Computational Statistics and Data Analysis 135 (2019) 70–94

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A classification point-of-view about conditional Kendall's tau

Alexis Derumigny *, Jean-David Fermanian

CREST-ENSAE, 5, avenue Henry Le Chatelier, 91764 Palaiseau Cedex, France

ARTICLE INFO

Article history:
Received 26 June 2018
Received in revised form 25 January 2019
Accepted 26 January 2019
Available online 5 February 2019

Keywords:
Conditional Kendall's tau
Conditional dependence measure
Machine learning
Classification task
Stock indices

ABSTRACT

It is shown how the problem of estimating conditional Kendall's tau can be rewritten as a classification task. Conditional Kendall's tau is a conditional dependence parameter that is a characteristic of a given pair of random variables. The goal is to predict whether the pair is concordant (value of 1) or discordant (value of −1) conditionally on some covariates. The consistency and the asymptotic normality of a family of penalized approximate maximum likelihood estimators is proven, including the equivalent of the logit and probit regressions in our framework. Specific algorithms are detailed, adapting usual machine learning techniques, including nearest neighbors, decision trees, random forests and neural networks, to the setting of the estimation of conditional Kendall's tau. Finite sample properties of these estimators and their sensitivities to each component of the data-generating process are assessed in a simulation study. Finally, all these estimators are applied to a dataset of European stock indices.

1. Introduction

Besides linear correlations, most dependence measures between two random variables are functions of the underlying copula only: Spearman's rho, Kendall's tau, Blomqvist's beta, Gini's measure of association, etc. As a consequence, they are independent of the corresponding margins. This is seen as a positive point. See Joe (2015) and Nelsen (2007), for instance, and, as a reminder, some basic definitions in Appendix A. Such measures are well-known and widely used by practitioners. When some covariates are available, natural extensions of these tools can be defined, providing so-called "conditional" measures of dependence. In theory, it is sufficient to replace copulas by conditional copulas to obtain the "conditional version" of any dependence measure. Surprisingly, this simple and fruitful idea has not yet been widely used in the literature. Nonetheless, in a series of papers, Gijbels et al. (2011a,b, 2012, 2015) have popularized this approach, with a focus on conditional Kendall's tau and Spearman's rho. Note that conditional dependence measures have been invoked in different frameworks, often without any explicit link with conditional copulas: truncated data (e.g. Tsai, 1990), multivariate dynamic models (Jondeau and Rockinger, 2006; Almeida and Czado, 2012, among others), vine structures (So and Yeung, 2014), etc.

Now, let us introduce our key dependence measure: for each $\mathbf{z} \in \mathbb{R}^p$, the conditional Kendall's tau of a bivariate random vector $\mathbf{X} := (X_1, X_2)$ given some covariates $\mathbf{Z} = \mathbf{z}$ may be defined as

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

where $\mathbf{X}_1 = (X_{1,1}, X_{1,2})$ and $\mathbf{X}_2 = (X_{2,1}, X_{2,2})$ are two independent versions of $\mathbf{X}$. To simplify, we will assume that the law of $\mathbf{X}$ given $\mathbf{Z} = \mathbf{z}$ is continuous w.r.t. the Lebesgue measure, for every $\mathbf{z}$. This implies

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

* Corresponding author.
E-mail addresses: alexis.derumigny@ensae.fr (A. Derumigny), jean-david.fermanian@ensae.fr (J.-D. Fermanian).

https://doi.org/10.1016/j.csda.2019.01.013
0167-9473/

A conditional Kendall's tau belongs to the interval $[-1, 1]$ and reflects a positive ($\tau_{1,2|\mathbf{Z}=\mathbf{z}} > 0$) or negative ($\tau_{1,2|\mathbf{Z}=\mathbf{z}} < 0$) dependence between X_1 and X_2, given $\mathbf{Z} = \mathbf{z}$. Unlike correlations, this measure has the advantage of being always defined, even if some X_k, $k = 1, 2$, has no finite second moments (when it follows a Cauchy distribution, for example).

Some estimators of conditional Kendall's tau have already been proposed in the literature, either as a by-product of the estimation of conditional copulas – see Gijbels et al. (2011a) and Fermanian and Lopez (2018) – or directly, as in Derumigny and Fermanian (2018a,b). Nonetheless, to the best of our knowledge, nobody has yet noticed the relationship between conditional Kendall's tau and classification methods.

Let us explain this simple idea. Denote $W := 2 \times \mathbb{1}\{(X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0\} - 1$ and

$$\mathbb{P}\big((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) = \mathbb{P}\big(W = 1 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big) =: p(\mathbf{z}).$$

Actually, the prediction of concordance/discordance among pairs of observations $(\mathbf{X}_1, \mathbf{X}_2)$ given $\mathbf{Z}$ can be seen as a classification task of such pairs. If a model is able to evaluate the conditional probability of observing concordant pairs of observations, then it is able to evaluate conditional Kendall's tau, and the former quantity is one of the outputs of most classification techniques. Therefore, most classifiers can potentially be invoked (for example linear classifiers, decision trees, random forests, neural networks and so on Friedman et al. (2001)), but applied here to pairs of observations.

Indeed, for every $1 \leq i, j \leq n$, $i \neq j$, define $W_{(i,j)}$ as

$$W_{(i,j)} := 2 \times \mathbb{1}\{(X_{j,1} - X_{i,1})(X_{j,2} - X_{i,2}) > 0\} - 1 = \begin{cases} 1 & \text{if } (i,j) \text{ is a concordant pair,} \\ -1 & \text{if } (i,j) \text{ is a discordant pair.} \end{cases} \tag{1}$$

A classification technique will allocate a given couple (i, j) into one of the two categories $\{1, -1\}$ (or "concordant versus discordant", equivalently), with a certain probability, given the value of the common covariate $\mathbf{Z}$.

Section 2 introduces a general regression-type approach for the estimation of conditional Kendall's tau. Some asymptotic results of consistency and asymptotic normality are stated. In Section 3, we explain how some machine learning techniques can be adapted to deal with our particular framework, and we detail the corresponding algorithms. A small simulation study compares the small-sample properties of all these algorithms in Section 4. In Section 5, these techniques are applied to European stock market data. We evaluate to what extent the dependence between pairs of European stock indices may evolve with respect to different covariates. All proofs have been postponed into appendices.

2. Regression-type approach

Typically, a regression-type model based on conditional Kendall's tau may be written as

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = g_0(\mathbf{z}, \beta^*), \quad \forall \mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^p, \tag{2}$$

for some finite dimensional parameter $\beta^* \in \mathbb{R}^{p'}$ and some function g_0. As a particular case, a single-index approach would be

$$\tau_{1,2|\mathbf{Z}=\mathbf{z}} = g(\boldsymbol{\psi}(\mathbf{z})^T \beta^*), \quad \forall \mathbf{z} \in \mathcal{Z}, \tag{3}$$

where $\boldsymbol{\psi} : \mathbb{R}^p \to \mathbb{R}^{p'}$ is known, and g may be known (parametric model) or not (semi-parametric model), as in Derumigny and Fermanian (2018a). In this section, we propose an inference procedure of β^* under (3) when the link function g is analytically known. This procedure will be based on the signs of pairs only, and not on the specific values of the vectors $\mathbf{X}_i$. Then, since inference will be based on the observations of $W \in \{1, -1\}$, our model belongs to the family of limited-dependent variable methods. One difficulty will arise from the pointwise conditioning events $\mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}$, that will necessitate localization techniques. Actually, we will consider couples of observations $\mathbf{X}_i$ and $\mathbf{X}_j$ for which the associate covariates are close to a given value $\mathbf{z}$. Indeed, the relationship (3) does not define the dependence levels between every couple $(\mathbf{X}_i, \mathbf{Z}_i)$ and $(\mathbf{X}_j, \mathbf{Z}_j)$, $i \neq j$, but only between those that share the same covariate. If the variables $\mathbf{Z}$ were discrete, we would consider a subset of couples such that $\mathbf{Z}_i = \mathbf{Z}_j$. In our case of continuous variables $\mathbf{Z}$ (see below), the latter event does not occur almost surely, and some smoothing/localization techniques have to be invoked.

Let K be a p-dimensional kernel and (h_n) be a bandwidth sequence. The bandwidth will simply be denoted by h and we set $K_h(\mathbf{z}) = K(\mathbf{z}/h)/h^p$. The log-likelihood associated to the observation $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$ given $\mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}$ is

$$\ell_\beta(W_{(i,j)}, \mathbf{z}) := \left(\frac{1 + W_{(i,j)}}{2}\right) \log \mathbb{P}_\beta\left(W_{(i,j)} = 1 \Big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\right) + \left(\frac{1 - W_{(i,j)}}{2}\right) \log \mathbb{P}_\beta\left(W_{(i,j)} = -1 \Big| \mathbf{Z}_i = \mathbf{Z}_j = \mathbf{z}\right).$$

In practice, when the underlying law of $\mathbf{Z}$ is continuous, there is virtually no couple for which $\mathbf{Z}_i = \mathbf{Z}_j$. Therefore, we will consider a localized "approximated" log-likelihood, based on $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$ for all pairs (i, j), $i \neq j$. It will be defined as the double sum

$$L_n(\beta) := \frac{1}{n(n-1)} \sum_{i,j;\, i \neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j) \ell_\beta(W_{(i,j)}, \tilde{\mathbf{Z}}_{i,j}),$$

for any choice of $\tilde{\mathbf{Z}}_{i,j}$ that belongs to a neighborhood of $\mathbf{Z}_i$ or $\mathbf{Z}_j$. We will assume that K is a compactly supported p-dimensional kernel of order $m \geq 2$.

The most obvious choices would be to select $\tilde{\mathbf{Z}}_{i,j}$ among $\{\mathbf{Z}_i, \mathbf{Z}_j, (\mathbf{Z}_i + \mathbf{Z}_j)/2\}$. Here, we propose

$$\begin{aligned} L_n(\beta) &:= \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j)\ell_\beta(W_{(i,j)}, \mathbf{Z}_i) \\ &= \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j) \left\{ \left(\frac{1+W_{(i,j)}}{2}\right) \log\left(\frac{1}{2} + \frac{1}{2} g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right) \right. \\ &+ \left. \left(\frac{1-W_{(i,j)}}{2}\right) \log\left(\frac{1}{2} - \frac{1}{2} g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right) \right\}, \end{aligned}$$

under (3). We can therefore derive an estimator of β^* based on the maximization of the latter function, with a ℓ_1 penalty (Lasso-type estimator), as

$$\hat{\beta} := \arg\max_{\beta\in\mathbb{R}^{p'}} L_n(\beta) - \lambda_n|\beta|_1, \tag{4}$$

where λ_n (also simply denoted as λ) is a tuning parameter to be chosen. Note that $L_n(\beta)$ is not really a likelihood function since the observations $(W_{(i,j)}, \mathbf{Z}_i, \mathbf{Z}_j)$ for every couple (i, j), $i \neq j$, are not mutually independent.

If $K \geq 0$, the objective function is a concave function of β if it satisfies

$$\delta g''(t)(1 + \delta g)(t) \leq (g')^2(t), \ \forall t, \tag{5}$$

for $\delta \in \{1, -1\}$. When $\beta \mapsto L_n(\beta)$ is concave, the penalized criterion above is concave too and the calculation of $\hat{\beta}$ can be led in practical terms through convex optimization routines, even with a large number of regressors ($p' \gg 1$). Since this will be our framework, we will show that (5) holds for some usual classification techniques. When it is not the case, we have to rely on other optimization schemes and to avoid considering too many regressors.

Moreover, note that, when g is odd (i.e. $g(-t) = -g(t)$), the estimator simply becomes

$$\hat{\beta} := \arg\max_{\beta\in\mathbb{R}^{p'}} \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j) \log\left(\frac{1}{2} + \frac{1}{2} g(W_{(i,j)}\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right) - \lambda|\beta|_1. \tag{6}$$

The implementation of an algorithm to solve problem (4) or its simplified version (6) may seem difficult due to the non-differentiability of the l_1 norm. Nevertheless, as in the case of the ordinary Lasso, it can be solved in a very efficient way using the Alternative Direction Method of Multipliers (ADMM) for general l_1 minimization, following Boyd et al. (2011, Section 6.1). More precisely, assume $L_n(\beta)$ is a concave and differentiable function of β (this is the case in both Examples 1 and 2). Then the optimization task (4) can be rewritten as finding the solution $(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^{2p'}$ of

$$\begin{cases} \text{minimize } f(\mathbf{x}) + g(\mathbf{z}) \\ \text{subject to } \mathbf{x} - \mathbf{z} = 0, \end{cases} \tag{7}$$

where $f(\mathbf{x}) := -L_n(\mathbf{x})$ and $g(\mathbf{z}) := \lambda_n|\mathbf{z}|_1$. The solution is given by iterating the following algorithm, denoting by $\mathbf{u} \in \mathbb{R}^{p'}$ the dual variable of the problem (7) and by $\rho > 0$ the step size (similarly to the usual gradient descent algorithm),

$$\begin{aligned} \mathbf{x}^{k+1} &:= \arg\min_{\mathbf{x}} \left(f(\mathbf{x}) + (\rho/2)\|\mathbf{x} - \mathbf{z}^k + \mathbf{u}^k\|_2^2\right), \\ \mathbf{z}^{k+1} &:= S_{\lambda_n/\rho}(\mathbf{x}^{k+1} + \mathbf{u}^k), \\ \mathbf{u}^{k+1} &:= \mathbf{u}^k + \mathbf{x}^{k+1} - \mathbf{z}^{k+1}, \end{aligned}$$

where for any $\kappa > 0$, S_κ is the element-wise soft thresholding operator, i.e. for each component $S_\kappa(a) := (1 - \kappa/|a|)_+ \times a$, for $a \neq 0$, and $S_\kappa(0) := 0$. Note that we have reduced the non-differentiable problem (4) into a sequence of differentiable optimization steps for $\mathbf{x}$, and the computation of the proximal operator S_κ for the $\mathbf{z}$-updates. We refer to Parikh et al. (2014) for a detailed presentation about proximal operators and their use in optimization. ADMM can also be adapted for large-scale data, using standard libraries and frameworks for parallel computing such as MPI, MapReduce and Hadoop, see Boyd et al. (2011) for more details about the implementation of such methods.

Example 1 (*Logit*)**.** If we choose the Fisher transform $g(t) = (e^t - 1)/(e^t + 1)$, then g is odd and the optimization program becomes

$$\hat{\beta} := \arg\max_{\beta\in\mathbb{R}^{p'}} \frac{1}{n(n-1)} \sum_{i,j;i\neq j} K_h(\mathbf{Z}_i - \mathbf{Z}_j) \log\left(logit(W_{(i,j)}\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right) - \lambda|\beta|_1,$$

where the so-called logit link function is defined by $logit(x) = e^x/(1 + e^x)$. Therefore $\hat{\beta}$ can be seen as the maximizer of the log-likelihood of a weighted logistic regression with independent realizations of an explained variable $W_{(i,j)}$, given some explanatory variables $\mathbf{Z}_i$. On a practical side, when $K \geq 0$, the β-criterion is concave. This allows to use the existing software and optimization routines of logistic regression without many changes.

Example 2 (*Probit*)**.** Similarly, choosing $g(t) = 2\Phi(t) - 1$, where Φ denotes the cdf of the standard normal distribution, yields the equivalent of a (weighted) probit regression. Indeed, this function g is odd, (6) applies in this case and our criterion in (4) is concave w.r.t. β.

Let us assume that a family of models or some statistical procedure allow the calculation of the functional link $g(\epsilon\psi(\mathbf{z})^T\beta)$ and then $p(\mathbf{z})$, for any $\mathbf{z}, \epsilon \in \{-1, 1\}$ and any given value β: logit, probit, regression trees, neural networks, etc. Then, we can estimate the "true" parameter β^* by $\hat{\beta}$, as given by (4), in practical terms.

Now, we state the asymptotic properties of $\hat{\beta}$, under the assumption that $\beta \mapsto L_n(\beta)$ is concave. To this goal, we introduce some notation.

For any $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^p$, denote

$$p(\mathbf{x}, \mathbf{y}) := \mathbb{P}_{\beta^*}((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0 | \mathbf{Z}_1 = \mathbf{x}, \mathbf{Z}_2 = \mathbf{y}).$$

The latter expectations are calculated when the underlying parameter is assumed to be the true value β^*. Note that $p(\mathbf{x}) := p(\mathbf{x}, \mathbf{x})$ and $2p(\mathbf{z}) - 1 = \tau_{1,2|\mathbf{Z}=\mathbf{z}}$. Moreover, for any $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z} \in \mathbb{R}^p$, set

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) := \mathbb{P}_{\beta^*}((X_{2,1} - X_{1,1})(X_{2,2} - X_{1,2}) > 0, (X_{3,1} - X_{1,1})(X_{3,2} - X_{1,2}) > 0 | \mathbf{Z}_1 = \mathbf{x}, \mathbf{Z}_2 = \mathbf{y}, \mathbf{Z}_3 = \mathbf{z}).$$

This is the conditional probability that $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ are concordant, given their respective covariates. Denote, for any $\beta \in \mathbb{R}^{p'}$,

$$\phi(\mathbf{x}, \mathbf{y}, \beta) := p(\mathbf{x}, \mathbf{y}) \log(q(\mathbf{x}, \beta)) + (1 - p(\mathbf{x}, \mathbf{y})) \log(1 - q(\mathbf{x}, \beta)), \quad q(\mathbf{x}, \beta) := 1/2 + g(\boldsymbol{\psi}(\mathbf{x})^T\beta)/2.$$

Note that $q(\mathbf{z}, \beta^*) = p(\mathbf{z})$. Finally, for any real function f and $\varepsilon > 0$, denote by f_ε the function $x \mapsto \sup_{t,\ |x-t|<\varepsilon} |f(t)|$.

Regularity assumption R0: The density $f_{\mathbf{Z}}$ of $\mathbf{Z}$ is assumed to be m-times continuously differentiable. Moreover, the functions $\phi(\mathbf{x}, \cdot, \beta)$ and $q(\cdot, \beta)$ are continuous for any $\mathbf{x} \in \mathcal{Z}$ and any $\beta \in \mathbb{R}^{p'}$. To simplify, $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \mapsto p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ will be continuous on $\mathcal{Z}^3$.

Theorem 3. *Under R0,* (B.2) *and* (B.3) *in Appendix* B*, if $\lambda_n \to \lambda_\infty$ and $n^2h^p \to \infty$ when $n \to \infty$, if the true model is given by* (3) *and $\beta \mapsto L_n(\beta)$ is concave, then the solution $\hat{\beta}$ of* (4) *tends in probability towards $\beta^{**} := \arg\max_\beta L_\infty(\beta) - \lambda_\infty|\beta|_1$, where*

$$L_\infty(\beta) := \int \phi(\mathbf{z}, \mathbf{z}, \beta) f_{\mathbf{Z}}^2(\mathbf{z})\, d\mathbf{z}.$$

In particular, when $\lambda_\infty = 0$, the estimator $\hat{\beta}$ tends to $\arg\max_\beta L_\infty(\beta) = \beta^*$, because $\phi(\mathbf{z}, \mathbf{z}, \beta)$ is the expected log-likelihood associated to $W_{(1,2)}$ given $\mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}$. Thus, for every $\mathbf{z}$, the latter quantity is maximal when $\beta = \beta^*$.

Theorem 4. *Under the conditions of Theorem* 3 *and some additional conditions of regularity in Appendix* C *(notably* (C.1)–(C.4)*), if $n^{1/2}\lambda_n \to \mu$ and $nh^p \to \infty$ when $n \to \infty$, then $n^{1/2}(\hat{\beta} - \beta^*)$ weakly tends to*

$$\mathbf{u}^* = \arg\min_{\mathbf{u}\in\mathbb{R}^p} \mathbb{W}(\beta^*)\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbb{H}(\beta^*)\mathbf{u} - \mu \sum_{k;\beta_k^*=0} |u_k| - \mu \sum_{k;\beta_k^*\neq 0} sign(\beta_k^*)u_k,$$

where $\mathbb{W}(\beta^*) \sim \mathcal{N}(0_p, \Sigma_{\beta^*})$, $\Sigma_{\beta^*} = \int \partial_\beta\phi(\mathbf{z}, \mathbf{z}, \beta^*)\partial_\beta\phi(\mathbf{z}, \mathbf{z}, \beta^*)^T f_{\mathbf{Z}}^3(\mathbf{z})\, d\mathbf{z}$ and

$$\mathbb{H}(\beta^*) = \int \partial^2_{\beta,\beta^T}\phi(\mathbf{z}, \mathbf{z}, \beta^*) f_{\mathbf{Z}}^2(\mathbf{z})\, d\mathbf{z}.$$

Remark 5. All the previous results and those of the next sections are based on the kernel-weighted log-likelihood criterion $L_n(\beta)$, and then on the choice of the bandwidth h. We have not tried to find an "optimal" smoothing parameter h. This task is outside the scope of this paper and is left for further research. Instead, we have preferred to rely on the usual rule-of-thumb (Scott, 1992), even if, strictly speaking, it is relevant only for kernel estimators of densities. Nonetheless, we have not empirically found an "excessive sensitivity" of our simulation results w.r.t. h.

3. Classification algorithms and conditional Kendall's tau

In the latter section, we have studied a localized likelihood procedure to estimate β^* under (3), when we can explicitly write (and code) the link function g. This may be seen as a restrictive approach, because it is far from obvious to guess the right functional form of g. To improve the level of flexibility of our conditional Kendall's tau model, we recall the estimation of $\tau_{1,2|\mathbf{z}}$ is similar to the evaluation of $\mathbb{P}\big(W_{(1,2)} = 1 \big| \mathbf{Z}_1 = \mathbf{Z}_2 = \mathbf{z}\big)$, i.e. the probability $p(\mathbf{z})$ of classifying the couple $(1, 2)$ into one of two categories (concordant or discordant), given a common value $\mathbf{z}$ of their covariates. Formally, the answer of such a question can be directly yielded by some classification algorithms. This is the topic of this section. Therefore, instead of estimating an assumed parametric model by penalization, as in (4), a classification algorithm will "automatically" evaluate $p(\mathbf{z})$ by $\hat{p}(\mathbf{z})$. An estimator of the conditional Kendall's tau will simply be $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := 2\hat{p}(\mathbf{z}) - 1$.

Now, we show how different classification algorithms can be used and adapted to the estimation of $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ in practice. The first step is to transform the dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)$, called the *initial dataset*, into an object $\tilde{\mathcal{D}}$, that will be called the *dataset of pairs* (see Algorithm 1). Each element of this dataset of pairs is indexed by an integer $k \in \{1, \dots, n(n-1)/2\}$, which corresponds to any (unordered) pair (i, j), $i \neq j$, of observations in the initial dataset.

For any pair of observations, we compute the associated covariate $\tilde{Z}_k$ which is just the average of the two covariates $\mathbf{Z}_i$ and $\mathbf{Z}_j$ (contrary to Section 2 where we have chosen $\mathbf{Z}_i$). Note that we want $\mathbf{Z}_i$ and $\mathbf{Z}_j$ to be close to each other, so that the pair (i, j) is relevant. This means that a weight variable V_k is defined for any pair. It is related to the proximity between $\mathbf{Z}_i$ and $\mathbf{Z}_j$. Obviously, if $V_k = 0$ then the corresponding pair is not kept, finally. This selection induces also a computational benefit, by reducing the size of the dataset and the computation time. For example, suppose that $n = 4000$. Then, up to around 8×10^6 possible pairs can be constructed but only a small group of them (around 10^4 or 10^5 pairs, typically) will be relevant. The others are pairs for which the covariates are considered too far apart. Note that, in order to increase the proportion of k such that the weight V_k is zero, it is sufficient to use compactly supported kernels. For instance, for any arbitrary p-dimensional kernel K, we can consider $\tilde{K}(\mathbf{z}) := \gamma K(\mathbf{z}) \mathbb{1}\{\|\mathbf{z}\|_\infty \leq 1\}$, with some normalizing constant γ so that $\int \tilde{K} = 1$.

Algorithm 1: Algorithm for creating the dataset of pairs from the initial dataset.

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
$k \leftarrow 0$;
for $i \leftarrow 1$ **to** $(n-1)$ **do**
  for $j \leftarrow (i+1)$ **to** n **do**
    $\tilde{\mathbf{Z}}_k \leftarrow (\mathbf{Z}_i + \mathbf{Z}_j)/2$;
    $W_k \leftarrow W_{(i,j)}$ as defined in Eq. (1) ;
    $V_k \leftarrow K_h(\mathbf{Z}_i - \mathbf{Z}_j)$;
    $k \leftarrow k + 1$;
  end
end
Define $\mathcal{K} := \{k : V_k > 0\}$;
Output: A dataset of pairs $\tilde{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k\in\mathcal{K}} \in \left(\{-1, 1\} \times \mathbb{R}^p \times \mathbb{R}_+\right)^{n(n-1)/2}$.

3.1. The case of probit and logit classifiers

With the new dataset $\tilde{\mathcal{D}}$, we can virtually apply any classification method to predict the concordance value W_k of the pair k, given the covariate $\tilde{\mathbf{Z}}_k$ and the weight V_k. The logit and probit models yield some of the oldest and easiest methods in classification. They have straightforward adapted versions in our case: see Algorithm 2. These weighted penalized GLM procedures are estimated using the R package `ordinalNet` Wurm et al. (2017). Note that we are still estimating $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ under the parametric model given by (3). The tuning parameter λ can be chosen using a generalization of Algorithm 2 in Derumigny and Fermanian (2018a). The chosen λ is the one which minimizes the cross-validation criterion,

$$CV(\lambda) := \sum_{k=1}^{N} d\left(\mathbf{z} \mapsto \hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\mathbf{z}} ;\ \mathbf{z} \mapsto g\left(\boldsymbol{\psi}(\mathbf{z})^T \hat{\beta}^{(\lambda,-k)}\right)\right),$$

where $d(\cdot\,;\,\cdot)$ is a distance on a space of bounded functions of $\mathbf{z}$, for example the distance generated by the L_2 norm, $\hat{\tau}^{(k)}_{1,2|\mathbf{Z}=\cdot}$ is an estimator of Kendall's tau using the dataset $\mathcal{D}_k$, $\hat{\beta}^{(\lambda,-k)}$ is estimated on the dataset $\mathcal{D}\backslash\mathcal{D}_k$ using the tuning parameter λ, and the initial dataset $\mathcal{D}$ has been separated at random in N subsets $\mathcal{D}_1, \dots, \mathcal{D}_N$ of equal size.

Algorithm 2: Estimation of the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ using a logit (resp. probit) regression.

Input: A dataset of pairs $\tilde{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k\in\mathcal{K}}$
Input: A point $\mathbf{z} \in \mathcal{Z}$, a function $\boldsymbol{\psi}$ and a penalty level λ;
Compute the usual weighted penalized logit (resp. probit) estimator $\hat{\beta}$ on the dataset $(W_k, \boldsymbol{\psi}(\tilde{\mathbf{Z}}_k), V_k)_{k\in\mathcal{K}}$ with a tuning parameter λ ;
Output: An estimator $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := (e^{\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}} - 1)/(e^{\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}} + 1)$
(resp. $\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := 2\Phi(\boldsymbol{\psi}(\mathbf{z})^T\hat{\beta}) - 1$).

3.2. Decision trees and random forests

Now, let us discuss how partition-based methods can be used for the estimation of the conditional Kendall's tau. Strictly speaking, such techniques are parametric: the relationship (2) implicitly applies, but for some complex untractable function

g. And the parameter β^* is related to some covariate thresholds, typically. Nonetheless, a classical decision tree can be directly trained on the weighted dataset $\tilde{\mathcal{D}}$. We use the R package `tree` by Ripley (2018), following Breiman et al. (1984). Therefore, the application of decision trees to our framework is straightforward, and does not require any special adaptation, contrary to random forests. And the tree procedure allows the calculation of the probability of observing a concordant pair, given any common value of $\mathbf{Z}$.

In a classical classification setting, random forests are techniques of aggregation of decision trees that are built on a subset of samples and subsets of variables. More precisely, a typical random forest algorithm is the following: sample 80% of the rows of the dataset (without replacement), and 80% of the explanatory variables; estimate a tree on this, and repeat this procedure a certain number of times, with different sub-samples every time. In our framework, it is not clear at which level subsampling should take place.

The easiest solution would be to directly plug-in the dataset of pairs $\tilde{\mathcal{D}}$ into a classical random forest algorithm, but it does not obviously lead to the best solution. For comparison, we detail this solution in Algorithm 3. We propose now an improvement on Algorithm 3. Indeed, noting that aggregation of trees is useless if all trees are identical, it seems that the more variability in the input of the trees, the better. Following this idea, we have noticed that the observations in the dataset of pairs are not independent. Influence of this lack of independence is discussed in a general setting in Section 3.5. For example, the pair $(1, 2)$ is usually not independent of the pair $(1, 3)$, because they both share the first observation $(X_{1,1}, X_{1,2}, \mathbf{Z}_1)$. Therefore, to increase the diversity of inputs in the different trees, we suggest to lead a first sampling $\mathcal{S}_j$ on the initial dataset, and then to build a dataset of pairs on the sampled observations $\mathcal{D}_j := (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i\in\mathcal{S}_j}$ (see Algorithm 4). As a matter of fact, if for example the first observation does not belong to the sample $\mathcal{S}_j$, then the dataset $\mathcal{D}_j$ and the estimated tree $\mathcal{T}_j$ become both independent of this first observation $(X_{1,1}, X_{1,2}, \mathbf{Z}_1)$. This independence property makes the trees less dependent, and significantly improves the performance in our results compared to the original Algorithm 3.

Algorithm 3: Random forests un-adapted for the estimation of the conditional Kendall's tau

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
Compute the dataset of pairs $\tilde{\mathcal{D}}$ using Algorithm 1 on $\mathcal{D}$;
for $j \leftarrow 1$ **to** N_{tree} **do**
  Sample a set $\mathcal{S}_j \subset \{1, \dots, n(n-1)/2\}$ without replacement ;
  Compute the dataset of pairs $\tilde{\mathcal{D}}_j = (W_k, \tilde{Z}_k, V_k)_{k\in\mathcal{S}_j}$ using observations from $\tilde{\mathcal{D}}$;
  Sample a set $\mathcal{S}'_j \subset \{1, \dots, p'\}$ without replacement ;
  Estimate a classification tree $\mathcal{T}_j$ on the dataset $(W_k, (\psi_l(\tilde{Z}_k))_{l\in\mathcal{S}'_j}, V_k)_{k\in\mathcal{S}_j}$;
end
Output: An estimator $\hat{\tau}_{1,2|\mathbf{Z}=\cdot} := N_{tree}^{-1} \sum_{j=1}^{N_{tree}} \mathcal{T}_j(\cdot)$.

Algorithm 4: Random forests adapted for the estimation of the conditional Kendall's tau

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
for $j \leftarrow 1$ **to** N_{tree} **do**
  Sample a set $\mathcal{S}_j \subset \{1, \dots, n\}$ without replacement ;
  $\mathcal{D}_j \leftarrow (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i\in\mathcal{S}_j}$;
  Compute the dataset of pairs $\tilde{\mathcal{D}}_j = (W_k, \tilde{Z}_k, V_k)_{k\in\mathcal{K}_j}$ using Algorithm 1 on $\mathcal{D}_j$, providing $\mathcal{K}_j$;
  Sample a set $\mathcal{S}'_j \subset \{1, \dots, p'\}$ without replacement ;
  Estimate a classification tree $\mathcal{T}_j$ on the dataset $(W_k, (\psi_l(\tilde{Z}_k))_{l\in\mathcal{S}'_j}, V_k)_{k\in\mathcal{K}_j}$;
end
Output: An estimator $\hat{\tau}_{1,2|\mathbf{Z}=\cdot} := N_{tree}^{-1} \sum_{j=1}^{N_{tree}} \mathcal{T}_j(\cdot)$.

3.3. Nearest neighbors

The nearest neighbors also provide a very popular classification algorithm and can be directly used on the dataset $\tilde{\mathcal{D}}$ (see Algorithm 5). Here, we no longer assume (3) or even (2), and we live in a nonparametric framework. A pretty difficult problem is to choose a convenient number of nearest neighbors. As usual in nonparametric statistics, we must find a compromise between variance (tendency to undersmooth, i.e. to choose a too small N) and bias (tendency to oversmooth, i.e. to choose a too big N). Moreover, in our case, with $n(n-1)/2$ possible pairs, choosing a right value for N can be challenging. Indeed, in the usual (i.i.d.) nearest neighbor framework, the asymptotically optimal N is a power of the sample size. Here, this is different because there are three potential sample sizes: n, if we consider there are fundamentally n sources of randomness, $n(n-1)/2$ by considering that the new sample has a cardinality equal to the number of pairs, or even $|\mathcal{K}|$ that is random and depends on h. Thus, our problem is to choose a "relevant formula" for N based on the "convenient" sample size.

Algorithm 5: Estimation of the conditional Kendall's tau $\tau_{1,2|\mathbf{Z}=\mathbf{z}}$ using nearest neighbors.

Input: A dataset of pairs $\tilde{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k\in\mathcal{K}}$
Input: A point $\mathbf{z} \in \mathcal{Z}$, a number N of nearest neighbors and a distance d on $\mathbb{R}^{p'}$;
$\mathcal{K}_{\mathbf{z}} \leftarrow \arg\min_{E\subset\mathcal{K},|E|=N} \left(\sum_{k\in E} d\big(\boldsymbol{\psi}(\mathbf{z}), \boldsymbol{\psi}(\tilde{\mathbf{Z}}_k)\big)\right)$;
Output: An estimator $\hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\mathbf{z}} := \big(\sum_{k\in\mathcal{K}_{\mathbf{z}}} V_k W_k\big)/\sum_{k\in\mathcal{K}_{\mathbf{z}}} V_k$.

In applications, one might not be interested in the value of the conditional Kendall's tau at only one point, but also in the whole function $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$. The goodness of this estimation is linked to the underlying density $f_{\mathbf{Z}}$ of $\mathbf{Z}$: the estimation can be made more precise in regions where $f_{\mathbf{Z}}$ is high, allowing to use a higher number of neighbors with close covariates. At the opposite, in some regions where $f_{\mathbf{Z}}$ is low, a smaller N should be used. Note that, in general, $f_{\mathbf{Z}}$ is unknown and its estimation may be difficult as well, due to the curse of dimensionality. Therefore, it is highly desirable to build a local number of neighbors $N(\mathbf{z})$. Such a local choice $N(\mathbf{z})$ will help to avoid both under- and over-smoothing in all parts of the space $\mathcal{Z}$.

Cross-validation techniques are widely used for the choice of tuning parameters, but might not be here the best solution as one would like to find a local choice of N. This problem has similarities with classical non-parametric regression. We propose to use a procedure inspired by Lepski's method for choosing the bandwidth (Lepski and Spokoiny, 1997), once adapted to our setting. Lepski's method is built on a simple principle: when two non-parametric estimators are close, the best is the smoothest. When two non-parametric estimators are far apart, the best is the least smooth. Let $(\mathcal{Z}_i)_{i\in\mathcal{I}}$ be a partition of $\mathcal{Z}$. The goal will be to choose the best estimator on each $\mathcal{Z}_i$, which corresponds to the choice of a local number of nearest neighbors N_i. This procedure is called "local" since the diameters of the $\mathcal{Z}_i$ will be small. For example, if $p = 1$ and $\mathcal{Z}$ is a bounded interval then the $\mathcal{Z}_i$ can be chosen as small intervals. We denote by $\mathcal{N} \subset \mathbb{N}$ the finite set of possible numbers of neighbors. Following Lepski's approach, we choose $\mathcal{N}$ as a geometric progression, i.e. $\mathcal{N} = \{\lfloor a_1 \times a_2^i \rfloor,\ i = 1, \dots, i_{max}\}$ for some constants $a_1, a_2 > 0$, where $\lfloor x \rfloor$ denotes the integer part of a real x.

To measure how far the estimators are from each other, we introduce a distance d_i, $i \in \mathcal{I}$. As our estimators of conditional Kendall's tau are bounded (between -1 and 1) and measurable, several choices are possible. In applications, we will use

$$d_i(f,g) = \left(\frac{1}{j_{max}}\sum_{j=1}^{j_{max}}\Big[(f(\mathbf{z}_{i,j}) - g(\mathbf{z}_{i,j}))/M\Big]^2\right)^{1/2}, \mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,j_{max}} \in \mathcal{Z}_i, \tag{8}$$

where M is a normalization factor independent of i and the subsets $\mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,j_{max}}$ are arbitrarily chosen in $\mathcal{Z}_i$, $i = 1, \dots, i_{max}$. We will use $M = (\max - \min)\{\hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\mathbf{z}}, N \in \mathcal{N}, \mathbf{z} \in \mathcal{Z}\}$. Indeed, in the classical nonparametric regression model $Y = f(X)+\varepsilon$, with an unknown function f, M should be replaced by the standard deviation of the noise ε. In our case, we can define a (pseudo-)noise $\xi_{\mathbf{z},N} := \hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}}$, but it is unknown in practice and its distribution is complicated. Therefore M serves as a proxy of the amplitude of the variations in the estimated conditional Kendall's tau. This normalization by M ensures a kind of adaptivity of the estimation.

Algorithm 6: Lepski's method for a local choice of the number of nearest neighbors, and the corresponding estimator of the conditional Kendall's tau.

Input: A set $\mathcal{N} \subset \mathbb{N}$ of possible numbers of nearest neighbors and the corresponding estimates $\hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\cdot}$ given by Algorithm 5, for all $N \in \mathcal{N}$;
Input: A partition $(\mathcal{Z}_i)_{i\in\mathcal{I}}$ of $\mathcal{Z}$ and a distance d_i on a space of bounded measurable real functions defined on $\mathcal{Z}_i$, for every $i \in \mathcal{I}$;
foreach $i \in \mathcal{I}$ **do**
  $S_i \leftarrow \left\{N \in \mathcal{N} : d_i\left(\hat{\tau}^{(N)}_{1,2|\mathbf{Z}=\cdot},\ \hat{\tau}^{(N')}_{1,2|\mathbf{Z}=\cdot}\right) \le A\sqrt{(1/N')\log(\max(\mathcal{N})/N')}, \forall N' \in \mathcal{N} \cap [1, N]\right\}$; (9)
  $N_i \leftarrow \max(S_i)$;
end
Output: An estimator $\mathbf{z} \mapsto \hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} := \sum_{i\in\mathcal{I}} \mathbb{1}\{\mathbf{z} \in \mathcal{Z}_i\}\hat{\tau}^{(N_i)}_{1,2|\mathbf{Z}=\mathbf{z}}$.

We have observed that the sensitivity to $\mathcal{N}$ is not too large, if it is chosen in a reasonable way, for example between 5 or 10 possibilities. When $\mathbf{Z}$ is univariate, a simple partition $(\mathcal{Z}_i)_{i\in\mathcal{I}}$ can be given by the deciles of $\mathbf{Z}$. We choose $A = 1$ for simplicity since we believe there is no procedure for choosing it. A statistician who would like to play with the smoothness of the result is free to adjust the constant A, using an expert knowledge of the situation. Finally, the $\mathbf{z}_{i,j}$ can be chosen as quantiles of $\mathbf{Z}$, or as a regular grid on each $\mathcal{Z}_i$.

3.4. Neural networks

Nowadays, neural networks have become very popular with a wide range of applications. In classification problems, a neural network can be seen as an estimator that depends on some parameters, but in a very flexible and complex way. For every input $\mathbf{z}$, it yields the probability of belonging to any class. In our framework, we will train a network on the dataset of pairs $\tilde{\mathcal{D}}$. It is well-known that most neural networks do not induce convex programs, and the outputs therefore depend on some initial parameter values. One strategy is to independently train networks with different starting parameter values, that may be randomly chosen, for example.

This method of using independent estimators (conditionally on the initial sample $\mathcal{D}$) and then aggregating them is related to the random forest approach of the previous section and the discussion therein. Therefore, the same techniques are relevant and we have noticed an improvement in terms of performance by using an adapted version of Algorithm 4. More precisely, we fix a number of neural networks. For each neural network, we sample without replacement a part of the initial dataset from which the corresponding dataset of pairs is constructed and used as a training set. In order to improve stability, we aggregate the predictions of the different neural networks by using their median as the final predicted Kendall's tau. There is a trade-off between computation time and accuracy: a larger number of networks should improve the accuracy while taking obviously a longer time to be trained. The precise choice of the best architecture of the network is a complicated task, which is left for future research. As we are looking for functions $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ which are smooth almost everywhere and easy to interpret in applications, we choose a simple architecture with $N_{nnet} = 10$ neural networks, each having a single hidden layer of 3 neurons. Besides, bigger networks seem to deteriorate the performance of this estimator, see Section 4.6.

Algorithm 7: Neural networks with median bagging, adapted for the estimation of the conditional Kendall's tau

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
for $j \leftarrow 1$ **to** N_{nnet} **do**
  Sample a set $\mathcal{S}_j \subset \{1, \dots, n\}$ without replacement ;
  $\mathcal{D}_j \leftarrow (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i\in\mathcal{S}_j}$;
  Compute the dataset of pairs $\tilde{\mathcal{D}}_j = (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k\in\mathcal{K}_j}$ using Algorithm 1 on $\mathcal{D}_j$, providing $\mathcal{K}_j$;
  Estimate a neural net $\mathfrak{N}_j$ on the dataset $(W_k, \boldsymbol{\psi}(\tilde{\mathbf{Z}}_k), V_k)_{k\in\mathcal{K}_j}$;
end
Output: An estimator $\hat{\tau}_{1,2|\mathbf{Z}=\cdot} := Median\{\mathfrak{N}_j(\cdot), j = 1, \dots, N_{nnet}\}$.

3.5. Lack of independence and its influence on the proposed algorithms

The machine learning methods that are discussed in this section were all designed for i.i.d. data. But it is easy to see that some observations in the dataset of pairs $\tilde{\mathcal{D}}$ will not be independent. Indeed, assume that the observations in the original dataset $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n}$ are i.i.d., to simplify. The pair $(i = 1, j = 2)$ and the pair $(i = 1, j = 3)$ both involve the first observation $(X_{1,1}, X_{1,2}, \mathbf{Z}_1)$, and therefore are not independent. This is a theoretical problem, but numerical results in Section 4 show that this does not often seem to be a problem in practice.

As far as the logit and probit are concerned, it was proved in Section 2 that they are related to a family of estimators that can use $\tilde{\mathcal{D}}$ "as is". They yield consistent and asymptotically normal estimates, nonetheless, if the specification is correct. It is likely that the other methods presented here enjoy similar properties and are also largely unaffected by dependence between pairs. Note that, if all observations in $\mathcal{D}$ are identically distributed, then the observations in $\tilde{\mathcal{D}}$ are identically distributed as well. This is favorable to our methods.

Concerning the dependence inside $\tilde{\mathcal{D}}$, we will show that it is not too strong. For example, the pairs $(1, 2)$ and $(1, 3)$ are not independent, but the pairs $(1, 2)$ and $(3, 4)$ are indeed independent. This means that there is still "a large proportion of" independence left in $\tilde{\mathcal{D}}$. Formally, if two distinct pairs are randomly chosen in $\tilde{\mathcal{D}}$, the probability that they are really independent is high. Indeed, there are $N_{tot} := n(n-1)(n(n-1)-2)/8$ couples of distinct pairs. Besides, the number N_{ind} of couples of pairs which are independent is $N_{ind} := n(n-1)(n-2)(n-3)/8$. The factor $1/8$ appears in both N_{tot} and N_{ind} since we can always switch the two observations in the first pair, in the second pair, and switch the two pairs (every 4-tuple is counted $2^3 = 8$ times). It is easy to see that $N_{ind}/N_{tot} = 1 - O(1/n)$ as $n \to \infty$.

This means that the pairs are "almost all" independent from each other, as $n \to \infty$. In other words, the dependence between two pairs becomes negligible with averages. That is the reason why the machine learning methods used will perform well if the original dataset $\mathcal{D}$ is large enough. If the original dataset $\mathcal{D}$ is not i.i.d., for example as observations of a time series, we conjecture that such methods will work in a similar way as long as dependence is not too strong, for example if the data-generating process satisfies some usual assumptions, see Remark 6.

Whenever bootstrap, subsetting, resampling, or cross-validation is led on these classification-based estimators, we advise to perform them on the original dataset $\mathcal{D}$ rather than on the dataset of pairs $\tilde{\mathcal{D}}$, as we did in Sections 3.1, 3.2 and 3.4. This seems to yield a good improvement in performance. An example is given by the difference between Algorithms 3 and 4. This can be simply summed up as *"do the resampling on the original dataset $\mathcal{D}$, not on the transformed dataset $\tilde{\mathcal{D}}$"*. Nevertheless, a complete study and justification of this general principle is beyond the scope of this paper and is left for future work.

Table 1
Error criterion (10) for each choice of $\boldsymbol{\psi}$ and each method, multiplied by 1000.

Chosen $\boldsymbol{\psi}$	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
$\boldsymbol{\psi}^{(1)}$	48.1	48.1	7.5	4.89	2.26	0.561
$\boldsymbol{\psi}^{(2)}$	0.721	0.554	4.28	3.28	2.26	1.32
$\boldsymbol{\psi}^{(3)}$	0.663	0.528	4.13	3.41	2.23	1.73
$\boldsymbol{\psi}^{(4)}$	1.41	1.45	4.73	14.2	2.72	1.74
$\boldsymbol{\psi}^{(5)}$	1.05	1.06	4.76	10.3	2.79	2.67
$\boldsymbol{\psi}^{(6)}$	0.456	0.434	4.57	3.15	2.64	3.87

4. Simulation study

In this section, we have studied the relative performances of our estimators by simulation. For a given model and a given method of estimation, we sample 100 different experiments, and estimate the model for each sample. We fix the sample size as $n = 3000$. We remark that, for a given dimension $p > 0$ of $\mathbf{Z}$ and a given support $\mathcal{Z}$ of $\mathbf{Z}$, we have different "blocks" of the model which can be chosen in an independent way:

(i) the law $\mathbb{P}_{\mathbf{Z}}$ of $\mathbf{Z}$,
(ii) the function $\mathbf{z} \in \mathcal{Z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$,
(iii) the (conditional) copula family $(C_\tau)_{\tau\in(0,1)}$ or $(C_\tau)_{\tau\in(-1,1)}$ of $(X_1, X_2)|\mathbf{Z} = \mathbf{z}$, indexed by its conditional Kendall's tau — for example the Gaussian, Student, Clayton, Gumbel, etc., copula families. Such a family can also depend on $\mathbf{Z}$: for example, think of a Student copula with varying degrees of freedom -,
(iv) the conditional margins $X_1|\mathbf{Z}$ and $X_2|\mathbf{Z}$,
(v) the choice of the functions ψ_i, for $i = 1, \dots, p'$,
(vi) the choice of the estimator $\hat{\tau}_{1,2|\mathbf{Z}=\cdot}$.

Our so-called "reference setting" will be defined as $p = 1$, $\mathcal{Z} = [0, 1]$ and (i) $\mathbb{P}_Z = \mathcal{U}_{[0;1]}$; (ii) $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = 3z(1 - z)$; (iii) $(C_\tau)_{\tau\in(0,1)}$ is the Gaussian Copula family ; (iv) $\mathbb{P}_{X_1|Z=z} = \mathbb{P}_{X_2|Z=z} = \mathcal{N}(z, 1)$. For each tested model, the performance of the estimator will be evaluated by the mean integrated ℓ_2 error. With obvious notation, it will be estimated as

$$Err := \mathbb{E}\left[\int_{\mathcal{Z}} (\hat{\tau}_{1,2|\mathbf{Z}=\mathbf{z}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}})^2 d\mathbf{z}\right] \approx \frac{1}{N_{simu} N_{points}} \sum_{i=1}^{N_{simu}} \sum_{j=1}^{N_{points}} (\hat{\tau}^{(i)}_{1,2|\mathbf{Z}=\mathbf{z}^{(j)}} - \tau_{1,2|\mathbf{Z}=\mathbf{z}^{(j)}})^2, \tag{10}$$

where N_{simu}, N_{points} are positive integers, $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N_{points})}$ are fixed points in $\mathcal{Z}$, and $\hat{\tau}^{(i)}_{1,2|\mathbf{Z}=\mathbf{z}^{(j)}}$ is the estimated conditional Kendall's tau at point $\mathbf{z}^{(j)}$ trained on data from the ith simulation. We choose $N_{simu} := 100$ experiments, and in this reference setting, the integral is discretized with $N_{points} := 100$ equispaced points on the segment $[0.01, 0.99]$, to avoid numerical problems at the boundaries.

In the following simulations, "logit" and "probit" refer to Algorithm 2. "Tree" refers to the application of the method `tree()` of package `tree` by Ripley (2018) on the dataset $\tilde{\mathcal{D}}$ produced by Algorithm 1. "Random forests" refers to Algorithm 4. "Nearest neighbors" refers to the adapted version using Algorithm 5, once aggregated using Algorithm 6. Finally "Neural networks" refers to Algorithm 7. Such specifications are now part of our "reference setting".

4.1. Choice of the functions $\{\psi_i\}$, $i = 1, \dots, p'$.

We consider six different choices of $\boldsymbol{\psi}$, that are

1. No transformation, i.e. $\psi_1^{(1)}(z) = z$.
2. Polynomials of degree lower than 4: $\psi_i^{(2)}(z) = 2^{-i+1}(z - 0.5)^{i-1}$ for $i = 1, \dots, 5$.
3. Polynomials of degree lower than 10: $\psi_i^{(3)}(z) = 2^{-i+1}(z - 0.5)^{i-1}$ for $i = 1, \dots, 11$.
4. Fourier basis of order 2 with an intercept: $\psi_1^{(4)}(z) = 1$, $\psi_{2i}^{(4)}(z) = \cos(2\pi i z)$ and $\psi_{2i+1}^{(4)}(z) = \sin(2\pi i z)$ for $i = 1, 2$.
5. Fourier basis of order 5 with an intercept: $\psi_1^{(5)}(z) = 1$, $\psi_{2i}^{(5)}(z) = \cos(2\pi i z)$ and $\psi_{2i+1}^{(5)}(z) = \sin(2\pi i z)$ for $i = 1, \dots, 5$.
6. Concatenation of $\boldsymbol{\psi}^{(2)}$ and $\boldsymbol{\psi}^{(4)}$, which will be denoted by $\boldsymbol{\psi}^{(6)}$.

For each of the choices of $\boldsymbol{\psi}$ above, and each estimator, we compute the criterion (10). The results are displayed in Table 1.

With the choice of $\boldsymbol{\psi}^{(6)}$, logit and probit methods provide the best results. This good performance deteriorates with other choices of $\boldsymbol{\psi}$, especially when the model is misspecified. Neural networks provide the best results with $\boldsymbol{\psi}^{(1)}$, and their performance declines when further transformations of $\mathbf{z}$ are introduced in $\boldsymbol{\psi}$. Nearest neighbors have nearly the best behavior with $\boldsymbol{\psi}^{(1)}$, and it does not seem that other transformations can significantly increase its performance. On the contrary, for trees and random forests, it seems that bigger families $\boldsymbol{\psi}$ can yield improvements over $\boldsymbol{\psi}^{(1)}$.

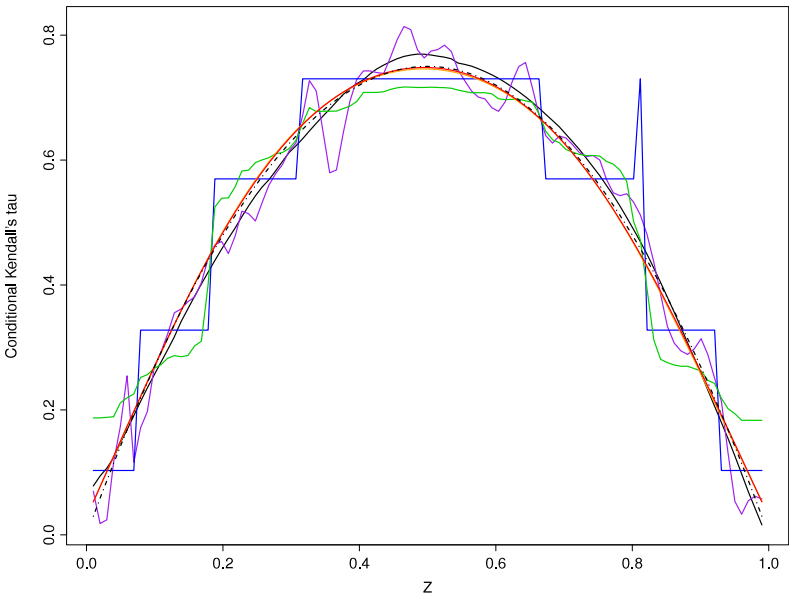

Fig. 1. An example of the estimation of the conditional Kendall's tau using different estimation methods (see Table 2). The black dash-dotted curve is the true conditional Kendall's tau that has been used in the simulation experiment.

Table 2
Summary of available estimation methods for the estimation of the conditional Kendall's tau and corresponding algorithm and curve color.

Method	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
Algorithm	Algorithm 2	Algorithm 2	`tree()` of Ripley (2018)	Algorithm 4	Algorithms 5–6	Algorithm 7
Color	Orange	Red	Blue	Green	Purple	Black

Table 3
Error criterion (10) for each copula family and each method, multiplied by 1000.

Copula family	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
Gaussian	0.456	0.434	4.57	3.15	2.26	0.561
Student 4 df	0.549	0.515	4.54	3.28	2.87	0.753
Student $(2 + 1/z)$ df	0.531	0.518	4.66	3.23	2.82	0.805
Clayton	0.498	0.472	4.52	3.36	2.67	0.742
Gumbel	0.45	0.431	4.56	3.23	2.66	0.775
Frank	0.448	0.42	4.5	3.28	2.13	0.615

From now on, we will choose $\boldsymbol{\psi}^{(6)}$ for the methods *logit*, *probit*, *tree* and *random forests*. Indeed, for these methods, this choice of $\boldsymbol{\psi}$ yields nearly the lowest error criterion and presents the advantages of proposing various shapes, which will help to combine the performances of both polynomials and oscillating functions. On the contrary, for the methods *nearest neighbors* and *neural networks*, we choose $\boldsymbol{\psi}^{(1)}$ as adding new functions does not seem to increase the performance of both of these methods. Fig. 1 displays a comparison of the different methods on a typical simulated sample.

For each estimator, we state in the second line of Table 2 the algorithm used to compute it, and in the third line the color of the corresponding curve in Figs. 1–13. For example, the estimator "probit" is computed using Algorithm 2 and corresponds to the red curves.

4.2. Comparing different copulas families

Now, we keep the reference setting and we change only its part (iii), i.e. the functional form of the conditional copula. The results are displayed in Table 3. We observe that such choice of a parametric copula families has nearly no effect on the performance of the estimators. Nonetheless, with the Student copula (either with fixed or variable degrees of freedom), most estimators have slightly worse performances than with other copulas. This can be explained by the fact that this copula allows asymptotic dependence, i.e. a strong tail association.

Table 4
Error criterion (10) for each choice of conditional margins and each method, multiplied by 1000.

Setting	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
1.	0.456	0.434	4.57	3.15	2.26	0.561
2.	0.809	0.818	4.65	3.72	2.65	0.838
3.	1.15	1.12	5.29	4.21	3.57	1.32
4.	0.493	0.471	4.43	3.44	2.54	0.662

Table 5
Error criterion (10) for different Kendall's tau models and each estimation method, multiplied by 1000.

Setting	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
f_1	11.2	11.6	4.12	4.03	3.89	1.48
f_2	0.456	0.434	4.57	3.15	2.26	0.561
f_3	3.77	3.22	5.95	4.76	2.35	2.17
f_4	12.8	12.8	16.8	10	3.71	1.97

4.3. Comparing different conditional margins

In this subsection, we still start from the reference setting and we change only its part (iv), i.e. the functional form of the conditional margins $(X_1|Z)$ and $(X_2|Z)$. We consider the following alternatives:

1. $\mathbb{P}_{X_1|Z=z} = \mathbb{P}_{X_2|Z=z} = \mathcal{N}(z, 1)$ (as in the reference case).
2. $\mathbb{P}_{X_1|Z=z} = \mathcal{N}(\cos(10\pi z), 1)$; $\mathbb{P}_{X_2|Z=z} = \mathcal{N}(z, 1)$. The idea is to make X_1 oscillate fast enough so that the algorithms will have difficulties to localize concordant and discordant pairs ;
3. $\mathbb{P}_{X_1|Z=z} = Exp(|z|)$; $\mathbb{P}_{X_2|Z=z} = \mathcal{U}_{[z,z+1]}$. This choice allows to see how estimation is affected by changes in the conditional support of (X_1, X_2) given $\mathbf{Z} = \mathbf{z}$;
4. $\mathbb{P}_{X_1|Z=z} = \mathcal{N}(0, z^2)$; $\mathbb{P}_{X_2|Z=z} = \mathcal{U}_{[0,|z|]}$. Then, we will see how estimation is affected by changes in the conditional variance of (X_1, X_2) given $\mathbf{Z} = \mathbf{z}$.

In a similar way as in the previous section, the results of these experiments, as displayed in Table 4 show that changes in terms of conditional marginal distributions generally have a mild impact on the overall performance of the estimators. Moreover, such changes have no effect on the ranking between estimators: the *logit* and *probit* methods are always the best, followed by the *neural networks*, the *nearest neighbors*, and the *random forests* are behind (in this order). The estimator *Tree* shows the lowest performance, but note that it also has the lowest computation time.

4.4. Comparing different forms for the conditional Kendall's tau

In this part, we keep the reference setting, but we change only its part (ii), i.e. the functional form of the conditional Kendall's tau itself. We consider the following choices:

1. $f_1(z) := 0.9 - 0.8\,\mathbb{1}\{z \geq 0.5\}$,
2. $f_2(z) := 3z(1-z)$,
3. $f_3(z) := 0.5 + 0.4\sin(4\pi z)$,
4. $f_4(z) := 0.1 + 1.6z\mathbb{1}\{z < 0.5\} + 1.6(z-0.5)\mathbb{1}\{u \geq 0.5\}$.

The results are presented in Table 5. If the estimated model is close to be well-specified, the best methods are parametric, i.e. the *logit* and *probit* regressions. In all the other cases, *neural networks* seem to perform very well. There appears a compromise between a minimization of the error and a minimization of the computation time. We refer to Table 8 for a quantitative comparison of the performance of such methods in terms of computation time as a function of the sample size n.

4.5. Higher dimensional settings

In the previous sections, we had chosen a univariate vector $\mathbf{Z}$, i.e. $p = 1$. Since this may sound a bit restrictive, we would like to obtain some finite-sample results in dimension $p = 2$. Note that the latter dimension cannot be too high because of the curse of dimensionality linked with the necessary kernel smoothing (done in Algorithm 1 when creating the dataset of pairs). We also choose a simple dictionary $\boldsymbol{\psi}$ of functions, which consists of the two projections on the coordinates of $\mathbf{Z}$. The performance of the estimators is still be assessed by the approximate error criterion (10). The corresponding $\mathbf{z}^{(j)}$ are chosen as a grid of 400 points equispaced on the square $[0.01, 0.99]^2$.

In this framework, we first choose block (iii) of the model : the conditional copula of X_1 and X_2 given $\mathbf{Z}$ will be Gaussian, and block (iv) : $\mathbb{P}_{X_1|\mathbf{Z}=\mathbf{z}} = \mathbb{P}_{X_2|\mathbf{Z}=\mathbf{z}} = \mathcal{N}(z_1, 1)$. We will try different combinations for the remaining blocks (i) and (ii), as described as follows:

Table 6
Error criterion (10) for each setting with 2-dimensional $\mathbf{Z}$ random vectors and each method, multiplied by 1000.

Setting	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
(1)	35.5	35.5	9.63	11.7	6.72	2.21
(2)	0.433	0.681	10.9	5.85	4.33	0.848
(3)	17.8	17.2	5.72	9.79	1.84	1.36

Table 7
Error criterion (10) multiplied by 1000, and average computation time in seconds for each architecture of the neural networks.

Number of neurons	3	5	10	30
Criteria	0.561	0.808	1.47	1.45
Time (s)	234	429	607	5.29e+03

Table 8
Error criterion (10) multiplied by 1000 and computation time in seconds for each method and each choice of n.

		Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
$n = 1000$	Criteria	1.58	1.52	5.85	4.45	4.01	2.01
	Time (s)	59.6	156	0.215	8.11	5.04	30.6
$n = 2000$	Criteria	0.666	0.64	4.9	3.39	2.95	1.79
	Time (s)	192	489	0.99	35.9	17.1	85.3
$n = 3000$	Criteria	0.456	0.434	4.57	3.15	2.26	0.561
	Time (s)	414	1010	2.37	87	36.9	234
$n = 5000$	Criteria	0.275	0.253	3.77	3.05	1.69	0.791
	Time (s)	957	2420	6.37	218	111	461
$n = 8000$	Criteria	0.22	0.204	3.6	3.39	1.27	0.225
	Time (s)	2178	5480	15.2	499	290	1268

(1) $Z_1 \sim \mathcal{N}(0,1)$, $Z_2 \sim \mathcal{U}_{[-1,1]}$, and the copula of (Z_1, Z_2) is Gaussian with a Kendall's tau equal to 0.5. Moreover, $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = z_2 \tanh(z_1)$. This model is interesting because the function $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ will be far away from a linear function of $\boldsymbol{\psi}(\mathbf{z})$, and machine learning techniques should work better than logistic/probit regressions.
(2) We keep the same model as previously, but by setting $g(\tau_{1,2|\mathbf{Z}=\mathbf{z}}) = z_1 + z_2$, using the function g in Example 1 so that we recover the parametric setting of Section 2.
(3) $Z_1 \sim Exp(1)$, $Z_2 \sim \mathcal{N}(0,1)$ and both variables are independent. Set $\tau_{1,2|\mathbf{Z}=\mathbf{z}} = \exp(-z_1|z_2|)$. Again, we have a misspecified nonlinear model that is far away from logit/probit models.

The results are given in Table 6. With the exception of the well-specified setting (2), the logit model performs worse than non-parametric methods. In all these settings, neural networks show better performances than all other methods, followed by nearest neighbors and tree-based methods. Finally, parametric methods are the worst, especially under misspecification of the model.

4.6. Choice of the number of neurons in the one-dimensional reference setting

We consider networks with different numbers of neurons, and study their performance, both statistically and computationally. The results are displayed in Table 7. We observe that increasing the number of neurons only seems to deteriorate the performance of the method.

4.7. Influence of the sample size n

In our one-dimensional reference setting, we fix all the parameters except n. For a grid of values of n we evaluate the performance of our estimators.

We observe that, for most methods, the computation time increases and the error criterion decreases when the sample size increases. We note that the number of pairs is $O(n(n-1))$ (at most) and, therefore, the computation time should increase as $O(n^2)$, which is coherent with the results of Table 8. The relative order of the performances does not seem to change with the sample size n: the same methods are the best ones with small or large n. Note that we have not tried to find an "optimal" fine-tuning of the parameters for each method and each choice of n. Indeed, finding optimal choices of tuning parameters is not an easy task (in a theoretical and practical sense). More accurate analysis is left for future research.

4.8. Influence of the lack of independence

In Section 3.5, we explain some theoretical considerations about the lack of independence in the dataset $\tilde{\mathcal{D}}$ and some consequences. The following simulation experiment complements this analysis with some empirical results.

Indeed, using Algorithm 1, we note that pairs of observations are not independent, and therefore, the elements of the dataset of pairs $\tilde{\mathcal{D}}$ are not independent from each other in general. This could damage the performance of our methods, compared to a situation where all elements would be independent. We now consider such a situation, in order to compare the performance of the methods in both cases. Note that the cardinality of $\tilde{\mathcal{D}}$ is $n(n-1)/2$. Therefore, we will compare the two following settings:

1. Reference situation: fix $n = 3000$, simulate n independent copies $\mathcal{D}_n := (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n}$, construct the dataset of pairs $\tilde{\mathcal{D}}_n$ using Algorithm 1. Use the estimators on the training set $\tilde{\mathcal{D}}_n$.
2. Independent situation: fix $n = 3000$, simulate $n(n-1) \simeq 9{,}000{,}000$ independent copies $\mathcal{D}_{n(n-1)} := (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n(n-1)}$. Create the dataset of consecutive pairs $\overline{\mathcal{D}}_{n(n-1)}$ on this sample using Algorithm 8. This means that we use only consecutive pairs, i.e. (1,2), (3,4), (5,6), and so on. Use the estimators on the training set $\overline{\mathcal{D}}_{n(n-1)}$.

Algorithm 8: Algorithm for creating the dataset of consecutive pairs from the initial dataset.

Input: Initial dataset $\mathcal{D} = (X_{i,1}, X_{i,2}, \mathbf{Z}_i)_{i=1,\dots,n} \in \left(\mathbb{R}^{2+p}\right)^n$;
for $k \leftarrow 1$ **to** $\lfloor n \rfloor / 2$ **do**
  $i, j \leftarrow 2k-1, 2k$;
  $\tilde{\mathbf{Z}}_k \leftarrow (\mathbf{Z}_i + \mathbf{Z}_j)/2$;
  $W_k \leftarrow W_{(i,j)}$ as defined in Eq. (1) ;
  $V_k \leftarrow K_h(\mathbf{Z}_i - \mathbf{Z}_j)$;
end
Define $\mathcal{K} := \{k : V_k > 0\}$;
Output: A dataset of pairs $\overline{\mathcal{D}} := (W_k, \tilde{\mathbf{Z}}_k, V_k)_{k \in \mathcal{K}} \in \left(\{-1, 1\} \times \mathbb{R}^p \times \mathbb{R}_+\right)^{\lfloor n \rfloor / 2}$.

Note that, by construction, the cardinalities of $\overline{\mathcal{D}}_{n(n-1)}$ and $\tilde{\mathcal{D}}_n$ are the same, i.e. both have exactly $n(n-1)/2$ pairs. This is the reason why we chose to simulate $n(n-1)$ points in the independent situation, so that these two numbers of pairs can match. Note that the elements in $\overline{\mathcal{D}}$ are independent from each other by construction while some elements in $\tilde{\mathcal{D}}$ may not be independent from each other in general. We can now compare the performances of the estimators trained on $\overline{\mathcal{D}}_{n(n-1)}$ and on $\tilde{\mathcal{D}}_n$ using the criterion (10). Some results are given in Table 9. Note that the simulation of the each $(X_{i,1}, X_{i,2}, \mathbf{Z}_i)$ is still made under the previous one-dimensional "reference setting".

As expected, all estimators show a better performance in the independent situation. Nonetheless, the independent situation has been simulated using $n(n-1) \simeq 9{,}000{,}000$ points whereas the reference situation uses only $n = 3,000$ points. Even if the numbers of pairs in both experiments are the same, the sample size of the dataset was much larger in the independent situation. This means that there is more information available, and explains also why the independent situation has a better performance: it just uses more data. Such a huge sample may not be available in practice though.

Nevertheless, the original procedure costs $O(n^2)$, which can be large for very large values of n. In this case, it is always possible to restrict oneself to consecutive pairs, with a cost of only $O(n)$. Such a procedure is possible if the dataset is very large and Algorithm 8 can be seen as an alternative to Algorithm 1 where only consecutive pairs are used. This would lower the computation cost at the expense of precision.

5. Applications to financial data

In this section, we study the changes of the conditional dependence between the daily returns of MSCI stock indices during two periods: the European debt crisis (from 18 March 2009 to 26 August 2012) and the after-crisis period (26 August 2012 to 2 March 2018). We will consider the couples (Germany, France), (Germany, Denmark), (Germany, Greece), respectively denoted by (X_1, X_2), (X_1, X_3), (X_1, X_4). We will separately consider two choices of conditioning variables $\mathbf{Z}$:

- a proxy variable for the intraday volatility $\sigma := (High - Low)/Close$, where $High$ denotes the maximum daily value of the Eurostoxx index, Low denotes its minimum and $Close$ is the index value at the end of the corresponding trading day.
- a proxy of so-called "implied volatility moves" $\Delta\sigma^I$. It will record the daily variations of the EuroStoxx 50 Volatility Index, whose quotes are available at https://www.stoxx.com/index-details?symbol=V2TX: $\Delta\sigma_i^I := V2TX(i) - V2TX(i-1)$ for each trading day i. The EuroStoxx 50 Volatility Index $V2TX$ measures the levels of future volatility, as anticipated by the market through option prices.

Table 9
Error criterion (10) multiplied by 1000 for each method and each situation. "Independent" means the independent situation with $\overline{\mathcal{D}}_{n(n-1)}$, and "Not independent" means the reference situation with $\tilde{\mathcal{D}}_n$.

	Logit	Probit	Tree	Random forests	Nearest neighbors	Neural network
Independent	0.127	0.114	3.02	2.52	0.12	0.0363
Not independent	0.456	0.434	4.57	3.15	2.26	0.561

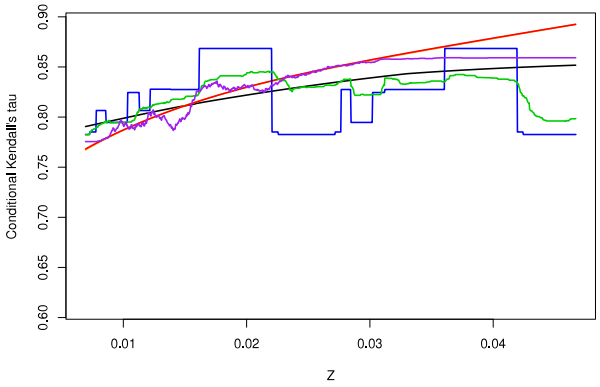

Fig. 2. Conditional Kendall's tau between (X_1, X_2) given σ during the European debt crisis.

Note that, for a given couple, the levels of the estimated conditional Kendall's tau are different (in general) for different conditioning variables. Indeed, the unconditional Kendall's tau $\tau_{1,2}$, the average conditional Kendall's tau with respect to σ, which is $\mathbb{E}_\sigma\left[\tau_{1,2|\sigma}\right]$ and the average conditional Kendall's tau with respect to $\Delta\sigma^I$, which is $\mathbb{E}_{\Delta\sigma^I}\left[\tau_{1,2|\Delta\sigma^I}\right]$ have no reason to be equal.

Both conditioning variables σ and $\Delta\sigma^I$ are of dimension 1. For each method and each conditioning variable, we will use the "best" choice of $\boldsymbol{\psi}$ as determined from the simulations in Section 4.1, that is $\boldsymbol{\psi}^{(6)}$ for the methods *logit*, *probit*, *tree* and *random forests* and $\boldsymbol{\psi}^{(1)}$ for the methods *Nearest neighbors* and *neural networks*. In the following figures, the matching between colors and corresponding estimators still follows Table 2.

Remark 6. It is well-known that sequences of asset returns are not i.i.d. In particular, their volatilities are time-dependent, as in GARCH-type or stochastic volatility models. Moreover, the tail behavior of their distributions is significantly varying, due to some periods of market stress. Several families of models (switching regime models, jumps, etc.) have tried to capture such stylized facts. We conjecture that such temporal dependencies will not affect our results too much. Indeed, dependence will be mitigated by considering all possible couples of random vectors, independently of their dates. It is easy to go one step beyond, for instance by keeping only the couples of returns indexed by i and j when $|i-j| > m$, for some "reasonably chosen" threshold m ($m = 20$, e.g.). In every case, it is highly likely that our inference procedures are still consistent and asymptotically normal, for most types of dependence between successive observations (mixing processes, weak dependence, m-dependence, mixingales, etc.), even if the asymptotic variances are different from ours.

5.1. Conditional dependence with respect to the Eurostoxx's volatility proxy σ

We will first consider the conditioning events given by σ, the proxy variable for the market intraday volatility. The results are displayed in Figs. 2–7. Intuitively, dependence should tend to increase with market volatility: when "bad news" are announced, they are source of stress for most dealers, especially inside the Eurozone that brings together economically connected countries. This phenomenon should be particularly sensitive during the European debt crisis, because a lot of such "bad news" were related to the Eurozone itself (economic/financial news of public debts in several European countries). Let us see whether this is the case.

On most figures, the estimated conditional Kendall's tau seems to exhibit some kind of concavity. The behavior of these functions can be roughly broken down into two main regimes:

1. The "moderate" volatility regime (also called the "normal regime") in the sense that the volatility stay mild, say in the lower half of its range. In this normal regime, conditional Kendall's tau is an increasing function of volatility. This is coherent with most empirical research where it is shown that dependence increases with volatility.

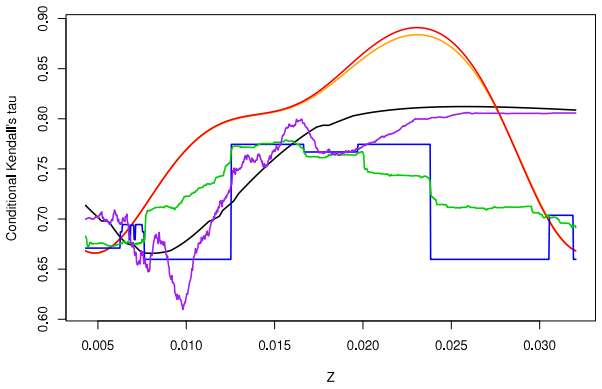

Fig. 3. Conditional Kendall's tau between (X_1, X_2) given σ during the after-crisis period.

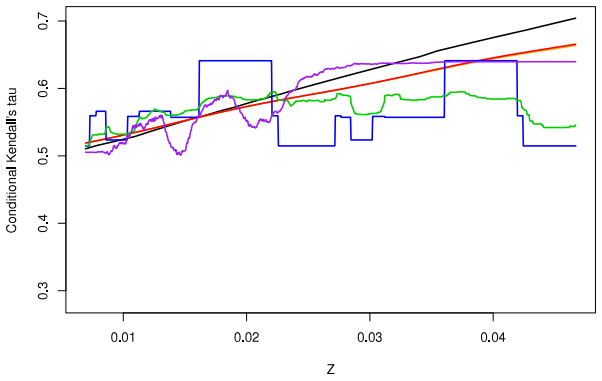

Fig. 4. Conditional Kendall's tau between (X_1, X_3) given σ during the European debt crisis.

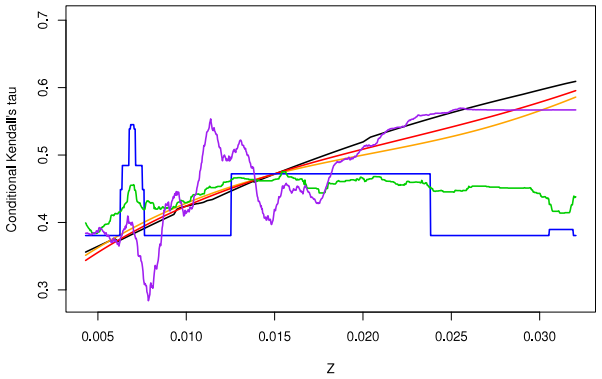

Fig. 5. Conditional Kendall's tau between (X_1, X_3) given σ during the after-crisis period.

2. The high volatility regime: this is a "stressed regime" where σ lies in the upper half of its range. In this less frequent regime, the influence of the European volatility σ on the conditional Kendall's tau appears to be less clear: the estimators become more "fluctuating" and more different from each other, as a consequence of the small number of observations in most stressed regimes.

During the European debt crisis (see Figs. 2, 4 and 6), the three couples seem to exhibit the same shape of conditional dependence with respect to σ, even if their average levels are different. These similarities can be a little bit surprising considering that the economic situations of the corresponding countries are different. It can be conjectured that the heterogeneity in the "mean" levels of conditional dependence is sufficient to reflect this diversity of situations. In this

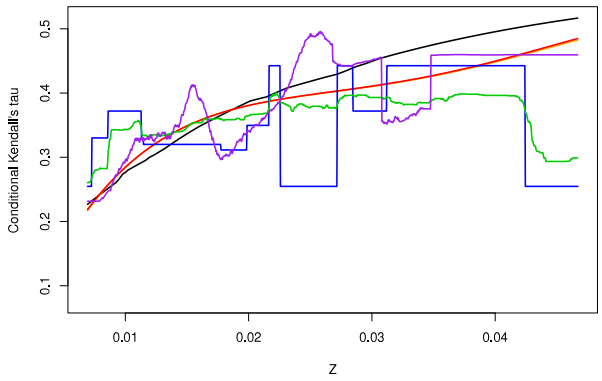

Fig. 6. Conditional Kendall's tau between (X_1, X_4) given σ during the European debt crisis.

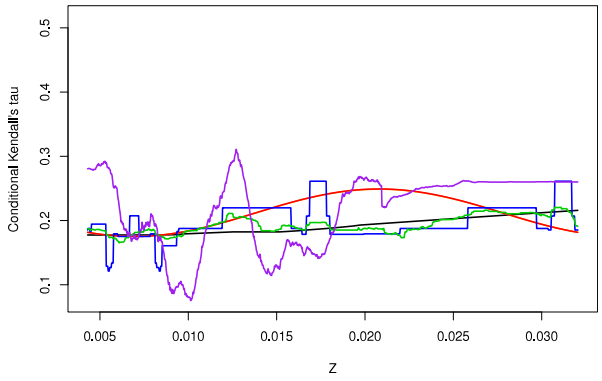

Fig. 7. Conditional Kendall's tau between (X_1, X_4) given σ during the after-crisis period.

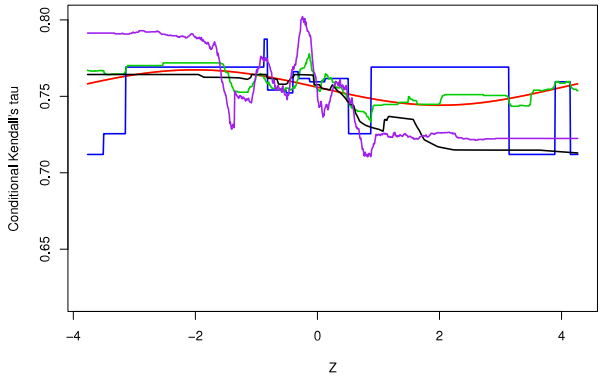

Fig. 8. Conditional Kendall's tau between (X_1, X_2) given $\Delta\sigma^I$ during the European debt crisis.

perspective, the increasing pattern of conditional dependence w.r.t. the "volatility" would be a pure characteristic of that period, regardless of the chosen pair of European countries. Indeed, we have observed this pattern for most couples of European countries in the Eurozone. An explanation might be that investors were focusing on the same international news, for example, about the future of the Eurozone, and, therefore, they were reacting in a similar way, irrespective of the country.

For each couple of countries, conditional Kendall's tau is nearly always lower during the After-crisis period than during the European debt crisis. Apparently, in the After-crisis period, factors and events that are specific to each country attract more attention from investors than during the crisis, which results in lower dependence. In this context, the shapes of conditional dependence are no longer similar for different couples. In particular, the conditional Kendall's tau between German and French returns shows a significant increase during the low volatility regime and a decrease during the high volatility regime:

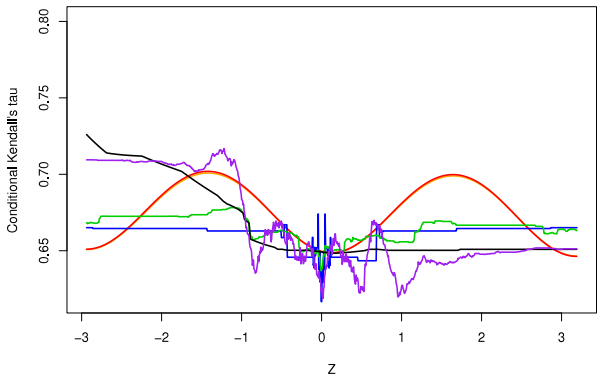

Fig. 9. Conditional Kendall's tau between (X_1, X_2) given $\Delta\sigma^I$ during the after-crisis period.

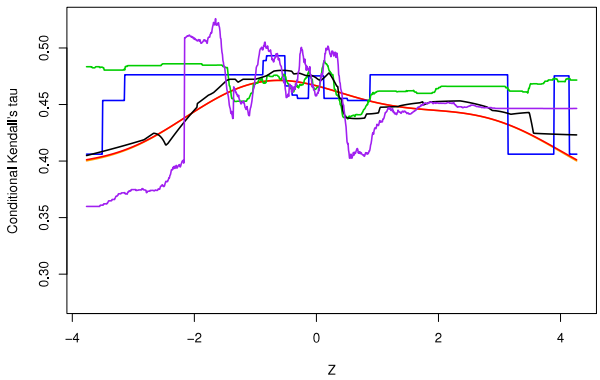

Fig. 10. Conditional Kendall's tau between (X_1, X_3) given $\Delta\sigma^I$ during the European debt crisis.

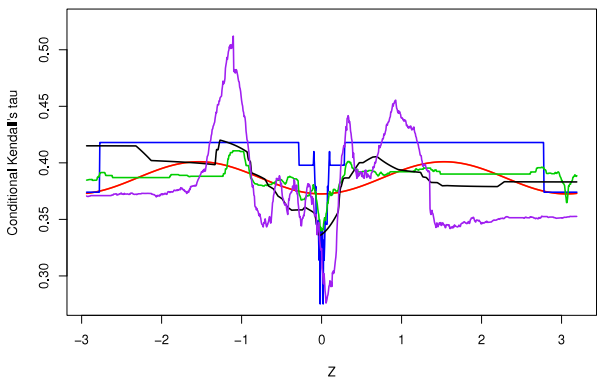

Fig. 11. Conditional Kendall's tau between (X_1, X_3) given $\Delta\sigma^I$ during the after-crisis period.

see Fig. 3. The conditional dependence between the German and the Danish returns is also increasing during the low volatility regime, but in the high volatility, their conditional Kendall's tau seems to be rather constant, even increasing according to the nearest neighbors and the neural networks estimators. Concerning Fig. 7, we do not seem any clear tendency. It is likely that σ has almost no impact on the conditional dependence between the German and Greek stock index returns.

5.2. Conditional dependence with respect to the variations $\Delta\sigma^I$ of the Eurostoxx's implied volatility index

The implied volatility is computed using option prices. In this sense, this financial quantity reflects investors' anticipation of future uncertainty. When important events happen, investors most often update their anticipations, which results in

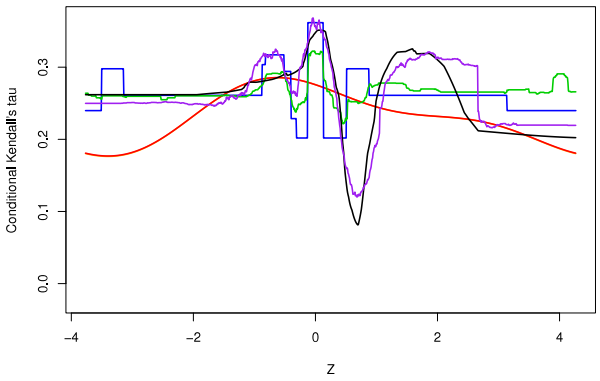

Fig. 12. Conditional Kendall's tau between (X_1, X_4) given $\Delta\sigma^I$ during the European debt crisis.

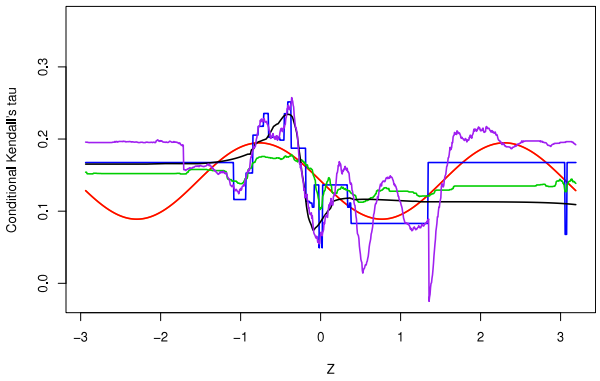

Fig. 13. Conditional Kendall's tau between (X_1, X_4) given $\Delta\sigma^I$ during the after-crisis period.

a change of implied volatilities. This change, denoted by $\Delta\sigma^I$ may be linked to variations of the conditional dependence between stock returns of different countries. Figs. 8 to 13 illustrate the variations of the conditional Kendall's tau between couples of stock returns with respect to the conditioning variable $\Delta\sigma^I$ during the two periods we study.

For each couple, the levels of the conditional Kendall's tau are higher during the European debt crisis than during the after-crisis period. This is coherent with our conclusions in the previous subsection. But here, conditional Kendall's taus look like concave functions of $\Delta\sigma^I$ during the crisis, while they exhibit "double bumps" features after the crisis. During the crisis, when $\Delta\sigma^I$ is small in absolute value, implied volatilities do not change much and the dependence is in general higher than during big changes of the market implied volatility, i.e. when $|\Delta\sigma^I|$ is high (see Figs. 10 and 12).

One exception is the couple (France, Germany), for which the conditional Kendall's tau is roughly a decreasing function of $\Delta\sigma^I$ during the crisis. France and Germany are close countries and have strong economic relationships, but Germany is seen as a country in the "center of Europe" while France shares a lot of similarities with countries of the periphery (in the South of Europe). Indeed, during the crisis, when implied volatility decreases (corresponding to a negative value of $\Delta\sigma^I$), the dependence is higher, which can be interpreted as investors seeing the two countries as close. On the contrary, when the market implied volatility increases, there are concerns in the market about the robustness of Eurozone and investors raise doubts about southern European countries – including France – which tend to decrease the conditional Kendall's tau between French and German returns.

After the crisis, the couples (Germany, France), and (Germany, Denmark) revert to a more usual shape of conditional dependence: when volatility does not change much, conditional Kendall's tau is low ; when volatility changes much, conditional Kendall's tau is higher, reflecting more stressed situations. In this period, an exception is the couple (Germany, Greece), whose conditional Kendall's tau has a particular shape, that looks like the one of the couple (Germany, France) during the crisis. This is coherent with the fact that, in stressed situations, when volatility increases, investors sometimes remember that Greece still has a fragile economy, which results in a lower conditional Kendall's tau. But three estimators suggest that, when volatility increases very much, conditional Kendall's tau between Germany and Greece increases again, following the classical tendencies that we had already observed.

Table 10
Strengths and weaknesses of the proposed estimation procedures.

Method	Performance in the sense of (10)	Computation time	Interpretation	Tuning parameters	
				Number	Difficulty of choice
Logit/Probit (well-specified)	Best	Very slow	Yes	1	Easy
Logit/Probit (mis-specified)	Low	Very slow	Possible	1	Easy
Tree	Average	Very fast	Possible	3 (see Ripley (2018))	Average
Random forests	Good	Average	No	At least 4	Average
Nearest neighbors	Very good	Fast	No	At least 5	Complicated
Neural network	Excellent	Slow	No	At least 2	Complicated

6. Conclusion

In a parametric setting, we have proposed a localized log-likelihood method to estimate conditional Kendall's tau. When the link function is analytically tractable and explicit, it is then possible to code and optimize the full penalized criterion. The consistency and the asymptotic normality of such estimators have been stated. In particular, this is the case for logit or probit-type link functions. We noticed that evaluating a Kendall's tau is equivalent to evaluating a probability of being classified as a concordant pair. Therefore, most classification procedures can be adapted to directly estimate conditional Kendall's tau. Classification trees, random forests, nearest neighbors and neural networks have been discussed. They generally provide more flexible parametric models than previously.

We note that multiple trade-offs arise when choosing one of these methods, as displayed in Table 10. Depending on the requirements of the situation, statisticians can choose some algorithms that best match their needs. To summarize, trees and random forests methods are the fastest ones, but exhibit the lowest performances. Parametric methods such as the logit and probit may perform very well under some "simple" functional forms of g and $\boldsymbol{\psi}$, but they deteriorate quickly when the true underlying model departs from their parametric specification. Note that they also show the longest computation time. Nonetheless, interpretability of the coefficient β can be useful in applications. Even if the model is misspecified, it can still be seen as an estimation of the best approximation of $\mathbf{z} \mapsto \tau_{1,2|\mathbf{Z}=\mathbf{z}}$ on the functional space generated by $\boldsymbol{\psi}$. Nearest neighbors methods are average in terms of computation time as well as performance. Neural networks are the slowest of all our nonparametric methods, but they behave nearly uniformly the best ones in terms of prediction. Finally, we have evaluated these different methods on several empirical illustrations.

Acknowledgments

This work is supported by the Labex Ecodec (France) under the grant ANR-11-LABEX-0047 from the French Agence Nationale de la Recherche. The first author thanks Solt Kovács for inspiring ideas and a discussion which lead to this article.

Appendix A. Some basic definitions about copulas

Here, we recall the main concepts around copulas and conditional copulas. First, a d-dimensional copula is a cdf on $[0,\,1]^d$ whose margins are uniform distributions. Sklar's theorem states that, for any d-dimensional distributions H, whose marginal cdfs' are denoted as $F_1, \ldots, F_d$, there exists a copula C s.t.

$$H(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big), \tag{A.1}$$

for every $(x_1, \ldots, x_d) \in \mathbb{R}^d$. If the law of H is continuous, the latter C is unique, and it is called *the copula* associated to H. Inversely, for a given copula and some univariate cdfs' F_k, $k = 1, \ldots, d$, Eq. (A.1) defines a d-dimensional cdf H.

The latter concept of copula is similarly related to any random vector $\mathbf{X}$ whose cdf is H, and there is no ambiguity by using the same term. Copulas are invariant w.r.t. strictly increasing transforms of the margins X_k, $k = 1, \ldots, d$. They provide very practical tools for modeling complex and/or highly dimensional distributions in a flexible way, by splitting the task into two parts: the specification of the marginal distributions on one side, and the specification of the copula on the other side. Therefore, a copula can be seen as a function that describes the dependence between the components of $\mathbf{X}$, independently of the marginal distributions. Several popular dependence measures are functionals of the underlying copula only: Kendall's tau, Spearman's rho, Blomqvist coefficient, etc. The classical textbooks by Joe (2015) or Nelsen (2007) provide numerous and detailed results.

Numerous parametric families of copulas have been proposed in the literature: Gaussian, Student, Archimedean, Marshall–Olkin, extreme-value, etc. Several inference methods have been adapted to evaluate an underlying copula, possible without estimating the marginal cdfs' (Canonical Maximum Likelihood). See Cherubini et al. (2004) for details. Nonparametric methods have been developed too, since the seminal papers of Deheuvels (1979, 1981) about empirical copula processes.

Second, conditional copulas have been formally introduced by Patton (2006b,a). They are rather straightforward extensions of the latter concepts, when dealing with conditional distributions. Formally, for a given sigma-algebra $\mathcal{F}$, let $H(\cdot|\mathcal{F})$

(resp. $F_k(\cdot|\mathcal{F})$) be the conditional distribution of $\mathbf{X}$ (resp. X_k, $k = 1, \dots, d$) given $\mathcal{F}$. The "conditional version of" Sklar's theorem now states that there exists a random copula $C(\cdot|\mathcal{F})$ s.t.

$$H(x_1, \dots, x_d|\mathcal{F}) = C\big(F_1(x_1|\mathcal{F}), \dots, F_d(x_d|\mathcal{F})|\mathcal{F}\big), \quad \text{a.e.} \tag{A.2}$$

for every $(x_1, \dots, x_d) \in \mathbb{R}^d$. If the law of $H(\cdot|\mathcal{F})$ is continuous, the latter $C(\cdot|\mathcal{F})$ is unique, and it is called *the conditional copula* associated to $H(\cdot|\mathcal{F})$, given $\mathcal{F}$. Inversely, given $\mathcal{F}$, a conditional copula $C(\cdot|\mathcal{F})$ and some univariate cdfs' $F_k(\cdot|\mathcal{F})$, $k = 1, \dots, d$, Eq. (A.2) defines a d-dimensional conditional cdf $H(\cdot|\mathcal{F})$. See Fermanian and Wegkamp (2012) for extensions of the latter concepts.

Appendix B. Proof of Theorem 3

Simple calculations provide: if $i \neq j$ and under (3),

$$\begin{aligned}
&\mathbb{E}[L_n(\beta)] = \mathbb{E}\left[K_h(\mathbf{Z}_i - \mathbf{Z}_j)\ell_\beta(W_{(i,j)}, \mathbf{Z}_i)\right] = \mathbb{E}\left[K_h(\mathbf{Z}_i - \mathbf{Z}_j)\mathbb{E}[\ell_\beta(W_{(i,j)}, \mathbf{Z}_i)|\mathbf{Z}_i, \mathbf{Z}_j]\right] \\
&= \mathbb{E}\left[K_h(\mathbf{Z}_i - \mathbf{Z}_j)\left(p(\mathbf{Z}_i, \mathbf{Z}_j)\log\left(\frac{1}{2} + \frac{1}{2}g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right) + (1 - p(\mathbf{Z}_i, \mathbf{Z}_j))\log\left(\frac{1}{2} - \frac{1}{2}g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)\right)\right)\right] \\
&= \mathbb{E}\left[K_h(\mathbf{Z}_i - \mathbf{Z}_j)\phi(\mathbf{Z}_i, \mathbf{Z}_j, \beta)\right] \\
&= \mathbb{E}\left[\int K(\mathbf{t})\phi(\mathbf{Z}_i, \mathbf{Z}_i - h\mathbf{t}, \beta)f_{\mathbf{Z}}(\mathbf{Z}_i - h\mathbf{t})\,d\mathbf{t}\right],
\end{aligned}$$

that tends to $\mathbb{E}[\phi(\mathbf{Z}_i, \mathbf{Z}_i, \beta)f_{\mathbf{Z}}(\mathbf{Z}_i)] = L_\infty(\beta)$ when $n \to \infty$, if $\int \left(\phi(\mathbf{z}, \cdot, \beta)f_{\mathbf{Z}}(\cdot)\right)_\varepsilon(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty$, for some $\varepsilon > 0$ (invoke the dominated convergence Theorem and the compact support of K).

Now, let us prove that, for any β, $L_n(\beta)$ tends towards $L_\infty(\beta)$ in probability, when $n \to \infty$. It is sufficient to prove that the variance of $L_n(\beta)$ tends to zero.

$$\begin{aligned}
&\mathbb{E}\left[\left(L_n(\beta) - \mathbb{E}[L_n(\beta)]\right)^2\right] = \frac{1}{n^2(n-1)^2}\sum_{i_1,j_1;i_1\neq j_1}\ \sum_{i_2,j_2;i_2\neq j_2} \\
&\quad\left(\mathbb{E}\left[K_h(\mathbf{Z}_{i_1} - \mathbf{Z}_{j_1})K_h(\mathbf{Z}_{i_2} - \mathbf{Z}_{j_2})\ell_\beta(W_{(i_1,j_1)}, \mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,j_2)}, \mathbf{Z}_{i_2})\right] - \mathbb{E}[L_n(\beta)]^2\right) \\
&=: \frac{1}{n^2(n-1)^2}\sum_{i_1,j_1;i_1\neq j_1}\ \sum_{i_2,j_2;i_2\neq j_2} v_{i_1,j_1,i_2,j_2},
\end{aligned}$$

with obvious notation. Obviously, v_{i_1,j_1,i_2,j_2} is zero when i_1 and j_1 are not equal to i_2 nor j_2. At the opposite, in the case of equalities between some of these four indices, we get non-zero terms.

To be specific, when $i_1 = i_2 = i$, and $j_1 \neq j_2$, we have

$$\begin{aligned}
&v_{i,j_1,i,j_2} = \mathbb{E}\left[K_h(\mathbf{Z}_i - \mathbf{Z}_{j_1})K_h(\mathbf{Z}_i - \mathbf{Z}_{j_2})\ell_\beta(W_{(i,j_1)}, \mathbf{Z}_i)\ell_\beta(W_{(i,j_2)}, \mathbf{Z}_i)\right] - \mathbb{E}[L_n(\beta)]^2 \\
&= \mathbb{E}\left[\int K(\mathbf{x})K(\mathbf{y})A(\mathbf{Z}_i, \mathbf{Z}_i - h\mathbf{x}, \mathbf{Z}_i - h\mathbf{y})f_{\mathbf{Z}}(\mathbf{Z}_i - h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_i - h\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}\right] - \mathbb{E}[L_n(\beta)]^2,
\end{aligned}$$

by setting

$$\begin{aligned}
&A(\mathbf{x}, \mathbf{y}, \mathbf{z}) := \mathbb{E}\left[\ell_\beta(W_{(i,j_1)}, \mathbf{Z}_i)\ell_\beta(W_{(i,j_2)}, \mathbf{Z}_i)|\mathbf{Z}_i = \mathbf{x}, \mathbf{Z}_{j_1} = \mathbf{y}, \mathbf{Z}_{j_2} = \mathbf{z}\right] \\
&= p(\mathbf{x}, \mathbf{y}, \mathbf{z})\log^2 q(\mathbf{x}, \beta) + (p(\mathbf{x}, \mathbf{y}) - p(\mathbf{x}, \mathbf{y}, \mathbf{z}))\log q(\mathbf{x}, \beta)\log(1 - q(\mathbf{x}, \beta)) \\
&+ (p(\mathbf{x}, \mathbf{z}) - p(\mathbf{x}, \mathbf{y}, \mathbf{z}))\log q(\mathbf{x}, \beta)\log(1 - q(\mathbf{x}, \beta)) + (1 - p(\mathbf{x}, \mathbf{y}) - p(\mathbf{x}, \mathbf{z}) + p(\mathbf{x}, \mathbf{y}, \mathbf{z}))\log^2(1 - q(\mathbf{x}, \beta)).
\end{aligned}$$

If $\int A(\mathbf{z}, \cdot, \cdot)_\varepsilon(\mathbf{z}, \mathbf{z})f^2_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty$ for some $\varepsilon > 0$, then v_{i,j_1,i,j_2} tends to a constant when $n \to \infty$ (independently of the choice of such indices).

A similar analysis can be led for the other terms. When $i_1 = j_2$ and $j_1 \neq i_2$, we get

$$\begin{aligned}
&v_{i_1,j_1,i_2,i_1} = \mathbb{E}\left[K_h(\mathbf{Z}_{i_1} - \mathbf{Z}_{j_1})K_h(\mathbf{Z}_{i_2} - \mathbf{Z}_{i_1})\ell_\beta(W_{(i_1,j_1)}, \mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,i_1)}, \mathbf{Z}_{i_2})\right] - \mathbb{E}[L_n(\beta)]^2 \\
&= \mathbb{E}\left[\int K(\mathbf{x})K(\mathbf{y})B(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_1} - h\mathbf{x}, \mathbf{Z}_{i_1} + h\mathbf{y})f_{\mathbf{Z}}(\mathbf{Z}_{i_1} - h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_{i_1} + h\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}\right] - \mathbb{E}[L_n(\beta)]^2,
\end{aligned}$$

by setting

$$\begin{aligned}
&B(\mathbf{x}, \mathbf{y}, \mathbf{z}) := \mathbb{E}\left[\ell_\beta(W_{(i_1,j_1)}, \mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,i_1)}, \mathbf{Z}_{i_2})|\mathbf{Z}_{i_1} = \mathbf{x}, \mathbf{Z}_{j_1} = \mathbf{y}, \mathbf{Z}_{i_2} = \mathbf{z}\right] \\
&= p(\mathbf{x}, \mathbf{y}, \mathbf{z})\log q(\mathbf{x}, \beta)\log q(\mathbf{z}, \beta) + (p(\mathbf{x}, \mathbf{y}) - p(\mathbf{x}, \mathbf{y}, \mathbf{z}))\log q(\mathbf{x}, \beta)\log(1 - q(\mathbf{z}, \beta))
\end{aligned}$$

$$+ (p(\mathbf{x},\mathbf{z}) - p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{z},\beta)\log(1-q(\mathbf{x},\beta))$$
$$+ (1 - p(\mathbf{x},\mathbf{y}) - p(\mathbf{x},\mathbf{z}) + p(\mathbf{x},\mathbf{y},\mathbf{z}))\log(1-q(\mathbf{x},\beta))\log(1-q(\mathbf{z},\beta)).$$

If $\int B(\mathbf{z},\cdot,\cdot)_\varepsilon(\mathbf{z},\mathbf{z})f^2_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty$, then v_{i_1,j_1,i,i_1} tends to a constant when $n \to \infty$.
When $j_1 = j_2 = j$ and $i_1 \neq i_2$, we obtain

$$v_{i_1,j_1,i_2,j_2} = \mathbb{E}\left[K_h(\mathbf{Z}_{i_1} - \mathbf{Z}_j)K_h(\mathbf{Z}_{i_2} - \mathbf{Z}_j)\ell_\beta(W_{(i_1,j)},\mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,j)},\mathbf{Z}_{i_2})\right] - \mathbb{E}[L_n(\beta)]^2$$
$$= \mathbb{E}\left[\int K(\mathbf{x})K(\mathbf{y})C(\mathbf{Z}_j + h\mathbf{x},\mathbf{Z}_j,\mathbf{Z}_j + h\mathbf{y})f_{\mathbf{Z}}(\mathbf{Z}_j + h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_j + h\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}\right] - \mathbb{E}[L_n(\beta)]^2,$$

by setting

$$C(\mathbf{x},\mathbf{y},\mathbf{z}) := \mathbb{E}\left[\ell_\beta(W_{(i_1,j)},\mathbf{Z}_{i_1})\ell_\beta(W_{(i_2,j)},\mathbf{Z}_{i_2})|\mathbf{Z}_{i_1} = \mathbf{x},\mathbf{Z}_j = \mathbf{y},\mathbf{Z}_{i_2} = \mathbf{z}\right]$$
$$= p(\mathbf{x},\mathbf{y},\mathbf{z})\log q(\mathbf{x},\beta)\log q(\mathbf{z},\beta) + (p(\mathbf{x},\mathbf{y}) - p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{x},\beta)\log(1-q(\mathbf{z},\beta))$$
$$+ (p(\mathbf{y},\mathbf{z}) - p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{z},\beta)\log(1-q(\mathbf{x},\beta))$$
$$+ (1 - p(\mathbf{x},\mathbf{y}) - p(\mathbf{y},\mathbf{z}) + p(\mathbf{x},\mathbf{y},\mathbf{z}))\log(1-q(\mathbf{x},\beta))\log(1-q(\mathbf{z},\beta)).$$

If $\int C(\cdot,\mathbf{z},\cdot)_\varepsilon(\mathbf{z},\mathbf{z})f^2_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty$, then $v_{i_1,j,i_2,j}$ tends to a constant when $n \to \infty$.
When $j_1 = i_2$ and $i_1 \neq j_2$:

$$v_{i_1,j_1,j_1,j_2} = \mathbb{E}\left[K_h(\mathbf{Z}_{i_1} - \mathbf{Z}_{j_1})K_h(\mathbf{Z}_{j_1} - \mathbf{Z}_{j_2})\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(j_1,j_2)},\mathbf{Z}_{j_1})\right] - \mathbb{E}[L_n(\beta)]^2$$
$$= \mathbb{E}\left[\int K(\mathbf{x})K(\mathbf{y})D(\mathbf{Z}_{j_1} + h\mathbf{x},\mathbf{Z}_{j_1},\mathbf{Z}_{j_1} - h\mathbf{y})f_{\mathbf{Z}}(\mathbf{Z}_{j_1} + h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_{j_1} - h\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}\right] - \mathbb{E}[L_n(\beta)]^2,$$

by setting

$$D(\mathbf{x},\mathbf{y},\mathbf{z}) := \mathbb{E}\left[\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(j_1,j_2)},\mathbf{Z}_{j_1})|\mathbf{Z}_{i_1} = \mathbf{x},\mathbf{Z}_{j_1} = \mathbf{y},\mathbf{Z}_{j_2} = \mathbf{z}\right]$$
$$= p(\mathbf{x},\mathbf{y},\mathbf{z})\log q(\mathbf{x},\beta)\log q(\mathbf{y},\beta) + (p(\mathbf{x},\mathbf{y}) - p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{x},\beta)\log(1-q(\mathbf{y},\beta))$$
$$+ (p(\mathbf{y},\mathbf{z}) - p(\mathbf{x},\mathbf{y},\mathbf{z}))\log q(\mathbf{y},\beta)\log(1-q(\mathbf{x},\beta))$$
$$+ (1 - p(\mathbf{x},\mathbf{y}) - p(\mathbf{y},\mathbf{z}) + p(\mathbf{x},\mathbf{y},\mathbf{z}))\log(1-q(\mathbf{x},\beta))\log(1-q(\mathbf{y},\beta)).$$

If $\int D(\cdot,\mathbf{z},\cdot)_\varepsilon(\mathbf{z},\mathbf{z})f^2_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty$, then v_{i_1,j_1,j_1,j_2} tends to a constant when $n \to \infty$.
There are two cases of two equalities. If $i_1 = i_2 = i$ and $j_1 = j_2 = j$, this yields

$$v_{i,j,i,j} = \mathbb{E}\left[K_h(\mathbf{Z}_i - \mathbf{Z}_j)^2\ell^2_\beta(W_{(i,j)},\mathbf{Z}_i)\right] - \mathbb{E}[L_n(\beta)]^2$$
$$= h^{-p}\mathbb{E}\left[\int K(\mathbf{x})^2 E(\mathbf{Z}_i,\mathbf{Z}_i - h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_i - h\mathbf{x})\,d\mathbf{x}\right] - \mathbb{E}[L_n(\beta)]^2,$$

by setting

$$E(\mathbf{x},\mathbf{y}) := \mathbb{E}\left[\ell^2_\beta(W_{(i,j)},\mathbf{Z}_i)|\mathbf{Z}_i = \mathbf{x},\mathbf{Z}_j = \mathbf{y}\right]$$
$$= p(\mathbf{x},\mathbf{y})\log^2 q(\mathbf{x},\beta) + (1 - p(\mathbf{x},\mathbf{y}))\log^2(1-q(\mathbf{x},\beta)).$$

If $\int E(\mathbf{z},\cdot)_\varepsilon(\mathbf{z})f_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty$, then $h^p v_{i,j,i,j}$ tends to a constant when $n \to \infty$.
Finally, if $i_1 = j_2$ and $j_1 = i_2$, we get

$$v_{i_1,j_1,i_1,j_1} = \mathbb{E}\left[K_h(\mathbf{Z}_{i_1} - \mathbf{Z}_{j_1})^2\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(j_1,i_1)},\mathbf{Z}_{j_1})\right] - \mathbb{E}[L_n(\beta)]^2$$
$$= h^{-p}\mathbb{E}\left[\int K(\mathbf{x})^2 F(\mathbf{Z}_i,\mathbf{Z}_i - h\mathbf{x})f_{\mathbf{Z}}(\mathbf{Z}_i - h\mathbf{x})\,d\mathbf{x}\right] - \mathbb{E}[L_n(\beta)]^2,$$

by setting

$$F(\mathbf{x},\mathbf{y}) := \mathbb{E}\left[\ell_\beta(W_{(i_1,j_1)},\mathbf{Z}_{i_1})\ell_\beta(W_{(j_1,i_1)},\mathbf{Z}_{j_1})|\mathbf{Z}_{i_1} = \mathbf{x},\mathbf{Z}_{j_1} = \mathbf{y}\right]$$
$$= p(\mathbf{x},\mathbf{y})\log q(\mathbf{x},\beta)\log q(\mathbf{y},\beta) + (1 - p(\mathbf{x},\mathbf{y}))\log(1-q(\mathbf{x},\beta))\log(1-q(\mathbf{y},\beta)).$$

If $\int F(\mathbf{z},\cdot)_\varepsilon(\mathbf{z})f_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty$, then $h^p v_{i_1,j_1,j_1,i_1}$ tends to a constant when $n \to \infty$.

Summarizing the previous terms, we have obtained $Var(L_n(\beta)) = O(n^{-1} + n^{-2}h^{-p})$, that tends to zero pointwise, when $n^2h^p \to \infty$. We deduce $L_n(\beta) - L_\infty(\beta) = L_n(\beta) - \mathbb{E}[L_n(\beta)] + \mathbb{E}[L_n(\beta)] - L_\infty(\beta) = o_P(1)$. Since $L_n(\cdot)$ and $L_\infty(\cdot)$ are concave, invoking the convexity lemma of Geyer (1996) (see Knight and Fu, 2000, alternatively), the maximizer $\hat{\beta}$ of L_n tends in probability towards the maximizer of L_∞. □

We summarize the latter technical assumptions that are sufficient to obtain the consistency of $\hat{\beta}$: for some $\varepsilon > 0$,

$$\int \left(\phi(\mathbf{z},\cdot,\beta)f_{\mathbf{Z}}(\cdot)\right)_\varepsilon(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty, \tag{B.1}$$

$$\int \Big(A(\mathbf{z},\cdot,\cdot)_\varepsilon(\mathbf{z},\mathbf{z}) + B(\mathbf{z},\cdot,\cdot)_\varepsilon(\mathbf{z},\mathbf{z}) + C(\cdot,\mathbf{z},\cdot)_\varepsilon(\mathbf{z},\mathbf{z}) + D(\cdot,\mathbf{z},\cdot)_\varepsilon(\mathbf{z},\mathbf{z}) \Big) f^2_{\mathbf{Z},\varepsilon}(\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} < \infty, \tag{B.2}$$

$$\int \Big(\phi(\mathbf{z},\cdot,\beta)_\varepsilon(\mathbf{z}) + E(\mathbf{z},\cdot)_\varepsilon(\mathbf{z}) + F(\mathbf{z},\cdot)_\varepsilon(\mathbf{z}) \Big) f_{\mathbf{Z},\varepsilon}(\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} < \infty. \tag{B.3}$$

Appendix C. Proof of Theorem 4

Set $\mathbf{u} := \sqrt{n}(\beta - \beta^*)$ and $\hat{\mathbf{u}} := \sqrt{n}(\hat{\beta} - \beta^*)$. Obviously,

$$\begin{aligned}\hat{\mathbf{u}} &= \arg\max_{\mathbf{u}\in\mathbb{R}^{p'}} L_n(\beta^* + n^{-1/2}\mathbf{u}) - \lambda_n |\beta^* + n^{-1/2}\mathbf{u}|_1, \\ &= \arg\max_{\mathbf{u}\in\mathbb{R}^{p'}} nL_n(\beta^* + n^{-1/2}\mathbf{u}) - nL_n(\beta^*) - n\lambda_n \big\{ |\beta^* + n^{-1/2}\mathbf{u}|_1 - |\beta^*|_1 \big\}.\end{aligned}$$

Note that

$$\begin{aligned} n\lambda_n|\beta^* + n^{-1/2}\mathbf{u}|_1 - n|\beta^*|_1 &= n^{1/2}\lambda_n \sum_{k;\beta_k^*=0} |u_k| + n^{1/2}\lambda_n \sum_{k;\beta_k^*\neq 0} sign(\beta_k^*)u_k \\ \longrightarrow\ \mu \sum_{k;\beta_k^*=0} |u_k| + \mu \sum_{k;\beta_k^*\neq 0} sign(\beta_k^*)u_k, \end{aligned}$$

when $n \to \infty$. Moreover,

$$nL_n(\beta^* + n^{-1/2}\mathbf{u}) - nL_n(\beta^*) = n^{1/2}\dot{L}_n(\beta^*).\mathbf{u} + \frac{1}{2}\mathbf{u}^T\ddot{L}_n(\overline{\beta})\mathbf{u} + \frac{1}{6\sqrt{n}}\dddot{L}_n(\overline{\beta}).\mathbf{u}^{(3)},$$

for some (random) $\overline{\beta}$ s.t. $|\beta^* - \overline{\beta}| < |\beta^* - \beta|$. We will successively prove that

(i) $n^{1/2}\dot{L}_n(\beta^*)$ weakly tends to a Gaussian random vector $\mathbb{W}$, $\mathbb{W} \sim \mathcal{N}(0_p, \Sigma_{\beta^*})$;
(ii) $\ddot{L}_n(\beta^*)$ tends in probability towards a constant matrix $\mathbb{H}(\beta^*)$;
(iii) $\dddot{L}_n(\overline{\beta})$ is $O_P(1)$.

Then, $\hat{\mathbf{u}} = \arg\max_{\mathbf{u}} \mathcal{L}_n(\mathbf{u})$, where $\mathcal{L}_n(\mathbf{u})$ weakly tends to

$$\mathcal{L}_\infty(\mathbf{u}) := \mathbb{W}.\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbb{H}(\beta^*)\mathbf{u} - \mu \sum_{k;\beta_k^*=0} |u_k| - \mu \sum_{k;\beta_k^*\neq 0} sign(\beta_k^*)u_k,$$

that is concave. The result will follow, applying Theorem 1 in Kato (2009).

First, let us prove (i), i.e. the asymptotic normality of $n^{1/2}\dot{L}_n(\beta)$ for any given parameter β. Consider the centered criterion

$$M_n(\beta) := \dot{L}_n(\beta) - \mathbb{E}[\dot{L}_n(\beta)] = \frac{1}{n(n-1)} \sum_{i,j;i\neq j} \ell_{ij}(\beta),$$

where $\ell_{ij} := K_h(\mathbf{Z}_i - \mathbf{Z}_j)\partial_\beta \ell_\beta(W_{(i,j)}, \mathbf{Z}_i) - \mathbb{E}[\dot{L}_n(\beta)]$. We symmetrize the localized likelihood:

$$M_n(\beta) = \frac{1}{2n(n-1)} \sum_{i,j;i\neq j} M_{ij}(\beta),$$

where $M_{ij}(\beta)$ (or simply M_{ij}) is $\ell_{ij}(\beta) + \ell_{ji}(\beta)$. Note that $M_{ij} = M_{ji}$ and that $\mathbb{E}[M_{ij}] = 0$.

By the dominated convergence theorem and a change of variable, we easily check that $\mathbb{E}[\dot{L}_n(\beta)] = \partial_\beta L_\infty(\beta) + o(1)$ if, for some $\varepsilon > 0$, we have $\int \big(\partial_\beta \phi(\mathbf{z},\cdot,\beta) f_{\mathbf{Z}}(\cdot)\big)_\varepsilon(\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} < \infty$. Moreover, by simple calculations, we get, if $i \neq j$,

$$\begin{aligned} \mathbb{E}[M_n|\mathbf{Z}_i] &= \frac{1}{2n}\mathbb{E}[M_{ij} + M_{ji}|\mathbf{Z}_i] = \frac{1}{n}\mathbb{E}[M_{ij}|\mathbf{Z}_i] \\ &= \frac{1}{n}\int \{K_h(\mathbf{Z}_i - \mathbf{z})\partial_\beta\phi(\mathbf{Z}_i,\mathbf{z},\beta) + K_h(\mathbf{z} - \mathbf{Z}_i)\partial_\beta\phi(\mathbf{z},\mathbf{Z}_i,\beta)\} f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} - \frac{2}{n}\mathbb{E}[\dot{L}_n(\beta)] \\ &= \frac{1}{n}\int K(\mathbf{t})\{\partial_\beta\phi(\mathbf{Z}_i,\mathbf{Z}_i - h\mathbf{t},\beta) f_{\mathbf{Z}}(\mathbf{Z}_i - h\mathbf{t}) + \partial_\beta\phi(\mathbf{Z}_i + h\mathbf{t},\mathbf{Z}_i,\beta) f_{\mathbf{Z}}(\mathbf{Z}_i + h\mathbf{t})\}\, d\mathbf{t} - \frac{2}{n}\mathbb{E}[\dot{L}_n(\beta)] \\ &= \frac{2}{n}\partial_\beta\phi(\mathbf{Z}_i,\mathbf{Z}_i,\beta) f_{\mathbf{Z}}(\mathbf{Z}_i) - \frac{2}{n}\dot{L}_\infty(\beta) + o(n^{-1}) + r_{n,i}, \end{aligned}$$

where, by a m-order limited expansion, we obtain

$$\|r_{n,i}\| \leq \frac{Cst.h^m \int |K|}{nm!} \|(f_{\mathbf{Z}}(\cdot)\partial_\beta\phi(\mathbf{Z}_i,\cdot,\beta))^{(m)} + (f_{\mathbf{Z}}(\cdot)\partial_\beta\phi(\cdot,\mathbf{Z}_i,\beta))^{(m)}\|_\varepsilon(\mathbf{Z}_i),$$

for any norm $\|\cdot\|$ on $\mathbb{R}^p$ and a positive constant Cst. We deduce that $n^{1/2}\sum_{i=1}^n \mathbb{E}[M_n|\mathbf{Z}_i]$ is asymptotically normal by invoking the usual CLT, under condition (C.2). To be specific, the Hájek projection of M_n is

$$\frac{\sqrt{n}}{2}\sum_{i=1}^n \mathbb{E}[M_n|\mathbf{Z}_i] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\partial_\beta\phi(\mathbf{Z}_i,\mathbf{Z}_i,\beta)f_{\mathbf{Z}}(\mathbf{Z}_i) - \dot{L}_\infty(\beta)\} + o_P(1) \rightsquigarrow \mathcal{N}(0,\,\Sigma_\beta),\ \text{with}$$

$$\Sigma_\beta := \int \partial_\beta\phi(\mathbf{z},\mathbf{z},\beta)\partial_\beta\phi(\mathbf{z},\mathbf{z},\beta)^T f_{\mathbf{Z}}^3(\mathbf{z})\,d\mathbf{z} - \dot{L}_\infty(\beta)\dot{L}_\infty(\beta)^T.$$

Note that $\dot{L}_\infty(\beta^*) = 0$.

Now consider the "remainder term" $\Delta_n := M_n(\beta) - \sum_{i=1}^n \mathbb{E}[M_n|\mathbf{Z}_i]/2$. Since $\mathbb{E}[M_n|\mathbf{Z}_i] = n^{-1}\mathbb{E}[M_{ij}|\mathbf{Z}_i]$, we deduce

$$\Delta_n = M_n(\beta) - \frac{1}{2}\sum_{i=1}^n \mathbb{E}[M_n|\mathbf{Z}_i] = \frac{1}{2n(n-1)}\sum_{i,j;i\neq j}\{M_{ij} - \mathbb{E}[M_{ij}|\mathbf{Z}_i]\}.$$

It is relatively easy to prove that Δ_n is negligible, i.e. $\Delta_n = o_P(n^{-1/2})$. Indeed, let us prove that the variance of $n^{1/2}\Delta_n$ tends to zero with n:

$$Var(n^{1/2}\Delta_n) = n\mathbb{E}[\Delta_n\Delta_n^T] = \frac{1}{4n(n-1)^2}\sum_{i_1,j_1:i_1\neq j_1}\sum_{i_2,j_2:i_2\neq j_2}\delta(i_1,i_2,j_1,j_2),$$

$$\delta(i_1,i_2,j_1,j_2) = \mathbb{E}\left[\{M_{i_1j_1} - \mathbb{E}[M_{i_1j_1}|\mathbf{Z}_{i_1}]\}\times\{M_{i_2j_2} - \mathbb{E}[M_{i_2j_2}|\mathbf{Z}_{i_2}]\}^T\right].$$

If there is no identity among the indices (i_1,j_1,i_2,j_2), with $i_1\neq j_1$ and $i_2\neq j_2$, then $\delta(i_1,i_2,j_1,j_2)$ is zero. Moreover, this is still the case when there is only a single identity. For instance, assume $i_1 = i_2 = i$ and $j_1\neq j_2$. Then,

$$\begin{aligned}
&\delta(i,i,j_1,j_2) = \mathbb{E}\left[\{M_{ij_1} - \mathbb{E}[M_{ij_1}|\mathbf{Z}_i]\}\times\{M_{ij_2} - \mathbb{E}[M_{ij_2}|\mathbf{Z}_i]\}^T\right]\\
&= \mathbb{E}\left[\{M_{ij_1} - \mathbb{E}[M_{ij_1}|\mathbf{Z}_i]\}\times\mathbb{E}[\{M_{ij_2} - \mathbb{E}[M_{ij_2}|\mathbf{Z}_i]\}^T|\mathbf{Z}_i,\mathbf{Z}_{j_1}]\right]\\
&= \mathbb{E}\left[\{M_{ij_1} - \mathbb{E}[M_{ij_1}|\mathbf{Z}_i]\}\times 0\right] = 0.
\end{aligned}$$

The other terms for which a single identity between the indices can be managed similarly.

At the opposite, non-zero terms appear when $i_1 = i_2 = i$ and $j_1 = j_2 = j$. In this case, we obtain

$$\delta(i,i,j,j) = \mathbb{E}\left[\{M_{ij} - \mathbb{E}[M_{ij}|\mathbf{Z}_i]\}\{M_{ij} - \mathbb{E}[M_{ij}|\mathbf{Z}_i]\}^T\right] = \mathbb{E}[M_{ij}M_{ij}^T] - \mathbb{E}[\mathbb{E}[M_{ij}|\mathbf{Z}_i]\mathbb{E}[M_{ij}|\mathbf{Z}_i]^T].$$

By a usual change of variable and by symmetry, we get

$$\begin{aligned}
&\mathbb{E}[M_{ij}M_{ij}^T] = 2\mathbb{E}\left[\{\ell_{ij}(\beta)\ell_{ij}(\beta)^T + \ell_{ij}(\beta)\ell_{ji}(\beta)^T\}\right]\\
&= 2\mathbb{E}\left[K_h^2(\mathbf{Z}_i-\mathbf{Z}_j)\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)^T\right.\\
&+\ \left. K_h(\mathbf{Z}_i-\mathbf{Z}_j)K_h(\mathbf{Z}_j-\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(j,i)},\mathbf{Z}_j)^T\right] + O(1)\\
&= 2h^{-p}\mathbb{E}\left[\int\left(K^2(\mathbf{x})H_1(\mathbf{Z}_i,\mathbf{Z}_i-h\mathbf{x}) + K(\mathbf{x})K(-\mathbf{x})H_2(\mathbf{Z}_i,\mathbf{Z}_i-h\mathbf{x})\right)f_{\mathbf{Z}}(\mathbf{Z}_i-h\mathbf{x})\,d\mathbf{x}\right] + O(1)\\
&= 2h^{-p}\int\left(K^2(\mathbf{x})H_1(\mathbf{z},\mathbf{z}-h\mathbf{x}) + K(\mathbf{x})K(-\mathbf{x})H_2(\mathbf{z},\mathbf{z}-h\mathbf{x})\right)f_{\mathbf{Z}}(\mathbf{z}-h\mathbf{x})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{x}\,d\mathbf{z} + O(1),
\end{aligned}$$

by setting

$$\begin{aligned}
&H_1(\mathbf{x},\mathbf{y}) = \mathbb{E}\left[\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)^T|\mathbf{Z}_i=\mathbf{x},\mathbf{Z}_j=\mathbf{y}\right]\\
&= p(\mathbf{x},\mathbf{y})\frac{\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})^T g'(\boldsymbol{\psi}(\mathbf{x})^T\beta)^2}{(1+g(\boldsymbol{\psi}(\mathbf{x})^T\beta))^2} + (1-p(\mathbf{x},\mathbf{y}))\frac{\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{x})^T g'(\boldsymbol{\psi}(\mathbf{x})^T\beta)^2}{(1-g(\boldsymbol{\psi}(\mathbf{x})^T\beta))^2},\ \text{and}\\
&H_2(\mathbf{x},\mathbf{y}) = \mathbb{E}\left[\partial_\beta\ell_\beta(W_{(i,j)},\mathbf{Z}_i)\partial_\beta\ell_\beta(W_{(j,i)},\mathbf{Z}_j)^T|\mathbf{Z}_i=\mathbf{x},\mathbf{Z}_j=\mathbf{y}\right]\\
&= p(\mathbf{x},\mathbf{y})\frac{\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{y})^T g'(\boldsymbol{\psi}(\mathbf{x})^T\beta)g'(\boldsymbol{\psi}(\mathbf{y})^T\beta)}{(1+g(\boldsymbol{\psi}(\mathbf{x})^T\beta))(1+g(\boldsymbol{\psi}(\mathbf{y})^T\beta))} + (1-p(\mathbf{x},\mathbf{y}))\frac{\boldsymbol{\psi}(\mathbf{x})\boldsymbol{\psi}(\mathbf{y})^T g'(\boldsymbol{\psi}(\mathbf{x})^T\beta)g'(\boldsymbol{\psi}(\mathbf{y})^T\beta)}{(1-g(\boldsymbol{\psi}(\mathbf{x})^T\beta))(1+g(\boldsymbol{\psi}(\mathbf{y})^T\beta))}.
\end{aligned}$$

Therefore, $\mathbb{E}[M_{ij}M_{ij}^T]$ is $O(h^{-p})$, if $\int(\|H_1\|+\|H_2\|)_\varepsilon(\mathbf{z},\cdot)f_{\mathbf{Z}}(\cdot)_\varepsilon(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty$.

The last possible case providing non-zero $\delta(i_1,i_2,j_1,j_2)$ is $i_1 = j_2 = i$ and $j_1 = i_2 = j$. Then, we obtain

$$\delta(i,j,j,i) = \mathbb{E}\left[\{M_{ij} - \mathbb{E}[M_{ij}|\mathbf{Z}_i]\}\times\{M_{ji} - \mathbb{E}[M_{ji}|\mathbf{Z}_j]\}^T\right] = \delta(i,i,j,j),$$

due to the symmetry of M_{ij}. Thus, we have proved that $Var(n^{1/2}\Delta_n) = O(n^{-1}h^{-p}) = o(1)$, which implies

$$n^{1/2}M_n(\beta) = \frac{n^{1/2}}{2}\sum_{i=1}^n \mathbb{E}[M_n|\mathbf{Z}_i] + o_P(1).$$

We deduce that $n^{1/2}M_n(\beta)$ weakly tends towards the Gaussian random vector $\mathcal{N}(0_p, \Sigma_\beta)$ for any β. When $\beta = \beta^*$, $\partial_\beta L_\infty(\beta^*) = 0$, and this yields (i).

Second, let us deal with (ii) above. It is easy to prove that $\ddot{L}_n(\beta^*)$ tends to $\mathbb{H}(\beta^*)$ in probability, when n tends to the infinity. Indeed, the arguments are exactly the same as in Appendix B, where we have proved that $L_n(\beta^*)$ is convergent in probability. We only have to replace $\ell_\beta(\cdot,\cdot)$ by its second derivatives w.r.t. β. To save space, the specific derivations of such conditions of regularity are left to the reader: simply replace the functions $A, B, \ldots, F$ of Appendix B by their second derivatives w.r.t. β, taken at $\beta = \beta^*$, and rewrite (B.2) and (B.3).

Third, to prove (iii), it is sufficient to state that $\mathbb{E}[\|\dddot{L}_n(\beta)\|]$ is bounded from above, uniformly w.r.t. β in a small neighborhood of β^*. By derivation, we get, for every indices a, b, c in $\{1, \ldots, p'\}$,

$$\mathbb{E}[|\frac{\partial^3}{\partial\beta_a\partial\beta_b\partial\beta_c}L_n(\beta)|]$$
$$\leq Cst \times \mathbb{E}\left[|K|_h(\mathbf{Z}_i - \mathbf{Z}_j)\big\{p(\mathbf{Z}_i,\mathbf{Z}_j)H(1,\mathbf{Z}_i,\beta,a,b,c,) + (1-p(\mathbf{Z}_i,\mathbf{Z}_j))H(-1,\mathbf{Z}_i,\beta,a,b,c)\big\}\right],$$

where, for every $\delta \in \{1,-1\}$,

$$H(\delta,\mathbf{Z}_i,\beta,a,b,c) = \left(\frac{|g'(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|^3}{|1+\delta g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|^3} + \frac{|g'g''(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|}{|1+\delta g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|^2} + \frac{|g'''(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|}{|1+\delta g(\boldsymbol{\psi}(\mathbf{Z}_i)^T\beta)|}\right)|\boldsymbol{\psi}(\mathbf{Z}_i)_a\boldsymbol{\psi}(\mathbf{Z}_i)_b\boldsymbol{\psi}(\mathbf{Z}_i)_c|.$$

and Cst denotes a real constant that depend on g and (a,b,c) only. Therefore, it is sufficient to assume that

$$\int |K|(\mathbf{t})\left(p(\mathbf{z}-h\mathbf{t})H(1,\mathbf{z},\beta,a,b,c) + (1-p(\mathbf{z},\mathbf{z}-h\mathbf{t}))H(-1,\mathbf{z},\beta,a,b,c)\right)f_{\mathbf{Z}}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z}-h\mathbf{t})\,d\mathbf{t}\,d\mathbf{z} < \infty.$$

This is guaranteed by Assumption (C.4). Then, under the latter assumption, (iii) is stated and this finishes the proof. □

For convenience, let us gather the main technical assumptions that have been requested to prove Theorem 4: for some $\varepsilon > 0$,

$$\int \left(\partial_\beta\phi(\mathbf{z},\cdot,\beta)f_{\mathbf{Z}}(\cdot)\right)_\varepsilon(\mathbf{z})f_{\mathbf{z}}(\mathbf{z})\,d\mathbf{z} < \infty. \tag{C.1}$$

$$\mathbb{E}\left[\|(f_{\mathbf{Z}}(\cdot)\partial_\beta\phi(\mathbf{Z}_i,\cdot,\beta))^{(m)} + (f_{\mathbf{Z}}(\cdot)\partial_\beta\phi(\cdot,\mathbf{Z}_i,\beta))^{(m)}\|_\varepsilon(\mathbf{Z}_i)\right] < \infty. \tag{C.2}$$

$$\int (\|H_1\| + \|H_2\|)_\varepsilon(\mathbf{z},\cdot)f_{\mathbf{Z}}(\cdot)_\varepsilon(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty. \tag{C.3}$$

For every indices $(a,b,c) \in \{1,\ldots,p'\}$ and for $\mathcal{V}(\beta^*)$, some (small) neighborhood around β^*,

$$\sup_{\beta\in\mathcal{V}(\beta^*)}\int\left((p(\mathbf{z},\cdot)f_{\mathbf{Z}}(\cdot))_\varepsilon(\mathbf{z})H(1,\mathbf{z},\beta,a,b,c) + ((1-p)(\mathbf{z},\cdot)f_{\mathbf{Z}}(\cdot))_\varepsilon(\mathbf{z})H(-1,\mathbf{z},\beta,a,b,c,)\right)f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty. \tag{C.4}$$

Remark 7. Note that $\|p(\cdot,\cdot)\|_\infty \leq 1$. If g and its derivatives are bounded, Condition (C.4) is satisfied if

$$\sup_{\beta\in\mathcal{V}(\beta^*)}\sup_{\delta\in\{-1,1\}}\int \|\boldsymbol{\psi}(\mathbf{z})\|^3|1+\delta g(\boldsymbol{\psi}(\mathbf{z})^T\beta)|^{-3}f_{\mathbf{Z},\varepsilon}(\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})\,d\mathbf{z} < \infty.$$

References

Almeida, C., Czado, C., 2012. Efficient bayesian inference for stochastic time-varying copula models. Comput. Statist. Data Anal. 56, 1511–1527.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. 3 (1), 1–122.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Republished by CRC Press, Wadsworth, Belmont, CA.

Cherubini, U., Luciano, E., Vecchiato, W., 2004. Copula Methods in Finance. Wiley.

Deheuvels, P., 1979. La fonction de dependence empirique et ses proprietes, un test non parametrique d'independance. Bull. Cl. Sci. Acad. R. Belg. 65, 274–292, 5e serie.

Deheuvels, P., 1981. A Kolmogorov-Smirnov type test for independence and multivariate samples. Rev. Roumaine Math. Pures Appl. 26 (2), 213–226.

Derumigny, A., Fermanian, J.-D., 2018a. About kendall's regression. ArXiv preprint, arXiv:1802.07613.

Derumigny, A., Fermanian, J.-D., 2018b. About kernel-based estimation of the conditional Kendall's tau: finite-distance bounds and asymptotic behavior. ArXiv preprint, arXiv:1810.06234.

Fermanian, J.-D., Lopez, O., 2018. Single-index copulas. J. Multivariate Anal. 165, 27–55.

Fermanian, J.-D., Wegkamp, M., 2012. Time-dependent copulas. J. Multivariate Anal. 110, 19–29.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. vol. 1, Springer series in statistics, New York.

Geyer, C.J., 1996. On the Asymptotics of Convex Stochastic Optimization. Tech. Rep., Dept. Statistics, Univ. Minnesota.

Gijbels, I., Omelka, M., Veraverbeke, N., 2015. Partial and average copulas and association measures. Electr. J. Statist. 9, 2420–2474.

Gijbels, I., Veraverbeke, N., Omelka, M., 2011a. Conditional copulas, association measures and their applications. Comput. Statist. Data Anal. 55 (5), 1919–1932.

Gijbels, I., Veraverbeke, N., Omelka, M., 2011b. Estimation of a conditional copula and association measures. Scand. J. Stat. 38, 766–780.

Gijbels, I., Veraverbeke, N., Omelka, M., 2012. Multivariate and functional covariates and conditional copulas. Electron. J. Stat. 6, 1273–1306.
Joe, H., 2015. Dependence Modeling with Copulas. Chapman & Hall.
Jondeau, E., Rockinger, M., 2006. The copula-garch model of conditional dependencies: An international stock market application. J. Int. Money Finance 25, 827–853.
Kato, K., 2009. Asymptotics for argmin processes: Convexity arguments. J. Multivariate Anal. 100 (8), 1816–1829.
Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. Ann. Statist. 1356–1378.
Lepski, O.V., Spokoiny, V.G., 1997. Optimal pointwise adaptive methods in nonparametric estimation. Ann. Statist. 2512–2546.
Nelsen, R.B., 2007. An introduction to copulas. Springer Science & Business Media.
Parikh, N., Boyd, S., et al., 2014. Proximal algorithms. Found. Trends Optim. 1 (3), 127–239.
Patton, A., 2006a. Estimation of multivariate models for time series of possibly different lengths. J. Appl. Econometrics 21 (2), 147–173.
Patton, A., 2006b. Modelling asymmetric exchange rate dependence. Internat. Econom. Rev. 47 (2), 527–556.
Ripley, B., 2018. Tree: classification and regression trees, R package version 1.0-39.
Scott, D., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley.
So, M.K., Yeung, C.Y., 2014. Vine-copula GARCH model with dynamic conditional dependence. Comput. Statist. Data Anal. 76, 655–671.
Tsai, W.-Y., 1990. Testing the assumption of independence of truncation time and failure time. Biometrika 77 (1), 169–177.
Wurm, M.J., Rathouz, P.J., Hanlon, B.M., 2017. Regularized ordinal regression and the ordinalnet R package. ArXiv preprint, arXiv:1706.05003.