Testing for equality between conditional copulas given discretized conditioning events

Derumigny, Alexis; Fermanian, Jean David; Min, Aleksey

DOI
10.1002/cjs.11742

Publication date
2022

Document Version
Final published version

Published in
Canadian Journal of Statistics

Testing for equality between conditional copulas given discretized conditioning events

Alexis DERUMIGNY[1]*, Jean-David FERMANIAN[2], and Aleksey MIN[3]

[1] *Department of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, Netherlands*
[2] *École Nationale de la Statistique et de l'Administration Économique (ENSAE) & Centre de Recherche en Économie et Statistique (CREST), 5 avenue Le Chatelier, 91120 Palaiseau Cedex, France*
[3] *Department of Mathematics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany*

Key words and phrases: Conditional copula; conditional Kendall's tau; decision tree; simplifying assumption.

MSC 2020: Primary 62G10; 62H20; Secondary 62G20; 62H05.

Abstract: Several procedures have been recently proposed to test the simplifying assumption for conditional copulas. Instead of considering pointwise conditioning events, we study the constancy of the conditional dependence structure when some covariates belong to general Borel conditioning subsets. We introduce several test statistics based on the equality of conditional Kendall's taus and derive their asymptotic distributions under the null hypothesis. In settings where such conditioning events are not fixed ex ante, we propose a data-driven procedure to recursively build such relevant subsets. This procedure is based on decision trees that maximize the differences between the conditional Kendall's taus, which correspond to the leaves of the trees. Empirical results for such tests are illustrated in the Supplementary Material. Moreover, a study of the conditional dependence between financial stock returns is presented and highlights specific contagion effects of past returns. The last application deals with conditional dependence between coverage amounts in an insurance dataset.

Résumé: Plusieurs procédures ont été proposées récemment pour tester l'hypothèse simplificatrice pour les copules conditionnelles. Au lieu de considérer des évènements conditionnants ponctuels, nous étudions le caractère constant de la structure de dépendance conditionnelle lorsque certaines covariables appartiennent à des ensembles boréliens conditionnants. Nous introduisons plusieurs statistiques de test basées sur l'égalité des taus de Kendall conditionnels, et nous explicitons leurs distributions asymptotiques sous l'hypothèse nulle. Dans les cas où de tels évènements conditionnants ne sont pas fixés à l'avance, nous proposons une procédure s'appuyant sur les données pour construire récursivement de tels ensembles conditionnants pertinents. Cette procédure est basée sur des arbres de décision qui maximisent les différences entre les taus de Kendall conditionnels, qui correspondent aux feuilles de l'arbre. Les résultats empiriques de ces tests sont présentés dans le contenu supplémentaire. De plus, nous étudions la dépendance conditionnelle entre plusieurs rendements d'actifs et illustrons certains effets particuliers de contagion par les rendements passés. La dernière application traite de la dépendance conditionnelle entre des montants de couverture dans une base de données d'assurance.

Additional Supporting Information may be found in the online version of this article at the publisher's website.
* *Corresponding author: a.f.f.derumigny@tudelft.nl*

1. INTRODUCTION

Copulas express the dependence structures of multivariate random vectors independent of their marginal distributions. Therefore, they often allow insightful two-step model specifications and inference. Copulas have motivated many academic papers during the last decades and have become popular in many applied fields. If several multivariate datasets of the same nature are available, then a natural question that arises is whether their dependence structures coincide. This problem was first tackled in Rémillard & Scaillet (2009) for the case of two samples. A general testing procedure for several samples was proposed in Bouzebda, Keziou & Zari (2011). Seo (2020) revisited this two-sample problem by introducing modified randomization procedures. In a related paper, Quessy (2016) proposed quadratic-type statistics for testing whether a given collection of induced lower-dimensional copulas from a multivariate distribution are identical. The k-sample problem for extreme-value copulas was tackled in Bücher, Kinsvater & Kojadinovic (2017). A similar question arises for conditional dependence structures in multivariate datasets. Currently, a key problem is in testing whether some conditional copulas differ or not, given different conditioning events. This is the motivation for our work.

To be specific, assume that we observe n independent and identically distributed (IID) replications $\left(\left(\mathbf{X}_{i,I}^\top, \mathbf{X}_{i,J}^\top\right)^\top\right)_{i=1,\dots,n}$ of a random vector $\mathbf{X} := \left(\mathbf{X}_I^\top, \mathbf{X}_J^\top\right)^\top \in \mathbb{R}^d$ where, without loss of generality, the subset of conditioned variables is indexed by $I = \{1, \dots, p\}$ and the subset of conditioning variables by $J = \{p+1, \dots, d\}$ for some integer $p \in \{1, \dots, d\}$. For $k \in \{1, \dots, p\}$, let $F_{k|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$ be the conditional law of X_k given $\mathbf{X}_J = \mathbf{x}_J$, where the conditioning event corresponds to a fixed point $\mathbf{x}_J \in \mathbb{R}^{d-p}$. Following Patton (2006a,b) and Fermanian & Wegkamp (2012), a conditional copula of $\mathbf{X}_I$ given $\mathbf{X}_J = \mathbf{x}_J$, denoted as $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$, exists and is defined by an equivalent version of Sklar's theorem: for every $\mathbf{x}_I \in \mathbb{R}^p$ and almost all $\mathbf{x}_J \in \mathbb{R}^{d-p}$,

$$\mathbb{P}\left(\mathbf{X}_I \leq \mathbf{x}_I | \mathbf{X}_J = \mathbf{x}_J\right) = C_{I|J}\left(F_{1|J}\left(x_1|\mathbf{X}_J = \mathbf{x}_J\right), \dots, F_{p|J}\left(x_p|\mathbf{X}_J = \mathbf{x}_J\right) \mid \mathbf{X}_J = \mathbf{x}_J\right)$$

with almost sure uniqueness of $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$ when the conditional margins of $\mathbf{X}_I$ given $\mathbf{X}_J = \mathbf{x}_J$ are continuous. Note that the maps $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$, for $\mathbf{x}_J \in \mathbb{R}^{d-p}$, are different in general. Nonetheless, a desirable property would be their constancy with respect to the choice of the pointwise conditioning event, in particular, for inferential purposes (Hobæk Haff, Aas & Frigessi, 2010). This is the famous "simplifying assumption" that is standard for vine models (Kurz & Spanhel, 2017; Czado, 2019). It can be written as

$$\mathcal{H}_0 : C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J) \text{ does not depend on } \mathbf{x}_J \text{ for almost all } \mathbf{x}_J \in \mathbb{R}^{d-p}.$$

Many tests of this simplifying assumption have appeared in the literature (see Section 2.1 for a complete discussion). Without restricting assumptions (semiparametric models, linear approximations, single-index, etc.), such techniques require smoothing. Besides the choice of additional tuning parameters, such smoothing-based tests are no longer numerically feasible when the dimension of $\mathbf{X}_J$ is larger than three due to the curse of dimensionality. To circumvent this problem, we deal with more general measurable conditioning subsets instead of pointwise conditioning events.

For $k = 1, \dots, p$, let $F_{k|J}(\cdot|\mathbf{X}_J \in A_J)$ be the conditional law of X_k given that $\mathbf{X}_J$ belongs to a Borel subset A_J of $\mathbb{R}^{d-p}$ with $\mathbb{P}\left(\mathbf{X}_J \in A_J\right) > 0$. Similar to the pointwise case, the conditional copula of $\mathbf{X}_I$ given $\mathbf{X}_J \in A_J$ is defined as

$$\mathbb{P}\left(\mathbf{X}_I \leq \mathbf{x}_I | \mathbf{X}_J \in A_J\right) = C_{I|J}^{A_J}\left(F_{1|J}\left(x_1|\mathbf{X}_J \in A_J\right), \dots, F_{p|J}\left(x_p|\mathbf{X}_J \in A_J\right) \mid \mathbf{X}_J \in A_J\right),$$

for every $\mathbf{x}_J$. It can be easily checked that the conditional copula $C^{A_J}_{I|J}(\cdot|\mathbf{X}_J \in A_J)$ is the cumulative distribution function of the random vector $\big(F_{1|J}(X_1|\mathbf{X}_J \in A_J), \dots, F_{p|J}(X_p|\mathbf{X}_J \in A_J)\big)$ given the event $(\mathbf{X}_J \in A_J)$ when all the latter conditional margins are continuous. This will be assumed in this article.

Box-type events provide a natural framework in many situations. In finance and insurance, the analysis of dependencies between asset returns when some covariates belong to specific subspaces is of particular interest. In Section 7, we will study the level of dependence between two stock indices given the range of values of their past returns. As another example, bank stress tests will focus on some quantiles of losses, that is, $A_J = (\text{Loss} > q_\alpha)$. When dealing with high-level quantiles, it is no longer possible to rely on marginal or joint estimators given pointwise conditioning events as $A_J = (\text{Loss} = q_\alpha)$ due to the sparsity of relevant data. Moreover, when the dimension of $\mathbf{X}_J$ is larger than three, discretizing is often the only feasible way of building nonparametric and semiparametric dependence models in practice. Finally, in the case of categorical explanatory variables $\mathbf{X}_J$, the new conditional copulas $\big(C^{A_J}_{I|J}, A_J \in \mathcal{A}_J\big)$, where $\mathcal{A}_J = \big\{\{\mathbf{x}_J\} : \mathbf{x}_J \in \mathbb{R}^{d-p}, \mathbb{P}(\mathbf{X}_J = \mathbf{x}_J) > 0\big\}$, are obviously appropriate and very natural.

We now consider several such sets in order to compare the dependencies knowing $(\mathbf{X}_J \in A_J)$ for general different conditioning subsets A_J in a way similar to that for the usual simplifying assumption. Let $\mathcal{A}_J := \{A_{1,J}, \dots, A_{m,J}\}$ be a collection of m Borel sets $A_{k,J}$ with $\mathbb{P}\big(\mathbf{X}_J \in A_{k,J}\big) > 0$ for $k \in \{1, \dots, m\}$. This family $\mathcal{A}_J$ may be disjoint or even a partition of $\mathbb{R}^{d-p}$, but this is not mandatory in our framework. As in Derumigny & Fermanian (2017), consider the null hypothesis

$$\overline{\mathcal{H}}_0 : A_{k,J} \mapsto C^{A_{k,J}}_{I|J}(\cdot|\mathbf{X}_J \in A_{k,J}) \text{ is constant over } \mathcal{A}_J.$$

Importantly, it is known that neither of the simplifying assumptions $\mathcal{H}_0$ or $\overline{\mathcal{H}}_0$ implies the other when $\mathcal{A}_J$ is finite: see Section 3.1 of Derumigny & Fermanian (2017). In other words, there is no relationship between a test of $\mathcal{H}_0$ and a test of $\overline{\mathcal{H}}_0$ strictly speaking (except when $\mathbf{X}_J$ is discrete), even if they are based on similar intuitions.

Several omnibus tests of $\overline{\mathcal{H}}_0$ have already been proposed in Derumigny & Fermanian (2017). All of them are based on empirical counterparts of conditional distributions, with some integration on possibly high-dimensional spaces. This may induce burdensome numerical problems and slow calculations. To the best of our knowledge, the latter is the only work in the literature that formally tackles the problem of testing $\overline{\mathcal{H}}_0$. In this article, we propose a simpler and quicker omnibus test procedure. Our procedure is related to a less-demanding null hypothesis $\overline{\mathcal{H}}_0^{\,\tau}$ involving the equality of all bivariate conditional Kendall's taus associated with the random vector $\mathbf{X}_I$.

To be more specific, let us first formulate our null hypothesis when $\mathbf{X}_I$ is a bivariate vector, that is, when $p = 2$. In this case, the null hypothesis $\overline{\mathcal{H}}_0^{\,\tau}$ is

$$\overline{\mathcal{H}}_0^{\,\tau} : \tau_{1,2|\mathbf{X}_J \in A_{1,J}} = \tau_{1,2|\mathbf{X}_J \in A_{2,J}} = \dots = \tau_{1,2|\mathbf{X}_J \in A_{m,J}}, \tag{1}$$

where $\tau_{1,2|\mathbf{X}_J \in A_{k,J}}$ denotes Kendall's tau between X_1 and X_2 given $\mathbf{X}_J \in A_{k,J}$. In other words, we test whether the conditional Kendall's taus associated with every subset in $\mathcal{A}_J$ are equal. We recall that

$$\begin{aligned}\tau_{1,2|\mathbf{X}_J \in A_{k,J}} &:= 4\int C^{A_{k,J}}_{I|J}\left(u_1, u_2|\mathbf{X}_J \in A_{k,J}\right)\, C^{A_{k,J}}_{I|J}\left(\mathrm{d}u_1, \mathrm{d}u_2|\mathbf{X}_J \in A_{k,J}\right) - 1 \\ &= \mathbb{P}\left(\left(X_{11} - X_{21}\right)\left(X_{12} - X_{22}\right) > 0 \mid \mathbf{X}_{1,J} \in A_{k,J} \ , \ \mathbf{X}_{2,J} \in A_{k,J}\right)\end{aligned}$$

$$- \mathbb{P}\left(\left(X_{11} - X_{21}\right)\left(X_{12} - X_{22}\right) < 0 \mid \mathbf{X}_{1,J} \in A_{k,J},\ \mathbf{X}_{2,J} \in A_{k,J} \right)$$
$$= 4p_{1,2|A_{k,J}} - 1,$$

where

$$p_{1,2|A_{k,J}} := \mathbb{P}\left(X_{11} < X_{21}, X_{12} < X_{22} \mid \mathbf{X}_{1,J} \in A_{k,J},\ \mathbf{X}_{2,J} \in A_{k,J} \right)$$

and $\mathbf{X}_1$ and $\mathbf{X}_2$ are two independent copies of $\mathbf{X}$. Clearly, $\overline{\mathcal{H}}_0$ implies $\overline{\mathcal{H}}_0^{\tau}$ but the converse is not true. Therefore, if $\overline{\mathcal{H}}_0^{\tau}$ is not rejected (by a statistical test), this does not imply the same conclusion for $\overline{\mathcal{H}}_0$. Nonetheless, a reasonable lack of power may be accepted for a gain in terms of simplicity (implementation or interpretation). We argue that this is the case for some test statistics based on conditional Kendall's taus in Section 2.3.

For the general case with $p \geq 2$, consider Kendall's tau for all possible pairs (X_i, X_j) with $(i,j) \in \{1, \dots, p\}^2$ and $i < j$. For each $A_{k,J} \in \mathcal{A}_J$, there are $p(p-1)/2$ conditional Kendall's taus: we order them in the vector

$$\boldsymbol{\tau}_{I|\mathbf{X}_J \in A_{k,J}} := \left(\tau_{1,2|\mathbf{X}_J \in A_{k,J}}, \tau_{1,3|\mathbf{X}_J \in A_{k,J}}, \dots, \tau_{1,p|\mathbf{X}_J \in A_{k,J}}, \tau_{2,3|\mathbf{X}_J \in A_{k,J}}, \dots, \tau_{p-1,p|\mathbf{X}_J \in A_{k,J}} \right)^{\top}.$$

The null hypothesis $\overline{\mathcal{H}}_0^{\tau}$ in Equation (1) can be generalized as the new null hypothesis

$$\overline{\mathcal{H}}_0^{\boldsymbol{\tau}} : \boldsymbol{\tau}_{I|\mathbf{X}_J \in A_{1,J}} = \boldsymbol{\tau}_{I|\mathbf{X}_J \in A_{2,J}} = \dots = \boldsymbol{\tau}_{I|\mathbf{X}_J \in A_{m,J}}.$$

In other words, $\overline{\mathcal{H}}_0^{\boldsymbol{\tau}} = \cap_{a,b,a \neq b} \overline{\mathcal{H}}_0^{\tau_{a,b}}$, which corresponds to the assumption that the function $A_J \mapsto \boldsymbol{\tau}_{I|\mathbf{X}_J \in A_J}$ is constant over $\mathcal{A}_J$. Equivalently, this means that for every distinct pair of indices $(a,b) \in \{1, \dots, p\}^2$, the function $A_J \mapsto \tau_{a,b|\mathbf{X}_J \in A_J}$ is constant over $\mathcal{A}_J$.

Our tests will therefore be based on equality between some of the conditional Kendall's taus, as defined above. These quantities are unknown and have to be estimated over the corresponding subsamples that are of random sizes. Inference and hypothesis testing using Kendall's tau is popular in dependence modelling. This is, for example, the approach chosen in Jaser & Min (2021) to test symmetry and radial symmetry between two independent samples of equal size. However, our framework can deal with more than two subsamples and does not require the equality of their sample sizes.

We organize the rest of this article as follows. In Section 2, we discuss the relationship between the hypothesis $\overline{\mathcal{H}}_0^{\boldsymbol{\tau}}$ and the simplifying assumption as well as the link between pointwise conditional Kendall's tau and Spearman's rho. In Section 3, we introduce a statistic for testing the hypothesis $\overline{\mathcal{H}}_0^{\tau}$ and we state its limiting distribution. We generalize this construction for the hypothesis $\overline{\mathcal{H}}_0^{\boldsymbol{\tau}}$ in Section 4. Section 5 explains how a typical nonparametric bootstrapping scheme can be invoked to calculate P-values. In Section 6, we propose an algorithm to generate a relevant collection of sets $\mathcal{A}_J$. We provide two applications on real datasets in Section 7. First, we study conditional dependence between the S&P500 and the Eurostoxx indices in a copula-GARCH model for which we highlight some contagion phenomena. The second application focuses on dependence between different coverages in an insurance dataset. We defer proofs to the Appendix. We discuss the empirical properties of all methods in an extensive simulation study given in the Supplementary Material, which also contains two counterexamples that illustrate relationships between $\overline{\mathcal{H}}_0^{\tau}$ and $\mathcal{H}_0^{\boldsymbol{\tau}}$.

2. RELATIONSHIPS TO RELATED HYPOTHESES AND TESTS ABOUT CONDITIONAL COPULAS

2.1. Relationship between $\overline{\mathcal{H}}_0$ and the Simplifying Assumption of Pointwise Conditional Copulas

To test the simplifying assumption of pointwise constant conditional copulas, a generalized likelihood ratio test was introduced by Acar, Craiu & Yao (2013) under a semiparametric framework where a conditional copula belongs to a one-dimensional parametric family. This idea has been extended in Gijbels et al. (2017) in the case of unknown margins. Several tests were proposed in Derumigny & Fermanian (2017) in more general nonparametric and semiparametric settings. Levi & Craiu (2019) consider the problem of establishing the validity of the simplifying assumption in a Bayesian setting. To the best of our knowledge, all papers in the literature, except Derumigny & Fermanian (2017), deal with pointwise conditioning events in a fully nonparametric framework. The idea of discretizing the support of the conditioning vector appears in Kurz & Spanhel (2017) for D-vine parametric copula models. In particular, Gijbels, Omelka & Veraverbeke (2017) propose some tests of the simplifying assumption with score-type statistics and comparisons of pointwise conditional Kendall's taus with their average.

Whenever the collection $\mathcal{A}_J$ forms a partition of the support of $\mathbf{X}_J$, it is possible to define a random variable Y as the random integer $k = k(\mathbf{X}_J) \in \{1, \dots, m\}$ such that $Y \in A_{k(\mathbf{X}_J),J}$, as noted in Derumigny & Fermanian (2017, Section 3). The assumption $\overline{\mathcal{H}}_0$ is then equivalent to the assumption that the conditional copula of $\mathbf{X}_I$ given $Y = k$ is constant with respect to $k \in \{1, \dots, m\}$. However, since most existing tests of the simplifying assumption require that the covariate $\mathbf{X}_J$ has a continuous distribution, they rely on smoothing (e.g., by kernel or local-linear methods). Therefore, they cannot be directly applied in our setting, since the theoretical properties would not hold and the implementation could not manage discrete conditioning variables. Conversely, when the conditioning variable has a discrete, finite support, the simplifying assumption can be equivalently rewritten using $\overline{\mathcal{H}}_0$ and $A_{k,J} := \{\mathbf{x}_{k,J}\}$, where the support of $\mathbf{X}_J$ is $\{\mathbf{x}_{k,J} : k = 1, \dots, m\}$. This allows us to directly apply the tests presented in this article.

2.2. The Simplifying Assumption and Conditional Kendall's Tau

When the conditioning event is pointwise, the corresponding assumption is

$$\mathcal{H}_0^{\boldsymbol{\tau}} : \boldsymbol{\tau}_{I|\mathbf{X}_J=\mathbf{x}_J} \text{ does not depend on } \mathbf{x}_J \in \mathbb{R}^{d-p},$$

where $\boldsymbol{\tau}_{I|\mathbf{X}_J=\mathbf{x}_J}$ is the vector stacking all pointwise conditional Kendall's taus for every pair of distinct indices. Unfortunately, neither of the simplifying assumptions $\mathcal{H}_0^{\boldsymbol{\tau}}$ or $\overline{\mathcal{H}}_0^{\boldsymbol{\tau}}$ implies the other when $\mathcal{A}_J$ is finite. In other words, there exists a distribution such that $\overline{\mathcal{H}}_0^{\boldsymbol{\tau}}$ is satisfied and $\mathcal{H}_0^{\boldsymbol{\tau}}$ is not, and there exists another distribution such that $\mathcal{H}_0^{\boldsymbol{\tau}}$ is satisfied and $\overline{\mathcal{H}}_0^{\boldsymbol{\tau}}$ is not. For completeness, two such counterexamples are given in Section 8 of the Supplementary Material. To fuel intuition, we note that, in the case with $I = \{1, 2\}$, where $f_{\mathbf{X}_J}(\cdot)$ denotes the density of $\mathbf{X}_J$ with respect to the Lebesgue measure,

$$\begin{aligned} p_{1,2|A} &= \mathbb{P}\left(X_{11} < X_{21}, X_{12} < X_{22} | \mathbf{X}_{1,J} \in A, \mathbf{X}_{2,J} \in A\right) \\ &= \int_{A\times A} p_{1,2|\mathbf{x}_J,\mathbf{x}'_J} f_{\mathbf{X}_J}(\mathbf{x}_J) f_{\mathbf{X}_J}\left(\mathbf{x}'_J\right) \, \mathrm{d}\mathbf{x}_J \, \mathrm{d}\mathbf{x}'_J / \mathbb{P}(A)^2, \end{aligned}$$

where, for every $(\mathbf{x}_J, \mathbf{x}'_J) \in \mathbb{R}^{d-2} \times \mathbb{R}^{d-2}$,

$$p_{1,2|\mathbf{x}_J,\mathbf{x}'_J} = \mathbb{P}\left(X_{11} < X_{21}, X_{12} < X_{22} | \mathbf{X}_{1,J} = \mathbf{x}_J, \mathbf{X}_{2,J} = \mathbf{x}'_J\right).$$

In other words, $p_{1,2|A}$ is a weighted average of the quantities $p_{1,2|\mathbf{x}_J,\mathbf{x}'_J}$ when $\mathbf{x}_J$ and $\mathbf{x}'_J$ both describe A, and it is not an average of the pointwise conditional probabilities $p_{1,2|\mathbf{x}_J} = p_{1,2|\mathbf{x}_J,\mathbf{x}_J}$ (imposing $\mathbf{x}_J = \mathbf{x}'_J$) that yield $\tau_{1,2|\mathbf{X}_J=\mathbf{x}_J}$.

2.3. Conditional Spearman's Rho

As an alternative to $\overline{\mathcal{H}}_0^{\tau}$, we could consider testing for equality between conditional Spearman's rhos given $\mathbf{X}_J \in A_{k,J}$ for different subsets $A_{k,J}$, in the same vein as Gaißer & Schmid (2010). Nonetheless, this would require working with pseudo-observations, while our approach with estimated conditional Kendall's taus involves simpler analytic functions of the original sample. Generally speaking, working with Kendall's tau avoids many technicalities induced by the theory of empirical copulas and keeps the limiting laws rather simple. In particular, the asymptotic variances of our test statistics have straightforward empirical counterparts, while this is almost never the case for empirical copula processes. The same remark applies to the competing tests of the constancy of the conditional copula given $\mathbf{X}_J \in A_{k,J}$ as proposed in Derumigny & Fermanian (2017, Section 3). Moreover, the latter approach is based on the calculation of p-dimensional integrals, a numerically demanding task when $p > 3$.

3. ASYMPTOTIC TEST IN TWO DIMENSIONS

In this section, we study some test statistics for the null hypothesis $\overline{\mathcal{H}}_0^{\tau}$ and we derive their asymptotic distributions. A generalization for $p > 2$ is based on the same idea and is presented in the next section.

The hypothesis $\overline{\mathcal{H}}_0^{\tau}$ can be rewritten as an m-sample problem by defining the subsamples $\mathcal{S}_k := \{i = 1, \dots, n : \mathbf{X}_{i,J} \in A_{k,J}\}$ for $k \in \{1, \dots, m\}$. In this case, we want to test whether the dependence between X_1 and X_2 is the same across all the samples $\mathcal{S}_k$. Contrary to the usual m-sample problem, the sample sizes $N_{k,n}$ are random for fixed or data-driven "boxes" $A_{k,J}$ for $k \in \{1, \dots, m\}$ (see Section 6). Moreover, we do not restrict ourselves to disjoint samples. In other words, the conditioning subsets $A_{k,J}$ for $k \in \{1, \dots, m\}$ may intersect.

Following Derumigny & Fermanian (2019b), for any $k \in \{1, \dots, m\}$, candidate estimators of conditional Kendall's tau are

$$\hat{\tau}^{(1)}_{1,2|\mathbf{X}_J \in A_{k,J}} := 4 \sum_{i=1}^{n} \sum_{j=1}^{n} w^{(k)}_{i,n} w^{(k)}_{j,n} \mathbf{1}\left\{X_{i,1} < X_{j,1}, X_{i,2} < X_{j,2}\right\} - 1,$$

$$\begin{aligned}\hat{\tau}^{(2)}_{1,2|\mathbf{X}_J \in A_{k,J}} := \sum_{i=1}^{n} \sum_{j=1}^{n} w^{(k)}_{i,n} w^{(k)}_{j,n} \big(&\mathbf{1}\left\{(X_{i,1} - X_{j,1})(X_{i,2} - X_{j,2}) > 0\right\} \\ &-\mathbf{1}\left\{(X_{i,1} - X_{j,1})(X_{i,2} - X_{j,2}) < 0\right\}\big),\end{aligned}$$

and

$$\hat{\tau}^{(3)}_{1,2|\mathbf{X}_J \in A_{k,J}} := 1 - 4 \sum_{i=1}^{n} \sum_{j=1}^{n} w^{(k)}_{i,n} w^{(k)}_{j,n} \mathbf{1}\left\{X_{i,1} < X_{j,1}, X_{i,2} > X_{j,2}\right\},$$

where $w^{(k)}_{i,n} := \mathbf{1}\{X_{i,J} \in A_{k,J}\}/N_{k,n}$ for $i = 1, \dots, n$. In contrast to Jaser & Min (2021), we do not require equal sample sizes for $k \in \{1, \dots, m\}$.

With $s_n^{(k)} := \sum_{i=1}^{n}\left(w_{i,n}^{(k)}\right)^2$, it can be easily proved that $\hat{\tau}^{(1)}_{1,2|\mathbf{X}_J\in A_{k,J}}$ belongs to the interval $\left[-1, 1-2s_n^{(k)}\right]$, $\hat{\tau}^{(2)}_{1,2|\mathbf{X}_J\in A_{k,J}}$ stays in $\left[-1+s_n^{(k)},\ 1-s_n^{(k)}\right]$, and $\hat{\tau}^{(3)}_{1,2|\mathbf{X}_J\in A_{k,J}}$ is in $[-1+2s_n^{(k)},\ 1]$. Moreover, there exists a direct relationship between these three estimators. Indeed, as noted in Derumigny & Fermanian (2019b),

$$\hat{\tau}^{(1)}_{1,2|\mathbf{X}_J\in A_{k,J}} + s_n^{(k)} = \hat{\tau}^{(2)}_{1,2|\mathbf{X}_J\in A_{k,J}} = \hat{\tau}^{(3)}_{1,2|\mathbf{X}_J\in A_{k,J}} - s_n^{(k)}$$

almost surely. We can rescale the previous estimators so that they take values in the whole interval $[-1, 1]$:

$$\hat{\tau}_{1,2|\mathbf{X}_J\in A_{k,J}} := \frac{\hat{\tau}^{(1)}_{1,2|\mathbf{X}_J\in A_{k,J}}}{1-s_n^{(k)}} + \frac{s_n^{(k)}}{1-s_n^{(k)}} = \frac{\hat{\tau}^{(2)}_{1,2|\mathbf{X}_J\in A_{k,J}}}{1-s_n^{(k)}} = \frac{\hat{\tau}^{(3)}_{1,2|\mathbf{X}_J\in A_{k,J}}}{1-s_n^{(k)}} - \frac{s_n^{(k)}}{1-s_n^{(k)}}. \qquad (2)$$

The quantity $\hat{\tau}_{1,2|\mathbf{X}_J\in A_{k,J}}$ is our empirical Kendall's tau given $\mathbf{X}_J \in A_{k,J}$ for any $k \in \{1, \ldots, m\}$. It coincides with the usual Kendall's tau based on the subsample $\mathcal{S}_k$.

Under $\overline{\mathcal{H}}_0^{\tau}$, all the conditional Kendall's taus are the same and many test statistics could be proposed. In particular, we build a test of $\overline{\mathcal{H}}_0^{\tau}$ based on a random vector whose components have the form

$$\Delta_{k,l} := \sqrt{n}\left(\hat{\tau}_{1,2|\mathbf{X}_J\in A_{k,J}} - \hat{\tau}_{1,2|\mathbf{X}_J\in A_{l,J}}\right),$$

for some (k, l) in $\{1, \ldots, m\}^2$ with $k < l$. Since $\mathbb{P}(\mathbf{X}_J \in A_{k,J}) = \mu_k > 0$, the estimator $\hat{\tau}_{1,2|\mathbf{X}_J\in A_{k,J}}$ is constructed on a subsample that is roughly (but not exactly) a fraction μ_k of the whole sample. Therefore, the random size $N_{k,n}$ will have an influence on the joint limiting law of $\Delta_{k,l}$ for any (k, l). To find this law, we will first consider the law of the random vectors

$$\hat{\mathbf{W}}_{1,2} := \sqrt{n}\left(\hat{\tau}_{1,2|\mathbf{X}_J\in A_{1,J}} - \tau_{1,2|\mathbf{X}_J\in A_{1,J}}, \ldots, \hat{\tau}_{1,2|\mathbf{X}_J\in A_{m,J}} - \tau_{1,2|\mathbf{X}_J\in A_{m,J}}\right)^{\top}$$

and

$$\hat{\mathbf{W}}^{(j)}_{1,2} := \sqrt{n}\left(\hat{\tau}^{(j)}_{1,2|\mathbf{X}_J\in A_{1,J}} - \tau_{1,2|\mathbf{X}_J\in A_{1,J}}, \ldots, \hat{\tau}^{(j)}_{1,2|\mathbf{X}_J\in A_{m,J}} - \tau_{1,2|\mathbf{X}_J\in A_{m,J}}\right)^{\top},$$

for $j \in \{1, 2, 3\}$. Their laws will be deduced from the limiting law of

$$\hat{\mathbf{V}} := \left(\hat{D}_1 - D_1, \ldots, \hat{D}_m - D_m, \hat{p}_1 - p_1, \ldots, \hat{p}_m - p_m\right)^{\top},$$

where

$$\hat{D}_k := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\mathbf{1}\left\{X_{i,1} < X_{j,1}, X_{i,2} < X_{j,2}, \mathbf{X}_{i,J} \in A_{k,J}, \mathbf{X}_{j,J} \in A_{k,J}\right\},$$

$$D_k := \mathbb{E}[\hat{D}_k] = \mathbb{P}\left(X_{i,1} < X_{j,1}, X_{i,2} < X_{j,2}, \mathbf{X}_{i,J} \in A_{k,J}, \mathbf{X}_{j,J} \in A_{k,J}\right),$$

$\hat{p}_k := \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{\mathbf{X}_{i,J} \in A_{k,J}\}$, and $p_k = \mathbb{P}\big(\mathbf{X}_J \in A_{k,J}\big)$ for $k \in \{1, \ldots, m\}$.

Denote by $\mathbb{P}_k$ the law of $\mathbf{X}$ given $\mathbf{X}_J \in A_{k,J}$, i.e. $\mathbb{P}_k(\mathrm{d}\mathbf{x}) = \mathbf{1}\{\mathbf{x}_J \in A_{k,J}\}\ \mathbb{P}(\mathrm{d}\mathbf{x})/p_k$. Moreover, set

$$\pi(\mathbf{x}_1, \mathbf{x}_2) := \big(\mathbf{1}\{x_{1,1} < x_{2,1}, x_{1,2} < x_{2,2}\} + \mathbf{1}\{x_{2,1} < x_{1,1}, x_{2,2} < x_{1,2}\}\big)/2,$$

$$I_{k,l} := \int \mathbf{1}\{\mathbf{x}_{3,J} \in A_{k,J} \cap A_{l,J}\} \pi(\mathbf{x}_1, \mathbf{x}_3) \pi(\mathbf{x}_2, \mathbf{x}_3)\ \mathbb{P}_k(d\mathbf{x}_1)\ \mathbb{P}_l(d\mathbf{x}_2)\ \mathbb{P}(d\mathbf{x}_3),$$

$$J_{k,l} := \int \mathbf{1}\{\mathbf{x}_{2,J} \in A_{k,J} \cap A_{l,J}\} \pi(\mathbf{x}_1, \mathbf{x}_2)\ \mathbb{P}_k(d\mathbf{x}_1)\ \mathbb{P}(d\mathbf{x}_2),$$

and $p_{k,l} := \mathbb{P}(X_J \in A_{k,J} \cap A_{l,J})$ for every $k, l \in \{1, \dots, m\}$. The above notations imply $D_k = p_k^2 \int \pi(\mathbf{x}_1, \mathbf{x}_2)\ \mathbb{P}_k(d\mathbf{x}_1)\ \mathbb{P}_k(d\mathbf{x}_2) = p_k J_{k,k}$.

Theorem 1. *When n tends to infinity, $\sqrt{n}\ \hat{\mathbf{V}}$ tends in law to the $2m$-dimensional Gaussian distribution vector $\mathcal{N}(0, \Sigma)$, where*

$$\Sigma := \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{1,2}^\top & \Sigma_{2,2} \end{bmatrix},$$

$\Sigma_{1,1} := \left[\sigma_{k,l}\right]_{k,l=1}^m$ *with* $\sigma_{k,l} := 4 p_k p_l I_{k,l} - 4 D_k D_l$, $\Sigma_{1,2} := \left[2 p_k J_{k,l} - 2 D_k p_l\right]_{k,l=1}^m$, *and* $\Sigma_{2,2} := \left[p_{k,l} - p_k p_l\right]_{k,l=1}^m$.

We now state the limiting law of $\hat{\mathbf{W}}_{1,2}^{(j)}$ for $j \in \{1, 2, 3\}$, and $\hat{\mathbf{W}}_{1,2}$. By virtue of Equation (2) and the relation $s_n^{(k)} = 1/(n\hat{p}_k) = O(n^{-1})$, all four statistics have the same limiting law. This asymptotic law is presented in the next proposition, whose proof can be found in the Appendix.

Proposition 1. *When n tends to infinity, $\sqrt{n}\ \hat{\mathbf{W}}_{1,2}$ and $\sqrt{n}\hat{\mathbf{W}}_{1,2}^{(j)}$ for $j \in \{1, 2, 3\}$ tend in law to the m-dimensional Gaussian distribution $\mathcal{N}(0, \Delta)$, where*

$$\Delta := 64 \left[\frac{I_{kl}}{p_k p_l} + \frac{D_k D_l p_{k,l}}{p_k^3 p_l^3} - \frac{D_l J_{k,l}}{p_k p_l^3} - \frac{D_k J_{l,k}}{p_l p_k^3} \right]_{k,l=1}^m.$$

When the subsets $(A_{k,J})_{k=1,\dots,m}$ are disjoint, we simply get the diagonal matrix

$$\Delta = \mathrm{Diag}(\Delta_k)_{k=1}^m := 16\ \mathrm{Diag}\Big(4 I_{k,k}/p_k^2 - \big(1 + \tau_{1,2|\mathbf{X}_J \in A_{k,J}}\big)^2/(4p_k)\Big)_{k=1}^m.$$

It is easy to consistently estimate the limiting variance–covariance matrix Δ by replacing every unknown expectation with its empirical counterpart. For instance, in the disjoint case, replace the conditional Kendall's taus with their estimators, replace p_k with $\hat{p}_k$ and estimate $I_{k,k}$ as

$$\hat{I}_{k,k} := \frac{1}{n^3 \hat{p}_k^2} \sum_{i_1=1}^n \sum_{i_2=1}^n \sum_{i_3=1}^n \pi\big(\mathbf{X}_{i_1}, \mathbf{X}_{i_3}\big) \pi\big(\mathbf{X}_{i_2}, \mathbf{X}_{i_3}\big) \mathbf{1}\big\{\mathbf{X}_{i_1,J} \in A_{k,J}, \mathbf{X}_{i_2,J} \in A_{k,J}, \mathbf{X}_{i_3,J} \in A_{k,J}\big\}.$$

This yields the estimator

$$\hat{\Delta} := \mathrm{Diag}\big(\hat{\Delta}_k\big)_{k=1}^m := 16\ \mathrm{Diag}\Big(4 \hat{I}_{k,k}/\hat{p}_k^2 - \big(1 + \hat{\tau}_{1,2|\mathbf{X}_J \in A_{k,J}}\big)^2/(4\hat{p}_k)\Big)_{k=1}^m.$$

However, the cost of computing $\hat{I}_{k,k}$ grows as $O(n^3)$, which can be quite high for large subsample sizes.

To build a test statistic for $\overline{\mathcal{H}}_0^{\tau}$, one can consider a subset $\mathcal{S}$ of q pairs of indices (k_i, l_i) in $\{1, \dots, m\}^2$ where $k_i \neq l_i$, $i \in \{1, \dots, q\}$ and $q < m$. A $q \times m$ contrast matrix T will describe this set $\mathcal{S}$ in the following way: on every row of T, say the ith, all components are zero except the k_ith and the l_ith ones since $(k_i, l_i) \in \mathcal{S}$: these elements are 1 and −1, respectively. By construction, the rank of T is q. We can test any linear null hypothesis of the form

$$T\left(\tau_{1,2|X_J \in A_{1,J}}, \tau_{1,2|X_J \in A_{2,J}}, \dots, \tau_{1,2|X_J \in A_{m,J}}\right)^{\top} = \mathbf{0}_q,$$

where $\mathbf{0}_q$ is a q-dimensional vector of zeros. For instance, for the null hypothesis $\overline{\mathcal{H}}_0^{\tau}$, the contrast matrix T with rank $m-1$ may be chosen as

$$T := \left[\mathbf{1}_{m-1} \quad -I_{m-1}\right], \tag{3}$$

where $\mathbf{1}_{m-1}$ is an $(m-1)$-dimensional vector of ones and I_m is the $(m-1)$-dimensional identity matrix. For $m = 4$, this yields the matrix

$$T = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix}$$

whose rank is three. This reduces to testing the null hypothesis $\overline{\mathcal{H}}_0^{\tau} : \tau_{1,2|X_J \in A_{1,J}} = \tau_{1,2|X_J \in A_{k,J}}$ for all $k \in \{2, \dots, m\}$.

Due to Proposition 1, the random vector $\sqrt{n}\, T\hat{\mathbf{W}}_{1,2}$ asymptotically follows the nondegenerate Gaussian distribution $\mathcal{N}\left(0, T\Delta T^{\top}\right)$ and so the following statement holds.

Corollary 1. *Under the null hypothesis* $\overline{\mathcal{H}}_0^{\tau}$, $\mathcal{T}_n := \hat{\mathbf{W}}_{1,2}^{\top} T^{\top}\left(T\hat{\Delta}T^{\top}\right)^{-1} T\hat{\mathbf{W}}_{1,2}$ *tends in law to a chi-squared distribution with* $m-1$ *degrees of freedom.*

The columns of the contrast matrix T in Equation (3) can be arbitrarily permuted without changing the limiting law of $\mathcal{T}_n$ under $\overline{\mathcal{H}}_0^{\tau}$. If the column of ones is the jth column (with $j \neq 1$) of T, then the corresponding equivalent formulation of the null hypothesis is

$$\overline{\mathcal{H}}_0^{\tau} : \tau_{1,2|X_J \in A_{j,J}} = \tau_{1,2|X_J \in A_{k,J}} \text{ for } k \in \{1, \dots, m\} \backslash \{j\}.$$

Therefore, without loss of generality, we consider only the contrast matrix in Equation (3) from here on.

4. ASYMPTOTIC TEST IN HIGHER DIMENSIONS

In this section, we deal with a p-dimensional subvector $\mathbf{X}_I$ with $p > 2$. As before, we still test whether the conditioning subsets $A_{k,J}$ for $k \in \{1, \dots, m\}$ influence the underlying conditional copula of $\mathbf{X}_I$ given $\mathbf{X}_J \in A_{k,J}$ (i.e., $\overline{\mathcal{H}}_0$). A natural approach would be to rely on bivariate (conditional) Kendall's taus as before, but with all possible pairs (X_a, X_b) for $(a, b) \in \{1, \dots, p\}^2$ and $a < b$. For the given family $\mathcal{A}_J$, the limiting law of the stacked random vector of interest,

$$\hat{\mathbf{W}} := \left(\hat{\mathbf{W}}_{1,2}^{\top}, \hat{\mathbf{W}}_{1,3}^{\top}, \dots, \hat{\mathbf{W}}_{1,p}^{\top}, \hat{\mathbf{W}}_{2,3}^{\top}, \dots, \hat{\mathbf{W}}_{p-1,p}^{\top}\right)^{\top}$$

or

$$\hat{\mathbf{W}}^{(j)} := \left(\hat{\mathbf{W}}_{1,2}^{(j)\top}, \hat{\mathbf{W}}_{1,3}^{(j)\top}, \dots, \hat{\mathbf{W}}_{1,p}^{(j)\top}, \hat{\mathbf{W}}_{2,3}^{(j)\top}, \dots, \hat{\mathbf{W}}_{p-1,p}^{(j)\top}\right)^\top,$$

for $j \in \{1, 2, 3\}$, each of size $mp(p-1)/2$, is needed. Here, we use the notation

$$\hat{\mathbf{W}}_{a,b} := \sqrt{n}\left(\hat{\tau}_{a,b|\mathbf{X}_J \in A_{1,J}} - \tau_{a,b|\mathbf{X}_J \in A_{1,J}}, \dots, \hat{\tau}_{a,b|\mathbf{X}_J \in A_{m,J}} - \tau_{a,b|\mathbf{X}_J \in A_{m,J}}\right)^\top$$

and

$$\hat{\mathbf{W}}_{a,b}^{(j)} := \sqrt{n}\left(\hat{\tau}^{(j)}_{a,b|\mathbf{X}_J \in A_{1,J}} - \tau_{a,b|\mathbf{X}_J \in A_{1,J}}, \dots, \hat{\tau}^{(j)}_{a,b|\mathbf{X}_J \in A_{m,J}} - \tau_{a,b|\mathbf{X}_J \in A_{m,J}}\right)^\top,$$

for every pair of indices (a, b) in $\{1, \dots, p\}^2$ with $a < b$.

The corresponding limiting laws are Gaussian. A test of $\cap_{a,b,a \neq b} \overline{\mathcal{H}}_0^{\tau_{a,b}}$, and then of $\overline{\mathcal{H}}_0$, can be based on a linear transformation of $\hat{\mathbf{W}}$. For example, such a test could be based on the average value of conditional Kendall's tau over all possible pairs (X_a, X_b) with $a, b \in \{1, \dots, p\}$ and $a < b$ in the spirit of Kendall & Smith (1940).

To derive the asymptotic distribution of $\hat{\mathbf{W}}$ and $\hat{\mathbf{W}}^{(j)}$, we need to generalize Theorem 1. First define

$$\hat{\mathbf{V}} := \left(\hat{\mathbf{V}}_{1,2}^\top, \hat{\mathbf{V}}_{1,3}^\top, \dots, \hat{\mathbf{V}}_{p-1,p}^\top, \hat{p}_1 - p_1, \dots, \hat{p}_m - p_m\right)^\top,$$

$$\hat{\mathbf{V}}_{a,b} := \left(\hat{D}_{a,b,1} - D_{a,b,1}, \dots, \hat{D}_{a,b,m} - D_{a,b,m}\right)^\top,$$

$$\hat{D}_{a,b,k} := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbf{1}\left\{X_{i,a} < X_{j,a}, X_{i,b} < X_{j,b}, X_{i,J} \in A_{k,J}, X_{j,J} \in A_{k,J}\right\},$$

and

$$D_{a,b,k} := \mathbb{E}[\hat{D}_{a,b,k}] = \mathbb{P}\left(X_{i,a} < X_{j,a}, X_{i,b} < X_{j,b}, X_{i,J} \in A_{k,J}, X_{j,J} \in A_{k,J}\right),$$

for every pair of indices (a, b) in $\{1, \dots, p\}^2$ with $a < b$. Moreover, set

$$\pi_{a,b}(\mathbf{x}_1, \mathbf{x}_2) := \left(\mathbf{1}\{x_{1,a} < x_{2,a}, x_{1,b} < x_{2,b}\} + \mathbf{1}\{x_{2,a} < x_{1,a}, x_{2,b} < x_{1,b}\}\right)/2,$$

$$I_{a,b,a',b',k,l} := \int \mathbf{1}\{\mathbf{x}_{3,J} \in A_{k,J} \cap A_{l,J}\} \pi_{a,b}(\mathbf{x}_1, \mathbf{x}_3) \pi_{a',b'}(\mathbf{x}_2, \mathbf{x}_3)\, \mathbb{P}_k(d\mathbf{x}_1)\, \mathbb{P}_l(d\mathbf{x}_2)\, \mathbb{P}(d\mathbf{x}_3),$$

and

$$J_{a,b,k,l} := \int \mathbf{1}\{\mathbf{x}_{2,J} \in A_{k,J} \cap A_{l,J}\} \pi_{a,b}(\mathbf{x}_1, \mathbf{x}_2)\, \mathbb{P}_k(d\mathbf{x}_1)\, \mathbb{P}(d\mathbf{x}_2).$$

Note that $D_{a,b,k,k} = p_k^2 \int \pi_{a,b}(\mathbf{x}_1, \mathbf{x}_2)\, \mathbb{P}_k(d\mathbf{x}_1)\, \mathbb{P}_k(d\mathbf{x}_2) = p_k J_{a,b,k,k}$.

Theorem 2. *When n tends to infinity, $\sqrt{n}\, \hat{\mathbf{V}}$ tends in law to the $(mp(p-1)/2 + p)$-dimensional Gaussian distribution $\mathcal{N}(0, \Sigma_e)$. The block matrix Σ_e is*

$$\Sigma_e := \begin{bmatrix} \Sigma_{(1,2),(1,2)} & \Sigma_{(1,2),(1,3)} & \cdots & \Sigma_{(1,2),(p-1,p)} & \Sigma_{(1,2),0} \\ \Sigma_{(1,3),(1,2)} & \Sigma_{(1,3),(1,3)} & \cdots & \Sigma_{(1,3),(p-1,p)} & \Sigma_{(1,3),0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Sigma_{(p-1,p),(1,2)} & \Sigma_{(p-1,p),(1,3)} & \cdots & \Sigma_{(p-1,p),(p-1,p)} & \Sigma_{(p-1,p),0} \\ \Sigma_{0,(1,2)} & \Sigma_{0,(1,3)} & \cdots & \Sigma_{0,(p-1,p)} & \Sigma_{0,0} \end{bmatrix},$$

where $\Sigma_{(a,b),(a',b')} = \Sigma^{\top}_{(a',b'),(a,b)}$ *for every* (a,b) *and* (a',b')*. Similarly,* $\Sigma_{(a,b),0} = \Sigma^{\top}_{0,(a,b)}$*. Each blockwise component of* Σ *is an* $m \times m$ *matrix with the form*

$$\Sigma_{(a,b),(a',b')} := \left[4p_k p_l I_{a,b,a',b',k,l} - 4D_{a,b,k}D_{a',b',l}\right]_{k,l=1}^{m},$$

$$\Sigma_{(a,b),0} := \left[2p_k J_{a,b,k,l} - 2D_{a,b,k}p_l\right]_{k,l=1}^{m},$$

or

$$\Sigma_{0,0} := \left[p_{k,l} - p_k p_l\right]_{k,l=1}^{m}.$$

We now state the limiting law of $\hat{\mathbf{W}}^{(j)}$, for $j \in \{1,2,3\}$, and $\hat{\mathbf{W}}$. As before, all four statistics have the same limiting law.

Proposition 2. *When n tends to infinity,* $\sqrt{n}\,\hat{\mathbf{W}}$ *and* $\sqrt{n}\,\hat{\mathbf{W}}^{(j)}$*, for* $j = 1,2,3$*, tends in law to the* $(mp(p-1)/2)$*-dimensional Gaussian distribution* $\mathcal{N}(0,\Delta_e)$*, where* Δ_e *is a square matrix of size* $mp(p-1)/2$ *that can be written blockwise as*

$$\Delta_e := \begin{bmatrix} \Delta_{(1,2),(1,2)} & \Delta_{(1,2),(1,3)} & \cdots & \Delta_{(1,2),(p-1,p)} \\ \Delta_{(1,3),(1,2)} & \Delta_{(1,3),(1,3)} & \cdots & \Delta_{(1,3),(p-1,p)} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{(p-1,p),(1,2)} & \Delta_{(p-1,p),(1,3)} & \cdots & \Delta_{(p-1,p),(p-1,p)} \end{bmatrix},$$

where, for any (a,b) *and* (a',b')*,* $\Delta_{(a,b),(a',b')}$ *is the matrix of size* $m \times m$ *defined by*

$$\Delta_{(a,b),(a',b')} := 64\left[\frac{I_{a,b,a',b',k,l}}{p_k p_l} + \frac{D_{a,b,k}D_{a',b',l}p_{k,l}}{p_k^3 p_l^3} - \frac{D_{a',b',l}J_{a,b,k,l}}{p_k p_l^3} - \frac{D_{a,b,k}J_{a',b',l,k}}{p_l p_k^3}\right]_{k,l=1}^{m}.$$

In particular, when the subsets $A_{k,J}$ are disjoint over $k \in \{1,\dots,m\}$, then every submatrix $\Delta_{(a,b),(a',b')}$ is diagonal and

$$\Delta_{(a,b),(a',b')} := 16\,\mathrm{Diag}\left(\frac{4I_{a,b,a',b',k,k}}{p_k^2} - \frac{\left(1+\tau_{a,b|\mathbf{X}_J\in A_{k,J}}\right)\left(1+\tau_{a',b'|\mathbf{X}_J\in A_{k,J}}\right)}{4p_k}\right)_{k,l=1}^{m}$$

since $4D_{a,b,k}/p_k^2 = 1 + \tau_{a,b|\mathbf{X}_J\in A_{k,J}}$ for every (a,b) and every k.

As in Section 3, the components of Δ_e can be empirically estimated. For any $a < b \in \{1,\dots,p\}$ with $a < b$, introduce

$$\hat{I}_{a,b,a',b',k,k} := \frac{1}{n^3\hat{p}_k^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{i_3=1}^{n}\pi_{a,b}\left(\mathbf{X}_{i_1},\mathbf{X}_{i_3}\right)\pi_{a',b'}\left(\mathbf{X}_{i_2},\mathbf{X}_{i_3}\right) \times \mathbf{1}\left\{\mathbf{X}_{i_1,J}\in A_{k,J},\mathbf{X}_{i_2,J}\in A_{k,J},\mathbf{X}_{i_3,J}\in A_{k,J}\right\}.$$

Then, in the case of disjoint boxes, a consistent estimator of $\Delta_{(a,b),(a',b')}$ is

$$\hat{\Delta}_{(a,b),(a',b')} := 16\,\mathrm{Diag}\left(\frac{4\hat{I}_{a,b,a',b',k,k}}{\hat{p}_k^2} - \frac{\left(1+\hat{\tau}_{a,b|\mathbf{X}_J\in A_{k,J}}\right)\left(1+\hat{\tau}_{a',b'|\mathbf{X}_J\in A_{k,J}}\right)}{4\hat{p}_k}\right)_{k,l=1}^{m},$$

which suggests an estimator $\hat{\Delta}_e$ of the limiting covariance matrix Δ_e.

Recall that the matrix $T = \left[\mathbf{1}_{m-1} \ : \ -I_{m-1}\right]$ has a rank of $m-1$. Let T_e be the $(m-1)p(p-1)/2 \times mp(p-1)/2$ block matrix

$$T_e := I_{p(p-1)/2} \otimes T = \mathrm{Diag}(T, \dots, T) \tag{4}$$

whose rank is $(m-1)p(p-1)/2$. Then $\sqrt{n}T_e\hat{\mathbf{W}}$ tends in law to $\mathcal{N}\left(0, T_e\Delta_e T_e^\top\right)$ under the null hypothesis. Similar to Corollary 1, we can build a Wald-type test statistic as follows.

Corollary 2. *Under the null hypothesis* $\overline{\mathcal{H}}_0^{\tau}$, *the test statistic* $\mathcal{T}_n^{(e)} := n\hat{\mathbf{W}}^\top T_e^\top \left(T_e\Delta_e T_e^\top\right)^{-1} T_e\hat{\mathbf{W}}$ *converges in distribution to a chi-squared distribution with* $(m-1)p(p-1)/2$ *degrees of freedom.*

5. BOOTSTRAPPING AND OTHER TEST STATISTICS

In practice, the covariance matrix Δ_e can be consistently estimated using U-statistics of order three. The accuracy of the distributional approximation of the test statistic $\mathcal{T}_n^{(e)}$ can suffer in this estimation stage if the sample size n is not large enough. Besides, its computational complexity is $O(n^3)$ for fixed p and m. Therefore, we also consider two statistics that do not require the estimation of Δ_e and whose corresponding P-values can easily be bootstrapped.

The first test statistic is based on the maximal absolute deviation between two conditional Kendall's taus over all compared pairs of the conditioning subsets and is defined by

$$\mathcal{T}_{\infty,n} := \left|\sqrt{n}T_e\hat{\mathbf{W}}\right|_\infty.$$

The second test statistic is

$$\mathcal{T}_{2,n} := n\hat{\mathbf{W}}^\top T_e^\top T_e\hat{\mathbf{W}},$$

which is the sum of squared differences between pairs of Kendall's taus over all specified pairs of the conditioning subsets. Under the null hypothesis $\overline{\mathcal{H}}_0^{\tau}$, the asymptotic distributions of $\mathcal{T}_{\infty,n}$ and $\mathcal{T}_{2,n}$ are more complex than those of $\mathcal{T}_n^{(e)}$ and still depend on the unknown covariance matrix Δ_e. However, their computation does not require an estimation of the covariance matrix Δ_e and their asymptotic distributions can quickly be estimated using bootstrap techniques.

Remark 1. Both of the test statistics $\mathcal{T}_{\infty,n}$ and $\mathcal{T}_{2,n}$ depend on a choice of the contrast matrix T_e. This contrast matrix T_e can be chosen in a random way (such as a random permutation of a given contrast matrix). In this case, random contrast matrices have to be resampled in each bootstrap replication from the same distribution of contrast matrices. This will change the limiting law established in Corollaries 1 and 2. In the following, we prefer to choose a fixed contrast matrix that is reused for all bootstrap replications.

We consider the classical nonparametric bootstrap scheme introduced in Efron (1979). Here, the bootstrapped sample is obtained by resampling n observations $\mathbf{X}_i^*$ from the initial sample $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ with replacement. Denote by $\hat{\mathbf{W}}^*$ the bootstrapped version of $\hat{\mathbf{W}}$ built on the bootstrapped sample $\left(\mathbf{X}_1^*, \dots, \mathbf{X}_n^*\right)$ instead of $(\mathbf{X}_1, \dots, \mathbf{X}_n)$. The bootstrapped test statistics are, respectively,

$$\mathcal{T}_{\infty,n}^* := \left|\sqrt{n}T_e\hat{\mathbf{W}}^* - \sqrt{n}T_e\hat{\mathbf{W}}\right|_\infty$$

and

$$\mathcal{T}_{2,n}^* := n\left(\hat{\mathbf{W}}^* - \hat{\mathbf{W}}\right)^\top T_e^\top T_e\left(\hat{\mathbf{W}}^* - \hat{\mathbf{W}}\right).$$

These two statistics share the same asymptotic distributions as $\mathcal{T}_{\infty,n}$ and $\mathcal{T}_{2,n}$ respectively, under the null hypothesis $\overline{\mathcal{H}}_0^{\tau}$ when n tends to infinity. Their empirical P-values are computed as the empirical frequency of the events $\left(\mathcal{T}^*_{\infty,n} > \mathcal{T}_{\infty,n}\right)$ and $\left(\mathcal{T}^*_{2,n} > \mathcal{T}_{2,n}\right)$.

This empirical bootstrap technique is valid due to the consistency of the empirical bootstrap process (Van der Vaart & Wellner, 1996, Theorem 3.6.1). This can be applied without any hurdle because our test statistics can be approximated by IID expansions due to our Hájek projection technique (see $\tilde{D}_k$ and Equation (A1) in the proof of Theorem 1 in the Appendix). This directly implies that $\sqrt{n}\hat{\mathbf{W}}$ and $\sqrt{n}\left(\hat{\mathbf{W}}^* - \hat{\mathbf{W}}\right)$ share the same limiting law. As a consequence, $\mathcal{T}^*_{\infty,n}$ (respectively, $\mathcal{T}^*_{2,n}$) has the same weak limit as $\mathcal{T}_{\infty,n}$ (respectively, $\mathcal{T}_{2,n}$).

Remark 2. In this article, our test statistics are functions of realizations of $\mathbf{Y} := \left(\mathbf{X}_I, \mathbf{1}\left(\mathbf{X}_J \in A_1\right), \dots, \mathbf{1}\left(\mathbf{X}_J \in A_m\right)\right)$. This means that the random vector of interest is not $\mathbf{X}$ but, rather, $\mathbf{X}_I$ plus the index of the box that contains $\mathbf{X}_J$. As a consequence, the nonparametric bootstrap scheme is equivalent to the so-called "conditional bootstrap" scheme proposed in Derumigny & Fermanian (2017). Indeed, the probability of drawing $\mathbf{Y}_i$ is

$$q_i := \mathbb{P}(\text{the box } k(i) \text{ is drawn}) \times \mathbb{P}(\mathbf{X}_{i,I} \text{ is drawn} \mid k(i) \text{ has been drawn}),$$

where $k(i) \in \{1, \dots, m\}$ is the index of the box that contains $\mathbf{X}_{i,J}$. Obviously, $\mathbb{P}(\text{the box } \; k(i)$ is drawn$) = N_{k(i),n}/n$ and $\mathbb{P}(\mathbf{X}_{i,I}$ is drawn $\mid k(i)$ has been drawn$) = 1/N_{k(i),n}$. Therefore, $q_i = 1/n$, for every $i \in \{1, \dots, n\}$, corresponds to the resampling probability of $\mathbf{Y}_i$ within Efron's bootstrap scheme.

6. BUILDING RELEVANT BOXES

A problem may occur in practice when the dimension $d - p$ of the conditioning random vector $\mathbf{X}_J$ is larger than three or four. Indeed, except when the boxes are imposed by the particular geometry of the problem or by some specific prior information, it is not obvious what the most relevant boxes are. In other words, without knowing whether the dependence between the components of $\mathbf{X}_I$ depend on $\mathbf{X}_J$, it is of interest to build some boxes $A_{1,J}, \dots, A_{m,J}$ so that the dependence structures of $\mathbf{X}_I$ given $\mathbf{X}_J \in A_{k,J}$ for $k \in \{1, \dots, m\}$ are "as different as possible" from each other. This practical problem is particularly relevant in vine structures for which we want to weaken the standard simplifying assumption. In other words, it makes sense to build a realistic vine model for which the dependence copulas of any pair (X_1, X_2) given $\mathbf{X}_J = \mathbf{x}_J$ are not constant (the usual simplifying assumption) or a continuous function of $\mathbf{x}_J$ (a difficult task in terms of model specification in general) but, rather, an intermediate solution: the copulas would be chosen among a finite number of conditional copulas of (X_1, X_2) given $\mathbf{X}_J \in A_{k,J}$ for $k \in \{1, \dots, m\}$.

To this end, it is necessary to build the boxes $A_{k,J}$ for $k \in \{1, \dots, m\}$. Assume that m is fixed. Intuitively, the best sets of boxes will be able to discriminate among the m corresponding conditional copulas in a clear-cut way. A simple solution is to rely on classification trees to build m boxes after successively splitting some components of $\mathbf{X}_J$ into two intervals. The loss function can be defined by a distance $d(\cdot,\cdot)$ between the conditional copulas at every stage. For instance, set $p = 2$ and consider a tree algorithm similar to the classification and regression tree (CART) algorithm (Friedman, Hastie & Tibshirani, 2001). In the first step, one searches for an index $k_1 \in \{p+1, \dots, d\}$ and a threshold t_1 so that

$$(k_1, t_1) = \arg\max_{k,t} d\left(C_{1,2|X_k \leq t}, C_{1,2|X_k > t}\right) + \text{pen}(k,t),$$

where the penalty pen function may be related to the size of the obtained boxes: for a non-negative tuning parameter α, set $\text{pen}(k,t) = \alpha \min\left(\mathbb{P}(X_k > t), \mathbb{P}(X_k \le t)\right)$. As a variant, we could impose a minimum size ν for all the boxes by choosing

$$\text{pen}(k,t) = \alpha \min\left(\mathbb{P}(X_k > t), \mathbb{P}(X_k \le t)\right) - M\mathbf{1}\left\{\min\left(\mathbb{P}(X_k > t), \mathbb{P}(X_k \le t)\right) < \nu\right\}$$

for some given, large constant $M \gg 1$ and a given, small $\nu < 1$. In practice, the conditional copulas have to be estimated from an n-sample of IID realizations of $\mathbf{X}$. The empirical criterion is

$$(k_1, t_1) = \arg\max_{k,t} d\left(\hat{C}_{1,2|X_k \le t}, \hat{C}_{1,2|X_k > t}\right) + \widehat{\text{pen}}(k,t),$$

where $\widehat{\text{pen}}(k,t) = \alpha \min\left(\mathbb{P}_n(X_k > t), \mathbb{P}_n(X_k \le t)\right)$ and where $\hat{C}_{1,2|X_k \le t}$ and $\hat{C}_{1,2|X_k > t}$ denote the estimated conditional copulas of X_1 and X_2 given $(X_k \le t)$ and $(X_k > t)$, respectively, and $\mathbb{P}_n$ denotes the empirical measure. The same procedure can then be recursively applied to the observations. See Section 9 of Friedman, Hastie & Tibshirani (2001) for details.

This procedure may be computationally expensive in general, especially due to the inference of the conditional copulas and the calculation of a distance between multivariate distribution functions. As an alternative, we now propose to replace the function $C_{1,2|X_k \le t}$ by some conditional dependence measure, namely, by conditional Kendall's tau. Indeed, the estimation of Kendall's tau (conditional or not) is related to a classification task, as noted in Derumigny & Fermanian (2019a). The new program has

$$(k_1, t_1) := \arg\max_{k,t} |p_{X_1,X_2|X_k \le t} - p_{X_1,X_2|X_k > t}|^{\gamma} + \text{pen}(k,t),$$

where

$$p_{X_1,X_2|X_k \le t} := \mathbb{P}\left(X_{1,1} \le X_{2,1}, X_{1,2} \le X_{2,2} | X_{1,k} \le t, X_{2,k} \le t\right)$$

and

$$p_{X_1,X_2|X_k > t} := \mathbb{P}\left(X_{1,1} \le X_{2,1}, X_{1,2} \le X_{2,2} | X_{1,k} > t, X_{2,k} > t\right),$$

for some $\gamma > 0$ and independent $\mathbf{X}_1$ and $\mathbf{X}_2$. The empirical version of the above is

$$(k_1, t_1) := \arg\max_{k,t} |\hat{p}_{X_1,X_2|X_k \le t} - \hat{p}_{X_1,X_2|X_k > t}|^{\gamma} + \widehat{\text{pen}}(k,t),$$

where

$$\hat{p}_{X_1,X_2|X_k \le t} := \frac{1}{n(n-1)\mathbb{P}_n(X_k \le t)^2} \sum_{i=1}^{n} \sum_{j=1, j \ne i}^{n} \mathbf{1}\left\{X_{i,1} \le X_{j,1}, X_{i,2} \le X_{j,2}, X_{i,k} \le t, X_{j,k} \le t\right\}$$

and

$$\hat{p}_{X_1,X_2|X_k > t} := \frac{1}{n(n-1)\mathbb{P}_n(X_k > t)^2} \sum_{i=1}^{n} \sum_{j=1, j \ne i}^{n} \mathbf{1}\left\{X_{i,1} \le X_{j,1}, X_{i,2} \le X_{j,2}, X_{i,k} > t, X_{j,k} > t\right\}.$$

The recursive procedure is repeated on the two datasets corresponding to the conditioning subsets $(X_{k_1} \le t_1)$ and $(X_{k_1} > t_1)$, respectively.

Several termination rules can be implemented for this procedure. The simplest is to stop the procedure when the number of obtained categories (boxes) is larger than m. When m is even, we can obtain exactly m categories. It is also possible to specify a minimum number of observations for each box and a minimum difference between the conditional Kendall's taus from two estimated boxes. A more sophisticated approach uses a "pruning" rule once a large tree has been built (Friedman et al. 2001, p. 270).

When $p > 2$ conditioned variables are available, the first step becomes

$$(i_1, j_1, k_1, t_1) := \arg\max_{i,j,k,t} |\hat{p}_{X_i,X_j|X_k \le t} - \hat{p}_{X_i,X_j|X_k > t}|^{\gamma} + \widehat{\text{pen}}(i, j, k, t). \tag{5}$$

The complete algorithm is presented in Algorithm 1, where we fix $\gamma = 1$ and we use the notation $[a, b]_k := \mathbb{R}^{k-p-1} \times [a, b] \times \mathbb{R}^{d-k}$ to denote the hyper-rectangle of (conditioning) points $\mathbf{x}_J$ satisfying $x_k \in [a, b]$. The algorithm consists of one recursive function CutCKT that chooses the best boxes $(-\infty, t]_k$ and $(t, \infty)_k$ for separating the conditional Kendall's taus between the pairs of the conditioned variables as much as possible. Each of these two sets is recursively partitioned in the same way until the sample size corresponding to each box is sufficiently small or until the difference in the conditional Kendall's taus is sufficiently small.

Algorithm 1. Recursive algorithm for building a set of relevant boxes for conditional Kendall's taus

```
def CutCKT(a dataset $\mathcal{D} \in \mathbb{R}^{n \times d}$, a subset $A \overset{default}{=} \mathbb{R}^{|J|}$, minCut $\geq 0$,
 minSize $\geq 0$):
    for $i \leftarrow 1$ to $p-1$ do
        for $j \leftarrow i+1$ to $p$ do
            for $k \leftarrow p+1$ to $d$ do
                foreach $t \in \mathbb{R}$ do
                    Diff$[i, j, k, t] \leftarrow \left|\hat{\tau}_{i,j|\mathbf{X}_J \in A \cap (-\infty, t]_k} - \hat{\tau}_{i,j|\mathbf{X}_J \in A \cap (t, +\infty)_k}\right|$;
    $(i^*, j^*, k^*, t^*) \leftarrow \arg\max_{i,j,k,t}$ Diff$[i, j, k, t]$;
    Box1 $\leftarrow A \cap (-\infty, t^*]_{k^*}$;
    Box2 $\leftarrow A \cap (t^*, +\infty)_{k^*}$;
    if $\min\left(\mathbb{P}_n(X_J \in \text{Box1}), \mathbb{P}_n(X_J \in \text{Box2})\right) <$ minSize or
     Diff$[i^*, j^*, k^*, t^*] <$ minCut then
        return $\left(\hat{\boldsymbol{\tau}}_{I|X_J \in A}\right)$.
    else
        Child_ $\leftarrow$ CutCKT($\mathcal{D}, A =$ Box1, minCut, minSize);
        Child$_+$ $\leftarrow$ CutCKT($\mathcal{D}, A =$ Box2, minCut, minSize);
        return $\left(\hat{\boldsymbol{\tau}}_{I|X_J \in A}, (i^*, j^*, k^*, t^*), \text{Child\_}, \text{Child}_+\right)$
```

The object returned by the function CutCKT is a proper binary tree. Its Root attribute stores a vector of the (estimated) unconditional Kendall's taus, $\hat{\boldsymbol{\tau}}_I$. If Root has two children, then it also stores the indices (i_1, j_1) of the pair of conditioned variables selected as having the maximum difference in conditional Kendall's tau following Equation (5). In this case, it also stores the index k_1 of the selected conditional variable as well as the threshold t_1. Recursively, a `Child` of any `Node` of the tree is either

- a final leaf of the tree, with a conditional Kendall's tau vector of $\hat{\boldsymbol{\tau}}_{I|\mathbf{X}_J \in A}$, where A is the box corresponding to the conditioning subset passed to the `Child`, or
- an internal leaf of the tree.

An internal leaf is composed of the conditional Kendall's tau vector $\hat{\boldsymbol{\tau}}_{I|\mathbf{X}_J \in A}$; the indices i^*, j^* and k^*; the threshold defining the split; and the two children corresponding to, respectively, the lower box $A \cap (-\infty, t^*]_{k^*}$ (adding the event $\{X_{k^*} \leq t^*\}$) and to the upper box $A \cap \left(t^*, +\infty\right)_{k^*}$ (adding the event $\{X_{k^*} > t^*\}$). The type of such a proper binary tree can be recursively defined as

$$\begin{aligned}
\texttt{Tree} = \hat{\boldsymbol{\tau}}_I \; &\Big\| \; \left(\hat{\boldsymbol{\tau}}_I, i_1, j_1, k_1, t_1, \texttt{Child}\left((-\infty, t_1]_{k_1}\right), \texttt{Child}\left((t_1, +\infty)_{k_1}\right)\right), \\
\texttt{Child}(A) = \hat{\boldsymbol{\tau}}_{I|\mathbf{X}_J \in A} \; &\Big\| \; \Big(\hat{\boldsymbol{\tau}}_{I|\mathbf{X}_J \in A}, i^*, j^*, k^*, t^*, \\
&\qquad \texttt{Child}\left(A \cap (-\infty, t^*]_{k^*}\right), \texttt{Child}\left(A \cap (t^*, +\infty)_{k^*}\right)\Big),
\end{aligned}$$

where the symbol || refers here to the union of types. Examples of such trees are displayed in Figures 1 and 2.

The use of the same sample for the construction of the boxes using Algorithm 1 and for testing $\overline{\mathcal{H}}_0^{\tau}$ may not be theoretically justified. Indeed, using the same sample for both tasks would certainly lead to the over-rejection of $\overline{\mathcal{H}}_0^{\tau}$. We would exactly calibrate our conditional Kendall's tau to get the largest difference among them in the population, which would not happen for fixed boxes, and this would yield a misleading P-value. Another solution would be to invoke the bootstrap for the construction of the tree at the same time to take into account that effect, that is, whenever we have a bootstrapped sample, we computed a new tree based on it, as well as a new test statistic (based on the partition given by this new, random tree that will differ for each bootstrapped sample). In this case, the asymptotic theory is different and is left for future research.

We propose to use a sample-splitting strategy: a fraction κ of the sample is given to Algorithm 1 in order to construct a set of relevant boxes while the rest of the sample is used to

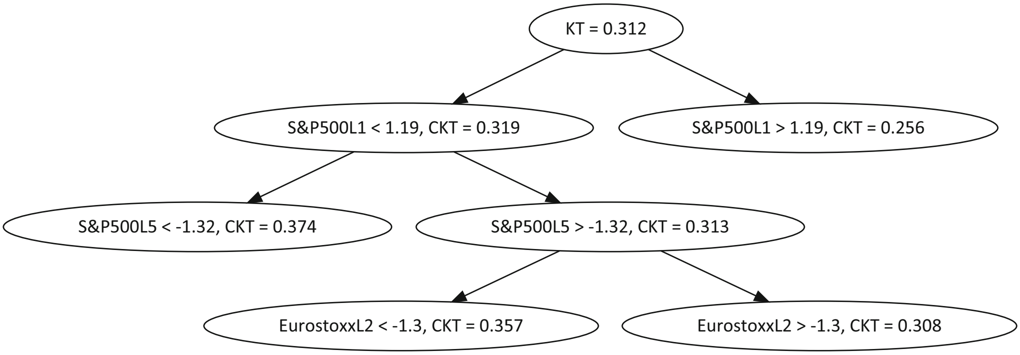
KT = 0.312
S&P500L1 < 1.19, CKT = 0.319
S&P500L1 > 1.19, CKT = 0.256
S&P500L5 < -1.32, CKT = 0.374
S&P500L5 > -1.32, CKT = 0.313
EurostoxxL2 < -1.3, CKT = 0.357
EurostoxxL2 > -1.3, CKT = 0.308

FIGURE 1: Data-driven tree for the conditional dependence between the Eurostoxx and S&P500 innovations. The conditioning variables are the lagged innovations $\mathbf{X}_{t,J2}$ including up to five lags. "KT" (respectively "CKT") denotes the unconditional Kendall's tau (respectively conditional Kendall's tau with the conditioning event corresponding to the node). For example, given $X_{t-1,2} < 1.19$ and $X_{t-5,2} > -1.32$, the conditional Kendall's tau between $X_{t,1}$ and $X_{t,2}$ is estimated as $\hat{\tau}_{X_{t,1},\, X_{t,2} \,|\, X_{t-1,2}<1.19,\, X_{t-5,2}>-1.32} = 0.313$.

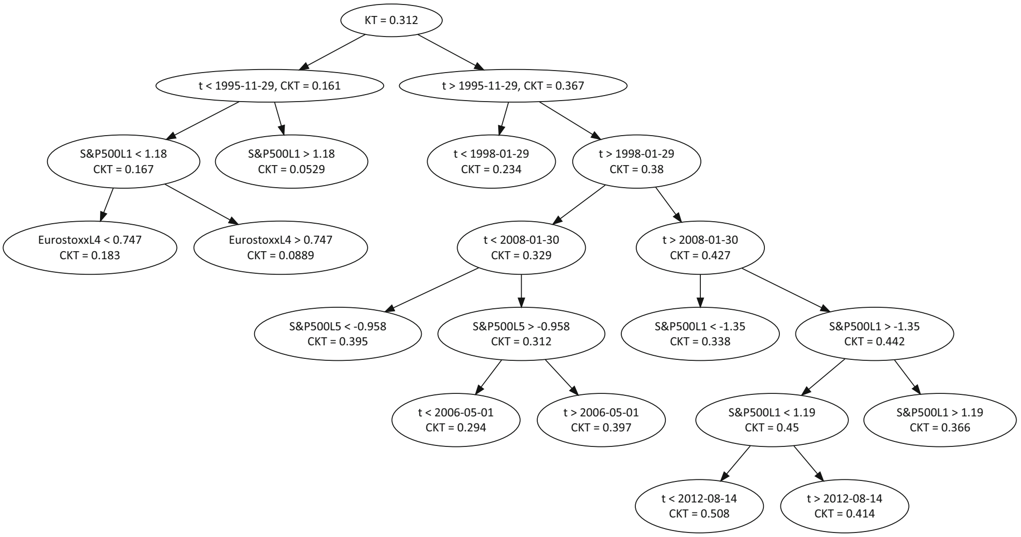

FIGURE 2: Data-driven tree for the conditional dependence between the Eurostoxx and S&P500 innovations. The conditioning variables are the lagged innovations $\mathbf{X}_{t,J3}$ including up to five lags and time. “KT” (respectively “CKT”) denotes the unconditional Kendall’s tau (respectively the conditional Kendall’s tau with the conditioning event corresponding to the node). For example, given $t <$ 1995-11-29 and $X_{t-1,2} <$ 1.18, conditional Kendall’s tau between $X_{t,1}$ and $X_{t,2}$ is estimated as $\hat{\tau}_{X_{t,1},\ X_{t,2} \mid\ t<\text{1995-11-29},\ X_{t-1,2}<1.18} = 0.167$.

compute the test statistic as well as its bootstrapped counterpart. In other words, bootstrapping is performed using only a fraction $1 - \kappa$ of the sample and the previously determined (i.e., fixed) boxes. This allows us not to contaminate the computation of the P-values with the information used to construct the tree and ensures that both parts of the process are independent. By default, we suggest $\kappa = 1/2$. The influence of κ is explored in Section 7 of the Supplementary Material.

Even when using sample splitting, it is important to note that the theoretical justification of our test procedures is only valid for fixed boxes (and not for random, i.e., data-driven, ones). Strictly speaking, the P-values obtained with the sample-splitting strategy are only valid conditional on the first part of the sample. Each collection of Borel sets $\mathcal{A}_J = \{A_{1,J}, \dots, A_{m,J}\}$ corresponds to a different null hypothesis that should be denoted by $\overline{\mathcal{H}}_0(\mathcal{A}_J)$ or $\overline{\mathcal{H}}_0^{\tau}(\mathcal{A}_J)$. Our splitting procedure can therefore be seen in the following way: the first part of the sample gives the statistician a relevant collection of Borel sets that correspond to the largest differences in terms of the estimated conditional Kendall’s tau, and the rest of the sample determines whether such differences are significantly different from zero.

7. TWO EMPIRICAL APPLICATIONS

The following empirical applications were performed in the statistical environment R using the functions `bCond.treeCKT` (which corresponds to Algorithm 1) and `bCond.simpA.CKT` (for the statistical test of constant dependence) from the package `CondCopulas` (Derumigny, 2022).

7.1. Financial Dataset

In this section, we consider time series for two stock indices, the Eurostoxx50 and the S&P500. The data are composed of $n = 8265$ observations of daily returns from January 5, 1987 to March 27, 2020. First, a preprocessing step is used to obtain an ARMA–GARCH filtering of each of the marginal processes. The orders are selected by minimizing the BIC using the R package `fGarch` (Wuertz et al., 2020). For the Eurostoxx (respectively, S&P500) returns, an ARMA(0,0)-GARCH(1,1) (respectively, ARMA(1,1)-GARCH(1,1)) model is selected. We denote by $X_{t,1}$ and $X_{t,2}$ the standardized residuals of the Eurostoxx and S&P500 returns, respectively, at time t. The vector of interest is $\mathbf{X}_{t,I} := \big(X_{t,1}, X_{t,2}\big)$.

First, we study the past residuals $\mathbf{X}_{t,J1} := L\mathbf{X}_{t,I} = \big(X_{t-1,1}, X_{t-1,2}\big)$, where L denotes the lag operator: we can apply Algorithm 1. Under the classical nonparametric bootstrap test procedure, the differences between our estimated conditional Kendall's taus are not significant (P-value = 0.244) and the assumption of a constant dependence structure cannot be rejected.

We then include more lags, with $\mathbf{X}_{t,J2} := \big(L\mathbf{X}_{t,I}, \dots, L^5\mathbf{X}_{t,I}\big)$. Using Algorithm 1, we obtain the tree displayed in Figure 1. Contrary to the case of a single lag, this partition proves to be relevant: it induces significantly different conditional Kendall's taus under the classical nonparametric bootstrap test procedure. The P-value is very close to zero.

By closer examination of the tree in Figure 1, we can distinguish several regimes: in the normal regime, when the first lag of the innovation of the S&P500 (denoted by "S&P500L1") is greater than 1.19, the conditional Kendall's tau between the two innovations is 0.256. More generally, we denote by "[name]L[i]" the ith lag of the innovation of the variable "[name]". We have an intermediate regime with S&P500L1 $<$ 1.19, EurostoxxL2 > -1.3, and S&P500L5 > -1.32 for which the conditional Kendall's tau is 0.308. The conditional Kendall's tau for the left leaf is even higher at 0.357. The last case corresponds to the event in which S&P500L1 and S&P500L5 are both lower than usual. This surely represents a crisis-like situation: here, the conditional Kendall's tau reaches its maximum value of 0.374.

It is interesting to note that, at each branch of the tree, the left-hand node (corresponding to a lower lagged innovation) always has a greater conditional Kendall's tau than that for the right-hand node. This illustrates the well-known contagion effect: when market conditions are bad, the dependencies between stock returns strengthen. Another way of seeing this contagion effect is by noticing that the tree displayed in Figure 1 is a binary search tree: every node has zero or two leaves and the value stored at each branch is smaller than the value on the left and bigger than the value on the right. This means that, for every subset A in the tree, for every conditioning variable X_k, and for every real x,

$$\tau_{1,2|\mathbf{X}_J \in A, X_k \le x} \ge \tau_{1,2|\mathbf{X}_J \in A} \ge \tau_{1,2|\mathbf{X}_J \in A, X_k > x}. \tag{6}$$

In other words, adding the information that $(X_k \le x)$ leads to an increase in the dependence.

In our last model, we also include time (even though it is not a random variable, strictly speaking) so that $\mathbf{X}_{t,J} := \big(L\mathbf{X}_{t,I}, \dots, L^5\mathbf{X}_{t,I}, t\big)$. This allows us to detect time-varying effects in the dependence between stock indices. Using Algorithm 1, we obtain the tree displayed in Figure 2. This partition induces strongly significant differences between the Kendall's taus (P-value = 0, under the classical nonparametric bootstrap test procedures). Three general effects can be noted.

- The dependence between the S&P500 and the Eurostoxx generally increases in time. This may be explained by financial globalization, by which major world indices become increasingly correlated with each other.
- The conditional dependence between the S&P500 and the Eurostoxx indices, given periods with strong, negative innovations, is generally higher than the conditional dependence between them during periods with strong, positive innovations. This corresponds to the contagion effect in Equation (6) that we have identified before.
- The contagion effect seems to be constant over time: Kendall's tau is around 0.10 higher in the "bad situations" than it is in the "good situations". In other words, $\tau_{1.2|\mathbf{X}_J \in \text{bad}, t \in T} - \tau_{1.2|\mathbf{X}_J \in \text{good}, t \in T}$ is nearly the same for each period of time.

More precisely, we see that the main split of the tree separates two periods of time. The second period is again split. In total, four main periods can be seen:

- 1987–1995 with a Kendall's tau of 0.161 between the S&P500 and the Eurostoxx,
- 1995–1998 with a Kendall's tau of 0.234 between the S&P500 and the Eurostoxx,
- 1998–2008 with a Kendall's tau of 0.329 between the S&P500 and the Eurostoxx, and
- 2008–2020 with a Kendall's tau of 0.427 between the S&P500 and the Eurostoxx.

These splits along the time variable illustrate the fact that the most important changes in the dependence between both financial indices are linked to time trends and not to past returns. This is coherent with intuition: long-term phenomena such as globalization have more influence than short-term events such as stock return variation a few days before. Note that the highest dependence levels are observed during the "hot times", 2008–2012, where financial markets suffered a financial crisis (Lehman's bankruptcy in particular) followed by the European sovereign debt crisis in 2010–2012.

In the branch of the tree corresponding to the period 1987–1995, there are three leaves, corresponding to three conditioning subsets describing bad, good and intermediate situations with increasing values of conditional Kendall's tau that satisfy Equation (6). The period 1995–1998 does not have enough observations to be split further as it is already very short. It represents a transition (Kendall's tau = 0.234) between the previous period (Kendall's tau = 0.161) and the next (Kendall's tau = 0.329). In the period 1998–2008, the Kendall's taus are higher than before, but the branches still satisfy the general contamination principle in Equation (6).

Interestingly, this principle is not satisfied in the most recent period (2008–2020) as the dependence during the "stressed event" $\{X_{t-1,2} < -1.35\}$ is in fact smaller (0.338) than during the complementary event (0.442). It could be possible that such extreme events are linked to purely American news that did not affect the Eurostoxx very much. Nevertheless, in the "normal" branch $\{X_{t-1,2} > -1.35\}$, the classical behaviour appears again, which suggests that the previous event corresponds to a very special situation.

7.2. Insurance Dataset

Frees, Lee & Yang (2016) presented an extensive analysis of an insurance dataset from the Wisconsin Local Government Property Insurance Fund using multivariate frequency–severity regression models. Their training sample covers the time period from 2006 to 2010 and consists of 41 variables and 5677 observations. Each observation corresponds to a local government entity, which is either a county, city, town, village, school or miscellaneous entity. The information about the type of a local entity, its number of claims and coverage sizes for a given year, insurance type, etc., is recorded through these 41 variables.

The training data from Frees, Lee & Yang (2016) consist of nominal (`Type`), categorical (`Year`), discrete and continuous variables. Continuous-type variables are non-negative and many of them have an atom at zero. We consider only the variables `Year` and `Type` as well as the three continuous variables listed in Table 1. Further, we restrict ourselves to the observations for which the three claim coverages of interest have a positive logarithm. The nominal variable `Type` classifies entities and consists of six categories `City`, `County`, `School`, `Town`, `Village` and `Miscellaneous`. We exclude observations of miscellaneous entities and deal with entities of the same type within each of the five remaining categories. We finally obtain 1435 observations overall. Figure 3 displays scatter plots of the three continuous variables, which are truncated for better illustration. The conclusions of our statistical analysis hold only for these three continuous variables conditioned on the event that the covered claims for the corresponding entity are larger than one million US dollars.

We apply the proposed framework to the three continuous variables to test whether their dependence structure varies across the five years (2006–2010). The conditioning variables

TABLE 1: Description of variables.

Variable	Description
Year	Claim year with values 2006, 2007, 2008, 2009, 2010
Type	Type of a local government entity with nominal values City, County, School, Town, Village, Miscellaneous
CoveragePN	Log-coverage amount of comprehensive new vehicles (PN), where coverage is in millions of dollars (non-negative or null)
CoveragePO	Log-coverage amount of comprehensive old vehicles (PO), where coverage is in millions of dollars (non-negative or null)
CoverageIM	Log-coverage amount of inland marine (IM), where coverage is in millions of dollars (non-negative or null)

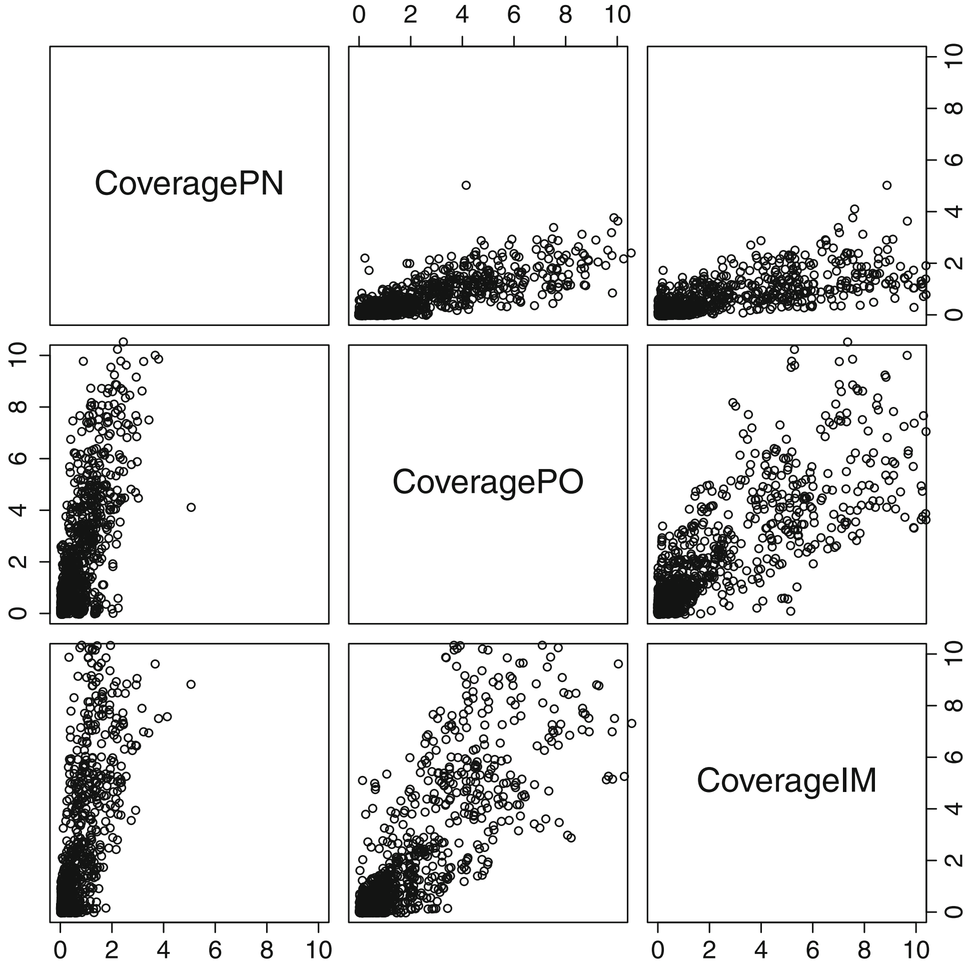

FIGURE 3: Scatter plots of CoveragePN, CoveragePO and CoverageIM.

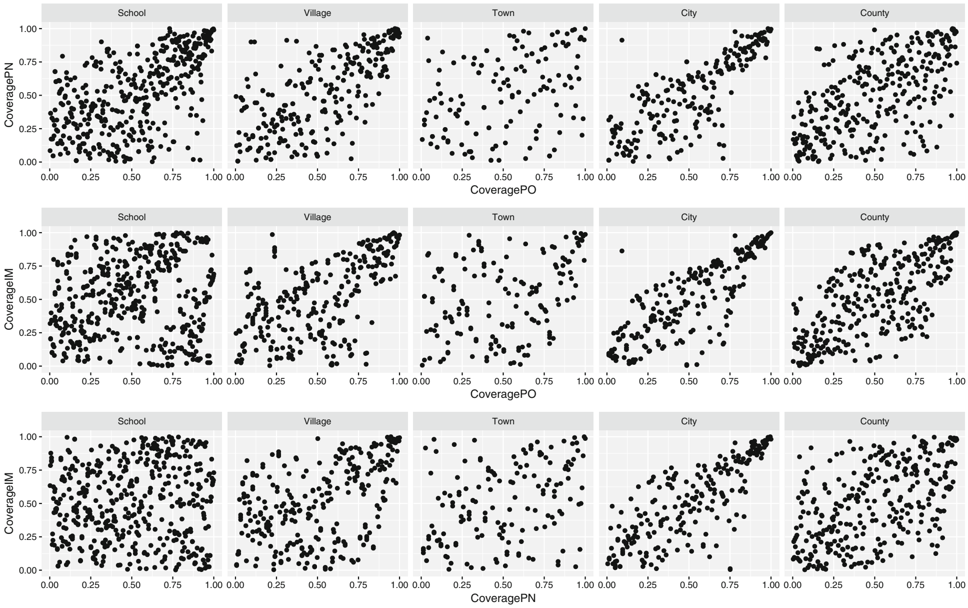

FIGURE 4: Scatter plots on the unit square of `CoveragePN`, `CoveragePO` and `CoverageIM` for each type of entity when transformed to have (conditionally) uniform margins.

specify boxes and do not have to be continuous. The null hypothesis is that the three Kendall's taus between the variables `CoveragePN`, `CoveragePO` and `CoverageIM` do not change over time. Under this null hypothesis, the test statistic $\mathcal{T}_n^{(e)}$ follows a chi-squared distribution with 12 degrees of freedom ($p = 3$ with five boxes), which has a level-0.95 quantile of 21.026. With Kendall's tau for the year 2006 as the reference value for comparisons with the other values of Kendall's tau, that is, we use the contrast matrix T_e from Equation (4), the statistic $\mathcal{T}_n^{(e)}$ is equal to 7.382 (P-value = 0.831) and the null hypothesis cannot be rejected at a 5% significance level. With 1000 bootstrap replicates of the dataset, the two tests based on the bootstrapped test statistics $\mathcal{T}^*_{\infty,n}$ and $\mathcal{T}^*_{2,n}$ also cannot reject the null hypothesis with P-values of 0.573 and 0.267, respectively.

We now can pool the data from the different years together to test the null hypothesis $\overline{\mathcal{H}}_0^{\tau}$ of constant conditional dependence expressed by the Kendall's taus for the five different entity types. Here, the null hypothesis can be rejected since the sample value of the test statistic $\mathcal{T}_n^{(e)}$ is equal to 232.363 (P-value = 0). With 1000 bootstrap replicates of the dataset, the two tests based on the bootstrapped test statistics also reject the null hypothesis as their P-values are equal to zero. Therefore, we can conclude that the dependence structure of nonzero log-coverage amounts for `CoveragePN`, `CoveragePO` and `CoverageIM` should be separately modelled for each entity type, additionally to their marginal, univariate modelling, which should also be separated for each entity type. Figure 4 visualizes our conclusion. In order to exclude the influence of conditional marginal distributions, the figure shows copula realizations obtained from the original data using the (conditional) marginal empirical distribution functions in each box. Thus, one can see in Figure 4 that the dependence between `CoverageIM` and `CoveragePO` for the entity `School` is significantly lower than for the entity `City`.

8. CONCLUSION

In this article, we propose to test the assumption of constant conditional dependence for a set of several conditioning events using conditional Kendall's tau. Our testing approach is very simple, does not rely on the theory of empirical processes on the theoretical side, and has favourable numerical performance. The asymptotic distribution of the proposed Wald-type test statistic is a chi-squared distribution independent of any conditional marginal distribution. To avoid estimating a high-dimensional covariance matrix, we additionally consider two alternative test statistics whose asymptotic distributions can be bootstrapped efficiently using the classical nonparametric bootstrap. In the Supplementary Material, we investigate the empirical level and power of the proposed tests with an extensive simulation study. An application to an insurance dataset illustrates the proposed test methods.

In most applications, conditioning events are not known ex ante. We construct them "blindly" and recursively in a way that maximizes the differences between conditional Kendall's taus and adapts the CART algorithm to the dependence framework. The output is a binary tree representing the decision paths to explain dependencies given some conditioning events. The leaves of the tree correspond to the final partition of the conditioning events. An application to a dataset of financial returns shows that the estimated binary search tree reflects increasing dependencies during crisis periods compared with noncrisis periods.

The proposed framework and ideas can be adapted to alternative dependence measures such as Spearman's rho at the price of additional technicalities related to the computation of conditional pseudo-observations. Moreover, several different multivariate dependence measures (Schmid & Schmidt, 2007a,b; Schmid et al., 2010; Genest, Nešlehová & Ben Ghorbal, 2011, for instance) could be grouped to build richer and more powerful test statistics. This will be the subject of future research.

ACKNOWLEDGEMENTS

Aleksey Min thanks Jae Youn Ahn for pointing out the insurance data from the Wisconsin Local Government Property Insurance Fund. Jean-David Fermanian's work has been supported by the Labex Ecodec (reference project ANR-11-LABEX-0047). Part of this work was realized while Alexis Derumigny was employed at the University of Twente. The authors are grateful to the Associate Editor and the two anonymous referees for their useful and constructive comments.

REFERENCES

Acar, E. F., Craiu, R. V., & Yao, F. (2013). Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics*, 7, 2822–2850.

Bouzebda, S., Keziou, A., & Zari, T. (2011). K-sample problem using strong approximations of empirical copula processes. *Mathematical Methods of Statistics*, 20(1), 14–29.

Bücher, A., Kinsvater, P., & Kojadinovic, I. (2017). Detecting breaks in the dependence of multivariate extreme-value distributions. *Extremes*, 20(1), 53–89.

Czado, C. (2019). *Analyzing Dependent Data with Vine Copulas*, Lecture Notes in Statistics, Springer, Cham.

Derumigny, A. (2022). *CondCopulas: Estimation and Inference for Conditional Copulas Models*, R package version 0.1.1. Available at: https://cran.r-project.org/package=CondCopulas

Derumigny, A. & Fermanian, J.-D. (2017). About tests of the "simplifying" assumption for conditional copulas. *Dependence Modeling*, 5(1), 154–197.

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

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.

Fermanian, J.-D. & Wegkamp, M. H. (2012). Time-dependent copulas. *Journal of Multivariate Analysis*, 110, 19–29.

Frees, E. W., Lee, G., & Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks*, 4(1), 4.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning*, Springer Series in Statistics, Vol. 1, Springer-Verlag, New York.

Gaißer, S. & Schmid, F. (2010). On testing equality of pairwise rank correlations in a multivariate random vector. *Journal of Multivariate Analysis*, 101(10), 2598–2615.

Genest, C., Nešlehová, J., & Ben Ghorbal, N. (2011). Estimators based on Kendall's tau in multivariate copula models. *Australian & New Zealand Journal of Statistics*, 53(2), 157–177.

Gijbels, I., Omelka, M., Pešta, M., & Veraverbeke, N. (2017). Score tests for covariate effects in conditional copulas. *Journal of Multivariate Analysis*, 159, 111–133.

Gijbels, I., Omelka, M., & Veraverbeke, N. (2017). Nonparametric testing for no covariate effects in conditional copulas. *Statistics*, 51(3), 475–509.

Hobæk Haff, I., Aas, K., & Frigessi, A. (2010). On the simplified pair-copula construction–simply useful or too simplistic? *Journal of Multivariate Analysis*, 101(5), 1296–1310.

Jaser, M. & Min, A. (2021). On tests for symmetry and radial symmetry of bivariate copulas towards testing for ellipticity. *Computational Statistics*, 36, 1–26.

Kendall, M. G. & Smith, B. B. (1940). On the method of paired comparisons. *Biometrika*, 31(3/4), 324–345.

Kurz, M. S. & Spanhel, F. (2017). Testing the simplifying assumption in high-dimensional vine copulas. arXiv preprint, arXiv:1706.02338.

Levi, E. & Craiu, R. V. (2019). Assessing data support for the simplifying assumption in bivariate conditional copulas. arXiv preprint, arXiv:1909.12688.

Patton, A. J. (2006a). Estimation of multivariate models for time series of possibly different lengths. *Journal of Applied Econometrics*, 21(2), 147–173.

Patton, A. J. (2006b). Modelling asymmetric exchange rate dependence. *International Economic Review*, 47(2), 527–556.

Quessy, J.-F. (2016). A general framework for testing homogeneity hypotheses about copulas. *Electronic Journal of Statistics*, 10(1), 1064–1097.

Rémillard, B. & Scaillet, O. (2009). Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100(3), 377–386.

Schmid, F. & Schmidt, R. (2007a). Multivariate conditional versions of Spearman's rho and related measures of tail dependence. *Journal of Multivariate Analysis*, 98(6), 1123–1140.

Schmid, F. & Schmidt, R. (2007b). Multivariate extensions of Spearman's rho and related statistics. *Statistics & Probability Letters*, 77(4), 407–416.

Schmid, F., Schmidt, R., Blumentritt, T., Gaißer, S., & Ruppert, M. (2010). Copula-based measures of multivariate association. In *Copula Theory and its Applications*, Springer, Berlin, Heidelberg, 209–236.

Seo, J. (2020). Randomization tests for equality in dependence structure. *Journal of Business & Economic Statistics*, 39, 1–12.

Van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*, Springer, New York.

Wuertz, D., Setz, T., Chalabi, Y., Boudt, C., Chausse, P. & Miklovac, M. (2020). *fGarch: Rmetrics-Autoregressive Conditional Heteroskedastic Modelling*. R package version 3042.83.2.

APPENDIX

Proof of Theorem 1

Consider a deterministic vector $\mathbf{a} \in \mathbb{R}^{2m}$. We only need to prove the asymptotic normality of the random variable $\sqrt{n}\,\mathbf{a}^\top \hat{V}$ since the weak convergence of $\sqrt{n}\,\hat{V}$ will be obtained via the usual Cramer–Wold device.

We approximate the U-statistic $\hat{D}_k$ by its Hájek projection. We first symmetrize the relevant quantities as

$$\hat{D}_k := \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} \left(g_{ij,k} + g_{ji,k}\right)$$

and

$$g_{ij,k} := \mathbf{1}\{X_{i,1} < X_{j,1}, X_{i,2} < X_{j,2}, X_{i,J} \in A_{k,J}, X_{j,J} \in A_{k,J}\}.$$

Set $g^*_{ij,k} := (g_{ij,k} + g_{ji,k})/2$ and $\tilde{D}_k := 2n^{-1}\sum_{i=1}^n \mathbb{E}\left[g^*_{i0,k}|\mathbf{X}_i\right] - D_k$ and introduce another independent realization $\mathbf{X}_0$. It follows that

$$\mathbb{E}[\tilde{D}_k] = \mathbb{E}\left[2g^*_{i0,k}\right] - D_k = 2\mathbb{E}[g_{1,2,k}] - D_k = D_k.$$

Simple calculations yield that

$$\begin{aligned}\hat{D}_k - \tilde{D}_k &= \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j=1,j\neq i}^n g^*_{ij,k} - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[g^*_{i0,k}|\mathbf{X}_i\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[g^*_{0i,k}|\mathbf{X}_i\big] + D_k \\ &= \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j=1,j\neq i}^n \Big(g^*_{ij,k} - \mathbb{E}\big[g^*_{ij,k}|\mathbf{X}_i\big] - \mathbb{E}\big[g^*_{ji,k}|\mathbf{X}_i\big] + D_k\Big) \\ &=: \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j=1,j\neq i}^n \overline{g}_{ij,k}.\end{aligned}$$

Note that $\mathbb{E}\big[\overline{g}_{ij,k}|\mathbf{X}_i\big] = \mathbb{E}\big[\overline{g}_{ij,k}|\mathbf{X}_j\big] = 0$. By standard reasoning for U-statistics,

$$\mathrm{Var}\left(\hat{D}_k - \tilde{D}_k\right) = \frac{1}{n^2(n-1)^2}\sum_{i_1=1}^n \sum_{j_1=1,j_1\neq i_1}^n \sum_{i_2=1}^n \sum_{j_2=1,j_2\neq i_2}^n \mathbb{E}\left[\overline{g}_{i_1j_1,k}\overline{g}_{i_2j_2,k}\right] = O\left(n^{-2}\right).$$

Indeed, the previous cross-products are zero when some of the four indices in (i_1, j_1, i_2, j_2) differ. This yields $\hat{D}_k = \tilde{D}_k + O_P(n^{-1})$ and we deduce that

$$\begin{aligned}\sqrt{n}\,\mathbf{a}^\top\hat{V} &= \sqrt{n}\sum_{k=1}^m a_k(\tilde{D}_k - D_k) + \sqrt{n}\sum_{k=m+1}^{2m} a_k(\hat{p}_k - p_k) + O_P\big(n^{-1/2}\big) \\ &= n^{-1/2}\sum_{i=1}^n \left\{\sum_{k=1}^m 2a_k\left(\mathbb{E}\left[g^*_{i0,k}|\mathbf{X}_i\right] - D_k\right) + \sum_{k=m+1}^{2m} a_k\left(\mathbf{1}\{X_{i,J}\in A_k\} - p_k\right)\right\} + O_P\big(n^{-1/2}\big) \\ &= n^{-1/2}\sum_{i=1}^n \mathbf{a}^\top\tilde{\mathbf{v}}_i + O_P\big(n^{-1/2}\big),\end{aligned} \tag{A1}$$

where $\tilde{\mathbf{v}}_i$ is the random vector

$$\begin{aligned}\tilde{\mathbf{v}}_i := \Big(&2\Big(\mathbb{E}\big[g^*_{i0,1}|\mathbf{X}_i\big] - D_1\Big)\ ,\ \dots\ ,\ 2\Big(\mathbb{E}\big[g^*_{i0,m}|\mathbf{X}_i\big] - D_m\Big), \\ &\mathbf{1}\{X_{i,J}\in A_1\} - p_1\ ,\ \dots\ ,\ \mathbf{1}\{X_{i,J}\in A_m\} - p_m\Big]^\top.\end{aligned}$$

By the usual central limit theorem, we deduce that $\sqrt{n}\,\mathbf{a}^\top\hat{V}$ tends in law to $\mathcal{N}(0, \mathbf{a}^\top\Sigma\mathbf{a})$, where $\Sigma = \mathbb{E}[\tilde{\mathbf{v}}_i^\top\tilde{\mathbf{v}}_i]$. Since this is true for every vector $\mathbf{a}$, this means that $\sqrt{n}\,\hat{V} \rightsquigarrow \mathcal{N}(0,\Sigma)$. Note that

$$\begin{aligned}\mathbb{E}\left[g^*_{i0,k}|\mathbf{X}_i\right] &= \mathbf{1}\{\mathbf{X}_{i,J}\in A_{k,J}\}\int \pi(\mathbf{x},\mathbf{X}_i)\mathbf{1}\{\mathbf{x}_J\in A_{k,J}\}\ \mathbb{P}(\mathrm{d}\mathbf{x}) \\ &= p_k\mathbf{1}\{\mathbf{X}_{i,J}\in A_{k,J}\}\int \pi(\mathbf{x},\mathbf{X}_i)\ \mathbb{P}_k(\mathrm{d}\mathbf{x}).\end{aligned}$$

Simple calculations yield the components of Σ. For example,

$$
\begin{aligned}
\sigma_{k,l} &:= 4\mathbb{E}\left[\mathbb{E}\left[g^*_{i0,k}|\mathbf{X}_i\right]\mathbb{E}\left[g^*_{i0,l}|\mathbf{X}_i\right]\right] - 4D_k D_l \\
&= 4p_k p_l \mathbb{E}\left[\mathbf{1}\{\mathbf{X}_{i,J} \in A_{k,J}, \mathbf{X}_{i,J} \in A_{l,J}\} \int \pi(\mathbf{x}_1, \mathbf{X}_i)\, \mathbb{P}_k(\mathrm{d}\mathbf{x}_1) \right. \\
&\qquad \left. \times \int \pi_l(\mathbf{x}_2, \mathbf{X}_i)\, \mathbb{P}_l(\mathrm{d}\mathbf{x}_2)\right] - 4D_k D_l.
\end{aligned}
$$

For $k = l$,

$$
\begin{aligned}
\sigma_{k,k} &= 4p_k^2 \mathbb{E}\left[\mathbf{1}\{\mathbf{X}_{i,J} \in A_{k,J}\} \int \pi(\mathbf{x}_1, \mathbf{X}_i)\, \mathbb{P}(\mathrm{d}\mathbf{x}_1) \int \pi(\mathbf{x}_2, \mathbf{X}_i)\, \mathbb{P}(\mathrm{d}\mathbf{x}_2)\right] - 4D_k^2 \\
&= 4p_k^3 \int_{\mathbf{x}_3 \in A_{k,J}} \left[\int \pi(\mathbf{x}_1, \mathbf{x}_3)\, \mathbb{P}_k(\mathrm{d}\mathbf{x}_1) \int \pi(\mathbf{x}_2, \mathbf{x}_3)\, \mathbb{P}_k(\mathrm{d}\mathbf{x}_2)\right] \mathbb{P}_k(\mathrm{d}\mathbf{x}_3) - 4D_k^2.
\end{aligned}
$$

Concerning $\Sigma_{1,2} := [\rho_{k,l}]_{k,l=1}^m$, we have that

$$
\begin{aligned}
\rho_{k,l} &= 2\mathbb{E}\left[\mathbb{E}\left[g^*_{i0,k}|\mathbf{X}_i\right] \mathbf{1}\{X_{i,J} \in A_{l,J}\}\right] - 2D_k p_l \\
&= 2p_k \mathbb{E}\left[\mathbf{1}\{\mathbf{X}_{i,J} \in A_{k,J} \cap A_{l,J}\} \int \pi(\mathbf{x}, \mathbf{X}_i)\, \mathbb{P}_k(\mathrm{d}\mathbf{x})\right] - 2D_k p_l\ .
\end{aligned}
$$

When $k = l$,

$$
\begin{aligned}
\rho_{k,k} &= 2p_k \mathbb{E}\left[\mathbf{1}\{\mathbf{X}_{i,J} \in A_{k,J}\} \int \pi(\mathbf{x}, \mathbf{X}_i)\, \mathbb{P}_k(\mathrm{d}\mathbf{x})\right] - 2D_k p_k \\
&= 2p_k^2 \int \left[\int \pi(\mathbf{x}_1, \mathbf{x}_2)\, \mathbb{P}_k(\mathrm{d}\mathbf{x}_1)\right] \mathbb{P}_k(\mathrm{d}\mathbf{x}_2) - 2D_k p_k \\
&= 2D_k - 2D_k p_k.
\end{aligned}
$$

Proof of Proposition 2

We first prove the asymptotic normality of $n^{1/2}\hat{W}^{(1)}$. The desired result follows. Note that, for $k \in \{1, \dots, m\}$, the kth component of $\hat{W}^{(1)}$ is

$$
\begin{aligned}
&\sqrt{n}\left(\hat{\tau}^{(1)}_{1,2|\mathbf{X}_J \in A_{k,J}} - \tau_{1,2|\mathbf{X}_J \in A_{k,J}}\right) = 4\sqrt{n}\left(\frac{\hat{D}_k}{\hat{p}_k^2} - \frac{D_k}{p_k^2}\right) \\
&= 4\sqrt{n}\left(\frac{\hat{D}_k - D_k}{p_k^2}\left(1 + \frac{p_k^2 - \hat{p}_k^2}{\hat{p}_k^2}\right) + \frac{D_k\left(p_k^2 - \hat{p}_k^2\right)}{\hat{p}_k^2 p_k^2}\right) \\
&= 4\sqrt{n}\left(\frac{\hat{D}_k - D_k}{p_k^2} + O_P\left(\left(\hat{D}_k - D_k\right)\left(\hat{p}_k - p_k\right)\right) - \frac{2D_k\left(\hat{p}_k - p_k\right)}{\hat{p}_k^2 p_k} - \frac{D_k\left(\hat{p}_k - p_k\right)^2}{\hat{p}_k^2 p_k^2}\right)
\end{aligned}
$$

$$= 4\sqrt{n}\left(\frac{\hat{D}_k - D_k}{p_k^2} - \frac{2D_k\left(\hat{p}_k - p_k\right)}{p_k^3}\right) + O_P\left(n^{-1/2}\right)$$

$$=: 4\zeta_k + O_P(n^{-1/2})$$

due to Theorem 1. Direct calculations provide, for every $k, l \in \{1, \dots, m\}$, that

$$\mathbb{E}[\zeta_k\zeta_l] = \frac{\sigma_{kl}}{p_k^2 p_l^2} + \frac{4D_k D_l\left(p_{k,l} - p_k p_l\right)}{p_k^3 p_l^3}$$

$$- \frac{2D_l\left(2p_k J_{k,l} - 2D_k p_l\right)}{p_k^2 p_l^3} - \frac{2D_k\left(2p_l J_{l,k} - 2D_l p_k\right)}{p_l^2 p_k^3}$$

$$= \frac{4I_{k,l}}{p_k^2 p_l^2} + \frac{4D_k D_l p_{k,l}}{p_k^3 p_l^3} - \frac{4D_l J_{k,l}}{p_k p_l^3} - \frac{4D_k J_{l,k}}{p_l p_k^3}.$$

Since $4D_k/p_k^2 = 1 + \tau_{1,2|\mathbf{X}_J \in A_{k,J}}$, the desired result holds when $k = l$.

Proof of Proposition 4

Consider a deterministic vector $\mathbf{a} \in \mathbb{R}^{m+p(p-1)m/2}$. It will be decomposed as a block vector $\mathbf{a} := \left[\mathbf{a}_0^\top, \mathbf{a}_{1,2}^\top, \mathbf{a}_{1,3}^\top, \dots, \mathbf{a}_{p-1,p}^\top\right]^\top$ with $\mathbf{a}_0 := \left[a_{0,1}, \dots, a_{0,m}\right]^\top$ and $\mathbf{a}_{a,b} := \left[a_{a,b,1}, \dots, a_{a,b,m}\right]^\top$.

After proving the asymptotic normality of the random variable $\sqrt{n}\ \mathbf{a}^\top \hat{V}$, the weak convergence of $\sqrt{n}\ \hat{V}$ can be obtained by invoking the usual Cramer–Wold device. As in the proof of Theorem 1, for any pair (a, b), define

$$\hat{D}_{a,b,k} := \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left(g_{ij,abk} + g_{ji,abk}\right)$$

and

$$g_{ij,abk} := \mathbf{1}\left\{X_{i,a} < X_{j,a}, X_{i,b} < X_{j,b}, X_{i,J} \in A_{k,J}, X_{j,J} \in A_{k,J}\right\}.$$

Set $g^*_{ij,abk} := (g_{ij,abk} + g_{ji,abk})/2$ and $\tilde{D}_{a,b,k} := 2n^{-1}\sum_{i=1}^n \mathbb{E}\left[g^*_{i0,abk}|\mathbf{X}_i\right] - D_{a,b,k}$. Obviously, $\mathbb{E}[\tilde{D}_{a,b,k}] = D_{a,b,k}$. Note that

$$\mathbb{E}\left[g^*_{i0,abk}|\mathbf{X}_i\right] = p_k \mathbf{1}\left\{\mathbf{X}_{i,J} \in A_{k,J}\right\} \int \pi_{abk}(\mathbf{x}, \mathbf{X}_i)\ \mathbb{P}_k(\mathrm{d}\mathbf{x}).$$

We obtain $\hat{D}_{a,b,k} = \tilde{D}_{a,b,k} + O_P(n^{-1})$ by a standard argument for U-statistic and we deduce that

$$\sqrt{n}\ \mathbf{ad}\hat{\mathbf{V}} = \sqrt{n}\sum_{(a,b)} \sum_{k=1}^{m} a_{a,b,k}\left(\tilde{D}_{a,b,k} - D_{a,b,k}\right) + \sqrt{n}\sum_{k=1}^{m} a_{0,k}\left(\hat{p}_k - p_k\right) + O_P\left(n^{-1/2}\right)$$

$$= n^{-1/2}\sum_{(a,b)} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{m} 2a_{a,b,k}\left(\mathbb{E}\left[g^*_{i0,abk}|\mathbf{X}_i\right] - D_{a,b,k}\right) + \sum_{k=1}^{m} a_{0,k}\left(\mathbf{1}\{X_{i,J} \in A_k\} - p_k\right) \right\}$$

$$
\begin{aligned}
&+ O_P(n^{-1/2}) \\
&= n^{-1/2}\sum_{i=1}^{n} \mathbf{a}\mathbf{d}\tilde{\mathbf{V}}_i + O_P(n^{-1/2}),
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{\mathbf{V}}_i &= \left[\tilde{\mathbf{V}}_{i,0}^\top, \tilde{\mathbf{V}}_{i,1,2}^\top, \tilde{\mathbf{V}}_{i,1,3}^\top, \dots, \tilde{\mathbf{V}}_{i,p-1,p}^\top\right]^\top, \\
\tilde{\mathbf{V}}_{i,a,b} &= \left[2\left(\mathbb{E}\left[g^*_{i0,ab1}|\mathbf{X}_i\right] - D_{a,b,k}\right), \dots, 2\left(\mathbb{E}\left[g^*_{i0,abm}|\mathbf{X}_i\right] - D_{a,b,m}\right)\right]^\top
\end{aligned}
$$

and

$$
\tilde{\mathbf{V}}_{i,0} := \left[\mathbf{1}\{X_{i,J} \in A_1\} - p_1, \dots, \mathbf{1}\{X_{i,J} \in A_m\} - p_m\right]^\top.
$$

By the usual central limit theorem, we deduce that $\sqrt{n}\,\mathbf{a}^\top\hat{\mathbf{V}}$ tends in law to $\mathcal{N}\left(0, \mathbf{a}^\top\Sigma\mathbf{a}\right)$, where $\Sigma = \mathbb{E}\left[\tilde{\mathbf{V}}_i^\top\tilde{\mathbf{V}}_i\right]$. Since this is true for every vector $\mathbf{a}$, this means that $\sqrt{n}\,\hat{\mathbf{V}} \rightsquigarrow \mathcal{N}(0, \Sigma_e)$. The calculation of Σ_e follows the calculations in the proof of Theorem 1.

Proof of Proposition 5

We first prove the asymptotic normality of $n^{1/2}\hat{\mathbf{W}}^{(1)}$. The desired result follows. As in the proof of Proposition 2, for every pair $(a, b) \in \{1, \dots, p\}^2$ with $a \neq b$ and every $k \in \{1, \dots, m\}$, we have that

$$
\begin{aligned}
\sqrt{n}\left(\hat{\tau}^{(1)}_{a,b|\mathbf{X}_J \in A_{k,J}} - \tau_{a,b|\mathbf{X}_J \in A_{k,J}}\right) &= 4\sqrt{n}\left(\frac{\hat{D}_{a,b,k}}{\hat{p}_k^2} - \frac{D_{a,b,k}}{p_k^2}\right) \\
&= 4\sqrt{n}\left(\frac{\hat{D}_{a,b,k} - D_{a,b,k}}{p_k^2} - \frac{2D_{a,b,k}\left(\hat{p}_k - p_k\right)}{p_k^3}\right) + O_P\left(n^{-1/2}\right) \\
&=: 4\zeta_{a,b,k} + O_P\left(n^{-1/2}\right)
\end{aligned}
$$

due to Theorem 2. Direct calculations provide, for every (a, b), (a', b') and for every $k, l \in \{1, \dots, m\}$, that

$$
\begin{aligned}
\mathbb{E}\left[\zeta_{a,b,k}\zeta_{a',b',l}\right] &= \frac{4p_k p_l I_{a,b,a',b',k} - 4D_{a,b,k}D_{a',b',l}}{p_k^2 p_l^2} + \frac{4D_{a,b,k}D_{a',b',l}\left(p_{k,l} - p_k p_l\right)}{p_k^3 p_l^3} \\
&\quad - \frac{2D_{a',b',l}\left(2p_k J_{a,b,k,l} - 2D_{a,b,k}p_l\right)}{p_k^2 p_l^3} - \frac{2D_{a,b,k}\left(2p_l J_{a',b',l,k} - 2D_{a',b',l}p_k\right)}{p_l^2 p_k^3} \\
&= 4\left(\frac{I_{a,b,a',b',k,l}}{p_k p_l} + \frac{D_{a,b,k}D_{a',b',l}p_{k,l}}{p_k^3 p_l^3} - \frac{D_{a',b',l}J_{a,b,k,l}}{p_k p_l^3} - \frac{D_{a,b,k}J_{a',b',l,k}}{p_l p_k^3}\right),
\end{aligned}
$$

which yields the desired result.

Received 11 September 2020
Accepted 7 April 2022