On Ranking-based Tests of Independence

Myrto Limnios
Department of Mathematical Sciences
University of Copenhagen
Copenhagen, Denmark
myli@math.ku.dk

Stephan Clémençon
Telecom Paris, LTCI
Institut Polytechnique de Paris
Palaiseau, France
stephan.clemencon@telecom-paris.fr

Abstract

In this paper we develop a novel nonparametric framework to test the independence of two random variables $\mathbf{X}$ and $\mathbf{Y}$ with unknown respective marginals $H(\mathrm{d}x)$ and $G(\mathrm{d}y)$ and joint distribution $F(\mathrm{d}x\mathrm{d}y)$, based on *Receiver Operating Characteristic* (ROC) analysis and bipartite ranking. The rationale behind our approach relies on the fact that, the independence hypothesis $\mathcal{H}_0$ is necessarily false as soon as the optimal scoring function related to the pair of distributions $(H \otimes G,\ F)$, obtained from a bipartite ranking algorithm, has a ROC curve that deviates from the main diagonal of the unit square. We consider a wide class of rank statistics encompassing many ways of deviating from the diagonal in the ROC space to build tests of independence. Beyond its great flexibility, this new method has theoretical properties that far surpass those of its competitors. Nonasymptotic bounds for the two types of testing errors are established. From an empirical perspective, the novel procedure we promote in this paper exhibits a remarkable ability to detect small departures, of various types, from the null assumption $\mathcal{H}_0$, even in high dimension, as supported by the numerical experiments presented here.

1 INTRODUCTION

Let $(\mathbf{X}_1,\ \mathbf{Y}_1),\ \ldots,\ (\mathbf{X}_N,\ \mathbf{Y}_N)$ be $N \geq 1$ independent and identically distributed (*i.i.d.*) random pairs, defined on a space $(\Omega,\ \mathcal{F},\ \mathbb{P})$ and valued in the product space $\mathcal{X} \times \mathcal{Y}$, copies of the generic random pair $(\mathbf{X},\ \mathbf{Y})$. An important problem, occurring in many applications, consists in testing the independence of the two *r.v.*'s $\mathbf{X}$ and $\mathbf{Y}$ based on the observation of the $(\mathbf{X}_i,\ \mathbf{Y}_i)$'s. It is considered here from a nonparametric perspective, meaning that no assumptions are made about the distribution $F(\mathrm{d}x\mathrm{d}y)$ of the pair $(\mathbf{X},\ \mathbf{Y})$, nor about the marginal distributions $H(\mathrm{d}x)$ and $G(\mathrm{d}y)$ of $\mathbf{X}$ and $\mathbf{Y}$. The goal is to test the *composite hypothesis*:

$$\mathcal{H}_0 : \quad F = H \otimes G \quad \text{versus} \quad \mathcal{H}_1 : \quad F \neq H \otimes G\ . \tag{1}$$

The problem thus consists in testing whether two probability distributions on the product space $\mathcal{X} \times \mathcal{Y}$ are equal or not. Under additional (parametric) assumptions on the distribution F (*e.g.* discreteness, Gaussianity), various measures of dependence can be classically used to build pivotal test statistics (*e.g.* chi-square statistic, empirical linear correlation). In the nonparametric case, most techniques consists in computing a statistical version of a (pseudo-) distance between F and $H \otimes G$ (*e.g.* integral probability metrics, see Rachev et al. (2013)). Refer to *e.g.* Székely and Rizzo (2007, 2013) for covariance-based distances, generalized to metric spaces in Lyons (2013); Jakobsen (2017). Gretton et al. (2005a,b, 2007a) introduced kernel-based extensions relying on the *Hilbert-Schmidt Independence Criterion* (HSIC), where the covariance distance being shown to be a specific instance of the class of HSIC-type measures of dependence in Sejdinovic et al. (2013). Other measures for testing independence have been recently proposed, see in particular, Berrett and Samworth (2019); Gonzalez et al. (2021) using the notion of mutual information, Gretton and Györfi (2010); Heller et al. (2016) based on partitioning techniques, and Reshef et al. (2011, 2016, 2018) considering use of the maximal information criterion.

Rank statistics for testing independence. The approach developed here, of completely different nature, is inspired by *rank-based* methods (Hájek and Sidák (1967) or Kallenberg and Ledwina (1997), Kallenberg and Ledwina (1999) or Kallenberg et al. (1997))

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

tailored to the situations where $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $\mathcal{H}_0$ is tested against specific alternatives of *positive (regression) dependence*[1]. Assuming in addition that $\mathbf{X}$ and $\mathbf{Y}$ are continuous *r.v.*'s, a natural strategy (see Kendall (1975)) consists in ranking the pairs $(\mathbf{X}_i,\ \mathbf{Y}_i)$ according to increasing values of the $\mathbf{X}_i$'s: $(\mathbf{X}_{\sigma(1)},\ \mathbf{Y}_{\sigma(1)}),\ \ldots,\ (\mathbf{X}_{\sigma(N)},\ \mathbf{Y}_{\sigma(N)})$, where σ is the permutation of the index set $\{1,\ \ldots,\ N\}$ (*i.e.* the element of the symmetric group $\mathfrak{S}_N$) *s.t.* $\mathbf{X}_{\sigma(\mathbf{1})} < \ldots < \mathbf{X}_{\sigma(\mathbf{N})}$ and analyzing the ranks of the $\mathbf{Y}_{\sigma(i)}$'s through the rank correlation coefficient, see *e.g.* Chapter 6 in Lehmann and Romano (2005): conditioned upon $(\mathbf{X}_{\sigma(1)},\ \ldots,\ \mathbf{X}_{\sigma(N)})$, the latter being uniformly random under $\mathcal{H}_0$, while the rank of $\mathbf{Y}_{\sigma(i)}$ among the $\mathbf{Y}_{\sigma(j)}$'s exhibits an 'upward trend' under the positive dependence alternative (*i.e.* it is stochastically increasing with i). The approach to independence testing based on statistical learning we propose shares similarities with such rank-based techniques, it also consists ranking pairs in $\mathcal{X} \times \mathcal{Y}$. Extension of rank-based techniques for independence testing to multivariate data has been recently the subject of much attention in the literature. The sole approach enjoying distribution-freeness under nonparametric assumptions so far, is based on the notion of center-outward ranks/signs in Hallin (2017). It is used in Shi et al. (2022) to build *generalized symmetric (test) statistics*: it boils down to plugging into classic statistics, e.g. the distance covariance measure for independence testing, a center-outward generalization of rank statistics by mapping any absolute continuous distribution to the spherical uniform distribution on the d-dimensional unit ball, solution of the related optimal transport formulation, see Hallin et al. (2021). This encompasses the main modern rank-based and distance-based methods for testing the hypothesis of independence, see *e.g.* Deb and Sen (2021); Leung and Drton (2018). While these methods have appealing theoretical properties, see Shi et al. (2022), they are limited by the strong negative impact of d of the feature on their power, studied for kernel and distance based techniques, see section 3 in Ramdas et al. (2015). As shown in Huang et al. (2023), this is caused by the dependence of the kernel of the U-statistic of degree two *w.r.t.* the dimension d.

Our contributions. The nature of our approach is quite different. It involves a preliminary statistical learning step, namely bipartite ranking on $\mathcal{X} \times \mathcal{Y}$, and relies on *Receiver Operating Characteristic* (ROC) analysis. The ROC curve is the gold standard to differentiate between two univariate distributions. The rationale behind our methodology lies in the fact that, under the null hypothesis $\mathcal{H}_0$, the *optimal* ROC curve related to the bipartite ranking defined by the pair $(H \otimes G,\ F)$ of distributions on $\mathcal{X} \times \mathcal{Y}$, is known and coincides with the main diagonal of $[0,1]^2$. It is thus natural to quantify the departure from $\mathcal{H}_0$, by the deviations of the optimal ROC curve from the diagonal. The latter can be summarized by appropriate *two-sample (linear) rank statistics*, whose concentration properties have been investigated in Clémençon et al. (2021). Since the optimal ROC curve is unknown in practice, a bipartite ranking task on the product space $\mathcal{X} \times \mathcal{Y}$ must be completed first to rank the pairs. Our method is implemented in three steps: after splitting the sample in two parts (2-split trick) and shuffling the pairs, a first part of the sample is used to train a bipartite ranking function to output a scoring function. The second part is then ranked using the scoring function previously learned so as to compute a test statistic assessing the possible departure from independence. It may be applied in a general multivariate framework and has considerable advantages in the high-dimension case, compared to all its competitors, especially those based on probability metrics between statistical versions of F and $H \otimes G$, see *e.g.* Gretton and Györfi (2010). In contrast, provided that the model bias (*i.e.*, the error inherent in the choice of the set of ranking functions over which the learning step is performed) is 'small', the power of the test proposed is possibly affected by the dimension only through the choice of bipartite ranking algorithm. This is supported here by a sound (nonasymptotic) theoretical analysis based on the concentration results for two-sample R-processes proved in Clémençon et al. (2021) and promising empirical results. Our method is shown to work well, in the vicinity of independence especially, surpassing the existing methods.

Connection to the two-sample problem. We point out that the use of (an estimate of) the optimal ROC curve, on which the novel independence testing method promoted here relies, has been recently exploited for the purpose of statistical hypothesis testing in Clémençon et al. (2023) to solve the two-sample problem, *i.e.* to test the assumption that two *i.i.d.* samples share the same distribution. The major difference naturally lies in the nature of the alternatives to the null assumption, *i.e.* departure from independence *vs.* departure from homogeneity, but also in the statistical framework/analysis: whereas independent observations drawn from each of the two distributions to be tested equal are supposedly available in the two-sample problem, no sample of the distribution $H \otimes G$ is directly available under $\mathcal{H}_1$ in independence testing. A *shuffling* procedure (*i.e.* a random permutation of parts of the indices $\{1,\ \ldots,\ N\}$), that aims at building independent observations drawn from $H \otimes G$, is key in the testing method we propose and analyze here. To summarize, for the method proposed here, new to

[1] Two real-valued *r.v.*'s X and Y defined on the same space exhibit *positive dependence iff* $\mathbb{P}(X > x,\ Y > y) \geq \mathbb{P}(X > x) \times \mathbb{P}(Y > y)$ for any $(x, y) \in \mathbb{R}^2$.

the literature, we show that: 1) the test statistic is distribution-free resulting in the exact computation of the testing threshold, 2) a nearly optimal control of the type-II error with explicit parameters can be obtained for all types of alternative, 3) the method depends on the dimension of the underlying spaces only through the bipartite ranking algorithm, importantly avoiding any mispessification of the asymptotic distribution (Huang et al. (2023)) and harmfull high-dimensional setting.

The article is organized as follows. Section 2 recalls key notions pertaining to ROC analysis and bipartite ranking, providing an insight into the rationale of the method. It is described and theoretically analyzed from a nonparametric and nonasymptotic perspective in section 3. Numerical results are displayed in section 4, while concluding remarks are collected in section 5. Due to space constraints, some properties related to ROC analysis and bipartite ranking, all technical details and proofs, as well as additional numerical experiments, are postponed to the Supplementary Material.

2 PRELIMINARIES

We first briefly recall the main concepts related to ROC analysis and bipartite ranking, involved in the methodology subsequently proposed and analyzed. The rationale behind the latter is next explained. Here and throughout, by $\mathbb{I}\{\mathcal{E}\}$ is meant the indicator function of any event $\mathcal{E}$, by δ_a the Dirac mass at any point a, by $W^{-1}(u) = \inf\{t \in (-\infty,\ +\infty]\ :\ W(t) \geq u\}$, $u \in [0,1]$ the generalized inverse of any cumulative distribution function $W(t)$ on $\mathbb{R} \cup \{+\infty\}$. The floor and ceiling functions are denoted by $u \in \mathbb{R} \mapsto \lfloor u \rfloor$ and by $u \in \mathbb{R} \mapsto \lceil u \rceil$ respectively. For any bounded function $\psi : (0,1) \to \mathbb{R}$, we also set $||\psi||_\infty = \sup_{u\in(0,1)} |\psi(u)|$. We consider *r.v.* denoted in bold symbols as valued in a multivariate space $\mathcal{Z}$, *e.g.* subset of $\mathbb{R}^d$, with $d \geq 2$.

2.1 Bipartite Ranking and ROC Analysis

We explain the connection between bipartite ranking and the quantification of the discrepancy between two probability distributions on a same space.

ROC **analysis.** The ROC curve is a gold standard to measure the difference between two *univariate* distributions, F_1 and F_2 say. It is defined by the Probability-Probability plot $t \in \mathbb{R} \mapsto (1 - F_1(t),\ 1 - F_2(t))$, connecting possible jumps by line segments by convention. It can alternatively be seen as the graph of a càd-làg (*i.e.* right-continuous and left-limited) non-decreasing mapping defined by $u \in (0,1) \mapsto \mathrm{ROC}_{\mathrm{F_1,F_2}}(\mathrm{u}) := 1 - \mathrm{F_2} \circ \mathrm{F_1^{-1}}(1-\mathrm{u})$ at points α such that $F_2 \circ F_1^{-1}(1-u) = 1-u$. The curve $\mathrm{ROC}_{\mathrm{F_1,F_2}}$ coincides with the main diagonal of $[0,1]^2$ *iff* $F_1 = F_2$. Hence, the notion of ROC curve offers a visual tool to examine the differences between two univariate distributions. For instance, the univariate distribution F_2 is stochastically larger[2] than F_1 *iff* the curve $\mathrm{ROC}_{\mathrm{F_1,F_2}}$ is everywhere above the main diagonal. Of course, the curve $\mathrm{ROC}_{\mathrm{F_1,F_2}}$is unknown in practice, just like the F_i's. Hence, ROC analysis must be based on independent *i.i.d.* samples $(X_{1,1},\ \dots,\ X_{1,n_1})$ and $(X_{2,1},\ \dots,\ X_{2,n_2})$ with distributions F_1 and F_2 respectively and consists in plotting $\mathrm{ROC}_{\hat{\mathrm{F}}_1,\hat{\mathrm{F}}_2}$, where $\hat{F}_i = (1/n_i)\sum_{k\leq n_i} \delta_{X_{i,k}}$ is the corresponding empirical counterpart of F_i with $i \in \{1,\ 2\}$. A popular scalar summary is the Area Under the ROC Curve (AUC), defined by $\mathrm{AUC(F_1,F_2)} = \int_0^1 \mathrm{ROC}_{\mathrm{F_1,F_2}}(\mathrm{u})\mathrm{du}$. Its empirical version can be expressed as an affine transform of a (two-sample linear) rank statistic, the Mann-Whitney Wilcoxon (MWW) statistic $\hat{W}_{n_1,n_2} = \sum_{k\leq n_2} R(X_{2,k})$, where the ranks $R(X_{2,k}) = \sum_{l\leq n_1} \mathbb{I}\{X_{1,l} \leq X_{2,k}\} + \sum_{l\leq n_2} \mathbb{I}\{X_{2,l} \leq X_{2,k}\}$ denotes the rank of $X_{2,k}$ among the pooled sample:

$$n_1 n_2 \mathrm{AUC}(\hat{\mathrm{F}}_1, \hat{\mathrm{F}}_2) = \hat{\mathrm{W}}_{\mathrm{n_1,n_2}} - \frac{\mathrm{n_2(n_2+1)}}{2}\ . \tag{2}$$

It is thus a distribution-free statistic (concentrated around the value 1/2) when $F_1 = F_2$, that can be naturally used to test the hypothesis of equality in distribution based on the $X_{i,k}$'s with $i \in \{1,\ 2\}$.

Bipartite ranking. Consider now two distributions F_+ and F_- on a general measurable space $\mathcal{Z}$, referred to as positive and negative distributions. Let two independent *i.i.d.* samples $\mathbf{X}_{+,1},\ \dots,\ \mathbf{X}_{+,n_+}$ and $\mathbf{X}_{-,1},\ \dots,\ \mathbf{X}_{-,n_-}$ drawn from F_+ and F_- respectively. The goal of bipartite ranking is to learn a scoring function $s : \mathcal{Z} \to (-\infty,\ \infty]$, based on the two samples, to rank any new observation without prior knowledge, by inducing a total preorder on $\mathcal{Z}$ statistically ranking the positive instances $(+)$ at the top of the resulting list compared to the negative ones $(-)$, *i.e.*, $\forall (x, x') \in \mathcal{Z}^2$, $x \preccurlyeq_s x'$ *iff* $s(x) \leq s(x')$. Let $\mathcal{S}$ be the set of all scoring functions on $\mathcal{Z}$. One evaluates the ranking performance of a candidate $s(z)$ in $\mathcal{S}$ by plotting (a statistical version of) the ROC curve $\mathrm{ROC}((\mathrm{F_{s,-}}, \mathrm{F_{s,+}}),\ \alpha) = \mathrm{ROC}(\mathrm{s}, \alpha)$, denoting by $F_{s,\epsilon}$ the pushforward distribution of F_ϵ by the mapping $s(z)$ for $\epsilon \in \{-,\ +\}$. This defines a partial preorder on $\mathcal{S}$: for all $(s_1,\ s_2)$, s_2 is more accurate than s_1 when $\mathrm{ROC}(\mathrm{s_1}, \cdot) \leq \mathrm{ROC}(\mathrm{s_2}, \cdot)$ on $[0,1]$. The most accurate scoring functions are increasing transforms of the likelihood ratio $\Psi(z) = \mathrm{d}F_{s,+}/\mathrm{d}F_{s,-}(z)$, as can be deduced from a straightforward Neyman-Pearson argument (see *e.g.* Proposition 4 in Clémençon and Vayatis (2009)): $\mathcal{S}^* =$

[2] Recall that F_2 is said to be stochastically larger than F_1 *iff* $F_1(t) \geq F_2(t)$ for all $t \in \mathbb{R}$.

$\{s \in \mathcal{S}, \forall (z,\ z') \in \mathcal{Z}^2, \Psi(z) < \Psi(z') \Rightarrow s^*(z) < s^*(z')\}$. For all $(s,\ u) \in \mathcal{S} \times (0,1)$, we have: $\mathrm{ROC(s,\ u)} \leq \mathrm{ROC^*(u)}$, where $\mathrm{ROC}^*(\cdot) = \mathrm{ROC}(\Psi,\ \cdot) = \mathrm{ROC}(\mathrm{s}^*,\ \cdot)$ for any $s^* \in \mathcal{S}^*$. The optimal curve is always concave, increasing, above the main diagonal of the ROC space consequently, *cf* Clémençon and Vayatis (2009). A key to understanding the method in section 3 is to realize that $F_+ = F_-$ *iff* ROC^* coincides with the diagonal of $[0,1]^2$, see subsection 2.2.

ROC **curve optimization.** From a quantitative perspective, bipartite ranking aims at building a scoring function $s(z)$, based on the $\mathbf{X}_{\epsilon,k}$'s with a ROC curve as close as possible to ROC^*. A typical way of measuring the deviation between these curves is to consider their distance in sup norm. As ROC^* is unknown, just like $\mathcal{S}^*$, no straightforward statistical counterpart of this loss can be computed. In Clémençon and Vayatis (2009) and Clémençon and Vayatis (2010), it is proved that bipartite ranking can be viewed as nested cost-sensitive classification tasks. By discretizing them adaptively, empirical risk minimization can be sequentially applied, with statistical guarantees in the sup-norm sense at the cost of an approximation bias. Ranking performance can be also measured by means of the L_1-norm in the ROC space: $\int_0^1 |\mathrm{ROC(s,u)} - \mathrm{ROC^*(u)}|\mathrm{du} = \mathrm{AUC}^* - \mathrm{AUC(s)}$, where $\mathrm{AUC(s)} = \mathrm{AUC(F_{s,-}, F_{s,+})}$ and $\mathrm{AUC}^* = \mathrm{AUC}(\Psi)$. The minimization of the L_1-distance to ROC^* is equivalent to the maximization of the (scalar) AUC criterion. Maximizing the latter over a class $\mathcal{S}_0 \subset \mathcal{S}$, of controlled complexity, is a popular approach to bipartite ranking, and documented in various articles. Refer to *e.g.* Agarwal et al. (2005) or Clémençon et al. (2008) for upper confidence bounds for the AUC deficit of scoring rules obtained by solving $\max_{s\in\mathcal{S}_0}\ \mathrm{AUC}(\hat{\mathrm{F}}_{\mathrm{s},-}, \hat{\mathrm{F}}_{\mathrm{s},+})$, where $\widehat{F}_{s,\epsilon} = (1/n_\epsilon)\sum_{j=1}^{n_\epsilon} \delta_{s(\mathbf{X}_{\epsilon,j})}$ for $\epsilon \in \{-,+\}$. As noticed in (2), this boils down to maximizing the rank-sum criterion: $\hat{W}_{n_-,n_+}(s) = \sum_{i=1}^{n_+} R(s(\mathbf{X}_{+,i}))$, where $R(s(\mathbf{X}_{+,i})) = N\hat{F}_{s,N}(s(\mathbf{X}_{+,i}))$ for $i \in \{1,\ \ldots,\ n_+\}$, $\hat{F}_{s,N}(t) = (1/N)\sum_{\epsilon\in\{-,+\}}\sum_{i=1}^{n_\epsilon} \mathbb{I}\{s(\mathbf{X}_{\epsilon,i}) \leq t\}$ for $t \in \mathbb{R}$ and $N = n_+ + n_-$. As expected, appropriate ranking performance criteria take the form of *(two-sample linear) rank statistics*, see Clémençon and Vayatis (2007). In Clémençon et al. (2021), the empirical ranking performance measures

$$\hat{W}^{\phi}_{n_-,n_+}(s) = \sum_{i=1}^{n_+} \phi\left(\frac{R(s(\mathbf{X}_{+,i}))}{N+1}\right)\ , \qquad (3)$$

where $\phi : [0,1] \to \mathbb{R}$ is an increasing *score-generating function* that weights the positive ranks involved the functional, are considered. For $\phi(v) = v$, one recovers the MWW statistic and the AUC criterion, see (2). If $F_{s,+} = F_{s,-}$, the ranks of the 'positive scores' are uniformly distributed. The distribution $\mathcal{L}^{\phi}_{n_-,n_+}$ of (22) is thus independent from the distributions of the $\mathbf{X}_{\epsilon,i}$'s, and can be tabulated by means of elementary combinatorial computations. When $n_+ = \lfloor pN \rfloor$ and $n_- = \lceil (1-p)N \rceil$ for $p \in (0,1)$, the statistic $(1/N)\hat{W}^{\phi}_{n_-,n_+}(s)$ can be viewed as an empirical version of W_ϕ-ranking performance:

$$W_\phi(s) = \mathbb{E}\left[(\phi \circ F_s)(s(\mathbf{X}_+))\right] = \frac{1}{p}\int_0^1 \phi(v)\mathrm{d}v$$

$$-\frac{1-p}{p}\int_0^1 \phi\left(p(1-\mathrm{ROC(s,\ \alpha)}) + (1-\mathrm{p})(1-\mathrm{u})\right)\mathrm{d}u\ , \qquad (4)$$

where $F_s = pF_{s,+} + (1-p)F_{s,-}$ for any $s \in \mathcal{S}$. For any score-generating function ϕ that rapidly vanishes near 0 and takes much higher values near 1, such as $\phi(v) = v^q$ with $q > 1$, the quantity (2.1) reflects the behavior of the curve $\mathrm{ROC(s,\ \cdot)}$ near 0, *i.e.*, the probability that $s(\mathbf{X}_+)$ takes the highest values in other words. As stated in Proposition 6 of Clémençon et al. (2021), for any $s,\ s^* \in \mathcal{S} \times \mathcal{S}^*$, we have $W_\phi(s) \leq W^*_\phi := W_\phi(\mathrm{d}F_+/\mathrm{d}F_-) = W_\phi(s^*)$. If ϕ is strictly increasing, $\mathcal{S}^*$ coincides with the ensemble of maximizers of W_ϕ. In Clémençon et al. (2021), bounds for the maximal deviations between (22) and $NW_\phi(s)$ over appropriate classes $\mathcal{S}_0$ have been proved, and generalization results for maximizers of the empirical W_ϕ-ranking performance criterion based on the latter have been established. The theoretical analysis carried out subsequently relies on these results.

2.2 On Dependence through ROC Analysis

We now go back to the problem recalled in section 1 and explain why the analysis of ROC curves and their scalar summaries (2.1) provide natural tools to test the statistical hypothesis of independence $\mathcal{H}_0$. Consider the notations introduced in section 2.1, and set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $F_- = H \otimes G$ and $F_+ = F$. Our approach relies on the observation that deviations of the curve ROC^* from the main diagonal of $[0,1]^2$, as well as those of W^*_ϕ from $\int_0^1 \phi(v)\mathrm{d}v$, for appropriate score-generating functions ϕ, provide a natural way of measuring the departure from $\mathcal{H}_0$, as revealed by the theorem below.

Theorem 1. *The following assertions are equivalent.*

(i) *The hypothesis '$\mathcal{H}_0:\ H \otimes G = F$' holds true.*

(ii) *The optimal* ROC *curve relative to the bipartite ranking problem defined by the pair $(H\otimes G, F)$ coincides with the diagonal of* $[0,1]^2$*:* $\forall u \in (0,1),\ \ \mathrm{ROC^*(u) = u}$.

(iii) *For any function $\phi(v)$, we have $W^*_\phi = \int_0^1 \phi(v)\mathrm{d}v$.*

(iv) *There exists a strictly increasing score-generating function $\phi(u)$ s.t. $W^*_\phi = \int_0^1 \phi(v)\mathrm{d}v$.*

(v) We have $\mathrm{AUC}^* = 1/2$. *In addition, we have:*

$$\mathrm{AUC}^* - \frac{1}{2} = \int\int \left| \frac{\mathrm{dF}}{\mathrm{d(H \otimes G)}}(\mathbf{x}, \mathbf{y}) - 1 \right| \mathrm{H(d\mathbf{x})G(d\mathbf{y})} \; . \tag{5}$$

Hence, the optimal curve ROC^* quantifies the dissimilarity between the $H \otimes G$ and F, as depicted by Eq. (5).

Example 1. (Multivariate Gaussian Variables) Consider a centered Gaussian r.v. $(\mathbf{X},\ \mathbf{Y})$ with definite positive covariance Γ, valued in $\mathbb{R}^q \times \mathbb{R}^l$. Denote by $\Gamma_{\mathbf{X}}$ and $\Gamma_{\mathbf{Y}}$ the (definite positive) covariance matrices of the components $\mathbf{X}$ and $\mathbf{Y}$. As an increasing transform of the likelihood ratio, the quadratic scoring function $s : z \in \mathbb{R}^{q+l} \mapsto z^t(\Gamma^{-1} - diag(\Gamma_{\mathbf{X}}^{-1}, \Gamma_{\mathbf{Y}}^{-1}))z$ is optimal. When $\mathrm{Cov}(X^1, Y^k) = \rho$, for all $k \leq l$, with $\rho \in [0, 1)$, and $\Gamma_{i,j} = \delta_{ij}$ otherwise, the hypothesis $\mathcal{H}_0$ is naturally true *iff* $\rho = 0$. The optimal ROC curve is plotted in Fig. 1 for different values of the parameter ρ, such that Γ is positive definite, and $q = l = 5$. We further refer to section 6.3 for an advanced analysis in light of the proposed method.

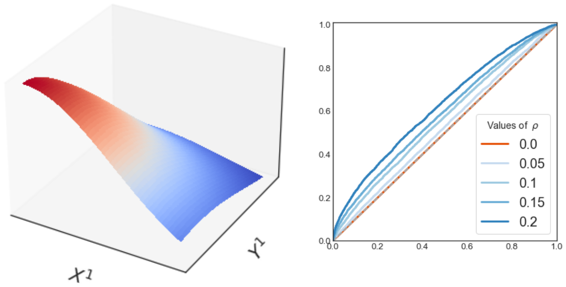


Figure 1: Left: Joint Gaussian density for $\rho = 0.20$ of (X^1, Y^1). Right: Plots of the optimal ROC curves for two Gaussian vectors with linear correlation $\rho \in \{0.0,\ 0.05,\ 0.10,\ 0.15,\ 0.20\}$ and $q = l = 5$.

Ranking-based rank tests of independence. Theorem 1 shows that the testing problem (1) can be reformulated in terms of properties of the optimal ROC curve, related to the bipartite ranking problem $(H \otimes G, F)$, as '$\mathcal{H}_0 :\ \ \mathrm{AUC}^* = 1/2$ *vs.* $\mathcal{H}_1 : \mathrm{AUC}^* > 1/2$', or, equivalently, as '$\mathcal{H}_0 :\ \ W_\phi^* = \int_0^1 \phi(u)\mathrm{d}u$ *vs.* $\mathcal{H}_1 :\ \ W_\phi^* > \int_0^1 \phi(u)\mathrm{d}u$', for any given strictly increasing score generating function $\phi(u)$. It is noteworthy that these formulations are *unilateral*, the optimal ROC curve being necessarily above the diagonal. From a practical perspective, the curve ROC^* as well as its scalar summaries, such as AUC^* or W_ϕ^*, are unknown. The approach we propose is thus implemented in three steps. After splitting the samples $\{(\mathbf{X}_1, \mathbf{Y}_1)\ \ldots,\ (\mathbf{X}_N, \mathbf{Y}_N)\}$ into two parts: 1) based on the first part, build two independent *i.i.d.* samples with respective distributions $H \otimes G$ and F, then 2) solve the corresponding bipartite ranking problem and produce a scoring function $\widehat{s}(z)$, as described above. Finally, 3) perform a univariate rank-based test based on a statistic of type (22) computed from the second part of the data, once scored using $\widehat{s}$, to detect possible statistically significant deviations between the ROC curve and the diagonal.

The subsequent sections provide both theoretical and empirical evidence that, beyond the fact that they are nearly unbiased, such testing procedures permit to detect very small departures, of various types, from the hypothesis of independence.

3 METHODOLOGY AND THEORY

We now describe at length the testing procedure previously sketched, and next establish the related theoretical guarantees by proving nonasymptotic bounds for the two types of testing error. Throughout this section, we set $F_- = H \otimes G$ and $F_+ = F$.

3.1 Ranking-based Rank Test Statistics

Following section 2, two steps are required to implement the procedure proposed. Let $n < N$ be an even integer. Hence, we use a classic two-split trick to independently divide the original *i.i.d.* sample $\{(\mathbf{X}_1, \mathbf{Y}_1)\ \ldots,\ (\mathbf{X}_N, \mathbf{Y}_N)\}$ into two:

$$\begin{aligned} \mathcal{D}_n &:= \{(\mathbf{X}_i, \mathbf{Y}_i) :\ i = 1,\ \ldots,\ n\} \\ \mathcal{D}'_{n'} &:= \{(\mathbf{X}_i, \mathbf{Y}_i) :\ i = n+1,\ \ldots,\ N\}\ , \end{aligned}$$

with $n < N$ and $n' = N - n$. Fix $p \in (0, 1)$, set $n_+ = \lfloor pn \rfloor = n - n_-$ and $n'_+ = \lfloor pn' \rfloor = n' - n'_-$. Consider two independent random variables σ and σ', defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as the $(\mathbf{X}_i, \mathbf{Y}_i)$'s and independent of the latter, uniformly distributed in $\mathfrak{S}_{n_-}$ and $\mathfrak{S}_{n'_-}$ respectively. From the first part $\mathcal{D}_n$, one considers the two samples: $\mathcal{D}_{n_-}^- = \{(\mathbf{X}_i, \mathbf{Y}_{\sigma(i)})_{1 \leq i \leq n_-}\}$, $\mathcal{D}_{n_+}^+ = \{(\mathbf{X}_i, \mathbf{Y}_i)_{1+n_- \leq i \leq n}\}$, whereas the samples below are formed from the second part $\mathcal{D}'^-_{n'_-} = \{(\mathbf{X}_i, \mathbf{Y}_{n+\sigma'(i-n)})_{1+n \leq i \leq n+n'_-}\}$, $\mathcal{D}'^+_{n'_+} = \{(\mathbf{X}_i, \mathbf{Y}_i)_{1+n+n'_- \leq i \leq N}\}$.

Proposition 1. *The following assertions hold true.*

(i) The samples $\mathcal{D}_{n_-}^-$, $\mathcal{D}_{n_+}^+$, $\mathcal{D}'^-_{n'_-}$, *and* $\mathcal{D}'^+_{n'_+}$ *are independent.*

(ii) For $\epsilon \in \{-, +\}$, $\mathcal{D}_{n_\epsilon}^\epsilon$ *and* $\mathcal{D}'^\epsilon_{n'_\epsilon}$ *are i.i.d. samples with distribution* F_ϵ.

Now that we are equipped with the two pairs of negative/positive samples constructed above, the procedure we propose requires two ingredients: a bipartite ranking algorithm $\mathcal{A}$ that permits to construct a scoring function $\hat{s} = \mathcal{A}(\mathcal{D}^-_{n_-}, \mathcal{D}^+_{n_+})$ based on the first part of the data (see the algorithms in *e.g.* Freund et al. (2003), Rakotomamonjy (2004), Rudin et al. (2005), Rudin (2006) or Burges et al. (2007)) and a strictly increasing score-generating function ϕ. As explained in section 2, the independence testing problem (1) for the couple $(H \otimes G, F)$ can be expressed as follows:

$$\mathcal{H}_0 : \; W^*_\phi = \int_0^1 \phi(u)\mathrm{d}u \;\text{ vs. }\; \mathcal{H}_1 : \; W^*_\phi > \int_0^1 \phi(u)\mathrm{d}u \; . \tag{6}$$

Notice that the formulation above is unilateral, the optimal curve ROC* being always above the first diagonal, or equivalently, the pushforward distribution of F by $\Psi(x,y)$ is always stochastically larger than that of $H \otimes G$. Relying on (6), one computes the values taken by the scoring function $\hat{s}(x,y)$ over the pooled data set $\mathcal{D}'^-_{n'_-} \cup \mathcal{D}'^+_{n'_+}$ and next the following version of the statistic (22):

$$\widehat{W}^{\phi}_{n'_-,n'_+}(\hat{s},\sigma') = \sum_{i=1+n+n'_-}^{N} \phi\left(\frac{R'_{\sigma'}(\hat{s}(\mathbf{X}_i,\mathbf{Y}_i))}{n'+1}\right) \; , \tag{7}$$

where the ranks are defined on $\mathcal{D}'_{n'}$ by $R'_{\sigma'}(t) = \sum_{i=1+n+n'_-}^{N} \mathbb{I}\{\hat{s}(\mathbf{X}_i,\mathbf{Y}_i) \le t\} + \sum_{i=1+n}^{n+n'_-} \mathbb{I}\{\hat{s}(\mathbf{X}_i,\mathbf{Y}_{n+\sigma'(i-n)}) \le t\}$. Under $\mathcal{H}_0$, the test statistic (7) has distribution $\mathcal{L}^{\phi}_{n'_-,n'_+}$, similar to the univariate rank statistic defined in (22) by Proposition 1. Fix now the desired level $\alpha \in (0,1)$ of the test of independence. Consider the $(1-\alpha)$-quantile $q^{\phi}_{n'_-,n'_+}(\alpha)$ of the pushforward distribution of $\mathcal{L}^{\phi}_{n'_-,n'_+}$, by the mapping $w \mapsto (1/n')w - \int_0^1 \phi(u)\mathrm{d}u$, depending only on ϕ, n'_+ and n'_-. Proposition 2 proves that constructing a test based on the statistic (7) and using this $(1-\alpha)$-quantile $q^{\phi}_{n'_-,n'_+}(\alpha)$ as testing threshold, has exact type-I error less than α by the bound (9). Figure 2 summarizes the procedure.

Proposition 2. (Type-I Error Bound.) *Let $\alpha \in (0,1)$ and let a scoring function $\hat{s} = \mathcal{A}(\mathcal{D}^-_{n_-}, \mathcal{D}^+_{n_+})$. The test statistic for testing (6), based on the second part of the data $\mathcal{D}'^-_{n'_-} \cup \mathcal{D}'^+_{n'_+}$, is defined by:*

$$\Phi^{\phi}_{\alpha} = \mathbb{I}\left\{\frac{1}{n'_+}\widehat{W}^{\phi}_{n'_-,n'_+}(\hat{s},\sigma') > \int_0^1 \phi(u)\mathrm{d}u + q^{\phi}_{n'_-,n'_+}(\alpha)\right\} \tag{8}$$

Under $\mathcal{H}_0$, we have for any pair of distributions $(H \otimes G, F)$ and for all $1 \le n'_- < n'$ and $1 \le n'_+ < n'$:

$$\mathbb{P}_{\mathcal{H}_0}\left\{\Phi^{\phi}_{\alpha}(\mathcal{D}'_{n'}(\hat{s})) = +1\right\} \le \alpha \; , \tag{9}$$

where $\mathcal{D}'_{n'}(s)$ denotes the dataset obtained by mapping the observations of $\mathcal{D}'_{n'}$ by any scoring function s.

The type-I error is exactly controlled, and essentially independent of the scoring function and holds true for any sample size n'.

3.2 Nonasymptotic Theoretical Guarantees Under the Alternative - Error Bound

We now investigate the theoretical properties of the test procedure previously described in the specific situation, where the bipartite ranking step is accomplished by maximizing, over a class $\mathcal{S}_0$ of scoring functions $s(x,y)$ on $\mathcal{X} \times \mathcal{Y}$, the empirical W_ϕ-ranking performance measure computed from $\mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}$:

$$\widehat{W}^{\phi}_{n_-,n_+}(s,\sigma) = \sum_{i=1+n_-}^{n} \phi\left(\frac{R_\sigma(s(\mathbf{X}_i,\mathbf{Y}_i))}{n+1}\right) \; , \tag{11}$$

where $R_\sigma(t) = \sum_{i=1+n_-}^{n} \mathbb{I}\{s(\mathbf{X}_i,\mathbf{Y}_i) \le t\} + \sum_{i=1}^{n_-} \mathbb{I}\{s(\mathbf{X}_i,\mathbf{Y}_{\sigma(i)}) \le t\}$. We thus consider

$$\hat{s} \in \operatorname*{arg\,max}_{s \in \mathcal{S}_0} \; \widehat{W}^{\phi}_{n_-,n_+}(s,\sigma) \; . \tag{12}$$

We focus on establishing a uniform nonasymptotic bound for the type-II error of the test statistic Φ^{ϕ}_{α}. It relies on the generalization properties of (12) *w.r.t.* the deficit of W_ϕ-ranking performance, investigated at length in Clémençon et al. (2021) (practical optimization issues are beyond the scope of the present paper, one may refer to Clémençon et al. (2021) for a dedicated study). The following technical assumptions are required to apply the related guarantees, and refer to the Suppl. Material for explicit definitions and details.

Assumption 1. *The score-generating function $\phi : [0,1] \mapsto \mathbb{R}$, is nondecreasing, of class $\mathcal{C}^2$.*

Assumption 2. *Let $M > 0$. For all $s \in \mathcal{S}_0$, the pushforward distributions of F and $H \otimes G$ by the mapping $s(x,y)$ are continuous, with density functions that are twice differentiable and have Sobolev $\mathcal{W}^{2,\infty}$-norms bounded by $M < +\infty$.*

Assumption 3. *The class of scoring functions $\mathcal{S}_0$ is a Vapnik-Chervonenkis (*VC*) class of finite* VC *dimension $\mathcal{V} < \infty$.*

Considering the quantity $W^*_\phi - \int_0^1 \phi(u)\mathrm{d}u$ to describe the departure from the null hypothesis $\mathcal{H}_0$ (see Theorem 1) and the bias model $W^*_\phi - \sup_{s\in\mathcal{S}_0} W_\phi(s)$ inherent in the bipartite ranking step (when formulated as empirical W_ϕ-ranking performance maximization), we introduce the two (nonparametric) classes of pairs of probability distributions on $\mathcal{X} \times \mathcal{Y}$.

Definition 1. *Let $\varepsilon > 0$. Denote by $\mathcal{H}_1(\varepsilon)$ the set of alternative hypotheses corresponding to all of probability distributions F on $\mathcal{X} \times \mathcal{Y}$ s.t. $W^*_\phi - \int_0^1 \phi(u)\mathrm{d}u \ge \varepsilon$,*

Ranking-based Independence Rank Testing

Input. Collection of $N \geq 1$ *i.i.d.* copies $\mathcal{D}_N = \{(\mathbf{X}_1, \mathbf{Y}_1)\ \ldots,\ (\mathbf{X}_N, \mathbf{Y}_N)\}$ of $(\mathbf{X}, \mathbf{Y})$; subsample sizes $n = n_+ + n_- < N$ and $n' = N - n = n'_+ + n'_-$; bipartite ranking $\mathcal{A}$ algorithm operating on the class $\mathcal{S}_0$ of scoring functions on $\mathcal{X} \times \mathcal{Y}$; score-generating function ϕ; target level $\alpha \in (0,1)$; quantile $q^{\phi}_{n'_-,n'_+}(\alpha)$.

1. **Splitting and Shuffling.** Divide the initial sample into two subsamples $\mathcal{D}_N = \mathcal{D}_n \cup \mathcal{D}'_{n'}$.
   Independently from the $(\mathbf{X}_i, \mathbf{Y}_i)'s$, draw uniformly at random two independent permutations σ and σ' in $\mathfrak{S}_{n_-}$ and $\mathfrak{S}_{n'_-}$ respectively, in order to build the independent samples: $\mathcal{D}^-_{n_-} = \{(\mathbf{X}_i, \mathbf{Y}_{\sigma(i)})_{1\leq i\leq n_-}\}$, $\mathcal{D}^+_{n_+} = \{(\mathbf{X}_i, \mathbf{Y}_i)_{1+n_-\leq i\leq n}\}$, and $\mathcal{D}'^-_{n'_-} = \{(\mathbf{X}_i, \mathbf{Y}_{n+\sigma'(i-n)})_{1+n\leq i\leq n+n'_-}\}$, $\mathcal{D}'^+_{n'_+} = \{(\mathbf{X}_i, \mathbf{Y}_i)_{1+n+n'_-\leq i\leq N}\}$.
2. **Bipartite Ranking.** Run the bipartite ranking algorithm $\mathcal{A}$ based on the pooled training dataset $\mathcal{D}_n = \mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}$ built at the previous step, in order to learn the scoring function $\hat{s} = \mathcal{A}(\mathcal{D}_n)$.
3. **Scoring and Two-sample Rank Statistic.** Build the univariate positive/negative subsamples using the scoring function $\hat{s}$ learned at the previous step $\{\hat{s}(\mathbf{X}_{n+1}, \mathbf{Y}_{n+\sigma'(1)}), \ldots, \hat{s}(\mathbf{X}_{n+n'_-}, \mathbf{Y}_{n+\sigma'(n'_-)})\}$ and $\{\hat{s}(\mathbf{X}_{n+n'_-+1}, \mathbf{Y}_{n+n'_-+1}), \ldots, \hat{s}(\mathbf{X}_N, \mathbf{Y}_N)\}$. Sort them by decreasing order to compute

$$\widehat{W}^{\phi}_{n'_-,n'_+}(\hat{s}, \sigma') = \sum_{i=1+n+n'_-}^{N} \phi\left(\frac{R'_{\sigma'}(\hat{s}(\mathbf{X}_i, \mathbf{Y}_i))}{n'+1}\right) . \tag{10}$$

Output. Compute the outcome of the test of level α based on the test statistic (10), *i.e.*, accept $\mathcal{H}_0$ if:

$$\frac{1}{n'_+}\widehat{W}^{\phi}_{n'_-,n'_+}(\hat{s}, \sigma') \leq \int_0^1 \phi(u)\mathrm{d}u + q^{\phi}_{n'_-,n'_+}(\alpha) \ , \quad \text{and reject it otherwise.}$$

Figure 2: Ranking-based independence rank test.

where we recall $W^*_\phi = W_\phi(s^*) = W^*_\phi(\mathrm{d}F/\mathrm{d}(H \otimes G))$ for any $s^* \in \mathcal{S}^*$.

Definition 2. *Let $\delta > 0$, $\mathcal{S}_0 \subset \mathcal{S}$. We denote by $\mathcal{B}(\delta)$ the set of all pairs $(H \otimes G, F)$ of probability distributions on $\mathcal{X} \times \mathcal{Y}$ such that $W^*_\phi - \sup_{s\in\mathcal{S}_0} W_\phi(s) \leq \delta$.*

The theorem below provides a rate bound for the type-II error of the ranking-based rank test (8) of size α. It depends on the sample sizes n used for bipartite ranking, and on $n' = N - n$ for performing the rank test based on the learned scoring function, see Fig. 2.

Theorem 2. (Type-II error bound.) *Let $\phi(u)$ be a score-generating function and $\varepsilon > \delta > 0$. Let σ, σ' two independent permutations drawn resp. from $\mathfrak{S}_{n_-}$ and $\mathfrak{S}_{n'_-}$, independent of the $\mathbf{X}_i$s, $\mathbf{Y}_j$s. Fix $\alpha \in (0,1)$. Suppose that Assumptions 1-3 are fulfilled. Let $p \in (0,1)$ such that $n \wedge n' \geq 1/p$. Set $n_+ = \lfloor pn \rfloor$ and $n_- = \lceil (1-p)n \rceil = n - n_+$, as well as $n'_+ = \lfloor pn' \rfloor$ and $n'_- = \lceil (1-p)n' \rceil = n' - n'_+$. Then, there exist constants C_1 and $C_2 \geq 24$, depending on $(\phi,\ \mathcal{V})$, such that the type-II error of the test (8) is uniformly bounded:*

$$\sup_{\substack{(H\otimes G,F)\\ \in\mathcal{H}_1(\varepsilon)\cap\mathcal{B}(\delta)}} \mathbb{P}_{\mathcal{H}_1}\left\{\Phi^{\phi}_{\alpha} = 0\right\} \leq 18\exp\left(-\frac{Cn'(\varepsilon-\delta)^2}{16}\right) + C_2\left(1 + \frac{\varepsilon-\delta}{32C_1\kappa_p}\right)^{-np\kappa_p(\varepsilon-\delta)/(8C_2)} \tag{13}$$

as soon as $n' \geq 4\log(18/\alpha)/(C(\varepsilon-\delta)^2)$ and $n \geq 16C_1^2/(p(\varepsilon-\delta)^2)$, with constants $\kappa_p = p \wedge (1-p)$, $C = 8^{-1}\min\left(p/\|\phi\|^2_\infty, (p\|\phi'\|^2_\infty)^{-1}, ((1-p)\|\phi'\|^2_\infty)^{-1}\right)$, the C_j's are explicitly detailed in the proof.

The first term results from the control of the type-II error of a univariate rank statistic. The second term relies on Theorem 5 established in Clémençon et al. (2021), inherited from the learning stage of the scoring function. If the bias δ induced by the learning step is guaranteed to be smaller that the departure ε from $\mathcal{H}_0$, such that $\varepsilon - \delta > 0$, and if this quantity is kept fixed, then both terms in (13) converge to zero when both $n,\ n' \to \infty$. Importantly, the error rate related to the hypothesis test is *independent* on the dimensions of the spaces $\mathcal{X}$ and $\mathcal{Y}$. The only term dependent on those dimensions comes from the learning step, through the choice of bipartite ranking related to the class of scoring functions $\mathcal{S}_0$. Precisely, only the constants C_1 and C_2 depend on the dimensions of $\mathcal{X}$ and $\mathcal{Y}$ as inherited by the VC dimension $\mathcal{V}$ of $\mathcal{S}_0$. We illustrate this bound and its parameters (ε, δ) in the context of Example 1 in the Suppl. Material.

This result is important and new to the literature for testing independence under nonparametric alternatives. It is, to the best of our knowledge, the first finite sample probabilistic uniform control of the type-II error. The power of test statistics from the literature comparatively suffers from the underlying dimensions, see Ramdas et al. (2015). The estimator of those statistic indeed take the form of U-statistics based on multivariate observations, for which it has been proved to be subject to misspecification of the asymptotic distribution under nonparametric alternatives, see Huang et al. (2023). Hence, our proposed method circumvents this limitation by computing the test statistic based on univariate samples that are mapped thanks to the scoring function solution of the bipartite ranking problem.

4 NUMERICAL EXPERIMENTS

This section presents the empirical performance of our proposed method (Fig. 2), by illustrating the theoretical testing guarantees of section 3 through: 1) high-dimensional settings and non-monotonic class of alternatives, and 2) application to fair learning by testing for *statistical parity* based on real data published in Jesus et al. (2022). We mainly consider synthetic datasets to exactly control the departure from independence. We refer to the Supp. Material, Section 8 for details on the implementation and additional experiments. These experiments can be reproduced using the Python code available at `https://github.com/MyrtoLimnios/independence_ranktest`.

Ranking-based independence rank tests. We implemented the Ranking Forest algorithm (`rForest`, Clémençon et al. (2013)) to solve *Step 2*, following the empirical results in Clémençon et al. (2023). We selected two score-generating functions to compute the rank statistic (8) for *Step 3*: $\phi(u) = u$ ($\texttt{rForest}_{MWW}$, Wilcoxon (1945)) and $\phi(u) = u\mathbb{I}\{u \geq u_0\}$ with $u_0 \in \{0.85,\ 0.90,\ 0.95\}$ ($\texttt{rForest}_{u_0}$, Clémençon and Vayatis (2007)) considering only the $1 - u_0$ higher ranks in the computation of the statistic corresponding to the beginning of the ROC curve.

Evaluation criteria and experimental parameters. Once all methods are calibrated for the range of significance levels $\alpha \in (0, 1)$, we compare the graphs of the the rate of rejecting $\mathcal{H}_0$ under $\mathcal{H}_1$, and also at fixed $\alpha = 0.05$ exposed in tables in the Supp. Material. These criteria are computed over $B = 100$ Monte-Carlo samplings, with 95% confidence interval, and plotted against the *dependence* parameter $\rho \in \mathbb{R}$, as function of the *departure* level $\varepsilon \in (0, 1)$, see Def. 1.

Probabilistic model and experimental parameters. We continue on Ex. 1 motivated by the results in Huang et al. (2023) refered to as model (GL). Consider $(\mathbf{X},\ \mathbf{Y}) \sim \mathcal{N}(e_d, \Gamma_\rho)$, where $e_d \in \mathbb{R}^d$ the null vector, $\mathrm{Cov}(X^1, Y^k) = \rho$, for all $k \leq l$ and $\Gamma_{\rho,i,j} = \delta_{ij}$ otherwise. We implement model (M1) for non-monotonic set of alternatives, wherein $X^1 = \rho \cos\Theta + \omega_1/4$, $Y^1 = \rho \sin\Theta + \omega_2/4$, with $\rho \in \{1, 2, 3\}$, $\omega_i \sim \mathcal{N}(0, 1)$, $i \in \{1, 2\}$, and $\Theta \sim \mathcal{U}([0, 2\pi])$ all variables being independent, and with $d \in \{4, 10, 26\}$, $N \in \{500, 2000\}$. (M1) is extended for high dimension to both a sparse (M1s) and dense (M1d) models, see the Supp. Material, Section 8 therein. The number of random permutations for our procedure is $K_p \in \{10, 50\}$ under $\mathcal{H}_1$. The pooled sample size N is fixed, with $n = 4N/5$ and $n' = N/5$, and set $q = l = d/2$.

Benchmark tests. We implement two state-of-the-art multivariate and nonparametric tests, namely the unbiased estimator of the Hilbert-Schmidt Independence Criterion (`HSIC`, Gretton et al. (2007b)), with the recommended Gaussian kernel with bandwidth the median heuristic of the distance between the points in the merged sample (*e.g.* Gretton et al. (2012)), and the centered estimator of the Distance correlation computed with either the L_1 or the L_2 distances ($\texttt{dCor}_{L_1}$, $\texttt{dCor}_{L_2}$, Székely and Rizzo (2007)). These methods require an additional implementation to estimate the null quantile, *e.g.* done by a permutation procedure. Due to their high computational complexity ($\mathcal{O}(N!)$), we restricted to a fixed number of permutations $K_0 = 200$.

Results and discussion. We focus on the ability of the ranking-based method to reject $\mathcal{H}_0$ for small dependence ρ and for increasing dimension d, depending on the choice of $\phi(u)$. First, the proposed method is distribution free under $\mathcal{H}_0$ for any bipartite ranking algorithm, hence its calibration only depends on n'_-, n'_+, ϕ and α. State-of-the-art (SoA) methods do not have this advantage in comparison. Other procedures than the implemented permutation-based one, approximate the asymptotic null distribution of the related statistics, namely using the Gamma distribution for the `HSIC`, see Gretton et al. (2007b) Section 3. However, this method is proved to be subject to misspecification under nonparametric assumptions, as proved in Huang et al. (2023), resulting in false estimation of the testing threshold and thus incorrect p-values. Notice that, for the proposed ranking-based tests, the number of random permutations K_p required to estimate the product of the marginal distributions, is lower than that for the estimation of the SoA's null threshold: we propose to only sample from both $\mathfrak{S}_{n_-}$ and $\mathfrak{S}_{n'_-}$, compared to $\mathfrak{S}_N$. The experiments show that for a well calibrated ranking-method, one achieves high empirical power with minimal number of permutations ($K_p \in \{10, 20, 50\}$), see Fig. 1 and 4. For small sample sizes, `RTB` is not competitive as it has lower power for increasing u_0: fewer observations are considered and yielding larger variance for the estimation of the rank statistic. RTB has, however, experimentally showed higher accuracy for estimating the beginning of the true ROC curve (ROC*) in Clémençon et al. (2021). RTB also achieves competitive rejection rates to MWW for larger N (Fig. 2), and to SoA for models (GL, M1), see Fig. 1. When the dimension increases $d \in \{N/10, N/5, N/2\}$, fixing N, the performance of MWW remains high, *e.g.* Fig. 1. Notice that the data generating processes are designed not to suffer from signal-to-noise low ratio, for high dimension d especially. However, there is a clear difference in the performances depending on the range of that ratio: the sparser the signal is and the smaller the rejection rates are for the SoA methods. For (GL), see Fig. 4 ($d = 4$) and 5 ($d = 10$) especially, wherein the ratio equals to *resp.* 1/4 and 1/10, rForest exhibits higher power for lower departures from $\mathcal{H}_0$, see also (M1s) Fig. 3 and plots 6, 7. On the contrary, for denser models, *e.g.* (M1d) Fig. 4, SoA methods have similar performances with MWW, however RTB shows no power for small departures ρ. The randomization related to rForest increases the chances to select important information, whereas for dense models, it might be ignored, see Clémençon et al. (2013) for further empirical analysis. Lastly, both HSIC and $\texttt{dCor}_{L2}$ show similar experimental performances as expected, see Sejdinovic et al. (2013). To conclude, for all ϕ and d, the rejection rates of the ranking-based tests increase with the departure ε. They empirically outperform the comparative SoA tests, studied for non-monotonic and sparse high dimensional models.

Interpretation of the null assumption rejection. We recall that certain bipartite ranking algorithms, such as those proposed in Clémençon et al. (2011) or Clémençon et al. (2013), produce scoring functions that can be interpreted to a certain extent. As explained in section 5 of Clémençon et al. (2011), the *relative importance* of each component of the argument (X, Y) of a scoring function $s(x, y)$ defined by a 'ranking tree' (or by a 'forest of ranking trees') can be easily quantified. When applied to the testing problem considered here, this interpretability tool may permit to identify the components mainly responsible for the departure from the independence assumption (or equivalently the departure of the ROC curve from the diagonal) possibly assessed from the data by means of the methodology we promote. We further refer to similar discussions on interpretability of the learned decision rule in the context of two-sample testing, when formulated as a classification learning problem in *e.g.* Lopez-Paz and Oquab (2017); Kübler et al. (2022).

Real data experiment: testing for statistical parity. In the context of 'responsible' statistical prediction, a significant number of works have studied *fair* statistical methods, aiming to be unbiased/fair *wrt.* *protected* attributes/subgroups considered as sensitive. In particular, *Statistical parity* is achieved when a decision rule producing a set of outcomes $\mathbf{X}$ based on an ensemble of covariates $\mathbf{Z}$, is independent of a set of protected attributes $\mathbf{Y}$. We propose to test for statistical parity formulated as a test for independence between $\mathbf{X}$ and $\mathbf{Y}$ as in (1). If $\mathbf{X}$ is univariate and discrete, typical methods in fairness propose to learn a classification model to predict $\mathbf{X}$, wherein both ($\mathbf{X}$, $\mathbf{Z}$) are used, and then to measure or test for statistical independence between the predicted $\mathbf{X}$ and the protected attributes $\mathbf{Y}$, see *e.g.* Fermanian and Guegan (2021). We propose to apply our proposed method in that context, to assess whether a typical algorithm learns to predict the outcome under statistical parity, when both outcomes $\mathbf{X}$ and protected variables $\mathbf{Y}$ are continuous and valued in spaces of possibly dimensions $q, l > 1$. We use the synthetic Bank Account Fraud (BAF) dataset developed by Jesus et al. (2022), and generated from real datasets of frauds in anonymized bank account openings. BAF has 31 explanatory variables plus one indicating the possible occurence of fraud. It has unbalanced representation of frauds and all features can be modeled as continuous observations. We selected three potentially protected variables related to the personal identity of the clients, namely the age of the client (Age), an indicator level of similarity between the name of the client and personal email address (Name), and the number of emails received for applicants with same date of birth four weeks prior to fraud (Date). We gather the distributions of the empirical p-values in Fig. 8, based on a 5-fold cross-validation. For each fold, a Random Forest algorithm is trained to predict the probability of Fraud $\mathbf{X}$, and our ranking-based procedure (Fig. 2) is used to estimate the associate p-value of (1). We subsampled at random from the original data set $N = 10^3$ while keeping the proportion of Fraud from the original dataset fixed. This plot shows that we cannot reject at level $\alpha = 0.05$ the statistical independence between the predicted probability of fraud and the protected variables $\mathbf{Y}$.

5 CONCLUSION

We have proposed a novel approach, involving a preliminary bipartite ranking stage, to test independence between random variables in a nonparametric and possibly high-dimensional setting. Nonasymptotic error bounds have been established for this method, and its theoretical optimality properties are confirmed by numerical experiments, showing that it generally detects small departures from the independence much better than its competitors and resists to the high dimension especially in sparse settings.

Acknowledgements

We thank the reviewers for the useful comments. Myrto Limnios was supported by Novo Nordisk Foundation Grant NNF20OC0062897.

References

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. (2005). Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425.

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022). Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2):858 – 879.

Berrett, T. B. and Samworth, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566.

Burges, C., Ragno, R., and Le, Q. (2007). Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Clémençon, S., Depecker, M., and Vayatis, N. (2011). Adaptive partitioning schemes for bipartite ranking. *Machine Learning*, 43(1):31–69.

Clémençon, S., Depecker, M., and Vayatis, N. (2013). An empirical comparison of learning algorithms for nonparametric scoring: the treerank algorithm and other methods. *Pattern Analysis and its Applications*, 16(4):475–496.

Clémençon, S., Depecker, M., and Vayatis, N. (2013). Ranking Forests. *Journal of Machine Learning Research*, 14:39–73.

Clémençon, S., Limnios, M., and Vayatis, N. (2021). Concentration inequalities for two-sample rank processes with application to bipartite ranking. *Electronic Journal of Statistics*, 15(2):4659 – 4717.

Clémençon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874.

Clémençon, S. and Vayatis, N. (2007). Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699.

Clémençon, S. and Vayatis, N. (2009). Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336.

Clémençon, S. and Vayatis, N. (2010). Overlaying classifiers: a practical approach to optimal scoring. *Constructive Approximation*, 32(3):619–648.

Clémençon, S., Limnios, M., and Vayatis, N. (2023). A bipartite ranking approach to the two-sample problem. *arXiv:2302.03592.*

Deb, N. and Sen, B. (2021). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, pages 1–16.

Fermanian, J.-D. and Guegan, D. (2021). Fair learning with bagging. Documents de travail du Centre d'Économie de la Sorbonne 2021.34 - ISSN : 1955-611X.

Freund, Y., Iyer, R. D., Schapire, R. E., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969.

Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 37(4):503–522.

Gonzalez, M. E., Silva, J. F., Videla, M., and Orchard, M. E. (2021). Data-driven representations for testing independence: Modeling, analysis and connection with mutual information estimation. *IEEE Transactions on Signal Processing*, 70:158–173.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B., and Smola, A. (2007a). A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B., and Smola, A. (2012). A kernel two-sample problem. *Journal of Machine Learning Research*, 13:723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a). Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory: 16th International Conference,ALT 2005*, volume 3734, pages 63–78.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007b). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Gretton, A. and Györfi, L. (2010). Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11(46):1391–1423.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. (2005b). Kernel constrained covariance for dependence measurement. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5, pages 112–119. Proceedings of Machine Learning Research.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707.

Hájek, J. and Sidák, Z. (1967). *Theory of Rank Tests.* Academic Press.

Hallin, M. (2017). On Distribution and Quantile Functions, Ranks and Signs in R_d. *Working Papers ECARES.*

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165.

Heller, R., Heller, Y., Kaufman, S., Brill, B., and Gorfine, M. (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17:1–54.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Huang, K. H., Liu, X., Duncan, A. B., and Gandy, A. (2023). A high-dimensional convergence theorem for u-statistics with applications to kernel-based testing. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3827–3918. Proceedings of Machine Learning Research.

Jakobsen, M. E. (2017). Distance covariance in metric spaces: Non-parametric independence testing in metric spaces (master's thesis).

Jesus, S., Pombal, J., Alves, D., Cruz, A., Saleiro, P., Ribeiro, R. P., Gama, J., and Bizarro, P. (2022). Turning the tables: Biased, imbalanced, dynamic tabular datasets for ML evaluation. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*

Kallenberg, W., Ledwina, T., and Rafajlowicz, E. (1997). Testing bivariate independence and normality. *Sankhya. Series A*, 59(1):42–59.

Kallenberg, W. C. M. and Ledwina, T. (1997). Data driven rank tests for independence. In *Proceedings of the 51th session of the International Statistical Institute*, number book 2 in Bulletin of the International Statistical Institute, pages 511–512, Netherlands. International Statistical Institute.

Kallenberg, W. C. M. and Ledwina, T. (1999). Data driven rank tests for independence. *Journal of the American Statistical Association*, 94(445):285–301.

Kendall, M. (1975). *Rank Correlation Methods.* 4th Edition, Charles Griffin, London.

Khavari, B. and Rabusseau, G. (2021). Lower and upper bounds on the pseudo-dimension of tensor network models. In *Advances in Neural Information Processing Systems*, volume 34, pages 10931–10943. Curran Associates, Inc.

Kübler, J. M., Jitkrittum, W., Schölkopf, B., and Muandet, K. (2022). A witness two-sample test. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1403–1419. PMLR.

Lehmann, E. L. and Romano, J. M. (2005). *Testing Statistical Hypotheses.* 3third Edition, Springer.

Leung, D. and Drton, M. (2018). Testing independence in high dimensions with sums of rank correlations. *The Annals of Statistics*, 46(1):280 – 307.

Lopez-Paz, D. and Oquab, M. (2017). Revisiting classifier two-sample tests. In *International Conference on Learning Representations.*

Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284 – 3305.

Nolan, D. and Pollard, D. (1987). U-Processes: Rates of Convergence. *The Annals of Statistics*, 15(2):780 – 799.

Rachev, S. T., Klebanov, L. B., Stoyanov, S. V., and Fabozzi, F. (2013). *The Methods of Distances in the Theory of Probability and Statistics.* Springer.

Rakotomamonjy, A. (2004). Optimizing area under roc curve with svms. In *Proceedings of the First Workshop on ROC Analysis in AI.*

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 3571–3577. AAAI Press.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524.

Reshef, D. N., Reshef, Y. A., Sabeti, P. C., and Mitzenmacher, M. (2018). An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics*, 12(1):123 – 155.

Reshef, Y. A., Reshef, D. N., Finucane, H. K., Sabeti, P. C., and Mitzenmacher, M. (2016). Measuring dependence powerfully and equitably. *Journal of Machine Learning Research*, 17(211):1–63.

Rudin, C. (2006). Ranking with a P-Norm Push. In *Proceedings of COLT 2006*, volume 4005 of *Lecture Notes in Computer Science*, pages 589–604.

Rudin, C., Cortes, C., Mohri, M., and Schapire, R. E. (2005). Margin-based ranking and boosting meet in the middle. In *Proceedings of COLT 2005*, volume 3559 of *Lecture Notes in Computer Science*, pages 63–78. Springer.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263 – 2291.

Shi, H., Hallin, M., Drton, M., and Han, F. (2022). On universally consistent and fully distribution-free rank tests of vector independence. *The Annals of Statistics*, 50(4):1933 – 1959.

Székely, G. J. and Rizzo, M. L. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag New York.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80–83.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, details are given in the repository]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, in the main, supplementary materials and URL]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable, no need for particular computational infrastructure]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

On Ranking-based Tests of Independence: Supplementary Materials

This material supplements the article *On Ranking-based Tests of Independence.* Section 6 gathers additional properties on ROC analysis, and on the assumptions required to prove the main Theorem 2, as well as developing Example 1. In Section 7, we derive the detailed proofs for all theoretical results stated in the main corpus of the article. Finally, section 8 presents additional numerical experiments on synthetic data. Importantly, we prove the nonasymptotic control of both type-I and type-II errors formulated as concentration inequalities and show empirical evidence for the competitiveness of our proposed method. We first recall the proposed ranking-based rank test of independence procedure in Figure 3 for the sake of clarity.

Ranking-based Independence Testing

Input. Collection of $N \geq 1$ *i.i.d.* copies $\mathcal{D}_N = \{(\mathbf{X}_1, \mathbf{Y}_1)\ \ldots,\ (\mathbf{X}_N, \mathbf{Y}_N)\}$ of $(\mathbf{X}, \mathbf{Y})$; subsample sizes $n = n_+ + n_- < N$ and $n' = N - n = n'_+ + n'_-$; bipartite ranking $\mathcal{A}$ algorithm operating on the class $\mathcal{S}$ of scoring functions on $\mathcal{X} \times \mathcal{Y}$; score-generating function ϕ; target level $\alpha \in (0, 1)$; quantile $q^{\phi}_{n'_-, n'_+}(\alpha)$.

1. **Splitting and Shuffling.** Divide the initial sample into two subsamples $\mathcal{D}_N = \mathcal{D}_n \cup \mathcal{D}'_{n'}$.

   Independently from the $(\mathbf{X}_i, \mathbf{Y}_i)'s$, draw uniformly at random two independent permutations σ and σ' in $\mathfrak{S}_{n_-}$ and $\mathfrak{S}_{n'_-}$ respectively, in order to build the independent samples: $\mathcal{D}^-_{n_-} = \{(\mathbf{X}_i, \mathbf{Y}_{\sigma(i)})_{1 \leq i \leq n_-}\}$, $\mathcal{D}^+_{n_+} = \{(\mathbf{X}_i, \mathbf{Y}_i)_{1+n_- \leq i \leq n}\}$, and $\mathcal{D}'^-_{n'_-} = \{(\mathbf{X}_i, \mathbf{Y}_{n+\sigma'(i-n)})_{1+n \leq i \leq n+n'_-}\}$, $\mathcal{D}'^+_{n'_+} = \{(\mathbf{X}_i, \mathbf{Y}_i)_{1+n+n'_- \leq i \leq N}\}$.

2. **Bipartite Ranking.** Run the bipartite ranking algorithm $\mathcal{A}$ based on the pooled training dataset $\mathcal{D}_n = \mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}$ built at the previous step, in order to learn the scoring function
$$\hat{s} = \mathcal{A}(\mathcal{D}_n)\ . \tag{14}$$

3. **Scoring and Two-sample Rank Statistic.** Build the univariate positive/negative subsamples using the scoring function $\hat{s}$ learned at the previous step $\{\hat{s}(\mathbf{X}_{n+1}, \mathbf{Y}_{n+\sigma'(1)}), \ldots, \hat{s}(\mathbf{X}_{n+n'_-}, \mathbf{Y}_{n+\sigma'(n'_-)})\}$ and $\{\hat{s}(\mathbf{X}_{n+n'_-+1}, \mathbf{Y}_{n+n'_-+1}), \ldots, \hat{s}(\mathbf{X}_N, \mathbf{Y}_N)\}$. Sort them by decreasing order to compute
$$\widehat{W}^{\phi}_{n'_-, n'_+}(\hat{s}, \sigma') = \sum_{i=1+n+n'_-}^{N} \phi\left(\frac{R'_{\sigma'}(\hat{s}(\mathbf{X}_i, \mathbf{Y}_i))}{n'+1}\right)\ , \tag{15}$$
   where $R'_{\sigma'}(t) = \sum_{i=1+n+n'_-}^{N} \mathbb{I}\{\hat{s}(\mathbf{X}_i, \mathbf{Y}_i) \leq t\} + \sum_{i=1+n}^{n+n'_-} \mathbb{I}\{\hat{s}(\mathbf{X}_i, \mathbf{Y}_{n+\sigma'(i-n)}) \leq t\}$.

Output. Compute the outcome of the test of level α based on the test statistic (10): accept the hypothesis $\mathcal{H}_0$ of independence if:
$$\frac{1}{n'_+}\widehat{W}^{\phi}_{n'_-, n'_+}(\hat{s}, \sigma') \leq \int_0^1 \phi(u)\mathrm{d}u + q^{\phi}_{n'_-, n'_+}(\alpha)\ , \quad \text{and reject it otherwise.}$$

Figure 3: Ranking-based independence testing.

6 PRELIMINARIES

This subsection gathers additional definitions and properties important to the main corpus. We first expose results related to ROC analysis. Then, we provide complementary material to Assumptions (1-3) required to prove the main Theorem 2, deriving the nonasymptotic uniform bound of the type-II error of the test statistic based on Eq. (10). We consider same notations as in the main corpus of the article.

6.1 ROC analysis

Lemma 3. *(Clémençon and Vayatis (2009)) Let* $\mathbf{Z}$ *denote either* $\mathbf{X}_+$ *or* $\mathbf{X}_-$*, and define the likelihood ratio* $\Psi(z) = \mathrm{d}F_+/\mathrm{d}F_-(z)$*. The property below holds true a.s.*

$$\Psi(\mathbf{Z}) = \frac{\mathrm{d}F_{\Psi,+}}{\mathrm{d}F_{\Psi,-}}(\Psi(\mathbf{Z})) \ ,$$

where $F_{\Psi,+}$ *(resp.* $F_{\Psi,-}$*) is the pushforward distribution* F_+ *(resp.* F_-*) by the likelihood ratio.*

Proposition 3. *(Clémençon and Vayatis (2009)) For any probability distributions* F_+ *and* F_-*, and any scoring function* $s : \mathcal{Z} \to \mathbb{R}$*, the following assertions hold true.*

(i) $\mathrm{ROC}(\mathrm{s}, 0) = 0$ *and* $\mathrm{ROC}(\mathrm{s}, 1) = 1$.

(ii) The ROC *curve is invariant by any nondecreasing transform* $c : \mathbb{R} \to \mathbb{R}$ *of a scoring function* $s(z)$ *on* $(0, 1)$*:* $\mathrm{ROC}(\mathrm{c} \circ \mathrm{s}, \cdot) = \mathrm{ROC}(\mathrm{s}, \cdot)$.

(iii) Let a scoring function $s(z)$*. Suppose both distributions* F_+ *and* F_- *are continuous. Then, the associated* ROC *curve of the function* $s(z)$ *is differentiable iff. the pushforward distributions* $F_{s,+}$ *and* $F_{s,-}$ *are continuous.*

6.2 Sobolev and VC-classes of functions

Sobolev space of functions. Assumption 2 requires that for all $s \in \mathcal{S}_0$, the pushforward distributions of F and $H \otimes G$ by the mapping $s(x, y)$ are continuous, with density functions that are twice differentiable and have Sobolev $\mathcal{W}^{2,\infty}$-norms bounded by a finite constant $M > 0$.

We recall that the Sobolev space $\mathcal{W}^{2,\infty}$ is composed of all Borelian functions $f : \mathbb{R} \to \mathbb{R}$, such that f and its first and second order weak derivatives f' and f'' are bounded almost-everywhere. It is a Banach space when equipped with the norm $||f||_{2,\infty} = \max\{||f||_\infty,\ ||f'||_\infty,\ ||f''||_\infty\}$, where $||.||_\infty$ is the norm of the Lebesgue space L_∞ of Borelian and essentially bounded functions.

VC-type classes of functions. We recall below the definition of VC-type class of functions formulated in Assumption 3. We further refer to van der Vaart and Wellner (1996), Chapter 2.6. therein, for additional generalizations, details and examples.

Definition 3. *A class* $\mathcal{F}$ *of real-valued functions defined on a measurable space* $\mathcal{Z}$ *is a bounded VC-type class with parameter* $(A, \mathcal{V}) \in (0,\ +\infty)^2$ *and constant envelope* $L_{\mathcal{F}} > 0$ *if for all* $\varepsilon \in (0, 1)$*:*

$$\sup_Q N(\mathcal{F}, L_2(Q), \varepsilon L_{\mathcal{F}}) \leq \left(\frac{A}{\varepsilon}\right)^{\mathcal{V}} \ , \tag{16}$$

where the supremum is taken over all probability measures Q *on* $\mathcal{Z}$ *and the smallest number of* $L_2(Q)$*-balls of radius less than* ε *required to cover class* $\mathcal{F}$ *(i.e. covering number) is meant by* $N(\mathcal{F}, L_2(Q), \varepsilon)$*.*

In particular, a bounded VC class of functions with finite VC dimension V is of VC-type, with $\mathcal{V} = 2(V - 1)$ and $A = (cV(16e)^V)^{1/(2(V-1))}$, where c is a universal constant, see *e.g.* van der Vaart and Wellner (1996), Theorem 2.6.7 therein.

6.3 Multivariate Gaussian framework - Example 1 continued

This section extends Example 1, *i.e.*, for testing independence between two multivariate Gaussian *r.v.*. We focus on deriving the explicit constants appearing in the bound that are related to: (i) the testing problem through the departure from the null $\varepsilon > 0$, and the bias $\delta > 0$, and (ii) the complexity of the selected class of scoring functions $\mathcal{S}_0$ (Assumption 3).

Framework and procedure. Consider a centered Gaussian *r.v.* $(\mathbf{X},\ \mathbf{Y})$ with definite positive covariance Γ, valued in $\mathbb{R}^{q+l}$. Denote by $\Gamma_{\mathbf{X}}$ and $\Gamma_{\mathbf{Y}}$ the (definite positive) covariance matrices of the components $\mathbf{X}$ and $\mathbf{Y}$. The oracle class of scoring functions $\mathcal{S}^*$ is composed of the increasing transforms of the likelihood ratio, taking the form of the quadratic scoring function:

$$s : z \in \mathbb{R}^{q+l} \mapsto z^t(\Gamma^{-1} - \text{diag}(\Gamma_{\mathbf{X}}^{-1}, \Gamma_{\mathbf{Y}}^{-1}))z\ .$$

and define $\theta^* = \Gamma^{-1} - \text{diag}(\Gamma_{\mathbf{X}}^{-1}, \Gamma_{\mathbf{Y}}^{-1})$. Following the procedure summarized in Figure 3, we thus propose to solve *Step 2* by learning the optimal scoring function in the class:

$$\mathcal{S}_0(\Theta) = \{s_\theta : z \in \mathbb{R}^{q+l} \mapsto z^t\theta z, \quad \theta \in \Theta\}\ ,$$

where Θ is a subset of real definite positive matrices of size $\mathbb{R}^{(q+l)\times(q+l)}$. Notice that, for any *r.v.* $\mathbf{Z}$ drawn either from $H \otimes G$ or F, the *r.v.* $s_\theta(\mathbf{Z})$ for any $\theta \in \Theta$, being a quadratic transform of multivariate Gaussian *r.v.*, is a weighted sum of $\chi^2(1)$ *r.v.*.

VC **dimension of** $\mathcal{S}_0(\Theta)$**.** We analyze the VC dimension of the class $\mathcal{S}_0(\Theta)$ to obtain explicit relations of the constants appearing in Theorem 2 with the dimensions of the spaces $\mathcal{X}$ and $\mathcal{Y}$. Notice that,

$$\mathcal{S}_0(\Theta) = \{s_\theta : z \in \mathbb{R}^{q+l} \mapsto \langle\theta, zz^t\rangle_F, \quad \theta \in \Theta\}$$

where $\langle\cdot,\cdot\rangle_F$ is the Frobenius inner product in $\mathbb{R}^{(q+l)\times(q+l)}$. This yields the collection of subgraphs taking the form:

$$\{\{(z,t) \in \mathbb{R}^{(q+l)} \times \mathbb{R} \mapsto \langle\theta, zz^t\rangle_F > t\}, \quad \theta \in \Theta\}\ .$$

It is a VC-class of functions by recognizing linear separator for matrix networks (taking the sign function), where θ is definite positive thus of full rank, the VC dimension can be upperbounded by $c(q+l)^2$, with $c = 2\log(24)$ constant. We refer to Khavari and Rabusseau (2021), Theorem 2 therein, stating general upperbounds applied to tensor networks. Therefore, the constants involved in Definition 3 for the class $\mathcal{S}_0$ are: $A = (cV(16e)^V)^{1/(2(V-1))}$, with $V = c(q+l)^2$. Applying the permanence properties proved in Clémençon et al. (2021), all resulting classes of functions implied in the analysis of the R-statistic in *Step 2* are therefore VC bounded and of parameters depending similarly to those of the basis class $\mathcal{S}_0$, see Lemma 14,19,20 in particular. By Proposition 2.1 Giné and Guillou (2001), we can see that the dominant constant C_1 appearing in Theorem 2, as function of the parameters of the class $\mathcal{S}_0$, is a linear combination of V and V^2, while C_2 is a linear combination of V and $\sqrt{V}$.

Definition 1: interpretation of the alternative hypothesis $\mathcal{H}_1$**.** Notice that for any distribution $H \otimes G$ and F, *i.e.* not necessarily Gaussian, choosing the score-generating function $\phi(u) = u$ trivially leads to $\mathcal{H}_1(\varepsilon)$: $\text{AUC(s)} - 1/2 \geq \varepsilon/(1-\text{p})$. The deviation from the null hypothesis thus depends linearly on ε.

Definition 2: bipartite ranking bias. In this setting $\delta = 0$ as $\mathcal{S}_0(\Theta) \subset \mathcal{S}^*$.

6.4 Nonlinear Dependence - Example 2

Gumbel (1960) proposed a construction of dependent absolute continuous univariate *r.v.* that allows for larger class of alternative hypotheses. Let $X,\ Y$ of *resp.* distribution functions $h(\mathrm{d}x)$ and $g(\mathrm{d}y)$, the class of joint distributions indexed by the dependence parameter $\rho \in [-1,1]$ can be defined by $f_\rho(x,y) = h(x)g(y)(1+\rho(2H(x)-1)(2G(y)-1))$, yielding the explicit oracle class $\mathcal{S}^*$ by noticing that $\Psi_\rho(x,y) = \rho(2H(x)-1)(2G(y)-1)$.

7 TECHNICAL PROOFS

7.1 Proof of Theorem 1

The equivalence between assertions (i) and (ii) results from Corollary 7 in Clémençon et al. (2011), applied to the pair $(H \otimes G,\ F)$ and combined with the equality $\text{ROC}^*(\cdot) = \text{ROC}(\Psi,\ \cdot) = \text{ROC}(\text{s}^*,\ \cdot)$ for any $s^* \in \mathcal{S}^*$ by Proposition 3. One establishes the remaining equivalences by using Equation (4).

7.2 Proof of Proposition 1

Assertions (i) and (ii) are inherent to the construction of the subsamples by mutual independence of all random pairs $(\mathbf{X}_i,\ \mathbf{Y}_i)$s, and their independence with the permutations σ, σ' independently drawn at random from $\mathfrak{S}_{n_-}$ and $\mathfrak{S}_{n'_-}$.

7.3 Proof of Proposition 2

Let $\alpha \in (0,1)$, $1 \leq n'_- < n'$ and $1 \leq\ n'_+ < n'$. Consider a scoring function $\hat{s} = \mathcal{A}(\mathcal{D}^-_{n_-}, \mathcal{D}^+_{n_+})$ solution of *Step 2*, see Fig. 3. By Proposition 1, $\hat{s}$ is independent of both $\mathcal{D}'^-_{n'_-}$ and $\mathcal{D}'^+_{n'_+}$, hence conditioning on the subsample $\mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}$ under the null hypothesis yields *a.s.*:

$$\mathbb{P}_{\mathcal{H}_0}\left\{\Phi^\phi_\alpha = +1 \;\middle|\; \mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}\right\} \;=\; \mathbb{P}_{\mathcal{H}_0}\left\{\frac{1}{n'_+}\widehat{W}^\phi_{n'_-,n'_+}(\widehat{s},\sigma') > \int_0^1 \phi(u)\mathrm{d}u + q^\phi_{n'_-,n'_+}(\alpha) \;\middle|\; \mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}\right\} \;\leq\; \alpha\ \ ,$$

where the first equality holds true by definition of the test statistic. The inequality results from the definition of the $(1-\alpha)$-quantile $q^\phi_{n'_-,n'_+}(\alpha)$ of the pushforward distribution of $\mathcal{L}^\phi_{n'_-,n'_+}$, by the mapping $w \mapsto (1/n')w - \int_0^1 \phi(u)\mathrm{d}u$, depending only on ϕ, n'_+ and n'_-. Then taking the expectation *w.r.t.* $\mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}$ concludes the proof.

7.4 Proof of Theorem 2

Let $\alpha \in (0,1)$, $\varepsilon > 0$, $\delta > 0$, and consider a scoring function $\hat{s} = \mathcal{A}(\mathcal{D}^-_{n_-}, \mathcal{D}^+_{n_+}) \in \mathcal{S}_0$, solution of the bipartite ranking step (*Step 2*) when formulated as the maximization of the empirical W_ϕ-performance criterion over the class $\mathcal{S}_0$, see Fig. 3. Observe that for all alternatives $(H \otimes G,\ F)$ in $\mathcal{H}_1(\varepsilon) \cap \mathcal{B}_1(\delta)$, the deviation of the rank statistic from the null decomposes *a.s.* as:

$$\frac{1}{n'_+}\widehat{W}^\phi_{n'_-,n'_+}(\widehat{s},\sigma') - \int_0^1 \phi(u)\mathrm{d}u = \left\{\frac{1}{n'_+}\widehat{W}^\phi_{n'_-,n'_+}(\widehat{s},\sigma') - W_\phi(\widehat{s})\right\} - \left\{W^*_\phi - W_\phi(\widehat{s})\right\} + \left\{W^*_\phi - \int_0^1 \phi(u)\mathrm{d}u\right\}\ , \tag{17}$$

and the generalization deviation of the W_ϕ-performance criterion satisfies, by Definition 2:

$$W^*_\phi - W_\phi(\widehat{s}) \leq 2\sup_{s\in\mathcal{S}_0}\left|\frac{1}{n_+}\widehat{W}_{n_-,n_+}(s,\sigma) - W_\phi(s)\right| + \delta\ . \tag{18}$$

We can bound the type-II error on the samples $\mathcal{D}'^-_{n'_-} \cup \mathcal{D}'^+_{n'_+}$ as follows:

$$\begin{aligned}\mathbb{P}_{H,G}\left\{\Phi^\phi_\alpha = 0\right\} &= \mathbb{P}_{H,G}\left\{\frac{1}{n'_+}\widehat{W}^\phi_{n'_-,n'_+}(\widehat{s},\sigma') - \int_0^1 \phi(u)\mathrm{d}u \leq q^\phi_{n'_-,n'_+}(\alpha)\right\}\\ &\leq \mathbb{P}_{H,G}\left\{2\sup_{s\in\mathcal{S}_0}\left|\frac{1}{n_+}\widehat{W}_{n_-,n_+}(s,\sigma) - W_\phi(s)\right| + \left|\frac{1}{n'_+}\widehat{W}^\phi_{n'_-,n'_+}(\widehat{s},\sigma') - W_\phi(\widehat{s})\right| \geq \varepsilon - \delta - \sqrt{\frac{\log(18/\alpha)}{Cn'}}\right\}\end{aligned} \tag{19}$$

where $C = 8^{-1}\min\left(p/\|\phi\|^2_\infty, (p\|\phi'\|^2_\infty)^{-1}, ((1-p)\|\phi'\|^2_\infty)^{-1}\right)$ and as soon as $n' \geq 4\log(18/\alpha)/(C(\varepsilon-\delta)^2)$. We sequentially used Eq. (17) and (18), and Proposition 4 to upperbound the quantile applied to samples of sizes n'_+, n'_-, proved in section 7.5.

We now apply Theorem 5 in Clémençon et al. (2021) to bound the uniform deviation of the W_ϕ-ranking performance criterion to its estimator based on the two-samples $\mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}$, such that for all $n \geq 16C_1^2/(p(\varepsilon-\delta)^2)$:

$$\begin{aligned}&\mathbb{P}_{H,G}\left\{2\sup_{s\in\mathcal{S}_0}\left|\frac{1}{n_+}\widehat{W}_{n_-,n_+}(s,\sigma) - W_\phi(s)\right| \geq \frac{\varepsilon-\delta}{2}\right\}\\ &\qquad\qquad \leq C_2\exp\left\{-\frac{np(p\wedge(1-p))}{4C_2}(\varepsilon-\delta)\log\left(1 + \frac{\varepsilon-\delta}{16C_1(p\wedge(1-p))}\right)\right\}\ ,\end{aligned} \tag{20}$$

as soon as $n \geq 16C_1^2/(p(\varepsilon-\delta)^2)$, constants $C_1 > 0$, $C_2 \geq 24$ depend on ϕ, $\mathcal{V}$ of values detailed in the dedicated proof, see Clémençon et al. (2021), Appendix section B.3 therein.

We can now upperbound the deviation of the two-sample rank statistic *w.r.t.* the W_ϕ-ranking performance criterion by conditioning on the first subsample $\mathcal{D}_n = \mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}$ and applying the inequality (27), to the two independent samples:

$$\{\hat{s}(\mathbf{X}_{n+1}, \mathbf{Y}_{n+\sigma'(1)}),\ \ldots,\ \hat{s}(\mathbf{X}_{n+n'_-}, \mathbf{Y}_{n+\sigma'(n'_-)})\} \cup \{\hat{s}(\mathbf{X}_{n+n'_-+1}, \mathbf{Y}_{n+n'_-+1}),\ \ldots,\ \hat{s}(\mathbf{X}_N, \mathbf{Y}_N)\}$$

$$\mathbb{P}_{H,G}\left\{\left|\frac{1}{n'_+}\widehat{W}^\phi_{n'_-,n'_+}(\widehat{s},\sigma') - W_\phi(\widehat{s})\right| \geq \frac{\varepsilon-\delta}{2} - \sqrt{\frac{\log(18/\alpha)}{Cn'}} \mid \mathcal{D}^-_{n_-} \cup \mathcal{D}^+_{n_+}\right\} \leq 18\exp\left\{-\frac{Cn'\left(\varepsilon-\delta\right)^2}{4}\right\} . \quad (21)$$

Finally, we obtain the desired bound by taking the expectation on the last inequality (21) and combining it with Eq. (20) using the union bound.

7.5 A nonasymptotic inequality for the testing threshold

Let $\{X_{\varepsilon,1},\ \ldots,\ X_{\varepsilon,n_\varepsilon}\}$ with $\varepsilon \in \{-,+\}$, be two independent *i.i.d.* random samples, drawn from univariate probability distributions F_ε. Recall that the univariate two-sample linear rank statistic based on these samples is defined by

$$\hat{W}^\phi_{n_-,n_+} = \sum_{i=1}^{n_+} \phi\left(\frac{R(X_{+,i})}{n+1}\right) , \quad (22)$$

where the ranks $R(X_{+,i}) = \sum_{\epsilon\in\{-,+\}}\sum_{j=1}^{n_\epsilon} \mathbb{I}\{s(X_{\epsilon,j}) \leq X_{+,i}\}$, for all $i \leq n_+$ The proposed class of linear rank statistics is distribution-free under the null, hence allows for the exact computation of the testing threshold for any sample sizes. Proposition 4 provides an upperbound for the $(1-\alpha)$-quantile $q^\phi_{n_-,n_+}(\alpha)$ of the pushforward distribution of $\mathcal{L}^\phi_{n_-,n_+}$ by the mapping $w \mapsto (1/n)w - \int_0^1 \phi(u)\mathrm{d}u$. It proves to be of order $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$ and only depending on ϕ, n_+, n_- and α.

Proposition 4. *Let $p \in (0,1)$ and $n \geq 1/p$. Let the score-generating function $\phi(u)$ satisfy Assumption 1. Set $n_+ = \lfloor pn \rfloor$ and $n_- = \lceil (1-p)n \rceil = n - n_+$. Then, for any $\alpha \in (0,1)$, the $(1-\alpha)$-quantile satisfies a.s.:*

$$q^\phi_{n_-,n_+}(\alpha) \leq \sqrt{\frac{\log(18/\alpha)}{Cn}} , \quad (23)$$

where $C = 8^{-1}\min\left(p/\|\phi\|^2_\infty, (p\|\phi'\|^2_\infty)^{-1}, ((1-p)\|\phi'\|^2_\infty)^{-1}\right)$.

Proof. The proof relies on the concentration results established in Clémençon et al. (2021), see Theorem 5 in particular, and builds upon the linearization technique exposed therein. Define by $F = pF_+ + (1-p)F_-$ the mixture *c.d.f.* of the pooled sample and of empirical estimator $\widehat{F}_n(t) = (1/n)\sum_{\varepsilon\in\{+,-\}}\sum_{i\leq n_\varepsilon} \mathbb{I}\{X_{\varepsilon,i} \leq t\}$. By considering $\phi(u)$ satisfying Assumption 1, writing its Taylor expansion of order 2 evaluated at $n\widehat{F}_n(X_{+,i})/(n+1)$ around $F(X_{+,i})$ for $1 \leq i \leq n_+$, and summed over $i \leq n_+$, results in a *a.s.* decomposition of the statistic Eq. (22). We refer to Eq. (B.3,4) in Clémençon et al. (2021) for the detailed arguments.

The terms of the resulting expansion of order one are composed of two U-statistics, for which the Hoeffding decomposition results in the linearization below:

$$\frac{1}{n_+}\widehat{W}^\phi_{n_-,n_+} - W_\phi = \widehat{W}_\phi - W_\phi + \frac{1}{n_+}\left(\widehat{V}^+_{n_+} - \mathbb{E}\left[\widehat{V}^+_{n_+}\right]\right) + \frac{1}{n_+}\left(\widehat{V}^-_{n_-} - \mathbb{E}\left[\widehat{V}^-_{n_-}\right]\right) + \frac{1}{n_+}\mathcal{R}_{n_-,n_+} , \quad (24)$$

where:

$$
\begin{aligned}
W_\phi &= \mathbb{E}[(\phi \circ F)\,(X_+)]\ ,\\
\widehat{W}_\phi &= \frac{1}{n_+}\sum_{i=1}^{n_+} (\phi \circ F)\,(X_{+,i})\ ,\\
\widehat{V}_{n_+}^{+} &= \frac{n_+-1}{n+1}\sum_{i=1}^{n_+}\int_{X_{+,i}}^{+\infty} (\phi' \circ F)(u)\mathrm{d}F_+(u)\ ,\\
\widehat{V}_{n_-}^{-} &= \frac{n_+}{n+1}\sum_{i=1}^{n_-}\int_{X_{-,i}}^{+\infty} (\phi' \circ F)(u)\mathrm{d}F_+(u)\ ,
\end{aligned}
$$

and $\mathcal{R}_{n_-,n_+}$ is the sum of the Taylor-Lagrange residual term $\widehat{T}_{n_-,n_+}$, and of the terms of order at most $\mathcal{O}_{\mathbb{P}}(n^{-1})$ inherited from the (two) Hoeffding decompositions. Precisely, it inherits from the linear statistics of order $\mathcal{O}_{\mathbb{P}}(n^{-1})$ defined by $\widehat{R}_{n_-,n_+}$, and both remainder terms being degenerate U-statistics. We detail hereafter the main steps for obtaining a nonasymptotic exponential deviation bound of the univariate rank statistic $(1/n_+)\widehat{W}^{\phi}_{n_-,n_+}$ based on Eq. (24). Following Clémençon et al. (2021), define the (nonsymmetric) bounded kernels defined on $\mathbb{R}^2$ by:

$$
k(z,z') \;\;=\;\; \mathbb{I}\{z' \leq z\}(\phi' \circ F)(z)\ .
$$

Then

$$
\mathcal{R}_{n_-,n_+} = \widehat{R}_{n_-,n_+} + \frac{n_+(n_+-1)}{n+1}U_{n_+}(k) + \frac{n_+n_-}{n+1}U_{n_-,n_+}(k) + \widehat{T}_{n_-,n_+}\ ,
$$

where $U_{n_+}(k)$ is the one-sample degenerate U-statistic of order 2 based on the positive sample with kernel k, $U_{n_-,n_+}(k)$ is the two-sample degenerate U-statistic of degree $(1,1)$ based on the two samples $\{X_{\varepsilon,1},\ \ldots,\ X_{\varepsilon,n_\varepsilon}\}$, with $\varepsilon \in \{-,+\}$, with kernel k.

Noticing that

$$
|\mathcal{R}_{n_-,n_+}| \leq |\widehat{R}_{n_-,n_+}| + p^2n|U_{n_+}(k)| + p(1-p)n|U_{n_-,n_+}(k)| + |\widehat{T}_{n_-,n_+}|\ ,
$$

one can sequentially upperbounded the tail of each term with threshold $t/16$, for any $t > 0$, in probability using: Hoeffding's classic exponential bound from Hoeffding (1963) with the union bound to $\widehat{R}_{n_-,n_+}$, Lemma 3 in Nolan and Pollard (1987) applied to U_{n_+}, Lemma 27 in Clémençon et al. (2021) to U_{n_+,n_-}, and finally for $\widehat{T}_{n_-,n_+}$, one has:

$$
\begin{aligned}
\frac{1}{n_+}|\widehat{T}_{n_-,n_+}| &\leq \|\phi''\|_\infty \left(\sup_{t\in\mathbb{R}}\left(\widehat{F}_n(t) - F(t)\right)^2 + \frac{1}{(n+1)^2}\right)\\
&\leq 3p^2\|\phi''\|_\infty \sup_{t\in\mathbb{R}}\left(\widehat{F}_{n_+}(t) - F_+(t)\right)^2 + 3(1-p)^2\|\phi''\|_\infty \sup_{t\in\mathbb{R}}\left(\widehat{F}_{n_-}(t) - F_-(t)\right)^2 + \frac{13\|\phi''\|_\infty}{n^2}\ .
\end{aligned}
$$

It remains to apply Dvoretzky–Kiefer–Wolfowitz inequality to each of the two first terms on the right hand side, while the third is negligeable *w.r.t.* the others. This concludes to, for all $nt \geq 512\|\phi'\|_\infty^2/(p\|\phi''\|_\infty)$:

$$
\mathbb{P}\left\{|\mathcal{R}_{n,m}| > \frac{t}{4}\right\} \leq 12\exp\left\{-\frac{Nt}{48\kappa_p\|\phi''\|_\infty}\right\}\ , \tag{25}
$$

and otherwise

$$
\mathbb{P}\left\{|\mathcal{R}_{n_-,n_+}| > \frac{t}{4}\right\} \leq 12\exp\left\{-\frac{\alpha_p n^2t^2}{512\|\phi'\|_\infty^2}\right\}\ , \tag{26}
$$

where $\alpha_p = \min(p,1-p)/(4(1-p))$, $\kappa_p = \max(p,1-p)$.

It remains to apply Hoeffding exponential inequality to the other terms of the decomposition Eq. (24) with threshold $t/4$ as follows:

$$\begin{aligned}
\mathbb{P}\left\{|\widehat{W}_\phi - W_\phi| > \frac{t}{4}\right\} &\leq 2\exp\left\{-\frac{pnt^2}{8\|\phi\|_\infty^2}\right\},\\
\mathbb{P}\left\{\frac{1}{n_+}\left|\widehat{V}_{n_+}^+ - \mathbb{E}\left[\widehat{V}_{n_+}^+\right]\right| > \frac{t}{4}\right\} &\leq 2\exp\left\{-\frac{nt^2}{8p\|\phi'\|_\infty^2}\right\},\\
\mathbb{P}\left\{\frac{1}{n_+}\left|\widehat{V}_{n_-}^- - \mathbb{E}\left[\widehat{V}_{n_-}^-\right]\right| > \frac{t}{4}\right\} &\leq 2\exp\left\{-\frac{nt^2}{8(1-p)\|\phi'\|_\infty^2}\right\}.
\end{aligned}$$

By virtue of the union bound, we obtain

$$\mathbb{P}\left\{\left|\frac{1}{n_+}\widehat{W}_{n_-,n_+}^\phi - W_\phi\right| > t\right\} \leq 18\exp\{-Cnt^2\}\ , \tag{27}$$

where $C = 8^{-1}\min\left(p/\|\phi\|_\infty^2, (p\|\phi'\|_\infty^2)^{-1}, ((1-p)\|\phi'\|_\infty^2)^{-1}\right)$, concluding the proof.

□

8 ADDITIONAL NUMERICAL EXPERIMENTS

This section details the technicalities related to the experiments exposed in the main corpus, as well as additional experiments on synthetic data.

Experimental parameters. All results are shown with 95% confidence interval based on $B \in \mathbb{N}^*$ Monte-Carlo samplings. The number of random permutations for the benchmark tests is chosen so that the test is calibrated $K_0 = 200$, the number of random permutations for our proposed procedure is fixed to $K_p \in \{10, 20, 50\}$. The significance level is chosen equal to $\alpha = 0.05$. We consider the pooled sample size $N \in \{500, 1000, 2000\}$, with $n = 4N/5$ and $n' = N/5$, where the subsamples are balanced $n_- = n_+ = n/2$, $n'_- = n'_+ = n'/2$, and denote by $d = 2q = 2l$. We choose the RTB parameter $u_0 \in \{0.85,\ 0.90,\ 0.95\}$.

Probabilistic models. We first consider different types of independence according to the following models. Define $\mathbf{X} = (X^1, X^2, \ldots, X^q)$ and $\mathbf{Y} = (Y^1, Y^2, \ldots, Y^l)$, the first two models sample $\mathbf{X}$ and $\mathbf{Y}$ according to the multivariate Gaussian distribution, in the continuity of Example 1.

(GL) $(\mathbf{X},\ \mathbf{Y}) \sim \mathcal{N}(e_d, (1/\sqrt{d}) \times \Gamma_\rho)$, where $e_d \in \mathbb{R}^d$ the null vector, $\text{Cov}(X^1, Y^k) = \rho$, for all $k \leq l$ and $\Gamma_{\rho,i,j} = \delta_{ij}$ otherwise, $d \in \{4, 10, 26, 50\}$ for $N = 500$ and $d \in \{4, 10\}$ for $N = 1000$.

(GL+) Covariance matrix from model (GL) extended for higher dimensions with $\text{Cov}(X^u, Y^k) = \rho$, for all $k \leq l$ and a $u \leq q$ only, and with $d \in \{100, 250, 500\}$, $N = 500$.

Also, for (GL), the range of the dependence parameter ρ are chosen such that the resulting Γ_ρ is positive definite to show directional dependency. The following data generation distributions model non-monotonic alternative hypothesis. The first subset of coordinates X^u, Y^v's are drawn according to the models below, and $X^i,\ Y^j$, for all $i, j \geq u, v$ are independently drawn from the Univariate distribution on $[0, 1]$ and are independent of the first coordinate.

(M1) $X^1 = \rho\cos\Theta + \omega_1/4$, $Y^1 = \rho\sin\Theta + \omega_2/4$, with $\rho \in \{1, 2, 3\}$, $\omega_i \sim \mathcal{N}(0, 1)$, $i \in \{1, 2\}$, and $\Theta \sim \mathcal{U}([0, 2\pi])$ all variables being independent, and with $d \in \{4, 10, 26\}$, $N \in \{500, 2000\}$.

(M1s) Sparse covariance matrix from model (M1) extended for higher dimensions by generating the X^u, Y^v's, for $u, v \leq q/2, l/2$ according to (M1) and the X^u, Y^v, for $u, v > q/2, l/2$ are drawn from the Univariate distribution on $[0, 1]$, with $d \in \{100, 250, 500\}$, $N = 500$.

(M1d) Dense covariance matrix from model (M1) extended for higher dimensions with by generating the X^u, Y^v, for all coordinates $u, v \leq q, l$ according to (M1), with $d \in \{100, 250, 500\}$, $N = 500$.

Model (M1) was proposed for both the univariate and bivariate settings by Berrett and Samworth (2019) and further studied by Albert et al. (2022) and for very small sample sizes. We compare our results for models (M1s) and (M1d) to the benchmark tests to see the resistance to high dimension d.

Model (GL)	$N = 500$, $d = 4$				$N = 500$, $d = 10$				$N = 500$, $d = 26$			
Rejection rate of the null	$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$	
Method	$\rho = 0.0$	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.0$	$\rho = 0.05$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.0$	$\rho = 0.02$	$\rho = 0.05$	$\rho = 0.07$
rForest$_{MWW}$	0.04 ± 0.19	**0.47 ± 0.50**	**0.92 ± 0.27**	**1.00 ± 0.00**	0.05 ± 0.22	**0.85 ± 0.36**	**0.92 ± 0.27**	**0.98 ± 0.14**	0.04 ± 0.19	**0.98 ± 0.14**	**0.99 ± 0.10**	**1.00 ± 0.00**
rForest$_{95}$	0.01 ± 0.10	0.02 ± 0.14	0.08 ± 0.27	0.17 ± 0.35	0.01 ± 0.10	0.04 ± 0.19	0.17 ± 0.38	0.16 ± 0.37	0.01 ± 0.10	0.29 ± 0.46	0.42 ± 0.59	0.39 ± 0.49
rForest$_{90}$	0.02 ± 0.14	0.10 ± 0.30	0.23 ± 0.42	0.49 ± 0.50	0.01 ± 0.10	0.32 ± 0.47	0.37 ± 0.46	0.46 ± 0.50	0.01 ± 0.10	0.64 ± 0.48	0.60 ± 0.49	0.77 ± 0.42
rForest$_{85}$	0.02 ± 0.14	0.17 ± 0.38	0.35 ± 0.48	0.71 ± 0.46	0.01 ± 0.10	0.41 ± 0.49	0.52 ± 0.50	0.63 ± 0.49	0.02 ± 0.14	0.79 ± 0.41	0.78 ± 0.42	0.87 ± 0.34
HSIC	0.06 ± 0.24	0.09 ± 0.29	0.06 ± 0.24	0.14 ± 0.35	0.06 ± 0.24	0.06 ± 0.24	0.09 ± 0.29	0.04 ± 0.19	0.06 ± 0.24	0.03 ± 0.17	0.06 ± 0.24	0.03 ± 0.17
dCor$_{L2}$	0.10 ± 0.30	0.06 ± 0.24	0.03 ± 0.17	0.12 ± 0.33	0.07 ± 0.26	0.03 ± 0.17	0.10 ± 0.30	0.11 ± 0.31	0.05 ± 0.22	0.03 ± 0.17	0.06 ± 0.24	0.10 ± 0.30
dCor$_{L1}$	0.08 ± 0.27	0.06 ± 0.24	0.09 ± 0.29	0.15 ± 0.36	0.05 ± 0.22	0.08 ± 0.27	0.11 ± 0.31	0.09 ± 0.29	0.04 ± 0.19	0.04 ± 0.20	0.04 ± 0.20	0.0è ± 0.26

Model (M1)	$N = 500$, $d = 4$				$N = 500$, $d = 10$				$N = 500$, $d = 26$			
Rejection rate of the null	$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$	
Method	$\rho = 0.0$	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 0.0$	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 0.0$	$\rho = 1$	$\rho = 2$	$\rho = 3$
rForest$_{MWW}$	0.04 ± 0.19	**0.78± 0.42**	**0.97± 0.17**	**0.99 ± 0.10**	0.04 ± 0.19	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	0.04 ± 0.20	**0.99 ± 0.10**	**1.00 ± 0.00**	**1.00 ± 0.00**
rForest$_{95}$	0.01 ± 0.10	0.02± 0.14	0.07 ± 0.26	0.13± 0.34	0.00 ± 0.00	0.16 ± 0.37	0.70 ± 0.46	0.88 ± 0.33	0.00 ± 0.00	0.33 ± 0.47	0.92 ± 0.27	0.99 ± 0.10
rForest$_{90}$	0.02 ± 0.14	0.16± 0.37	0.38 ± 0.47	0.52 ± 0.50	0.02 ± 0.14	0.67 ± 0.47	0.98 ± 0.14	**1.00 ± 0.00**	0.01 ± 0.10	0.80 ± 0.040	**1.00 ± 0.00**	**1.00 ± 0.00**
rForest$_{85}$	0.00 ± 0.00	0.23 ± 0.42	0.56 ± 0.50	0.71 ± 0.46	0.01 ± 0.10	0.88 ± 0.33	1.00 ± 0.00	**1.00 ± 0.00**	0.01 ± 0.10	0.89± 0.31	**1.00 ± 0.00**	**1.00 ± 0.00**
HSIC	0.03 ± 0.17	0.22 ± 0.42	0.60 ± 0.49	0.86 ± 0.35	0.05 ± 0.22	0.26 ± 0.44	0.74 ± 0.44	**1.0 ± 0.00**	0.04 ± 0.20	0.16 ± 0.37	0.98 ± 0.14	**1.00 ± 0.00**
dCor$_{L2}$	0.06 ± 0.24	0.20 ± 0.40	0.59 ± 0.50	0.83 ± 0.38	0.03 ± 0.17	0.26 ± 0.44	0.75 ± 0.44	**1.0 ± 0.00**	0.03 ± 0.17	0.17 ± 0.38	0.96 ± 0.20	**1.00 ± 0.00**
dCor$_{L1}$	0.02 ± 0.14	0.18 ± 0.39	0.49 ± 0.50	0.72 ± 0.45	0.05 ± 0.22	0.18 ± 0.39	0.55 ± 0.50	0.80 ± 0.40	0.08 ± 0.27	0.12 ± 0.33	0.79 ± 0.41	0.97 ± 0.17

Table 1: Empirical rejection rates for testing $\mathcal{H}_0$ of independence against $\mathcal{H}_1$, for models (GL, M1) ± 95% standard deviation at significance level $\alpha = 0.05$. Parameters: $\rho \in [0, 0.6]$ (GL), $\rho \in \{0, 1, 2, 3\}$ (M1), $d \in \{4, 10, 26\}$, $K_p = 50$, $B = 100$, $B_p = 200$. Results in bold specify the best performance among all methods.

Model (M1)	$N = 2000$, $d = 4$				$N = 2000$, $d = 10$				$N = 2000$, $d = 50$			
Rejection rate of the null	$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$	
Method	$\rho = 0.0$	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 0.0$	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 0.0$	$\rho = 1$	$\rho = 2$	$\rho = 3$
rForest$_{MWW}$	0.03 ± 0.17	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	0.06 ± 0.24	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	0.02 ± 0.14	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
rForest$_{95}$	0.02 ± 0.14	0.46 ± 0.50	0.89 ± 0.31	0.94 ± 0.24	0.00 ± 0.00	0.71 ± 0.46	0.96 ± 0.19	0.97 ± 0.17	0.00 ± 0.00	0.99 ± 0.1	**1.00 ± 0.00**	**1.00 ± 0.00**
rForest$_{90}$	0.02 ± 0.14	0.82 ± 0.38	**1.00 ± 0.00**	**1.00 ± 0.00**	0.00 ± 0.00	0.92 ± 0.27	**1.00 ± 0.00**	**1.00 ± 0.00**	0.01 ± 0.10	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
rForest$_{85}$	0.04 ± 0.20	0.93 ± 0.26	**1.00 ± 0.00**	**1.00 ± 0.00**	0.00 ± 0.00	0.96 ± 0.19	**1.00 ± 0.00**	**1.00 ± 0.00**	0.02 ± 0.14	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**

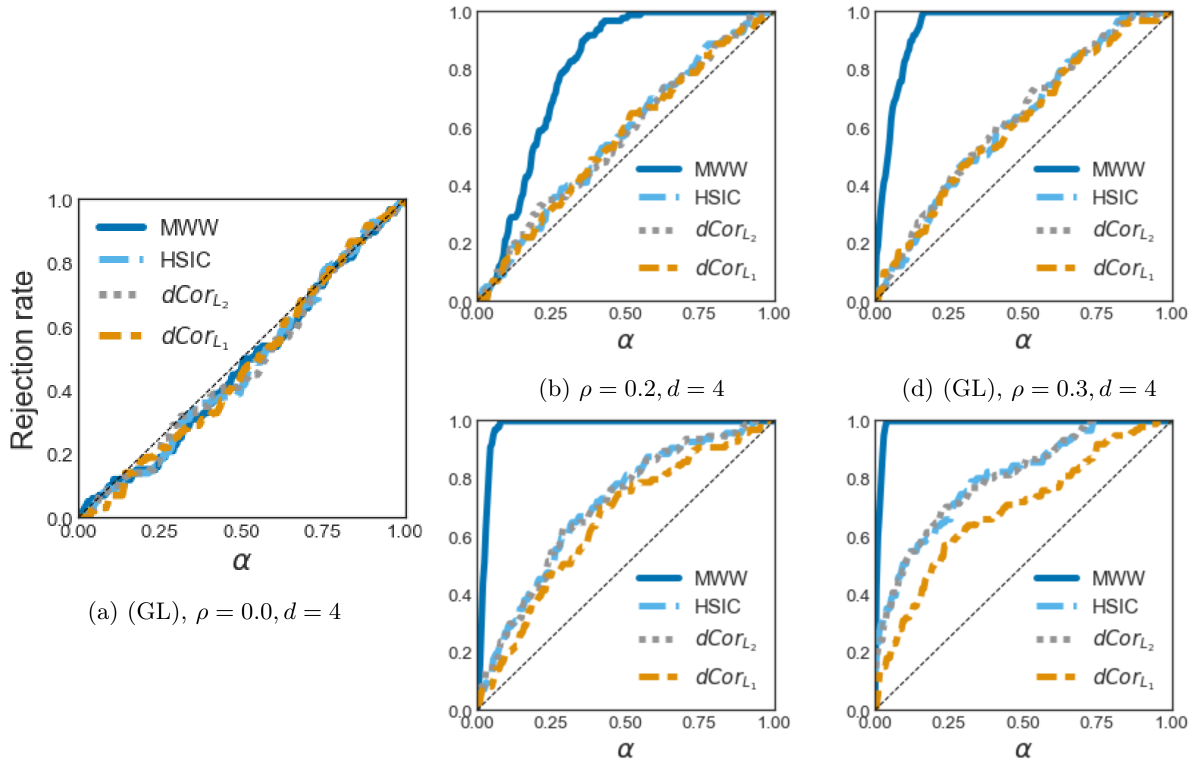
Table 2: Empirical rejection rates for the model (M1) ± 95% standard deviation at significance level $\alpha = 0.05$. The parameters are fixed to: $\rho \in \{0, 1, 2, 3\}$, $d \in \{4, 10, 50\}$, $K_p \in \{10, 20\}$, $B = 100$. Results in bold specify the best performance among all methods.

Model (M1s, $N = 500$**)**	$d = 50$				$d = 100$				$d = 250$			
Rejection rate of the null	$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$	
Method	$\rho = 0.0$	$\rho = 0.30$	$\rho = 0.40$	$\rho = 0.50$	$\rho = 0.0$	$\rho = 0.30$	$\rho = 0.40$	$\rho = 0.50$	$\rho = 0.0$	$\rho = 0.20$	$\rho = 0.30$	$\rho = 0.40$
rForest$_{MWW}$	0.03 ± 0.17	**0.94 ± 0.24**	**0.96 ± 0.20**	**0.97 ± 0.17**	0.03 ± 0.17	**0.82 ± 0.39**	**0.96 ± 0.20**	**1.00 ± 0.00**	0.07 ± 0.26	0.11 ± 0.31	**0.90 ± 0.30**	0.98 ± 0.14
rForest$_{95}$	0.01 ± 0.10	0.02 ± 0.14	0.03 ± 0.17	0.74 ± 0.44	0.01 ± 0.10	0.00 ± 0.00	0.05 ± 0.22	0.73 ± 0.45	0.01 ± 0.10	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
rForest$_{90}$	0.01 ± 0.10	0.32 ± 0.47	0.60 ± 0.49	0.95 ± 0.22	0.04 ± 0.19	0.00 ± 0.00	0.60 ± 0.49	0.98 ± 0.14	0.02 ± 0.14	0.00 ± 0.00	0.00 ± 0.00	0.06 ± 0.24
rForest$_{85}$	0.01 ± 0.10	0.71 ± 0.46	0.89 ± 0.31	**0.97 ± 0.17**	0.03 ± 0.17	0.00 ± 0.00	0.82 ± 0.39	0.99 ± 0.10	0.03 ± 0.17	0.00 ± 0.00	0.06 ± 0.24	0.24 ± 0.42
HSIC	0.05 ± 0.22	0.27 ± 0.45	0.41 ± 0.49	0.72 ± 0.45	0.03 ± 0.17	0.46 ± 0.50	0.80 ± 0.40	0.92 ± 0.27	0.04 ± 0.19	0.23 ± 0.42	0.74 ± 0.44	**1.00 ± 0.00**
dCor$_{L2}$	0.06 ± 0.24	0.28 ± 0.45	0.39 ± 0.49	0.73 ± 0.45	0.05 ± 0.22	0.41 ± 0.49	0.80 ± 0.40	0.93 ± 0.26	0.04 ± 0.19	**0.24 ± 0.43**	0.74 ± 0.44	**1.00 ± 0.00**
dCor$_{L1}$	0.04 ± 0.19	0.21 ± 0.41	0.27 ± 0.45	0.59 ± 0.49	0.02 ± 0.14	0.35 ± 0.48	0.58 ± 0.50	0.78 ± 0.42	0.03 ± 0.17	0.22 ± 0.42	0.58 ± 0.50	0.93 ± 0.27

Table 3: Empirical rejection rates for model (M1s) ± 95% standard deviation at significance level $\alpha = 0.05$. Parameters: $\rho \in \{0.0, 0.1, 0.2, 0.3\}$, $d \in \{50, 100, 250\}$, $K_p \in \{10, 50\}$, $K_0 = 200$, $B = 100$. Results in bold specify the best performance among all methods.

Model (M1d, $N = 500$**)**	$d = 50$				$d = 100$				$d = 250$			
Rejection rate of the null	$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$		$\mathcal{H}_0$		$\mathcal{H}_1$	
Method	$\rho = 0.0$	$\rho = 0.20$	$\rho = 0.30$	$\rho = 0.40$	$\rho = 0.0$	$\rho = 0.15$	$\rho = 0.20$	$\rho = 0.30$	$\rho = 0.0$	$\rho = 0.20$	$\rho = 0.30$	$\rho = 0.40$
rForest$_{MWW}$	0.03 ± 0.17	**0.37 ± 0.49**	**0.99± 0.10**	0.96 ± 0.20	0.03 ± 0.17	0.00 ± 0.00	0.02 ± 0.14	**0.99 ± 0.10**	0.07 ± 0.26	0.00 ± 0.00	0.97 ± 0.17	**1.00 ± 0.00**
rForest$_{95}$	0.01 ± 0.10	0.00 ± 0.00	0.10 ± 0.31	0.03 ± 0.17	0.01 ± 0.10	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.14	0.01 ± 0.10	0.00 ± 0.00	0.00 ± 0.00	0.85 ± 0.36
rForest$_{90}$	0.01 ± 0.10	0.00 ± 0.00	0.65 ± 0.48	0.60 ± 0.49	0.04 ± 0.19	0.00 ± 0.00	0.00 ± 0.00	0.32 ± 0.47	0.02 ± 0.14	0.00 ± 0.00	0.00 ± 0.00	0.98 ± 0.14
rForest$_{85}$	0.02 ± 0.14	0.00 ± 0.00	0.92 ± 0.39	0.89± 0.31	0.03 ± 0.17	0.00 ± 0.00	0.00 ± 0.00	0.71 ± 0.46	0.03 ± 0.17	0.00 ± 0.00	0.06 ± 0.24	0.99± 0.10
HSIC	0.05 ± 0.22	0.32 ± 0.47	0.93 ± 0.26	**0.99 ± 0.10**	0.03 ± 0.17	0.27± 0.45	0.43 ± 0.50	0.93 ± 0.26	0.04 ± 0.19	0.74± 0.44	**1.00 ± 0.00**	**1.00 ± 0.00**
dCor$_{L2}$	0.06 ± 0.24	0.33 ± 0.47	0.95 ± 0.22	**0.99 ± 0.10**	0.05 ± 0.2	**0.29± 0.46**	**0.47 ± 0.50**	0.95 ± 0.22	0.04 ± 0.19	0.74± 0.44	**1.00 ± 0.00**	**1.00 ± 0.00**
dCor$_{L1}$	0.04 ± 0.19	0.19 ± 0.39	0.80 ± 0.40	0.88 ± 0.33	0.02 ± 0.14	0.15 ± 0.36	0.32 ± 0.47	0.80 ± 0.40	0.03 ± 0.17	0.68 ± 0.47	**1.00 ± 0.00**	**1.00 ± 0.00**

Table 4: Empirical rejection rates for models (M1d) ± 95% standard deviation at significance level $\alpha = 0.05$. Parameters: $\rho \in \{0.0, 0.1, 0.2, 0.3\}$, $d \in \{50, 100, 250\}$, $K_p \in \{10, 20\}$, $K_0 = 200$, $B = 100$. Results in bold specify the best performance among all methods.



(a) (GL), $\rho = 0.0, d = 4$

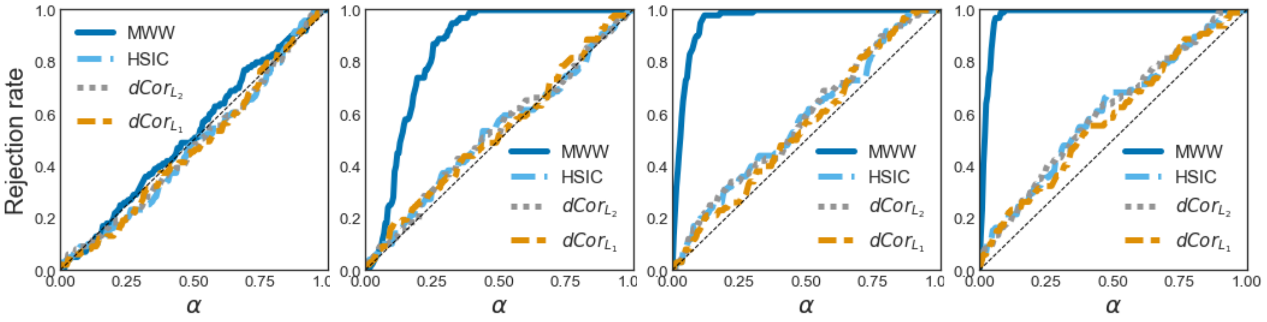
(b) $\rho = 0.2, d = 4$

(d) (GL), $\rho = 0.3, d = 4$

(c) (GL), $\rho = 0.4, d = 4$

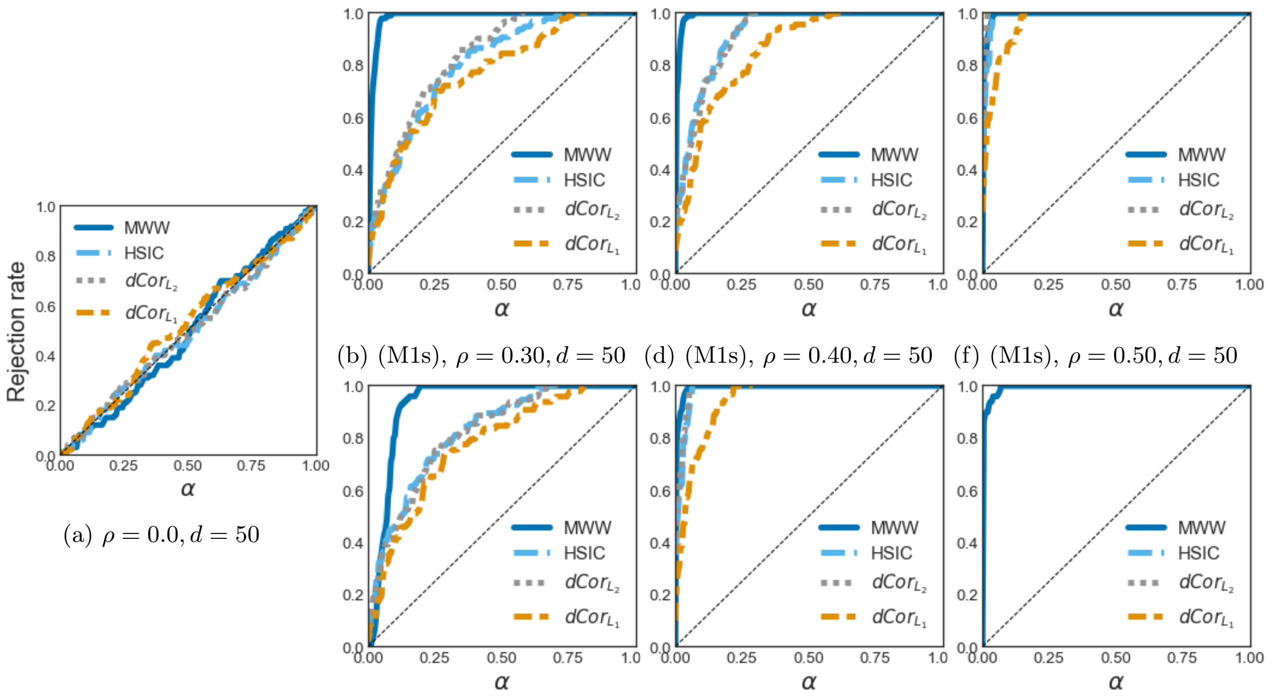
(e) (GL), $\rho = 0.5, d = 4$

Figure 4: Plots of the rejection rate under $\mathcal{H}_0$ (a) and $\mathcal{H}_1$ (b-e) against the significance level $\alpha \in (0, 1)$ for (GL) with $\phi(u) = u$ ($\texttt{rForest}_{MWW}$), $\rho = 0.0$ (a) $\rho = 0.2$ (b), $\rho = 0.3$ (c), $\rho = 0.4$ (d), $\rho = 0.5$ (e). The parameters are fixed to $N = 1000$, $d = 4$, $K_p = 10$, $K_0 = 200$, $B = 100$ for all experiments.



(a) (GL), $\rho = 0.0, d = 10$ (b) (GL), $\rho = 0.10, d = 10$ (c) (GL), $\rho = 0.15, d = 10$ (d) (GL), $\rho = 0.20, d = 10$

Figure 5: Plots of the rejection rate under $\mathcal{H}_0$ (a) and $\mathcal{H}_1$ (b-d) against the significance level $\alpha \in (0, 1)$ for (GL) with $\phi(u) = u$ ($\texttt{rForest}_{MWW}$), $\rho = 0.0$ (a) $\rho = 0.10$ (b), $\rho = 0.15$ (c), $\rho = 0.20$ (d). The parameters are fixed to $N = 1000$, $d = 10$, $K_p = 10$, $K_0 = 200$, $B = 100$ for all experiments.

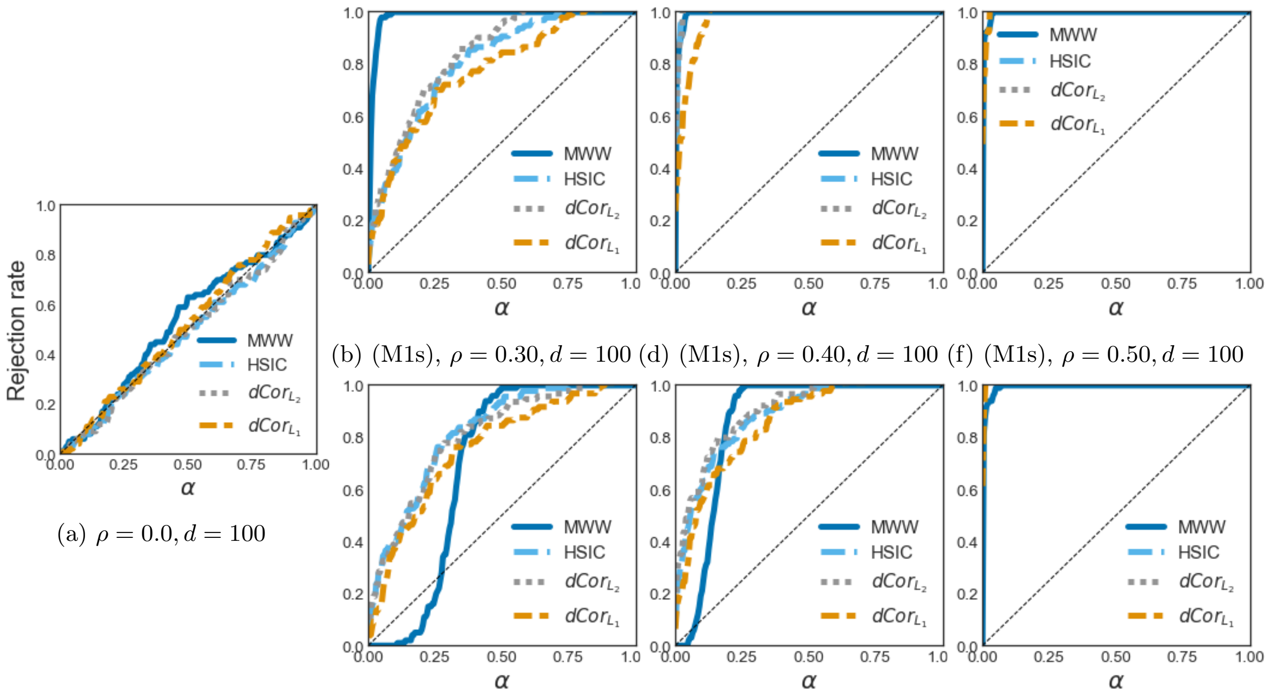


(a) $\rho = 0.0, d = 50$

(b) (M1s), $\rho = 0.30, d = 50$ (d) (M1s), $\rho = 0.40, d = 50$ (f) (M1s), $\rho = 0.50, d = 50$

(c) (M1d), $\rho = 0.20, d = 50$ (e) (M1d), $\rho = 0.30, d = 50$ (g) (M1d), $\rho = 0.40, d = 50$

Figure 6: Plots of the rejection rate under $\mathcal{H}_0$ (a) and $\mathcal{H}_1$ (b-g) against the significance level $\alpha \in (0,1)$ for (M1s) top row, and (M1d) bottom row, with $\phi(u) = u$ ($\texttt{rForest}_{MWW}$), $\rho = 0.0$ (a) $\rho \in (0.20, 0.50)$ (b-g). The parameters are fixed to $N = 500$, $d = 50$, $K_p = 10$, $K_0 = 200$, $B = 100$ for all experiments.



(a) $\rho = 0.0, d = 100$

(b) (M1s), $\rho = 0.30, d = 100$ (d) (M1s), $\rho = 0.40, d = 100$ (f) (M1s), $\rho = 0.50, d = 100$

(c) (M1d), $\rho = 0.15, d = 100$ (e) (M1d), $\rho = 0.20, d = 100$ (g) (M1d), $\rho = 0.30, d = 100$

Figure 7: Plots of the rejection rate under $\mathcal{H}_0$ (a) and $\mathcal{H}_1$ (b-g) against the significance level $\alpha \in (0,1)$ for (M1s) top row, and (M1d) bottom row, with $\phi(u) = u$ ($\texttt{rForest}_{MWW}$), $\rho = 0.0$ (a) $\rho \in (0.15, 0.50)$ (b-g). The parameters are fixed to $N = 500$, $d = 100$, $K_p = 10$, $K_0 = 200$, $B = 100$ for all experiments.

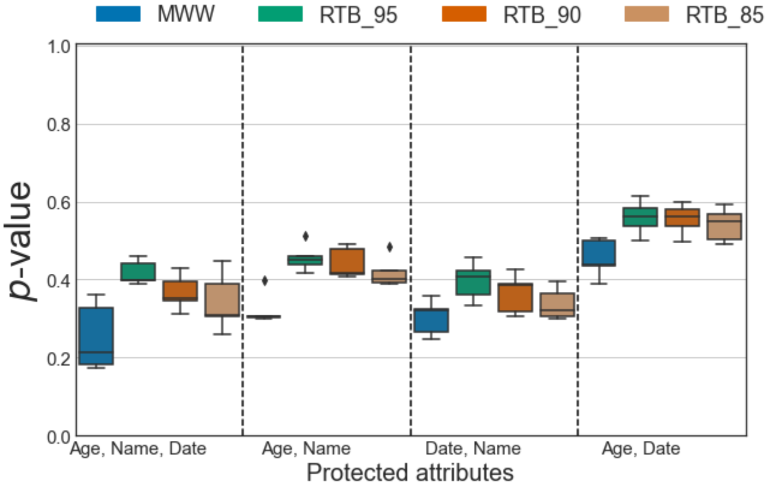


Figure 8: Boxplots of the p-values for different sets of protected attributes. The experimental parameters are fixed to $N = 10^3$, $K_p = 10$, $q = 1, l \in \{2, 3\}$, 5-fold cross-validation, 31 features, based on the open-source dataset available Jesus et al. (2022).