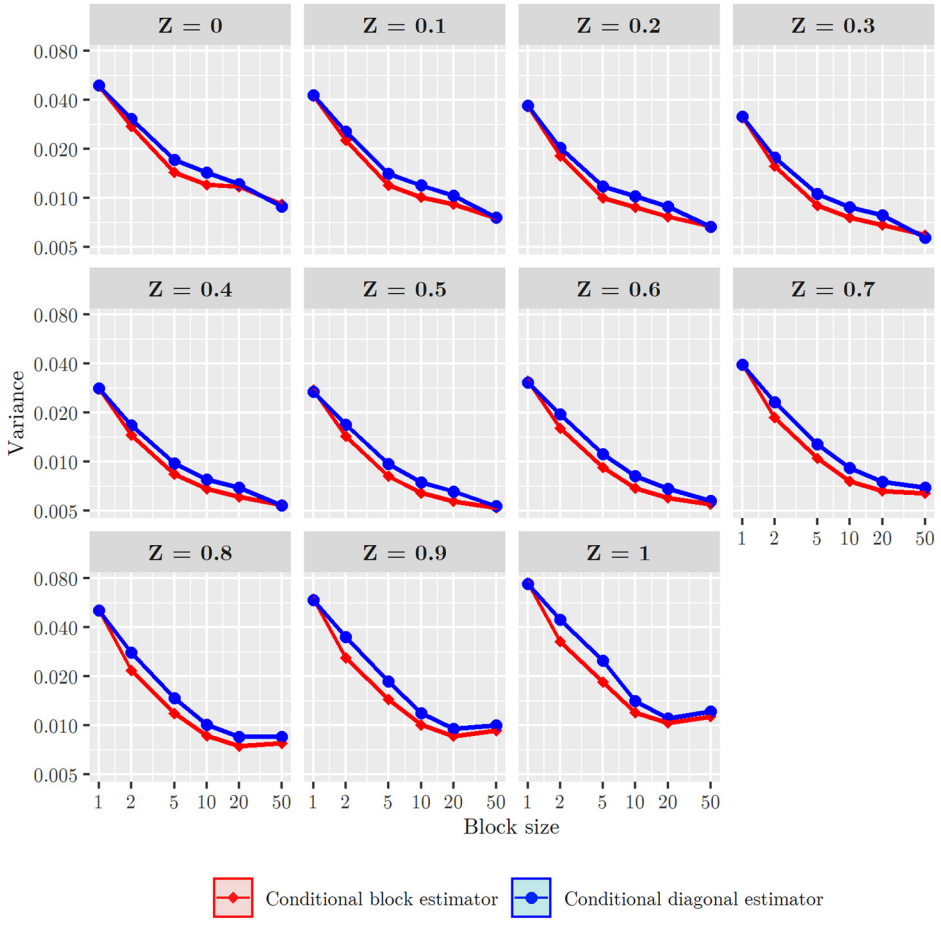

Figure A2: Log–log plots of the conditional estimators' variances as a function of the block size on several conditioning points including 95% confidence intervals, for a sample size of 20 and a bandwidth of 0.5.

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 is studied in Section 4.2.2. 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.

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 computation time than their unconditional counterparts, as was also seen in Figure 8. For the sake of completeness, we still include a plot of the average computation time against the block size, see Figure A3. The results correspond to estimating the off-diagonal block conditional Kendall's taus simultaneously on the 11 grid points and follow from 10,000 replications with a sample size of 150. As expected, the block estimator scales quadratically with the 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.

B Proofs

B.1 Proofs for Section 2.2

Proof of Proposition 1. In this proof, we will need the following notation: for an integer $i \in \{1, \ldots, p\}$, we define the vector $\mathbf{1}_i$ as the vector with a 1 at the ith component and 0 elsewhere. For a set $I \subset \{1, \ldots, p\}$, we define $\mathbf{1}_I := \sum_{i \in I} \mathbf{1}_i$, which the vector with 1 at the components in I and 0 elsewhere. Note that

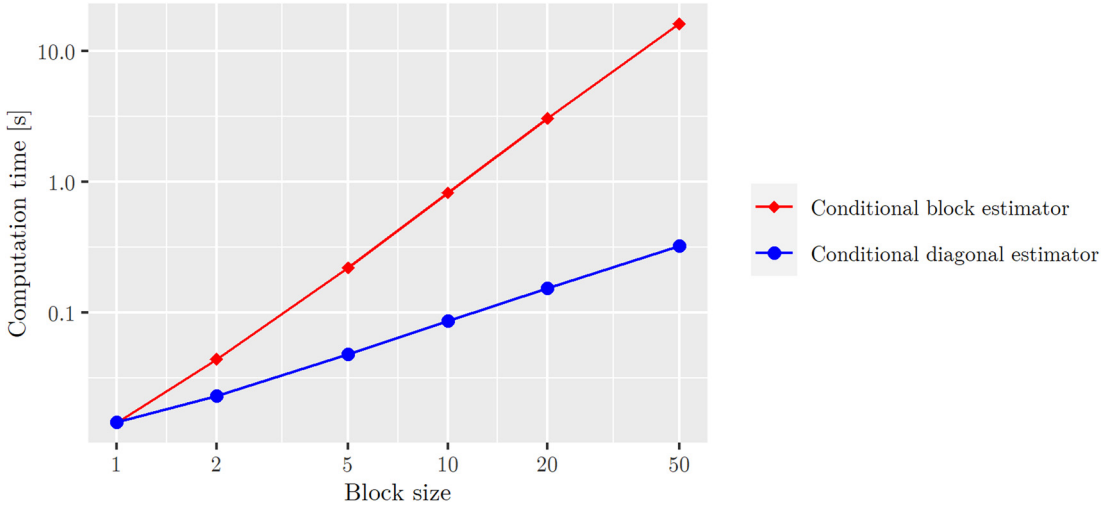

Figure A3: Log–log plot of the conditional estimators' mean computation time [s] as a function of the block size, for a sample size of 150.

$$M(\mathbf{1}_1 - \mathbf{1}_2) = (1-\rho_1, \rho_1 - 1, 0, \ldots, 0) = (1-\rho_1)(\mathbf{1}_1 - \mathbf{1}_2)$$
$$M(\mathbf{1}_{b_1+1} - \mathbf{1}_{b_1+2}) = (0, \ldots, 0, 1-\rho_2, \rho_2 - 1, 0, \ldots, 0) = (1-\rho_2)(\mathbf{1}_{b_1+1} - \mathbf{1}_{b_1+2}).$$

This gives us a number of $(b_1 - 1) + (b_2 - 1)$ eigenvectors with eigenvalues $1-\rho_1$ and $1-\rho_2$ that are positive since ρ_1, ρ_2 are smaller than 1. Moreover, remark that

$$M\mathbf{1}_{1:b_1} = (1 + (b_1 - 1))\rho_1 \mathbf{1}_{1:b_1} + b_1\rho_3 \mathbf{1}_{b_1+(1:b_2)}$$
$$M\mathbf{1}_{b_1+(1:b_2)} = b_2\rho_3 \mathbf{1}_{1:b_1} + (1 + (b_2 - 1)\rho_2)\mathbf{1}_{b_1+(1:b_2)},$$

so the eigenvalues of the matrix

$$\begin{pmatrix} (1 + (b_1 - 1))\rho_1 & b_1\rho_3 \\ b_2\rho_3 & 1 + (b_2 - 1)\rho_2 \end{pmatrix}$$

are also eigenvalues of M. These eigenvalues are

$$1 + \frac{(b_1-1)\rho_1}{2} + \frac{(b_2-1)\rho_2}{2} \pm \frac{\sqrt{((b_1-1)\rho_1 - (b_2-1)\rho_2)^2 + 4b_1\rho_3^2 b_2}}{2}.$$

The smallest eigenvalue is positive if and only if

$$\left(1 + \frac{(b_1-1)\rho_1}{2} + \frac{(b_2-1)\rho_2}{2}\right)^2 > \frac{((b_1-1)\rho_1 - (b_2-1)\rho_2)^2 + 4b_1\rho_3^2 b_2}{4}.$$

i.e.,

$$((b_2b_1 - b_1 - b_2 + 1)\rho_2 + b_1 - 1)\rho_1 - b_1\rho_3^2 b_2 + (b_2 - 1)\rho_2 + 1 > 0.$$

A sufficient condition is

$$\rho_1 > \frac{b_1b_2\rho_3^2 - (b_2-1)\rho_2 - 1}{(b_2b_1 - b_1 - b_2 + 1)\rho_2 + b_1 - 1},$$
$$\rho_2 > -\frac{b_1 - 1}{b_2b_1 - b_1 - b_2 + 1}.$$

□

Proof of Proposition 2. Assume that M is a correlation matrix, and let $\mathbf{X} \sim \mathcal{N}(0, M)$. Take one random variable from each block. Their correlation matrix is $(I + \rho J)$ and must therefore be positive semidefinite. This yields the constraint $\rho \geq -1/(K-1)$.

Conversely, this bound is reached by choosing the correlation matrices $\Sigma_k = \mathbf{1}$ for all $k = 1,\dots,K$ and considering M as the correlation matrix of $(X_1, \dots, X_1, X_2, \dots, X_2, \dots, X_K, \dots, X_K) \in \mathbb{R}^{b_1+b_2+\cdots+b_K}$, where $(X_1, \dots, X_K)$ follows an exchangeable normal distribution with correlation arbitrarily close to $-1/(K-1)$. □

B.2 Derivation of Equation (5)

We have

$$\begin{aligned} Q_{j_1,j_2} = {} & \mathbb{P}(\mathbf{X}_{1,(j_1,j_2)} < \mathbf{X}_{2,(j_1,j_2)}, \mathbf{X}_{1,(j_1,j_2)} < \mathbf{X}_{3,(j_1,j_2)}) + \mathbb{P}(\mathbf{X}_{1,(j_1,j_2)} < \mathbf{X}_{2,(j_1,j_2)}, \mathbf{X}_{1,(j_1,j_2)} > \mathbf{X}_{3,(j_1,j_2)}) \\ & + \mathbb{P}(\mathbf{X}_{1,(j_1,j_2)} > \mathbf{X}_{2,(j_1,j_2)}, \mathbf{X}_{1,(j_1,j_2)} > \mathbf{X}_{3,(j_1,j_2)}) + \mathbb{P}(\mathbf{X}_{1,(j_1,j_2)} > \mathbf{X}_{2,(j_1,j_2)}, \mathbf{X}_{1,(j_1,j_2)} < \mathbf{X}_{3,(j_1,j_2)}). \end{aligned}$$

Let us write these probabilities in terms of the copula C_{j_1,j_2} of $\mathbf{X}_{(j_1,j_2)} = (X_{j_1}, X_{j_2})$. This gives

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

as claimed.

B.3 Proof of Lemma 3

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

$$g^*(\mathbf{X}_{i_1,(j_1,j_2)}, \mathbf{X}_{i_2,(j_1,j_2)}) \coloneqq \operatorname{sign}((X_{i_1,j_1} - X_{i_2,j_1})(X_{i_1,j_2} - X_{i_2,j_2})).$$

Consequently,

$$\begin{aligned} \hat{\tau}^B &= \frac{1}{b_1 b_2} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} \hat{\tau}_{j_1,j_2} \\ &= \frac{1}{b_1 b_2} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} \frac{2}{n(n-1)} \sum_{i_1<i_2} g^*(\mathbf{X}_{i_1,(j_1,j_2)}, \mathbf{X}_{i_2,(j_1,j_2)}) \\ &= \frac{2}{n(n-1)} \sum_{i_1<i_2} \frac{1}{b_1 b_2} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} g^*(\mathbf{X}_{i_1,(j_1,j_2)}, \mathbf{X}_{i_2,(j_1,j_2)}), \end{aligned}$$

and it follows that $\hat{\tau}^B$ is a U-statistic with kernel

$$g^B(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) = \frac{1}{b_1 b_2} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} g^*(\mathbf{X}_{i_1,(j_1,j_2)}, \mathbf{X}_{i_2,(j_1,j_2)}).$$

In a similar manner, it is easily seen that $\hat{\tau}^R$, $\hat{\tau}^D$, and $\hat{\tau}^U$ are all U-statistics as well with respective kernels

$$g^R(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) = \frac{1}{N} \sum_{j=1}^{N} g^*(\mathbf{X}_{i_1,(1,j)}, \mathbf{X}_{i_2,(1,j)}),$$

$$g^D(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) = \frac{1}{N} \sum_{j=1}^{N} g^*(\mathbf{X}_{i_1,(j,b_1+j)}, \mathbf{X}_{i_2,(j,b_1+j)}),$$

$$g^U(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) = \frac{1}{b_1 b_2} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} W_{j_1,j_2} g^*(\mathbf{X}_{i_1,(j_1,j_2)}, \mathbf{X}_{i_2,(j_1,j_2)}).$$

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

B.4 Proof of Theorem 4

Recall from Lemma 3 that $\hat{\tau}_{j_1,j_2}$, $\hat{\tau}^B$, $\hat{\tau}^R$, $\hat{\tau}^D$, $\hat{\tau}^U$ can all be written as U-statistics of order 2 with the symmetric kernels defined earlier. To compute the variance of these U-statistics, we will use Hoeffding's formula [23] (see also [43, Section 5.2.1]) that we recall for reader's convenience. For a second-order U-statistic $U_n := \binom{n}{2}^{-1} \sum_{1 \le i_1 \ne i_2 \le n} g(\mathbf{X}_{i_1}, \mathbf{X}_{i_2})$ with symmetric kernel $g : \mathbb{R}^p \to \mathbb{R}$ satisfying $\mathbb{E}|g(\mathbf{X}_1, \mathbf{X}_2)| < +\infty$, the variance is given by

$$\mathbb{V}\mathrm{ar}[U_n] = \binom{n}{2}^{-1} \sum_{c=1}^{2} \binom{2}{c} \binom{n-2}{2-c} \zeta_c = \frac{2}{n(n-1)} (2(n-2)\zeta_1 + \zeta_2), \tag{A1}$$

where $\zeta_1 := \mathbb{V}\mathrm{ar}[g_1(\mathbf{X})]$, $\zeta_2 := \mathbb{V}\mathrm{ar}[g(\mathbf{X}_1, \mathbf{X}_2)]$, and $g_1(\mathbf{x}) := \mathbb{E}[g(\mathbf{X}_1, \mathbf{x})]$. Further, note that $\mathbb{E}[g_1(\mathbf{X})] = \mathbb{E}[g(\mathbf{X}_1, \mathbf{X}_2)] = \mathbb{E}[U_n]$.

We proceed to evaluate ζ_1 and ζ_2 for the different kernels, and then substitute them into (A1). Since the kernels g^*, g^B, g^R, and g^D are all deterministic, this leaves us with the variances of the corresponding estimators. First, we prove items (i)–(iv) of Theorem 4 and then proceed with the proof of (v), where we deal with the randomness of g^U.

B.4.1 Proof of (i)

Under Assumption 2, we have for every $(j_1, j_2) \in \{1, \ldots, b_1\} \times \{b_1 + 1, \ldots, p\}$,

$$\mathbb{E}[g^*(\mathbf{X}_{1,(j_1,j_2)}, \mathbf{X}_{2,(j_1,j_2)})] = \tau = 2P_{j_1,j_2} - 1.$$

Also,

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

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,

$$\begin{aligned}\zeta_1 &= \mathbb{V}\mathrm{ar}[g_1^*(X_{j_1}, X_{j_2})] \\ &= \mathbb{E}[(2P^c(X_{j_1}, X_{j_2}) - 1 - \tau)^2] \\ &= 4\mathbb{E}[P^c(X_{j_1}, X_{j_2})^2] - 4(1+\tau)\mathbb{E}[P^c(X_{j_1}, X_{j_2})] + (1+\tau)^2.\end{aligned}$$

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

$$\zeta_1 = 4(Q_{j_1,j_2} - P_{j_1,j_2}^2). \tag{A2}$$

For ζ_2, we find

$$\begin{aligned}\zeta_2 &= \mathbb{V}\mathrm{ar}[g^*(\mathbf{X}_{1,(j_1,j_2)}, \mathbf{X}_{2,(j_1,j_2)})] \\ &= \mathbb{E}[(2(\mathbb{1}\{X_{1,j_1} < X_{2,j_1}, X_{1,j_2} < X_{2,j_2}\} + \mathbb{1}\{X_{2,j_1} < X_{1,j_1}, X_{2,j_2} < X_{1,j_2}\}) - 1 - \tau)^2] \\ &= 4\mathbb{E}[(\mathbb{1}\{X_{1,j_1} < X_{2,j_1}, X_{1,j_2} < X_{2,j_2}\} + \mathbb{1}\{X_{2,j_1} < X_{1,j_1}, X_{2,j_2} < X_{1,j_2}\})^2] \\ &\quad - 4(1+\tau)\mathbb{E}[\mathbb{1}\{X_{1,j_1} < X_{2,j_1}, X_{1,j_2} < X_{2,j_2}\} + \mathbb{1}\{X_{2,j_1} < X_{1,j_1}, X_{2,j_2} < X_{1,j_2}\}] + (1+\tau)^2.\end{aligned}$$

Furthermore, note that

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

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

$$\begin{aligned}\zeta_2 &= 4\mathbb{E}[(\mathbb{1}\{X_{1,j_1} < X_{2,j_1}, X_{1,j_2} < X_{2,j_2}\} + \mathbb{1}\{X_{2,j_1} < X_{1,j_1}, X_{2,j_2} < X_{1,j_2}\})^2] \\ &\quad - 4(1+\tau)\mathbb{E}[\mathbb{1}\{X_{1,j_1} < X_{2,j_1}, X_{1,j_2} < X_{2,j_2}\} + \mathbb{1}\{X_{2,j_1} < X_{1,j_1}, X_{2,j_2} < X_{1,j_2}\}] + (1+\tau)^2 \\ &= -4\tau\mathbb{E}[\mathbb{1}\{X_{1,j_1} < X_{2,j_1}, X_{1,j_2} < X_{2,j_2}\} + \mathbb{1}\{X_{2,j_1} < X_{1,j_1}, X_{2,j_2} < X_{1,j_2}\}] + (1+\tau)^2 \\ &= -4(2P_{j_1,j_2} - 1)P_{j_1,j_2} + (1 + 2P_{j_1,j_2} - 1)^2 \\ &= 4(P_{j_1,j_2} - P_{j_1,j_2}^2),\end{aligned} \tag{A3}$$

where in the second step we have used that

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

Substitution of (A2) and (A3) into (A1) gives us the final expression

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

B.4.2 Proof of (ii)

We have

$$\mathbb{E}[g^B(\mathbf{X}_1, \mathbf{X}_2)] = 2P^{B,1} - 1,$$

and for any $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$,

$$g_1^B(\mathbf{x}) = \mathbb{E}\left[\frac{1}{b_1 b_2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p} g^*(\mathbf{X}_{i_1,(j_1,j_2)}, \mathbf{x}_{(j_1,j_2)})\right] = \frac{1}{b_1 b_2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p} 2P^c(x_{j_1}, x_{j_2}) - 1.$$

For ζ_1, we then obtain

$$\zeta_1 = \mathbb{V}\mathrm{ar}[g_1^B(\mathbf{X})] = \mathbb{E}\left[\left(\frac{1}{b_1 b_2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p} 2P^c(X_{j_1}, X_{j_2}) - \tau_{j_1,j_2} - 1\right)^2\right]$$

$$= \frac{1}{b_1^2 b_2^2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}\Bigg(\mathbb{E}[(2P^c(X_{j_1}, X_{j_2}) - 1 - \tau_{j_1,j_2})^2]$$

$$+ \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\varnothing}} \mathbb{E}[(2P^c(X_{j_1}, X_{j_2}) - 1 - \tau_{j_1,j_2})(2P^c(X_{j_3}, X_{j_4}) - 1 - \tau_{j_3,j_4})]$$

$$+ \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} \mathbb{E}[(2P^c(X_{j_1}, X_{j_2}) - 1 - \tau_{j_1,j_2})(2P^c(X_{j_3}, X_{j_4}) - 1 - \tau_{j_3,j_4})]\Bigg).$$

Now check that for any $j_1, j_2, j_3, j_4 \in \{1, \dots, p\}$,

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

We then have

$$\zeta_1 = \frac{1}{b_1^2 b_2^2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}\Bigg(4(Q_{j_1,j_2} - P_{j_1,j_2}^2) + \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\varnothing}} 4(Q_{j_1,j_2,j_3,j_4} - P_{j_1,j_2}P_{j_3,j_4})$$

$$+ \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4(Q_{j_1,j_2,j_3,j_4} - P_{j_1,j_2}P_{j_3,j_4})\Bigg).$$

Therefore, we have

$$\zeta_1 = \frac{4}{b_1 b_2}(Q_{B,2} - S_{B,2} + (b_1 + b_2 - 2)(Q_{B,1} - S_{B,1}) + (b_1 - 1)(b_2 - 1)(Q_{B,0} - S_{B,0})). \tag{A4}$$

For ζ_2, we have

$$\zeta_2 = \mathbb{V}\mathrm{ar}[g^B(\mathbf{X}_1, \mathbf{X}_2)]$$

$$= \mathbb{E}\left[\left(\frac{1}{b_1 b_2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p} g^*(\mathbf{X}_{1,(j_1,j_2)}, \mathbf{X}_{2,(j_1,j_2)}) - \tau_{j_1,j_2}\right)^2\right]$$

$$= \frac{1}{b_1^2 b_2^2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}(\mathbb{E}[(g^*(\mathbf{X}_{1,(j_1,j_2)}, \mathbf{X}_{2,(j_1,j_2)}) - \tau_{j_1,j_2})^2]$$

$$+ \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\varnothing}} \mathbb{E}[(g^*(\mathbf{X}_{1,(j_1,j_2)}, \mathbf{X}_{2,(j_1,j_2)}) - \tau_{j_1,j_2})(g^*(\mathbf{X}_{1,(j_3,j_4)}, \mathbf{X}_{2,(j_3,j_4)}) - \tau_{j_3,j_4})]$$

$$+ \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} \mathbb{E}[(g^*(\mathbf{X}_{1,(j_1,j_2)}, \mathbf{X}_{2,(j_1,j_2)}) - \tau_{j_1,j_2})(g^*(\mathbf{X}_{1,(j_3,j_4)}, \mathbf{X}_{2,(j_3,j_4)}) - \tau_{j_3,j_4})].$$

Remark that

$$\mathbb{E}[g^*(\mathbf{X}_{1,(j_1,j_2)}, \mathbf{X}_{2,(j_1,j_2)})g^*(\mathbf{X}_{1,(j_3,j_4)}, \mathbf{X}_{2,(j_3,j_4)})] = 4P_{j_1,j_2,j_3,j_4}.$$

Therefore,

$$\zeta_2 = \frac{1}{b_1^2 b_2^2} \sum_{j_1=1}^{b_1} \sum_{j_2=b_1+1}^{p} \Bigg(4(P_{j_1,j_2} - P_{j_1,j_2}^2) + \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\} \\ \{j_1,j_2\}\cap\{j_3,j_4\}=\varnothing}} 4(P_{j_1,j_2,j_3,j_4} - P_{j_1,j_2}P_{j_3,j_4}) + \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\} \\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4(P_{j_1,j_2,j_3,j_4} - P_{j_1,j_2}P_{j_3,j_4}) \Bigg),$$

and we obtain the expression

$$\zeta_2 = \frac{4}{b_1 b_2}(P_{B,2} - S_{B,2} + (b_1 + b_2 - 2)(P_{B,1} - S_{B,1}) + (b_1 - 1)(b_2 - 1)(P_{B,0} - S_{B,0})). \tag{A5}$$

Finally, by substituting (A4) and (A5) into (A1), we find

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\tau}^B] = \frac{8}{b_1 b_2 n(n-1)}&(2(n-2)(Q_{B,2} - S_{B,2} + (b_1 + b_2 - 2)(Q_{B,1} - S_{B,1}) + (b_1 - 1)(b_2 - 1)(Q_{B,0} - S_{B,0})) \\ &+ (P_{B,2} - S_{B,2} + (b_1 + b_2 - 2)(P_{B,1} - S_{B,1}) + (b_1 - 1)(b_2 - 1)(P_{B,0} - S_{B,0}))).\end{aligned}$$

B.4.3 Proof of (iii)

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

$$\zeta_1 = \frac{1}{N^2} \sum_{j_1=1}^{N} \Bigg(4(Q_{1,b_1+j_1} - P_{1,b_1+j_1}^2) + \sum_{j_2=1, j_2\neq j_1}^{N} 4(Q_{1,b_1+j_1,1,b_1+j_2} - P_{1,b_1+j_1}P_{1,b_1+j_2}) \Bigg).$$

Note that there is one less summation compared to the expressions in (ii) since $\{1, b_1 + j_1\} \cap \{1, b_1 + j_2\} \neq \varnothing$ for every j_1, j_2. Therefore,

$$\zeta_1 = \frac{4}{N}(Q_{R,2} - S_{R,2} + (N - 1)(Q_{R,1} - S_{R,1})).$$

Similarly, we find that

$$\zeta_2 = \frac{4}{N}(P_{R,2} - S_{R,2} + (N - 1)(P_{R,1} - S_{R,1})).$$

Hence,

$$\mathbb{V}\mathrm{ar}[\hat{\tau}^R] = \frac{8}{Nn(n-1)}(2(n-2)(Q_{R,2} - S_{R,2} + (N-1)(Q_{R,1} - S_{R,1})) + (P_{R,2} - S_{R,2} + (N-1)(P_{R,1} - S_{R,1}))).$$

B.4.4 Proof of (iv)

Again, in a similar manner to the proof of (ii), we obtain

$$\zeta_1 = \frac{1}{N^2} \sum_{j_1=1}^{N} (4(Q_{j_1,b_1+j_1} - P_{j_1,b_1+j_1}^2) + \sum_{j_2=1, j_2\neq j_1}^{N} 4(Q_{j_1,b_1+j_1,j_1,b_1+j_2} - P_{j_1,b_1+j_1}P_{j_1,b_1+j_2})).$$

Note that there is one less summation compared to the expressions in (ii), as in (iii). Therefore,

$$\zeta_1 = \frac{4}{N}(Q_{D,2} - S_{D,2} + (N - 1)(Q_{D,1} - S_{D,0})).$$

Similarly, we find that

$$\zeta_2 = \frac{4}{N}(P_{D,2} - S_{D,2} + (N-1)(P_{D,0} - S_{D,0})).$$

Hence,

$$\mathbb{V}\mathrm{ar}[\hat{\tau}^D] = \frac{8}{Nn(n-1)}(2(n-2)(Q_{D,2} - S_{D,2} + (N-1)(Q_{D,0} - S_{D,0})) + (P_{D,2} - S_{D,2} + (N-1)(P_{D,0} - S_{D,0}))).$$

B.4.5 Proof of (v)

This proof is very similar as the one of (ii), except that we replace all expectations by conditional expectations given $\mathbf{W} = \mathbf{w}$. Note that $\hat{\tau}^U_{k_1,k_2}|\mathbf{W}$ is a U-statistic with (deterministic) kernel $g^U_{k_1,k_2}|\mathbf{W}$. As mentioned earlier, we obtain the corresponding ζ_1 and ζ_2,

$$\zeta_1 = \frac{1}{b_1^2 b_2^2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}\Bigg(4W^2_{j_1,j_2}(Q_{j_1,j_2} - P^2_{j_1,j_2}) + \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\varnothing}} 4W_{j_1,j_2}W_{j_3,j_4}(Q_{j_1,j_2,j_3,j_4} - P_{j_1,j_2}P_{j_3,j_4})$$

$$+ \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4W_{j_1,j_2}W_{j_3,j_4}(Q_{j_1,j_2,j_3,j_4} - P_{j_1,j_2}P_{j_3,j_4})\Bigg),$$

and

$$\zeta_2 = \frac{1}{b_1^2 b_2^2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}\Bigg(4W^2_{j_1,j_2}(P_{j_1,j_2} - P^2_{j_1,j_2}) + \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\varnothing}} 4W_{j_1,j_2}W_{j_3,j_4}(P_{j_1,j_2,j_3,j_4} - P_{j_1,j_2}P_{j_3,j_4})$$

$$+ \sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4W_{j_1,j_2}W_{j_3,j_4}(P_{j_1,j_2,j_3,j_4} - P_{j_1,j_2}P_{j_3,j_4})\Bigg),$$

which lead to the result as claimed.

B.4.6 Proof of (vi)

Finally, for obtaining the variance of $\hat{\tau}^U$, we need to deal with the random kernel g^U. To this end, we use the law of total variance and find

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\hat{\tau}^U] &= \mathbb{V}\mathrm{ar}[\mathbb{E}[\hat{\tau}^U|\mathbf{W}]] + \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\tau}^U|\mathbf{W}]]\\ &= \mathbb{V}\mathrm{ar}_{J_1,J_2}[\tau_{J_1,J_2}] + \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\tau}^U|\mathbf{W}]]\\ &= \mathbb{V}\mathrm{ar}_{J_1,J_2}[\tau_{J_1,J_2}] + \mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\tau}^U|\mathbf{W}]].\end{aligned}$$

We can thus obtain the desired variance by first evaluating the variance of τ^U under a given $\mathbf{W}$ and then by taking the expectation with respect to $\mathbf{W}$. Therefore,

$$\mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\tau}^U|\mathbf{W}]] = \frac{2}{n(n-1)}(2(n-2)\mathbb{E}[\zeta_1] + \mathbb{E}[\zeta_2]). \tag{A6}$$

By inserting the expression of ζ_1 and ζ_2 found in (v) into (A6), we obtain

$$\begin{aligned}\mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\tau}^U|\mathbf{W}]] = \frac{8}{N^2 n(n-1)}\Bigg(&\frac{2(n-2)}{b_1^2 b_2^2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}\Bigg(4\mathbb{E}[W_{j_1,j_2}^2](Q_{j_1,j_2}-P_{j_1,j_2}^2)\\ &+\sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\varnothing}} 4\mathbb{E}[W_{j_1,j_2}W_{j_3,j_4}](Q_{j_1,j_2,j_3,j_4}-P_{j_1,j_2}P_{j_3,j_4})\\ &+\sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4\mathbb{E}[W_{j_1,j_2}W_{j_3,j_4}](Q_{j_1,j_2,j_3,j_4}-P_{j_1,j_2}P_{j_3,j_4})\Bigg)\\ &+\frac{1}{b_1^2 b_2^2}\sum_{j_1=1}^{b_1}\sum_{j_2=b_1+1}^{p}\Bigg(4\mathbb{E}[W_{j_1,j_2}^2](P_{j_1,j_2}-P_{j_1,j_2}^2)\\ &+\sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ \{j_1,j_2\}\cap\{j_3,j_4\}=\varnothing}} 4\mathbb{E}[W_{j_1,j_2}W_{j_3,j_4}](P_{j_1,j_2,j_3,j_4}-P_{j_1,j_2}P_{j_3,j_4})\\ &+\sum_{\substack{(j_3,j_4)\in\{1,\dots,b_1\}\times\{b_1+1,\dots,p\}\\ |\{j_1,j_2\}\cap\{j_3,j_4\}|=1}} 4\mathbb{E}[W_{j_1,j_2}W_{j_3,j_4}](P_{j_1,j_2,j_3,j_4}-P_{j_1,j_2}P_{j_3,j_4})\Bigg)\Bigg).\end{aligned} \tag{A7}$$

Recall that we select N pairs out of the $b_1 b_2$ possible pairs with uniform probability and without replacement. Therefore, for every distinct pairs (j_1, j_2) and (j_3, j_4), we can compute the following expectations:

$$\mathbb{E}[W_{j_1,j_2}^2] = \mathbb{E}[W_{j_1,j_2}] = \frac{N}{b_1 b_2},$$
$$\mathbb{E}[W_{j_1,j_2}W_{j_3,j_4}] = \frac{N(N-1)}{b_1 b_2(b_1 b_2 - 1)}.$$

Finally, by combining this with Equation (A7), we establish the desired formula

$$\begin{aligned}\mathbb{E}[\mathbb{V}\mathrm{ar}[\hat{\tau}^U|\mathbf{W}]] = \frac{8}{b_1 b_2 n(n-1)}\Bigg(&2(n-2)(Q_{B,2}-S_{B,2}+\frac{N-1}{b_1 b_2-1}(b_1+b_2-2)(Q_{B,1}-S_{B,1})\\ &+\frac{N-1}{b_1 b_2-1}(b_1-1)(b_2-1)(Q_{B,0}-S_{B,0}))\\ &+(P_{B,2}-S_{B,2}+\frac{N-1}{b_1 b_2-1}(b_1+b_2-2)(P_{B,1}-S_{B,1})+\frac{N-1}{b_1 b_2-1}(b_1-1)(b_2-1)(P_{B,0}-S_{B,0}))\Bigg).\end{aligned}$$

B.5 Proof of Theorem 8

Proof. In [10] (see p. 299), it was already shown that the conditional Kendall's tau estimator defined in (12) is asymptotically normal at different points of the conditioning variable. 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, and give a detailed proof for completeness. The asymptotic covariance matrices are then obtained by combining the results with the appropriate kernels under Assumption 4.

Remember that the averaged conditional Kendall's tau are conditional U-statistics with kernels given by Lemma 3, which treats the unconditional case. Furthermore, for any measurable function $g : \mathbb{R}^{2d} \to \mathbb{R}$, let us define the second-order U-statistic

$$U_{n,j'}(g) := \frac{1}{n(n-1)} \sum_{1 \le i_1 \ne i_2 \le n} g_{i_1,i_2}, \tag{A8}$$

where

$$g_{i_1,i_2} := \frac{g(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}) \mathcal{K}_h(\mathbf{z}'_{j'} - \mathbf{Z}_{i_1}) \mathcal{K}_h(\mathbf{z}'_{j'} - \mathbf{Z}_{i_2})}{\mathbb{E}[\mathcal{K}_h(\mathbf{z}'_{j'} - \mathbf{Z})]^2}.$$

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

$$\hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} = \frac{U_{n,j'}(g)}{U_{n,j'}(1) + \varepsilon_{n,j'}}, \tag{A9}$$

where we write $\hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}}$ for any of the estimators $\hat{\tau}^B_{|\mathbf{Z}=\mathbf{z}'_{j'}}$, $\hat{\tau}^R_{|\mathbf{Z}=\mathbf{z}'_{j'}}$, $\hat{\tau}^D_{|\mathbf{Z}=\mathbf{z}'_{j'}}$, $\hat{\tau}^U_{|\mathbf{Z}=\mathbf{z}'_{j'}}$ with g given by, respectively, g^B, g^R, g^D, and g^U. The residual term $\varepsilon_{n,j'}$ is given by

$$\varepsilon_{n,j'} := \frac{\sum_{i=1}^n \mathcal{K}_h^2(\mathbf{z}'_{j'} - \mathbf{Z}_i)}{n(n-1)\mathbb{E}[\mathcal{K}_h(\mathbf{z}'_{j'} - \mathbf{Z})]^2}.$$

Further, we set

$$\tilde{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} := \frac{U_{n,j'}(g)}{U_{n,j'}(1)}, \tag{A10}$$

to be the equivalent of (A9) with the term $\varepsilon_{n,j'}$ removed. This is a simpler version of (A9) that will be easier to analyze theoretically. We now show that $\hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}}$ and $\tilde{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}}$ are close. Under Assumption 5, we replace both $\hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} = U_{n,j'}(g)/U_{n,j'}(1)$ and $\tilde{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} = U_{n,j'}(g)/(U_{n,j'}(1) + \varepsilon_{n,j'})$ by their expressions to obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\varepsilon_{n,j'}}(\tilde{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} - \hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}})\right] &= \mathbb{E}\left[\frac{1}{\varepsilon_{n,j'}}\left(\frac{U_{n,j'}(g)}{U_{n,j'}(1)} - \frac{U_{n,j'}(g)}{U_{n,j'}(1) + \varepsilon_{n,j'}}\right)\right] \\
&= \mathbb{E}\left[\frac{U_{n,j'}(g)(U_{n,j'}(1) + \varepsilon_{n,j'}) - U_{n,j'}(g)U_{n,j'}(1)}{\varepsilon_{n,j'}U_{n,j'}(1)(U_{n,j'}(1) + \varepsilon_{n,j'})}\right] \\
&= \mathbb{E}\left[\frac{U_{n,j'}(g)U_{n,j'}(1) + U_{n,j'}(g)\varepsilon_{n,j'} - U_{n,j'}(g)U_{n,j'}(1)}{\varepsilon_{n,j'}U_{n,j'}(1)(U_{n,j'}(1) + \varepsilon_{n,j'})}\right] \\
&= \mathbb{E}\left[\frac{U_{n,j'}(g)\varepsilon_{n,j'}}{\varepsilon_{n,j'}U_{n,j'}(1)(U_{n,j'}(1) + \varepsilon_{n,j'})}\right] \\
&= \mathbb{E}\left[\frac{1}{U_{n,j'}(1)}\frac{U_{n,j'}(g)}{U_{n,j'}(1) + \varepsilon_{n,j'}}\right] \\
&= \mathbb{E}\left[\frac{1}{U_{n,j'}(1)}\hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}}\right] \\
&= O(1),
\end{aligned}$$

because $|\hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}}|$ is bounded by 1 and $U_{n,j'}(1)$ is asymptotically normal, hence convergent by Lemma 17 of [10]. Therefore, $\hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} - \tilde{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} = O_P(\varepsilon_{n,j'})$ using Markov's inequality. By Assumption 5(c) and by Bochner's lemma e.g., [44] we see that $\varepsilon_{n,j'} = O_P((nh^d)^{-1})$. It then follows by Assumption 7 that

$$(nh^d)^{1/2}(\hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} - \tilde{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}}) = O_P((nh^d)^{1/2}\varepsilon_{n,j'}) = o_P(1).$$

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

Now let us apply Lemma 17 again from [10] on the joint asymptotic law of U-statistics of the form $U_{n,j'}$. That is, under Assumptions 5–7 and for any two bounded measurable functions g_1 and g_2,

$$\begin{aligned}&(nh^d)^{1/2}((U_{n,j'}(g_1) - \mathbb{E}[U_{n,j'}(g_1)])_{j'=1,\dots,n'}, (U_{n,j'}(g_2) - \mathbb{E}[U_{n,j'}(g_2)])_{j'=1,\dots,n'})\\ &\xrightarrow{\text{law}} \mathcal{N}\left(0, \begin{bmatrix} M_\infty(g_1) & M_\infty(g_1,g_2)\\ M_\infty(g_1,g_2) & M_\infty(g_2)\end{bmatrix}\right),\end{aligned} \tag{A11}$$

as $n \to \infty$, where

$$[M_\infty(g_1,g_2)]_{j_1',j_2'} := \frac{4\int \mathcal{K}^2 \mathbb{1}_{\{\mathbf{z}'_{j_1'} = \mathbf{z}'_{j_2'}\}}}{f_{\mathbf{Z}}(\mathbf{z}'_{j_1'})} \int g_1(\mathbf{x}_1,\mathbf{x}) g_2(\mathbf{x}_2,\mathbf{x}) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j_1'}}(\mathbf{x}) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j_1'}}(\mathbf{x}_1) f_{\mathbf{X}|\mathbf{Z}=\mathbf{z}'_{j_1'}}(\mathbf{x}_2) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2. \tag{A12}$$

Let us investigate the expectation of $U_{n,j'}(g)$. We write by (A8)

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

Further, by a change of variable, we find

$$\begin{aligned}&\mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)\mathcal{K}_h(\mathbf{z}'_{j'} - \mathbf{Z}_1)\mathcal{K}_h(\mathbf{z}'_{j'} - \mathbf{Z}_2)]\\ &= \int g(\mathbf{x}_1,\mathbf{x}_2)\mathcal{K}_h(\mathbf{z}'_{j'} - \mathbf{Z}_1)\mathcal{K}_h(\mathbf{z}'_{j'} - \mathbf{Z}_2) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{Z}_1) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{Z}_2)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{Z}_1\mathrm{d}\mathbf{Z}_2\\ &= \int g(\mathbf{x}_1,\mathbf{x}_2)\mathcal{K}(\mathbf{u}_1)\mathcal{K}(\mathbf{u}_2) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z}'_{j'} + hu_1) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{z}'_{j'} + hu_2)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2\mathrm{d}\mathbf{u}_1\mathrm{d}\mathbf{u}_2,\end{aligned} \tag{A13}$$

with the change of variable $\mathbf{Z}_1 = \mathbf{z}'_{j'} + hu_1$ and $\mathbf{Z}_2 = \mathbf{z}'_{j'} + hu_2$. Let us define the function $\phi_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t) := f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_1,\mathbf{z}'_{j'} + th\mathbf{u}_1) f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_2,\mathbf{z}'_{j'} + th\mathbf{u}_2)$ for $t \in [0,1]$. By Assumption 6, $\phi_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t)$ is α times differentiable, allowing us to apply the Taylor-Lagrange formula. This gives

$$\phi_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t) = \sum_{k=0}^{\alpha-1} \frac{1}{k!}\phi^{(k)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(0) + \frac{1}{\alpha!}\phi^{(\alpha)}_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{u}_1,\mathbf{u}_2}(t^*), \tag{A14}$$

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

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

After substituting (A14) into (A13), we obtain

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

where in the first equality we have used the fact that $\int \mathcal{K}(\mathbf{u})u_{j_1}\dots u_{j_k}\mathrm{d}\mathbf{u} = 0$ for all $k = 1,\dots,\alpha-1$ as stated in Assumption 5(b), which results in the elimination of all the terms in the sum except the first one. In the second equality, we replaced ϕ and $\phi^{(\alpha)}$ by their expressions. In the last equality, we factor out $f^2_{\mathbf{Z}}(\mathbf{z}'_{j'})$ and recognize that the corresponding integral is a conditional expectation.

We now bound the second term of the previous display. By Assumption 6, we have

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

Therefore,

$$\mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)\mathcal{K}_h(\mathbf{z}'_{j'}-\mathbf{Z}_1)\mathcal{K}_h(\mathbf{z}'_{j'}-\mathbf{Z}_2)] = f^2_{\mathbf{Z}}(\mathbf{Z}'_j)\mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_{j'}] + O(h^\alpha),$$

and by the same reasoning, we obtain

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

Consequently, we find that

$$\begin{aligned}\mathbb{E}[U_{n,j'}(g)] &= \frac{1}{\mathbb{E}[\mathcal{K}_h(\mathbf{z}'_{j'}-\mathbf{Z})]^2}\mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)\mathcal{K}_h(\mathbf{z}'_{j'}-\mathbf{Z}_1)\mathcal{K}_h(\mathbf{z}'_{j'}-\mathbf{Z}_2)]\\ &= \mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_{j'}] + r_{n,j'},\end{aligned} \tag{A15}$$

where $|r_{n,j'}| \le C_0h^\alpha$ for some constant C_0 independent of j'. Then, by Assumption 7,

$$(nh^d)^{1/2}(\mathbb{E}[U_{n,j'}(g)] - \mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_{j'}]) = O((nh^d)^{1/2}h^\alpha) = o(1). \tag{A16}$$

Therefore, the asymptotic law of (A11) still holds after replacing $\mathbb{E}[U_{n,j'}(g)]$ with $\mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_{j'}]$. As such,

$$(nh^d)^{1/2}((U_{n,j'}(g)-\tau_{|\mathbf{Z}=\mathbf{z}'_{j'}})_{j'=1,\dots,n'},(U_{n,j'}(1)-1)_{j'=1,\dots,n'}) \xrightarrow{\text{law}} \mathcal{N}\left(0,\begin{bmatrix}M_\infty(g,g) & M_\infty(g,1)\\ M_\infty(g,1) & M_\infty(1,1)\end{bmatrix}\right), \tag{A17}$$

as $n\to\infty$, where we have used that $\mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{z}'_{j'}] = \tau_{|\mathbf{Z}=\mathbf{z}'_{j'}}$ under Assumption 4.

To derive the asymptotic law of $(nh^d)^{1/2}(\bar{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} - \tau_{|\mathbf{Z}=\mathbf{z}'_{j'}})$, we apply the Delta-method on (A17) with the function $\gamma(\mathbf{x},\mathbf{y}) \coloneqq \mathbf{x}/\mathbf{y}$, which divides two real vectors $\mathbf{x},\mathbf{y}$ of size n' component-wise. The corresponding Jacobian is given by the $n'\times 2n'$ matrix

$$J_\gamma(\mathbf{x},\mathbf{y}) = [\mathrm{Diag}(y_1^{-1},\dots,y_{n'}^{-1}),\mathrm{Diag}(-x_1y_1^{-2},\dots,-x_{n'}y_{n'}^{-2})].$$

Hence, as $n\to\infty$

$$(nh_n^d)^{1/2}(\hat{\tau}_{|\mathbf{Z}=\mathbf{z}'_{j'}} - \tau_{|\mathbf{Z}=\mathbf{z}'_{j'}})_{j'=1,\dots,n'} \xrightarrow{\text{law}} \mathcal{N}(0,\mathbf{H}),$$

setting

$$\mathbf{H} \coloneqq J_\gamma(\vec{\tau}_{|\mathbf{Z}},\mathbf{e})\begin{bmatrix}M_\infty(g) & M_\infty(g,1)\\ M_\infty(g,1) & M_\infty(1)\end{bmatrix}J_\gamma(\vec{\tau}_{|\mathbf{Z}},\mathbf{e})^T,$$

where $\vec{\tau}_{|\mathbf{Z}}$ and $\mathbf{e}$ denote n'-dimensional vectors filled with respectively $\tau_{|\mathbf{Z}=\mathbf{z}'_{j'}}$ and 1. This gives

$$\mathbf{H} = M_\infty(g,g) - \mathrm{Diag}(\vec{\tau}_{|\mathbf{Z}})M_\infty(g,1) - M_\infty(g,1)\mathrm{Diag}(\vec{\tau}_{|\mathbf{Z}}) + \mathrm{Diag}(\vec{\tau}_{|\mathbf{Z}})M_\infty(1,1)\mathrm{Diag}(\vec{\tau}_{|\mathbf{Z}})$$

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

$$[M_\infty(g,g)]_{j_1',j_2'} = \frac{4\int \mathcal{K}^2 \mathbb{1}_{\{\mathbf{z}'_{j_1'}=\mathbf{z}'_{j_2'}\}}}{f_{\mathbf{Z}}(\mathbf{z}'_{j_1'})} \mathbb{E}[g(\mathbf{X}_1,\mathbf{X})g(\mathbf{X}_2,\mathbf{X})|\mathbf{Z}=\mathbf{z}_1=\mathbf{Z}_2=\mathbf{z}'_{j_1'}],$$

$$\begin{aligned}[\mathrm{Diag}(\vec{\tau}_{|\mathbf{Z}})M_\infty(g,1)]_{j_1',j_2'} &= \frac{4\int \mathcal{K}^2 \mathbb{1}_{\{\mathbf{z}'_{j_1'}=\mathbf{z}'_{j_2'}\}}}{f_{\mathbf{Z}}(\mathbf{z}'_{j_1'})} \tau_{|\mathbf{Z}=\mathbf{z}'_{j'}} \mathbb{E}[g(\mathbf{X}_1,\mathbf{X})|\mathbf{Z}=\mathbf{z}_1=\mathbf{z}'_{j_1'}] \\ &= \frac{4\int \mathcal{K}^2 \mathbb{1}_{\{\mathbf{z}'_{j_1'}=\mathbf{z}'_{j_2'}\}}}{f_{\mathbf{Z}}(\mathbf{z}'_{j_1'})} \tau^2_{|\mathbf{Z}=\mathbf{z}'_{j'}} \\ &= [M_\infty(g,1)\mathrm{Diag}(\vec{\tau}_{|\mathbf{Z}})]_{j_1',j_2'} \\ &= [\mathrm{Diag}(\vec{\tau}_{|\mathbf{Z}})M_\infty(1,1)\mathrm{Diag}(\vec{\tau}_{|\mathbf{Z}})]_{j_1',j_2'},\end{aligned}$$

and thus,

$$[\mathbf{H}]_{j_1',j_2'} = \frac{4\int \mathcal{K}^2 \mathbb{1}_{\{\mathbf{z}'_{j_1'}=\mathbf{z}'_{j_2'}\}}}{f_{\mathbf{Z}}(\mathbf{z}'_{j_1'})} (\mathbb{E}[g(\mathbf{X}_1,\mathbf{X}_2)g(\mathbf{X}_1,\mathbf{X}_3)|\mathbf{Z}_1=\mathbf{Z}_2=\mathbf{Z}_3=\mathbf{z}'_{j_1'}] - \tau^2_{|\mathbf{Z}=\mathbf{z}'_{j'}}).$$

Finally, by substituting the appropriate kernels and by similar steps as in the derivation of the corresponding ζ_1's from the proof of Theorem 4, it is easily seen that under Assumption 4, we obtain the desired asymptotic covariance matrices. □

C 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)

References

[1] Abdous, B., Genest, C., & Rémillard, B. (2005). Dependence properties of meta-elliptical distributions. In: *Statistical modeling and analysis for complex data problems* (pp. 1–15). Boston, MA: Springer US.

[2] Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. *Review of Financial Studies*, *15*, 1137–1187.

[3] Ascorbebeitia, J., Ferreira, E., & Orbe, S. (2022). Testing conditional multivariate rank correlations: the effect of institutional quality on factors influencing competitiveness. *TEST*, *31*, 931–949.

[4] Barber, R. F., & Kolar, M. (2018). Rocket: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *The Annals of Statistics*, *46*(6B), 3422–3450.

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

[6] Cadima, J., Calheiros, F. L., & Preto, I. P. (2010). The eigenstructure of block-structured correlation matrices and its implications for principal component analysis. *Journal of Applied Statistics*, *37*(4), 577–589.

[7] Delft High Performance Computing Centre (DHPC) (2022). *DelftBlue Supercomputer (Phase 1)*. https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1.

[8] Derumigny, A. (2023). *CondCopulas: Estimation and Inference for Conditional Copulas Models. R package version 0.1.3*. Available at https://cran.r-project.org/package=CondCopulas.

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

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

[11] Derumigny, A., & Fermanian, J.-D. (2020). On Kendall's regression. *Journal of Multivariate Analysis*, *178*, 104610.

[12] Derumigny, A., & Fermanian, J.-D. (2022). Identifiability and estimation of meta-elliptical copula generators. *Journal of Multivariate Analysis*, *190*, 104962.

[13] Derumigny, A., & Fermanian, J.-D. (2023). *ElliptCopulas: Inference of Elliptical Copulas and Elliptical Distributions*. R package version 0.1.3. https://cran.r-project.org/package=ElliptCopulas.

[14] Erb, C., Harvey, C., & Viskanta, T. (1994). Forecasting international equity correlations. *Financial Analysts Journal*, *50*, 32–45.

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

[16] Fan, J., Liao, Y., & Wang, W. (2014). Projected principal component analysis in factor models. *SSRN Electronic Journal*, *44*, 219–254.

[17] Fang, H.-B., Fang, K.-T., & Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, *82*(1), 1–16.

[18] Genest, C., Favre, A.-C., Béliveau, J., & Jacques, C. (2007). Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resources Research*, *43*(9), W09401, 1–12.

[19] Genest, C., Nessslehová, J., & Ghorbal, N. (2011). Estimators based on Kendall's tau in multivariate copula models. *Australian & New Zealand Journal of Statistics*, *53*, 157–177.

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

[21] Gray, H., Leday, G. G., Vallejos, C. A., & Richardson, S. (2018). *Shrinkage estimation of large covariance matrices using multiple shrinkage targets*. ArXiv: arXiv:1809.08024.

[22] Hahsler, M., Buchta, C., & Hornik, K. (2022). *Seriation: Infrastructure for Ordering Objects Using Seriation. R package version 1.3.2*. https://cran.r-project.org/package=seriation.

[23] Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, *19*(4), 546–557.

[24] Kurowicka, D., & Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling.* England: John Wiley & Sons.

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

[26] Liu, H., Han, F., & Zhang, C.-H. (2012). Transelliptical graphical models. In *Advances in Neural Information Processing Systems* (*vol. 25*, pp. 800–808) .

[27] Longin, F., & Solnik, B. (2001). Extreme value correlation of international equity markets. *The Journal of Finance*, *56*, 649–676.

[28] Lu, J., Kolar, M., & Liu, H. (2018). Post-regularization inference for time-varying nonparanormal graphical models. *Journal of Machine Learning Research*, 18, 1–78.

[29] McNeil, A., Frey, R., & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools* (vol. 101), New Jersey, US: Princeton University Press.

[30] McNeil, A. J., Nessslehová, J. G., & Smith, A. D. (2022). On attainability of Kendall's tau matrices and concordance signatures. *Journal of Multivariate Analysis*, *191*, 105033.

[31] Nagler, T. (2023). *WDM: Weighted Dependence Measures*. R package version 0.2.4.

[32] Nelsen, R. B. (2007). *An Introduction to Copulas*. New York, NY, USA: Springer Science & Business Media.

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

[34] Perreault, S. (2020). *Structures de corrélation partiellement échangeables: inférence et apprentissage automatique*. PhD thesis, Université Laval.

[35] Perreault, S., Duchesne, T., & Nešlehová, J. (2019). Detection of block-exchangeable structure in large-scale correlation matrices. *Journal of Multivariate Analysis*, *169*, 400–422.
[36] Perreault, S., Nešlehová, J. G., & Duchesne, T. (2022). Hypothesis tests for structured rank correlation matrices. *Journal of the American Statistical Association*, *118*(544), 2889–2900.
[37] Pimenova, I. (2012). *Semi-parametric Estimation of Elliptical Distribution in Case of High Dimensionality*. Master's Thesis, Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät.
[38] R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
[39] Rothman, A., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, *104*(485), 177–186.
[40] Ryan, J. A., & Ulrich, J. M. (2024). *quantmod: Quantitative Financial Modelling Framework*. R package version 0.4.26.
[41] Ryan, V., & Derumigny, A. (2024). *On the choice of the two tuning parameters for nonparametric estimation of an elliptical distribution generator*. ArXiv: arXiv:2408.17087.
[42] Sadefo Kamdem, J. (2005). Value-at-risk and expected shortfall for linear portfolios with elliptically distributed risk factors. *International Journal of Theoretical and Applied Finance*, *8*, 537–551.
[43] Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics* (vol. 162), New York, NY, USA: John Wiley & Sons.
[44] Tsybakov, A. (2003). *Introduction à l'estimation non paramétrique* (vol. 41), Berlin Heidelberg, Germany: Springer Science & Business Media.
[45] Veraverbeke, N., Omelka, M., & Gijbels, I. (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics*, *38*, 766–780.